

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

**Improving compound synthesis efficiency
through laboratory automation
and artificial intelligence**

David Friedrich Erhard Nippa

aus

München, Deutschland

2024

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Ivan Huc betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 29.01.2024

David Nippa

Dissertation eingereicht am: 29.01.2024

1. Gutachter: Prof. Dr. Ivan Huc

2. Gutachter: Prof. Dr. Gisbert Schneider

Mündliche Prüfung am: 18.03.2024

Dissertation

IMPROVING COMPOUND SYNTHESIS
EFFICIENCY THROUGH LABORATORY
AUTOMATION AND ARTIFICIAL INTELLIGENCE



DAVID F. E. NIPPA

Ludwig-Maximilians-Universität München

2024

Dedicated to my parents and my brother.

SUMMARY

The synthesis of novel, complex drug molecules to establish structure-activity relationships (SAR) is often the limiting step in early drug discovery. To expedite SAR exploration and enhance the pharmacological profiles of lead structures within the design-make-test-analyze (DMTA) cycle, it is crucial to refine synthetic methodologies. Late-stage functionalization (LSF) offers an effective, step-saving approach for modifying advanced leads by directly substituting C–H bonds with other moieties, thereby facilitating chemical space exploration and modulating adsorption, distribution, metabolism and excretion (ADME) properties. However, the similarity of C–H bonds within structurally intricate drug and drug-like molecules necessitates a detailed understanding of their reactivity for targeted functionalization, which complicates the standardization of experimental protocols. This complexity often results in resource-intensive wet lab explorations, which may conflict with the stringent timelines and budgets of drug discovery projects.

High-throughput experimentation (HTE) has emerged as a key technology to streamline synthesis by efficiently evaluating reaction conditions in a plate format using automation equipment. Tackling certain remaining bottlenecks of HTE, specifically in the field of software/hardware integration and data governance, the technology has the potential to efficiently assess LSF reaction methodologies with the lowest possible material consumption. The LSF reaction data sets from HTE campaigns combined with big data analytics and machine learning (ML) are expected to enable the development of predictive models for C–H bond transformations. This would allow the estimation of reaction outcomes before carrying out resource and time-intensive experimentation in the laboratory facilitating the synthesis of target molecules in an environmentally conscious and material-efficient manner.

Despite the potential of making LSF a more efficient methodology to enable fast drug diversification and, consequently, speed up the development of novel medicines, a seamless connection between all three research fields, namely, LSF, HTE and reactivity prediction has not been made so far.

This thesis presents the development of a digital, semi-automated HTE system designed to systematically evaluate LSF methodologies on drug-like molecules. **DOLPHIN**, the **Data-orchestrated laboratory platform harnessing innovative neural network**, is an end-to-end platform tailored for LSF that incorporates automation, digitalization, and ML to enhance compound synthesis efficiency in early drug discovery. Advanced automated laboratory equipment, such as solid and liquid dosing robots, is employed to simultaneously initiate reactions and prepare controls, ensuring sample quality for subsequent analyses. A high level of software/hardware integration supports the workflow from literature analysis and reaction plate screening to scale-up planning and data management.

To allow the extraction, curation, storage and analysis of reaction data from the literature, in parallel with the development of **DOLPHIN**, efforts have been directed towards the development of a simple, user-friendly reaction format (**SURF**). After evaluating current data-sharing practices and identifying bottlenecks, **SURF** was designed to be both human- and machine-readable, streamlining the use of reaction data in ML applications. Application of this format to curate data from selected publications enabled systematic HTE plate design and provided high-quality data sets for ML model development.

Applying **DOLPHIN** and **SURF** in two case studies with different LSF reaction types enabled reactivity prediction. The first case study was centered around assessing the applicability of C–H borylation reactions for the late-stage diversification of complex molecules. Hundreds of HTE reactions were performed on systematically chosen commercial drugs under a wide array of conditions. The data generated from these experiments were captured in **SURF** and used to support the development of an ML algorithm capable of predicting binary reaction outcomes, yields, and regioselectivity for novel substrates. The influence of steric and electronic effects on model performance was quantified by featurization of the input molecular graphs with 2D, 3D and quantum mechanics (QM) augmented information. The reactivity of novel reactions with known and unknown substrates was classified with a balanced accuracy of 92% and 67%, respectively, while computational models predicted reaction yields for diverse reaction conditions with a mean absolute error (MAE) margin of 4–5%. The platform delivered numerous starting points for the structural diversification of commercial pharmaceuticals and

advanced drug-like fragments.

The second case study investigated a library-type screening approach for determining the substrate scope of late-stage Minisci-type C–H alkylations to explore new exit vectors. This approach aimed to facilitate the *in silico* prediction of suitable substrates that can undergo coupling with a diverse array of sp³-rich carboxylic acids. Again, DOLPHIN and SURF provided the experimental data sets to train ML models for the described task. The algorithms predicted reaction yields with an MAE of 11–12% and suggested starting points for scale-up reactions of 3180 advanced heterocyclic building blocks with various carboxylic acid building blocks. From those, a set of promising candidates was chosen, reactions were scaled up to the 50 to 100 mg range and products were isolated and characterized. This process led to the creation of 30 novel, functionally modified molecules that hold potential for further optimization. The results from both case studies positively advocate the application of ML based on high-quality HTE data for reactivity prediction in the LSF space and beyond.

In summary, this thesis established a semi-automated platform (DOLPHIN) and a new reaction format (SURF), facilitating the development of ML models for LSF reaction screening, thereby contributing to enhancing the compound synthesis efficiency in drug discovery through the strategic application of laboratory automation and artificial intelligence.

KURZFASSUNG

Die Synthese neuartiger, komplexer Arzneimoleküle zur Etablierung von Struktur-Aktivitäts-Beziehungen (structure-activity-relationships, SAR) ist oft der limitierende Schritt in der frühen Arzneimittelforschung. Um die Aufklärung von SAR zu beschleunigen und die pharmakologischen Profile von Leitstrukturen innerhalb des Design-Synthese-Test-Analyse (design-make-text-analyze, DMTA)-Zyklus zu verbessern, ist es von entscheidender Bedeutung, neue, synthetische Methoden zu explorieren. Die späte Funktionalisierung (late-stage functionalization, LSF) bietet einen effektiven, schrittsparenden Ansatz für die Modifizierung fortgeschrittener Leitstrukturen durch die direkte Substitution von C-H-Bindungen durch andere Reste oder funktionale Gruppen Komponenten. Dadurch kann die Erforschung des chemischen Raums und die Modulation der Adsorption, Verteilung, Metabolismus und Ausscheidung (ADME) Eigenschaften erleichtert werden. Allerdings erfordert die Ähnlichkeit der C-H-Bindungen in komplexen arzneistoff- und wirkstoffähnlichen Molekülen für eine gezielte Funktionalisierung, ein detailliertes Verständnis ihrer Reaktivität, wodurch sich die standardisierte Applikation von Reaktionsvorschriften schwierig gestaltet. Diese Komplexität führt häufig zu umfangreichen Laborexperimenten, die mit den strengen Zeit- und Budgetplänen von Arzneimittelentwicklungsprojekten in Konflikt geraten können.

Hochdurchsatz-Experimente (high-throughput experimentation, HTE) haben sich als Schlüsseltechnologie etabliert, um die Synthese von Molekülen durch paralleles Screening von Reaktionsbedingungen im Plattenformat unter Verwendung von Laborautomatisierung effizienter zu gestalten. Indem bestehende Limitierungen im Gebiet der HTE, insbesondere die Bereiche Software-/Hardware-Integration und Datenverwaltung, adressiert werden, hat die Technologie das Potenzial, die Anwendbarkeit von LSF-Reaktionen mit minimalem Verbrauch von Startmaterialien zu analysieren. Es wird erwartet, dass die aus diesen Experimenten gewonnenen qualitativ hochwertigen Reaktionsdatensätze, kombiniert mit Datenanalyse und maschinellem Lernen (ML) die Entwicklung von computergestützten Modellen zur Vorhersage von LSF

Transformationen ermöglichen könnten. Dies würde die Abschätzung von Reaktionsergebnissen ermöglichen, bevor ressourcen- und zeitintensive Experimente im Labor durchgeführt werden, wodurch die Synthese von Zielmolekülen in der medizinischen Chemie umweltbewusster und effizienter gestaltet werden könnte.

Trotz des Potenzials, LSF zu einer effizienteren Methode zu machen, die eine schnelle Derivatisierung von arzneimittel-ähnlichen Molekülen ermöglicht und damit die Entwicklung neuer Medikamente beschleunigt, wurde bisher keine nahtlose Verbindung zwischen den drei Forschungsbereichen, LSF, HTE und der computergestützten Vorhersage von Reaktionsprodukten, hergestellt.

Aus diesem Grund hat die vorliegende Dissertation ein digitales, halbautomatisiertes HTE-System mit dem Namen DOLPHIN (**D**ata-**o**rchestrated **l**aboratory **p**latform **h**arnessing **i**nnovative **n**eural **n**etworks, deut. daten-getriebene Laborplattform, die innovative neuronale Netzwerke nutzt) entwickelt. DOLPHIN ist darauf ausgelegt, die Anwendbarkeit von LSF-Methoden an wirkstoffähnlichen Molekülen systematisch zu analysieren. Dabei integriert die Plattform Automatisierung, Digitalisierung und ML, um die Effizienz der Synthese von Verbindungen in der frühen Arzneistoffforschung zu verbessern. Moderne, automatisierte Laborgeräte, wie zum Beispiel Feststoff- und Flüssigkeitsdosierroboter, werden eingesetzt, um Reaktionen gleichzeitig anzusetzen und den Reaktionsfortschritt zu kontrollieren. Ein hohes Maß an Software-Hardware-Integration unterstützt den Prozess von der Literaturanalyse über die Planung und Ausführung von Screening und Scale-up Experimenten bis hin zum Datenmanagement.

Um die Extraktion, Kuratierung, Speicherung und Analyse von Reaktionsdaten aus der Literatur zu ermöglichen, wurden parallel zur Entwicklung von DOLPHIN die Bemühungen auf die Entwicklung eines einfachen, benutzerfreundlichen Reaktionsformats (simple user-friendly reaction format, SURF) gerichtet. Nach einer Bewertung der derzeitigen Praktiken für die gemeinsame Nutzung von Daten und der Ermittlung von bestehenden Limitierungen wurde SURF so konzipiert, dass es sowohl von Menschen als auch von Maschinen verstanden werden kann und damit die Verwendung von Reaktionsdaten in ML-Modellen vereinfacht wird. Die Anwendung dieses Formats zur Kuratierung von Daten aus ausgewählten Veröffentlichungen ermöglichte das systematische Design von HTE-Platten und lieferte hochwertige Datensätze für die Entwicklung von ML-Algorithmen.

Die Anwendung von DOLPHIN und SURF in zwei Fallstudien mit verschiedenen LSF-Reaktionstypen wurde genutzt, um ML Modelle zur Vorhersage der chemischen Reaktivität zu entwickeln. Die erste Fallstudie konzentrierte sich auf die Bewertung der Anwendbarkeit von C-H-Borylierungsreaktionen für die LSF von komplexen Molekülen. Hunderte von HTE-Reaktionen wurden unter einer Vielzahl von Bedingungen an systematisch ausgewählten kommerziellen Arzneistoffen durchgeführt. Die aus diesen Experimenten gewonnenen Daten wurden in SURF erfasst und für die Entwicklung eines ML-Algorithmus verwendet, der in der Lage ist, binäre Reaktionsergebnisse, Ausbeuten und Regioselektivität für neue Substrate vorherzusagen. Der Einfluss sterischer und elektronischer Effekte auf die Genauigkeit der Modelle wurde durch die Kodeierung der Startmaterialien mit 2D-, 3D- und quantenmechanischen (QM) Informationen quantifiziert. Die Reaktivität neuartiger Reaktionen mit bekannten und unbekanntem Substraten wurde mit einer ausgewogenen Genauigkeit von 92% bzw. 67% klassifiziert, während die Algorithmen die Reaktionsausbeuten für verschiedene Reaktionsbedingungen mit einer mittleren absoluten Fehlermarge (mean absolute error, MAE) von 4-5% vorher sagten. Die Plattform lieferte zahlreiche Startpunkte für die strukturelle Diversifizierung kommerzieller Pharmazeutika und fortgeschrittener arzneistoffähnlicher Fragmente.

Die zweite Fallstudie untersuchte einen bibliotheksbasierten Screening-Ansatz zur Bestimmung des Substratspektrums von späten C-H-Alkylierungen des Minisci Reaktionstyps, um neue Exitvektoren zu erforschen. Diese Forschung zielte darauf ab, die *in silico* Vorhersage geeigneter Substrate zu erleichtern, welche mit einer vielfältigen Palette von sp³-reichen Carbonsäuren gekoppelt werden können. Auch hier lieferten DOLPHIN und SURF die experimentellen Datensätze, um ML-Modelle für die beschriebene Aufgabe zu trainieren. Die Algorithmen sagten Reaktionsausbeuten mit einem MAE von 11-12% voraus und schlugen Startpunkte für Reaktionen in größerem Massstab ausgehend von einem Datensatz mit 3180 fortgeschrittenen heterozyklischen Bausteinen und verschiedenen Carbonsäurebausteinen vor. Aus den Vorhersagen wurden vielversprechende Kandidaten ausgewählt, die Reaktionen wurden auf einen Bereich von 50 bis 100 mg hochskaliert, und die Produkte isoliert und charakterisiert. Auf diese Weise entstanden 30 neuartige, funktionell veränderte Moleküle, die sich für eine weitere Optimierung eignen. Die Ergebnisse beider Fallstudien befürworten die Anwendung von ML auf der Grundlage hochwertiger HTE-Datensätze für die Reaktivitätsvorhersage von LSF Reaktionen und weiteren Reaktionstypen.

Zusammenfassend hat diese Dissertation eine halbautomatisierte Plattform (DOLPHIN) und ein neues Reaktionsformat (SURF) entwickelt, welche die Entwicklung von ML-Modellen für das *in silico* Screening von LSF-Reaktionen ermöglicht haben. Damit hat diese Forschung dazu beigetragen, die Effizienz der chemischen Synthese in der Arzneistoffforschung durch die strategische Anwendung von Laborautomatisierung und künstlicher Intelligenz zu steigern.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my thanks to Prof. Dr. Ivan Huc for making this interdisciplinary and collaborative thesis possible. I appreciate the effort and time investment needed to support this research endeavor. The same holds for Dr. David B. Konrad, who went above and beyond to build a framework that enabled this work. I am very thankful for your advice, support and help that you gifted over the last three years, David.

I am very thankful to Prof. Dr. Gisbert Schneider for accepting to act as the co-examiner of this thesis, and, hence, taking the time to evaluate the work. Thinking back to our first encounter, it is a great pleasure that our paths crossed again during my doctoral research and led to joint publications. Gisbert, I appreciate your invaluable scientific guidance and admire your positive spirit, which continuously motivates me to push boundaries. I am very much looking forward to joint future scientific endeavors.

Next, my heartfelt appreciation goes to my exceptional supervisors and mentors at Roche, Dr. Uwe Grether and Dr. Rainer E. Martin. Without your persistence, diplomacy, positiveness and courage, this research would not have surfaced. Over the last three years, you gave me the freedom to explore science, identify my strengths and weaknesses, delve into the business operations of a pharmaceutical company and above all, you always had my back. Both of you are role models for scientific leadership in the industry and beyond. I am grateful that we will be continuing to work together at Roche and very much looking forward to what lies ahead.

I would like to thank Prof. Dr. Franz Bracher, Prof. Dr. Silvija Markic, Prof. Dr. Franz Paintner and Prof. Dr. Oliver Trapp for devoting their expertise and time to evaluate this work by accepting to join the thesis evaluation committee.

Dr. Kenneth Atz is thanked for his invaluable contribution and scientific excellence to making this thesis successful, for being a fantastic collaborator and co-first author, for proof-reading and helping with the formatting of the thesis, for his suggestions and ideas, but most importantly, for the friendship we have established. Kenneth, I still remember the day when we panned out the plan for the first manuscript. What a ride it has been since then and I am confident that the best is yet to come. Let's see where 4A, 5K, 11A or 11K will take us next.

I am also very thankful for the support of Dr. Alex T. Müller, who has been instrumental in educating me in data science, enriching ideas, establishing collaborations and elevating manuscripts. Alex, I am very much looking forward to continuing working with you, which will hopefully also include many more morning runs at stunning places around the globe.

I like to thank Dr. Antonia F. Stepan for her trust, ideas and leadership. Antonia, without your continuous support, this project would not have become a reality, let alone a success.

Furthermore, I would like to appreciate the support of the Therapeutic Modalities, Small Molecule Research and Medicinal Chemistry leadership teams at Roche, especially Dr. Hayley Binch, Dr. Chris Claiborne, Dr. Christian Kramer, Dr. Sylke Poehling, Dr. Stefan Weigand and Dr. Thomas Woltering, for funding my thesis as well as supporting my education and personal development.

Jens Wolfard is thanked for supporting this research from early on. Jens, thank you for your dedication, your openness to new technologies and digital tools, your feedback, your ideas, the project support and for making the LCMS the powerful workhorse that it is today.

I would further like to thank the talented bachelor and master students that I had the privilege of supervising. Thank you Nadja Flückiger, Remo Hohler, Donatella Perrone, Yannick Stenzhorn, and Sebastian Strobel for your hard work.

Special thanks go to Dr. Georg Wuitschik for being an inspiration, mentor and astonishing scientist. Georg, thanks for supporting my research with your ideas, constructive criticism and scientific excellence. I would also like to express my heartfelt appreciation to Vera Jost for her relentless support of the project with the "Robi". Thanks for forgiving me for the initial mistakes, for always being positive and for our conversations, Vera.

This research would not have been possible without the support of colleagues in the analytics and purification departments, who helped to purify and characterize complex molecular structures. Thank you to Eric Bald, Christian Bartelmus, Martin Binder, Sophie Brogly, Philippe Cron, Markus Hohler, Philippe Jablonski, Stephane Kritter, Martin Kuratli, Dr. Inken Plitzko, Dr. Alfred Ross, Oliver Scheidegger, David Wechsler, Dr. Caroline Wyss Gramberg, and Daniel Zimmerli.

Moreover, I would like to acknowledge the support of current and former colleagues from various departments across Roche and Genentech, who contributed with their expertise and time to make this project a success: Dr. Jean-Michel Adam, Joel Aigner, Ann-Kathrin Baier, Dr. Björn Bartels, Jens Barthold, Sophia Baumgärtel, Dr. Stefanie Bendels, Meinrad Birrer, Dr. Serena Bisagni, Dr. Jeff Blaney, Dr. Alex Boddy, Oliver Braun, Virginia Brom, Dr. Paola Caramenti, Vanessa Casonato, Hazel Clarke, Dr. Jamie Clifton, Dr. Diana Darowski, Dr. Emilia Di Francesco, Dr. Luca Docci, Cosimo Dolente, Dr. Ruth Dorel, Dr. Peter Dragovich, Nina-Louisa Efrem, Andrea Eichelmann, Louise Enzendorfer, Dr. Martin Fitzner, Dr. Stephen Fowler, Cornelius Frank, Dr. Guido Galley, Olivier Gavelle, Dr. Laura Guasch, Dr. Wolfgang Haap, Urs Hanke, Dr. Steven Hanlon, Francis Hartmann, Dr. Satoshi Hashimoto, Dr. Dieter Heindl, Malcolm Huestis, Dr. James Hunter, Dr. Daniel Hunziker, Ramona Ibele, Isabelle Kaufmann, Alexander Knaupp, Dr. Oliver Korb, Danny Krumm, Dr. Hannes Kuchelmeister, Dr. Peter Kueng, Dr. Bernd Kuhn, Anke Kurt, Dr. Thomas Lübbers, Raymond Lieu, Pawel Linke, Dr. Kyle Mack, Andreas Marx, Dr. Irene Marzuoli, Colin Masui, Ines Mazurek, Dr. Matthias Nettekoven, Dr. Christian Neuhaus, Nam Nguyen, Huy Nguyen, Dr. Miroslav Nikolov, Dr. Ulrike Obst Sander, Dr. Fionn O'Hara, Kacper Jan Patej, Dr. Joe Pease, Dr. Diana Pippig, Dr. Giuseppe Prencipe, Bernd Puellmann, Dr. Kurt Puentener, Dr. Michael Reutlinger, Antonio Ricci, Dr. Joachim Rudolph, Valerie Runtz-Schmitt, Tali Rupp, Dr. Torsten Schindler, Philipp Schmid, Sebastien Schmitt, Dominic Schwarz, Tatjana Sela, Dr. Benjamin Sellers, Melanie Siebold, Beat Spinnler, Joel Strub, Dr. Jack Terrett,

Cruz Torrecilla Para, Dr. Andreas Tosstorff, Monique Ucar, Dr. Lorenz Urner, Walter Vifian, Björn Wagner, Dr. Richard Walroth, Susanne Weissenborn, Dr. Matthias Wittwer, Newton Wu, Dr. Daniel Zell, Dr. Jitao David Zhang, Dr. Nicolas Zorn, and Dr. Bill Zuercher.

I would like to thank the whole AK Konrad for always providing a warm welcome during my visits to LMU and for the excellent scientific discussions. Special thanks go to Anna Milton, Benedikt Nissl, Constantin Nuber, and Helene Schricker.

Rene Bergner, Thomas Klein, and Dr. Jens Viehweg are thanked for initially sparking my interest in chemistry and biology. I would also like to thank all the other teachers and peers during my time in Sankt Afra. You have prepared me to tackle the toughest challenges in science and beyond.

Finally, I would like to express my sincere gratitude to my family and friends for their continuous support.

Contents

SUMMARY	vii
KURZFASSUNG	xi
ACKNOWLEDGEMENTS	xv
PUBLICATIONS AND CONFERENCE CONTRIBUTIONS	xxi
ABBREVIATIONS	xxiii
1 INTRODUCTION	1
1.1 Drug discovery	1
1.2 Late-stage functionalization	6
1.2.1 Concept	6
1.2.2 LSF application in drug discovery	7
1.2.3 Reaction types	8
1.2.4 Challenges and opportunities	13
1.3 High-throughput experimentation	14
1.3.1 Background	14
1.3.2 Concept, requirements and advantages	17
1.3.3 Application scope	22
1.3.4 Challenges and opportunities	26
1.4 Reactivity prediction	34
1.4.1 Background	34
1.4.2 Molecular representation of chemical reactions	35
1.4.3 Chemical language models	36
1.4.4 Binary reaction outcome assessment	38
1.4.5 Reaction yield prediction	39
1.4.6 Regioselectivity forecasting	41
1.4.7 Graph neural networks for reactivity prediction	41
1.4.8 Reaction data availability	43
2 AIMS OF THE THESIS	45
3 SEMI-AUTOMATED LSF SCREENING PLATFORM	49

3.1	Approach and concept	49
3.2	Reaction screening	54
	3.2.1 Plate design	54
	3.2.2 Plate testing	60
	3.2.3 Screening workflow	61
	3.2.4 Data backbone	63
	3.2.5 Data interplay	66
	3.2.6 Visualization and analysis	70
3.3	Scale-up	74
3.4	Discussion	77
4	THE SIMPLE USER-FRIENDLY REACTION FORMAT (SURF)	79
5	LATE-STAGE DRUG DIVERSIFICATION THROUGH C-H BORYLATION	87
5.1	Introduction and background	88
5.2	Publication	91
5.3	Experimental and supplementary information	105
6	LATE-STAGE MINISCI-TYPE C-H ALKYLATION CHEMISTRY	171
6.1	Introduction and background	172
6.2	Publication	175
6.3	Experimental and supplementary information	187
7	CONCLUSION AND OUTLOOK	261
	BIBLIOGRAPHY	264

PUBLICATIONS AND CONFERENCE CONTRIBUTIONS

The following publications represent the core of the presented research:

- **Nippa, D. F.**, Hohler, R., Stepan, A. F., Grether, U., Konrad, D. B. & Martin, R. E., Late-stage Functionalization and its Impact on Modern Drug Discovery, *Chimia*, **76**, 258 (2022).
- **Nippa, D. F.**[†], Atz, K.[†], Hohler, R., Müller, A. T., Marx, A., Bartelmus, C., Wuitschik, G., Marzuoli, I., Jost, V., Wolfard, J., Binder, M., Stepan, A. F., Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning, *Nat. Chem.*, **16**, 2, 239-248 (2024).
- **Nippa, D. F.**[†], Atz, K.[†], Müller, A. T., Wolfard, J., Isert, C., Binder, M., Scheidegger, O., Stepan, A. F., Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Identifying opportunities for late-stage C-H alkylation with in silico reaction screening and high-throughput experimentation *Commun. Chem.*, **6**, 256 (2023).
- **Nippa, D. F.**[†], Müller, A. T.[†], Atz, K.[†], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Simple User-Friendly Reaction Format, *ChemRxiv* (2023), DOI: 10.26434/chemrxiv-2023-nfq7h.

[†] denotes equal contribution.

The following publications are related to the presented research, but are not covered in this dissertation:

- Atz, K., Cotos, L., Isert, C., [...], **Nippa, D. F.**, [...], Grether, U. & Schneider, G., Prospective deep interactome learning for *de novo* drug design, *Nat. Commun.* (2024), accepted.

The following contributions at conferences covering the core of the research were delivered:

25/02 - 01/02/2023, **SLAS 2023**, San Diego, United States

- ✧ **Nippa, D. F.**, Atz, K., [...], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning.
- ✧ **Nippa, D. F.**, Atz, K., [...], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning.

13/08 - 17/08/2023, **ACS Fall 2023**, San Francisco, United States

- ✧ **Nippa, D. F., Atz, K., Müller, A. T., [...], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G.**, Identifying Minisci-type alkylation opportunities with deep learning-based *in silico* reaction screening and high-throughput experimentation.
- ✧ **Nippa, D. F., Müller, A. T., Atz, K., Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G.**, SURF: Simple User-Friendly Reaction Format: A Proposal to Foster FAIR Chemical Reaction Data.
- ◆ **Nippa, D. F., Atz, K., Müller, A. T., [...], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G.**, Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning.

26/09 - 28/09/2023, **Swiss Chemical Society at ILMAC Conference**, Basel, Switzerland

- ✧ **Nippa, D. F., Atz, K., Müller, A. T., [...], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G.**, Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning: Efficient application of late-stage functionalization in Medicinal Chemistry.

27/11 - 28/11/2023, **Automated Synthesis Forum**, Basel, Switzerland

- ✧ **Nippa, D. F., Atz, K., Müller, A. T., [...], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G.**, TBD.

✧ denotes oral and ◆ poster contributions.

ABBREVIATIONS

2D	Two-Dimensional
3D	Three-Dimensional
ADME	Absorption, Distribution, Metabolism, Excretion
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
AE-MS	Acoustic Ejection-Mass Spectrometry
aGNN	Atomistic Graph Neural Network
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Active Pharmaceutical Ingredients
AUC	Area Under Receiver Operating Characteristic Curve
B₂pin₂	Bis(pinacolato)diboron
BERT	Bidirectional Encoder Representation Transformer
BBr₃	Boron tribromide
Boc	<i>tert</i> -Butyloxycarbonyl
Bpin	Boronic Acid Pinacol Ester
br	broad
byp1A	1-(2-([2,2'-bipyridin]-5-yl)phenyl)-3-cyclohexylurea
byp1B	5-(2-(4-(trifluoromethyl)-1,3,2-dioxaborolan-2-yl)phenyl)-2,2'-bipyridine
CAS	Chemical Abstracts Service
CASP	Computer-Aided Synthesis Planning
CGR	Condensed Graph of Reaction
CLM	Chemical Language Model
CPME	Cyclopentyl Methyl Ether
csv	comma-separated values
CyHex	Cyclohexane
d	doublet
DCM	Dichloromethane
dd	doublet of doublet
DECIMER	Deep lEarning for Chemical Image Recognition
DELT	DNA-encoded library technology
DESI	Desorption Electrospray Ionization
DFT	Density Functional Theory
DMSO	Dimethylsulfoxide
DMTA	Design-make-test-analyze
dt	doublet of triplet
DOLPHIN	Data-orchestrated laboratory platform harnessing innovative neural networks
dtbyp	4,4-Di- <i>tert</i> -butyl-2,2-dipyridyl
ECFP	Extended Connectivity Fingerprint
ELN	Electronic Lab Journal

FAIR	Findable, Accessible, Interoperable and Reusable
FIA	Flow Injection Analysis
GC	Gas Chromatography
GCMS	Gas-Chromatography Mass Spectrometry
GNN	Graph Neural Network
GSK	GlaxoSmithKline
GTNN	Graph Transformer Neural Network
HAT	Hydrogen Atom Transfer
HBpin	Pinacolborane
HCOOH	Formic Acid
HMPA	Hexamethylphosphoramide
HPLC	High-Performance Liquid Chromatography
HRMS	High-Resolution Mass Spectrometry
HTE	High-Throughput Experimentation
HTS	High-Throughput Screening
ID	Identification
IP	Intellectual Property
IPC	Internal Process Control
IT	Information Technology
[Ir(COD)Cl]₂	(1,5-Cyclooctadiene)(methoxy)iridium(I)dimer
[Ir(COD)OMe]₂	Bis(1,5-cyclooctadiene)diiridium(I)dichloride
LCMS	Liquid Chromatography–Mass Spectrometry
LLM	Large Language Model
LO	Lead optimization
LSF	Late-Stage Functionalization
LSTM	Long-Short-Term Memory
m	multiplet
MAE	Mean Absolute Error
MALDI	Matrix-Assisted Laser Desorption/Ionization
MeCN	Acetonitrile
Me-THF	2-Methyltetrahydrofuran
MISER	Multiple Injections in a Single Experimental Run
ML	Machine Learning
MLP	Multi-Layer Perceptron
MS	Mass Spectrometry
MSD	Merck Sharp & Dohme Corporation
MSE	Mean Squared Error
NCE	New Chemical Entity
NME	New Molecular Entity
NMP	<i>N</i> -Methyl-2-pyrrolidone
NMR	Nuclear Magnetic Resonance
NN	Neural Network
(NH₄)₂S₂O₈	Ammonium persulfate
OFAT	One-Factor-At-a-Time
ORD	Open Reaction Database
PC	Principal Component
PCA	Principal Component Analysis
Pd(OAc)₂	Palladium(II) acetate
PDF	Portable Data Format

PET	Positron Emission Tomography
phen	1,10-Phenanthroline
PPV	Positive Predictive Value
PROTAC	Proteolysis-Targeting Chimaeras
PTC	Phase Transfer Catalysis
QM	Quantum Mechanics
QSAR	Quantitative Structure-Activity Relationship
R&D	Research & development
RNN	Recurrent Neural Networks
ROI	Return of investment
RP-HPLC	Reversed-Phase High-Pressure Liquid Chromatography
rpt	Report
s	singlet
SACT	Systematic Analysis of Chemical Transformations
SAR	Structure-Activity Relationship
SD	Structure Data
SELFIES	Self-Referencing Embedded Strings
SFC	Supercritical Fluid Chromatography
SI	Supplementary Information
SIM	Selected Ion Monitoring
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line Entry System
SURF	Simple User-Friendly Reaction Format
t	triplet
td	triplet of doublet
THF	Tetrahydrofuran
tmphen	3,4,7,8-Tetramethyl-1,10-phenanthroline
TPR	True Positive Rate
tsv	tab-separated values
UDM	Unified Data Model
UPLC	Ultra-High-Performance Liquid Chromatography
USD	U.S. Dollar
UV	Ultra-Violet
XGBoost	eXtreme Gradient Boosting

Passion is the fuel that ignites the fire within.

- Jan Frodeno

1

INTRODUCTION

1.1 Drug discovery

The discovery and development of new pharmaceuticals is an intricate and multifaceted challenge that persists despite considerable advancements in the understanding of disease mechanisms and breakthroughs in technological capabilities. The early pre-clinical drug discovery process consists of a series of stages classically starting from target identification through lead identification and lead optimization (LO) to candidate selection, which underlines the complexity of developing an efficient drug compound with suitable pharmacokinetic parameters and a strong safety profile. [1] Demonstrating both, safety and efficacy in humans, during clinical trials, which resemble the next step of the pharmaceutical development process, plays a vital role. Despite multiple iterative design cycles and precise optimization of properties, a high proportion of potential therapies still fail in clinical studies generating high costs without delivering any return on investment (ROI). [2] This decline in research & development (R&D) efficiency, often related to pre-clinical assumptions lacking human validation leading to late-stage failures, has inflated the overall cost of drug development. [3]

Over the years, several studies have analyzed the development of R&D efficiency across the global pharmaceutical industry. [4–9] In line with their conclusions, recent research revealed that the overall average R&D cost to develop a compound from discovery to launch has approximately doubled over the last decade. [10] While on average 1.3 billion USD were required to get a compound to market in 2013, by 2022 the drug discovery process demanded over 2.2 billion USD (Figure 1.1). Unsurprisingly, the average expected ROI dropped from 6.5 per cent to 1.2 per cent in the same period. Lower costs and higher ROIs reported during the pandemic in 2021, mainly due to the fast approvals of vaccinations and treatments, have to be seen as outliers due to unexpected external circumstances and do not brighten up the overall outlook for the pharmaceutical industry.

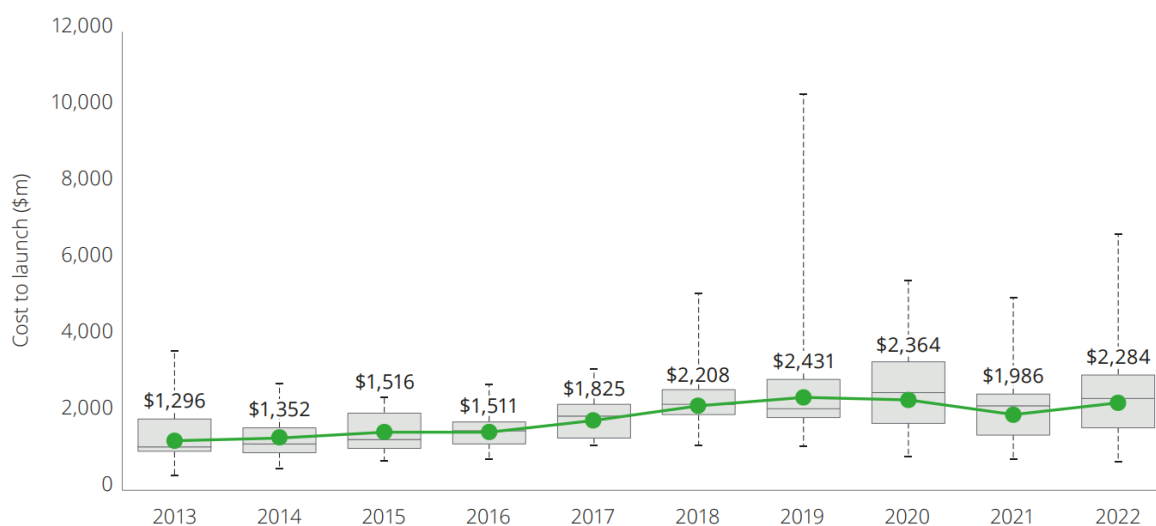


Figure 1.1: Average R&D cost to develop a compound from discovery to launch, 2013-2022. Data and figure derived from Terry & Lesser (2023). [10]

The discrepancy between investment and output has not only failed to meet investor expectations but has also raised questions about the sustainability and speed of the current R&D model to address unmet medical needs across the globe, prompting the industry to deliberately focus on efficiency improvements. [7, 8] Benchmarks from the industry have shown that not only the clinical trials, which represent the core of the development phase of a potential drug, are expensive and possess long cycle times, but that the LO phase in discovery also requires substantial resource allocation (Figure 1.2). [6] More importantly, this phase is responsible for defining the molecular structure of the future pharmaceutical industry, thereby largely determining pharmacokinetics and pharmacodynamics. [11] Further, during this part of the drug discovery process, a large amount of intellectual property (IP) is generated, which is paramount to the pharmaceutical industry's business model. [12, 13] Therefore, identifying

ways to accelerate the LO phase and improving the outcomes of this stage of drug discovery to generate high-quality and safe drug candidates have been a major focus of researchers in academia and industry. [14–17]

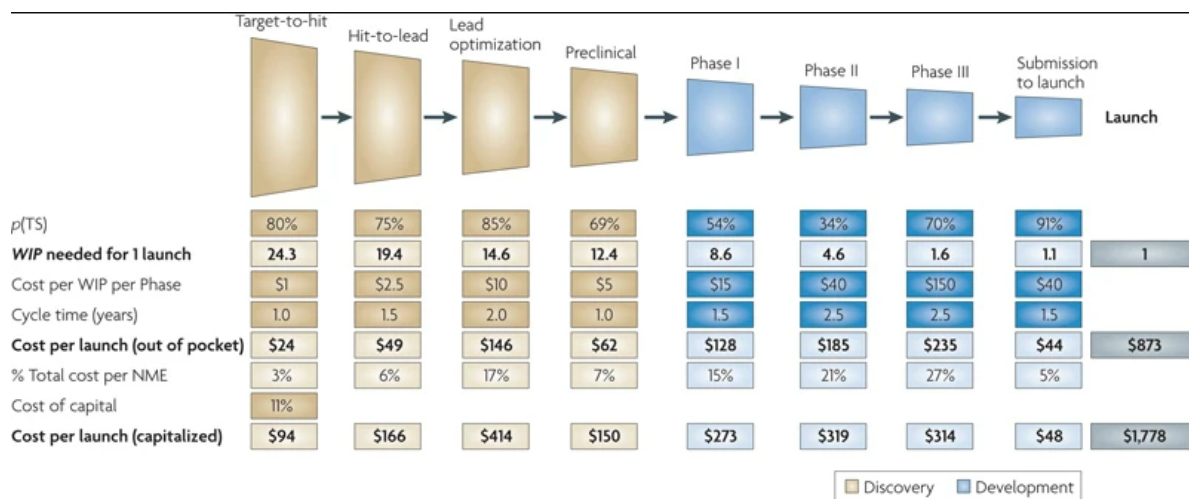


Figure 1.2: Distinct phases of drug discovery and development from the initial stage of target identification to launch. The model is based on a set of industry-appropriate R&D assumptions (industry benchmarks and data from Eli Lilly and Company) defining the performance of the R&D process at each stage of development. R&D parameters include the probability of successful transition from one stage to the next ($p(TS)$), the phase cost for each project, the cycle time required to progress through each stage of development and the cost of capital, reflecting the returns required by shareholders to use their money during the lengthy R&D process. Abbreviations: WIP: Work in Process, NME: New Molecular Entity. Data and figure derived from Paul *et al.* (2010). [6]

LO is the cornerstone of medicinal chemistry research and, in general, follows the concept of the design-make-test-analyze (DMTA) cycle (Figure 1.3). In this iterative process, the obtained lead structure from hit identification is altered through chemical modifications to identify molecules with maximal therapeutic efficacy and minimal undesired effects. [18–20] In the first step, based on biological data, new ligand ideas are generated that could enhance the pharmacological properties of the lead compound (*i.e.*, design). Next, based on the ranking of the designed molecules they are chemically synthesized in the laboratory, purified and characterized (make). The compounds are then undergoing biological testing in a variety of assays to identify their pharmacological and physicochemical properties (test). In the final stage of the DMTA cycle, the obtained experimental data is analyzed and interpreted to aid the next round of the process (*i.e.*, analyze). [21] Identifying ways to make the cycles faster and cheaper by advancing the different phases with technology, thereby contributing to reducing the required time and costs to get a drug from concept to the market, has been widely discussed. [2, 22–24]

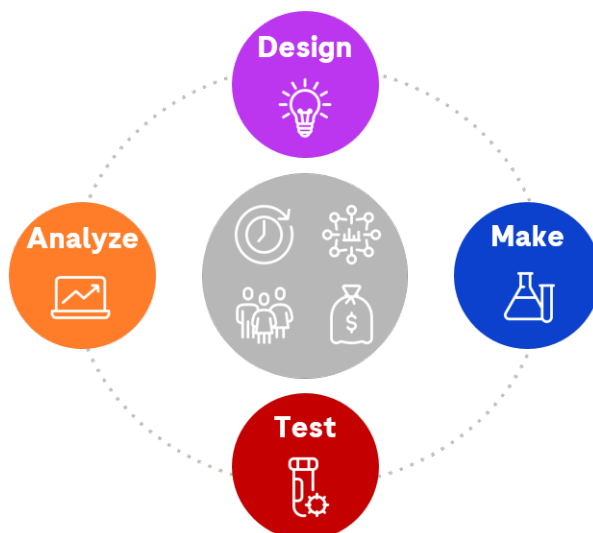


Figure 1.3: The design-make-test-analyze (DMTA) cycle. In an iterative process, new molecules are conceptualized (design) and synthesized in the laboratory (make). Purified chemical matter is subjected to biological assays (test) and the data is interpreted (analyze), delivering the foundation for the next design cycle. In addition to time, DMTA loops require resources in terms of manpower, material and machines to conduct the scientific experiments. Importantly, the generated experimental data needs to be structured systematically, easily accessible and analyzable to make informed decisions.

The accumulation of compound profiling data and the advances in computing power have paved the way for several computational methods that support and accelerate the design process of new chemical entities (NCEs). Quantitative structure–activity relationships (QSARs) determination, [25, 26] the prediction of physicochemical properties, [27–29] virtual screening, [30, 31] binding affinity predictions, [32, 33], and free energy perturbation calculations [34, 35] are some of many methodologies that belong to this computational drug discovery toolbox. Consequently, to date, a plethora of molecule ideas can be generated within short time frames requiring a limited amount of resources. [22] Based on the technological innovation happening for and in high throughput screening (HTS), automated testing capabilities for the determination of drug metabolism and pharmacokinetic properties were established. [36, 37]. Those capabilities reduced the turnaround times within the "test" phase to a large extent, allowing for screening more molecules in a shorter time while broadening the testing scope in parallel. [22] Through the faster generation of experimental data, storage and rapid analysis infrastructures were needed, leading to significant investments of the pharmaceutical industry into computational support of the discovery departments. [38–40] Specifically, the access to and visualization of large data sets plays a critical role in advancing drug discovery programs by validating hypotheses and enabling informed decision-making. [41]

Further, machine learning (ML) has become an invaluable tool for analyzing large amounts of data, summarizing findings and predicting a broad range of compound properties, thereby closing the loop to the design stage. [42–46]

Despite these advances, the duration of a DMTA cycle remains lengthy, often exceeding 4–8 weeks until one loop is closed. As the identification of a clinical candidate generally demands several iterations of the cycle, the LO phase requires a substantial investment into resources, materials and long timelines. [2] As discussed above, the ideation of novel molecules (design), the running of biological assays of different types, (*e.g.*, potency and adsorption, distribution, metabolism, excretion and toxicity (ADMET) (test), and the analysis of large data amounts (analysis) has already been expedited. The "make" phase, however, is more difficult to optimize and slows down the whole cycle to a large extent since the multi-step organic synthesis of new complex chemical matter typically requires weeks. [1] Thus, reducing the length of the "make" step could considerably decrease the overall resource and time demand for a single DMTA cycle, consequently contributing to the accelerated development of an optimized drug candidate. The utilization of laboratory automation equipment to speed up reaction set-up, analytics and purification is starting to become increasingly important for organic synthesis. [47–50] However, these automated systems are not the industry standard and, consequently, even common and simple reaction types are still carried out manually in classic fashion. [51, 52] As a consequence, the synthesis of structurally novel, complex drug molecules and their analogs to build structure-activity relationships (SAR) often remains the rate-limiting step in medicinal chemistry. [53]

To overcome this bottleneck and accelerate rapid SAR exploration to efficiently improve the pharmacological activity and physicochemical properties of lead structures in the DMTA cycle, it is paramount to make synthetic strategies more efficient. The late-stage functionalization (LSF) of complex scaffolds, where abundant C–H bonds serve as a starting point for the incorporation of functional groups to aid derivatization, has emerged as a powerful and step-economical approach as it bypasses the necessity for *de novo* synthesis or the incorporation of specific functional handles. [54] Therefore, the methodology can contribute to efficiently exploring closely related chemical space of the parent structure to potentially serve as an accelerator in the LO phase of drug discovery programs. [55]

1.2 Late-stage functionalization

The foundation of this chapter was published in: **Nippa, D. F.**, Hohler, R., Stepan, A. F., Grether, U., Konrad, D. B. & Martin, R. E., Late-stage Functionalization and its Impact on Modern Drug Discovery, *Chimia*, **76**, 258 (2022). [55] Figure 1.4 is reprinted from the original manuscript.

Author contributions: Conceptualization of the article, data search and analysis, figure preparation and manuscript writing.

1.2.1 Concept

The activation of C–H bonds to enable chemical transformations that generate direct analogs from complex molecular structures, such as advanced drug-like compounds or natural products, is seen as an important tool in synthetic organic chemistry. [56–59] LSF, in particular, describes the chemoselective direct substitution of C–H bonds by other functional groups in a single step without necessitating any pre-functionalization on structurally intricate molecules. [60] This offers the opportunity to achieve higher efficiency during the diversification of lead structures compared to the *de novo* synthesis of new analogs (Figure 1.4). [61] Despite the progress in methodology development in academia, the transferability of the methodologies for utilization in drug discovery campaigns remains challenging. [62] Nevertheless, the research in the field delivered some novel chemo-selective protocols with functional group compatibility that allow for wider applicability of LSF, which have been discussed in several comprehensive reviews. [54, 63–67]

Apart from avoiding molecule decomposition and maintaining functional group tolerance, the main difficulty of applying the LSF approach efficiently for drug discovery problems lies within the challenge of achieving site-selectivity among the numerous C–H bonds with similar bond dissociation energies in a complex organic molecule. [66] To circumvent these selectivity challenges, two principal strategies have been predominantly employed: innate and directed C–H functionalizations. [68] Innate C–H functionalization involves the substitution of a C–H bond by a new functional group at the most reactive site of the compound, determined by the intrinsic reactivity of the molecule itself. This reactivity is influenced by several factors, including bond dissociation energies, steric hindrance, electronic influences and kinetic acidity. The innate functionalization can be steered through variation of the reaction conditions leading to complementary functionalization at alternate C–H sites within one molecule. [69] In contrast, directed C–H functionalization leverages Lewis basic functionalities to selectively position the

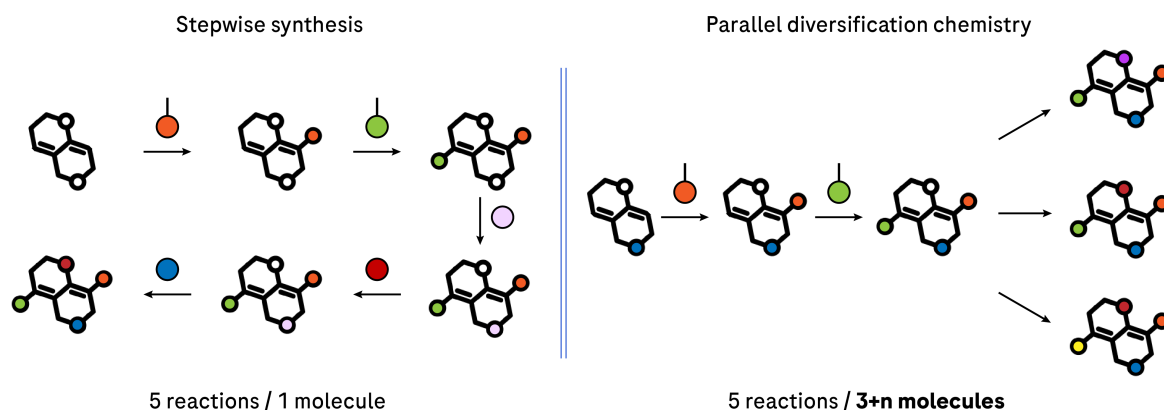


Figure 1.4: Stepwise synthesis and parallel diversification chemistry. Classic stepwise synthesis (left) represents the predominant strategy in medicinal chemistry for the construction of novel chemical entities from commercially available precursors. The assembly of the desired structure necessitates multiple, consecutive chemical reactions, with each derivative requiring a distinct synthetic route. In contrast, the parallel diversification approach (right) involves the rapid construction of a core structure, which is subsequently diversified through a variety of techniques, obviating the requirement for *de novo* synthesis of each analog. This approach facilitates the rapid generation of a multitude of new chemical entities (NCEs), potentially with an equal or reduced number of synthesis steps conducted in the laboratory. Consequently, an increase in the overall efficiency of compound synthesis can be achieved, which also aligns with the objectives of sustainable chemistry.

catalyst near a targeted C–H bond, facilitating C–H bond cleavage through a chelation effect. Typical directing groups include heterocycles, amides, and carboxylic acids, functionalities that also frequently appear in drug-like structures. [68] The distinction between the two LSF concepts is not always possible, as multiple factors contribute to the site selectivity observed. The full potential of both strategies is generated through their synergistic application to allow for strategic and efficient LSF of complex molecular structures. In general, the LSF concept is in alignment with the shift towards more sustainable chemical methods, which aim to work less toxic, with better atom economy and more cost-effective. [69]

Consequently, the LSF approach has been considered for drug discovery campaigns, especially within the LO phase, and the main application scope is described in the following section.

1.2.2 LSF application in drug discovery

LSF has emerged as a pivotal technique in drug discovery, addressing several critical aspects of the development process and the four main applications are shown in Figure 1.5. It enables the early identification and characterization of metabolites, thereby revealing potential safety risks such as toxicity, which is essential for assessing the viability of lead compounds. [70] By facilitating the introduction of functional groups, LSF allows for the precise modulation of ADMET properties, including solubility, permeability, stability, half-life, bioavailability, and

distribution, thereby optimizing the pharmacokinetic profile of drug candidates. [53, 64]

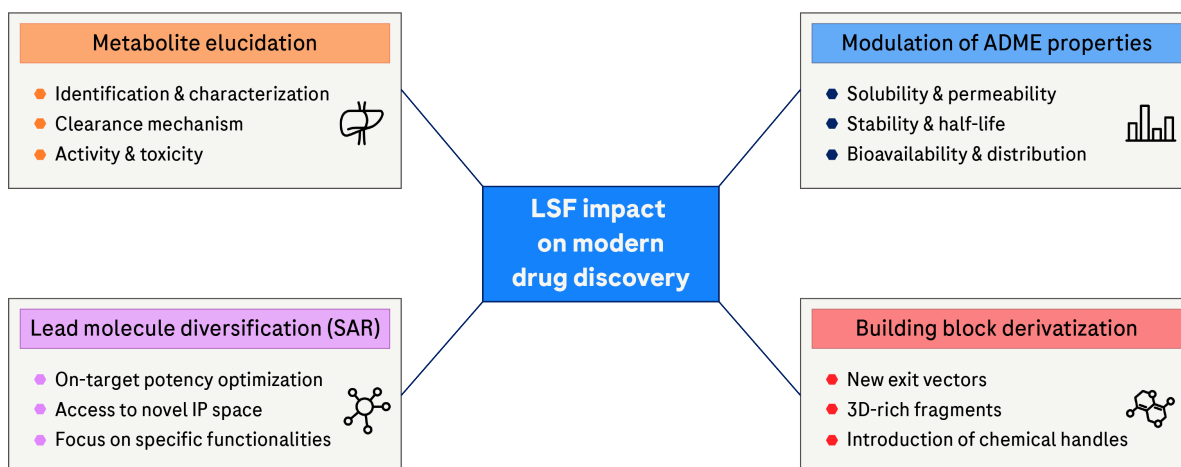


Figure 1.5: The impact of late-stage functionalization (LSF) on modern drug discovery. LSF enables access to metabolites, allowing the identification and characterization of potential safety risks, including toxicity, at an early stage in the discovery process (t.l.). Further, the introduction of functional groups allows the specific modulation of adsorption, distribution, metabolism and excretion (ADME) properties efficiently (t.r.). Those properties include solubility, permeability, stability, half-life, bioavailability and distribution, among others. LSF also contributes to the exploration of structure-activity relationships (SARs) by generating close analogues to the parent molecule without requiring *de novo* synthesis of each compound (b.l.). This facilitates access to novel intellectual property (IP) space and on-target potency optimization. LSF can also be seen as a tool to derivatize advanced intermediates or building blocks, which can help to introduce new exit vectors or three-dimensional-rich fragments (b.r.).

Moreover, LSF is instrumental in the exploration of SARs, as it permits the generation of analogues closely related to the parent molecule, thus bypassing the need for their complete *de novo* synthesis and accelerating the hit-to-lead and LO stages. [66, 71] This approach not only expedites the optimization of on-target potency but also opens new avenues for IP space. Additionally, LSF serves as a strategic method for the derivatization of advanced intermediates or building blocks, introducing new exit vectors or three-dimensional fragments that can significantly enhance the structural diversity and complexity of lead molecules. [61, 62]

Over the past decade, the number of methodology publications in the LSF field has witnessed a strong increase across various reaction types and a short overview with selected examples will be given in the following subchapter.

1.2.3 Reaction types

The results from a comprehensive literature search conducted in Scopus, Web of Science and SciFinder[®] are depicted in Figure 1.6. The analysis of the literature search revealed that the LSF toolbox of today encompasses a large number of chemical and enzymatic methods such

as fluorination, [72] amination, [73] hydroxylation [70] and methylation. [74] Over the last decade, tremendous progress towards providing reaction conditions for the functionalization of almost any (C-sp²)-H or (C-sp³)-H bond without the need of installing a synthetic handle was made. In the following, a short overview highlighting selected LSF reactions and their application to drug discovery projects using representative examples will be given.

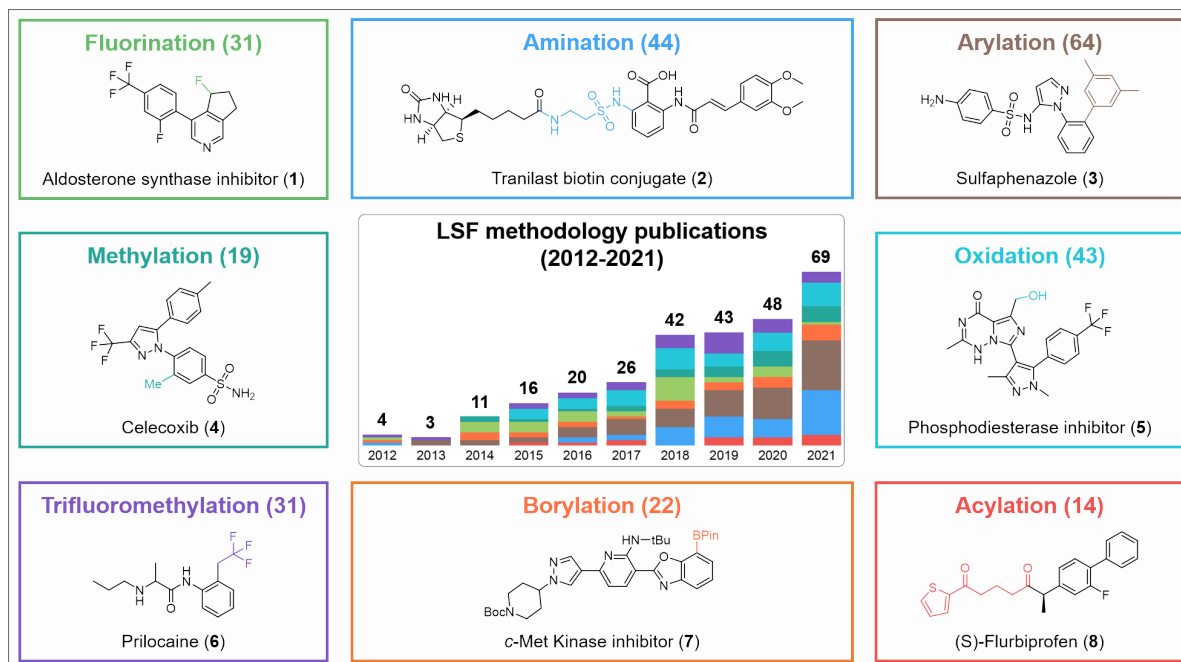


Figure 1.6: Overview of the LSF literature landscape (2012-2021). Increase in publications of the most common LSF methodologies from 2012 to 2021 (center). Selected applications of LSF fluorination, amination, arylation, methylation, oxidation, trifluoromethylation, borylation and acylation found in the literature. The number in brackets next to the methodology name states the publication count as of 2021. Published with permission from [55].

The substantial number of studies (31) addressing fluorination techniques and their elevated mean citation count (60) underscore the significance of C–F bond formation in the realm of drug discovery (Figure 1.6). Late-stage fluorination strategies are instrumental in enhancing the metabolic stability of labile C–H bonds and in augmenting protein–ligand interactions. The Britton group has established protocols that facilitate the selective fluorination of pyridinic C–H bonds, even in the presence of benzylic groups. This protocol was effectively employed in the modification of an aldosterone synthase inhibitor **1**. [72] Additionally, the same team has illustrated that ¹⁸F, a radionuclide frequently utilized in positron emission tomography (PET) imaging agents, can be incorporated in a site-specific manner into the methine position of leucine residues within unprotected peptides. [75] Several other LSF methodologies that aid the introduction of ¹⁸F were disclosed as well, including a two-step site-selective approach via aryl sulfonium salts, [76] a procedure on benzylic procedure with manganese, [77] and

a copper-mediated transformation of electron-rich heteroarenes [78]. Roque *et al.* presented a methodology employing deconstructive fluorination for the synthesis of mono- and difluorinated amine derivatives, achieved through the cleavage of C(sp³)-C(sp³) bonds within saturated nitrogen-containing heterocycles, including piperidines and pyrrolidines. [79]

Amines are extensively utilized in medicinal chemistry, serving various purposes such as enhancing target interactions or improving drug solubility. Additionally, they act as versatile intermediates for subsequent chemical transformations and provide pivotal points of attachment for conjugation with entities such as chemical biology tools and antibodies. Given their broad range of applications, it is unsurprising that, as of 2021, late-stage amination techniques are classified as the second most prevalent category of reactions in LSF. Exploiting their novel conditions, Weis *et al.* showcased the utility of late-stage amination through the successful synthesis of a Tranilast biotin conjugate **2**, a compound with the potential to be applied as a chemical biology probe. [73] Further, Wan *et al.* disclosed the development of a novel flow photoreactor that facilitates regioselective and scalable C(sp³)-H aminations through decatungstate photocatalysis. [80] Other examples cover the γ -selective C(sp³)-H amination of unactivated alkenes with varying alkyl chain lengths [81] or the light-driven C-H amination applying iron porphyrin catalysts [82].

Aromatic systems often enhance the potency of pharmaceutical agents through non-covalent interactions with proteins. Although cross-coupling reactions are the primary synthetic means to incorporate aryl groups, there is a growing need for more economical and efficient synthetic approaches, as evidenced by the plethora of late-stage arylation methods reported in the literature. Within this context, Simonetti *et al.* introduced a strategy that enables the arylation of various commercial pharmaceuticals, such as Sulfaphenazole (**3**). [83] In addition, a site-selective late-stage C(sp³)-H arylation of peptides utilizing the native side chain of asparagine, thereby enabling the synthesis of di-, tri-, and tetrapeptides without the need for external directing groups was published. [84] Further work on peptides, delivered a versatile, late-stage palladium-catalyzed C(sp³)-H arylation of peptides using thiazole motifs as internal directing groups, enabling regio- and site-selective functionalization of peptide side chains to facilitate diverse, bioactive peptidomimetic libraries with thiazole-modified backbones. [85]

The "magic methyl effect," a term coined within medicinal chemistry, refers to the significant enhancement of pharmacological attributes such as potency, metabolic stability, and reducing cytochrome P450 inhibition, achieved by the mere addition of a methyl group to a lead com-

pound. [86] Consequently, the development of efficient late-stage methylation techniques has garnered considerable interest among researchers specializing in LSF. A notable contribution to this field is the cobalt-catalyzed C-H methylation discovered by Friis *et al.*, which has been successfully applied to Celecoxib (**4**) and other small molecule drugs. [74] Moreover, a regioselective and chemoselective oxidative C(sp³)-H methylation method utilizing a manganese catalyst at low loadings enabled the use of a mildly nucleophilic organoaluminium methylating agent to modify a set of different heterocyclic cores and drug-like compounds. [87] Another report highlights the selective methylation of arenes employing a non-chelation-assisted approach that integrates C-H functionalization with nickel-catalyzed cross-coupling, facilitating the access to modified arenes with high selectivity for monosubstituted examples. [88]

Oxidative modification of molecules can exert a comparable influence. Beyond solubility enhancement, late-stage oxidation processes facilitate early identification, characterization, and evaluation of prospective metabolites during the initial phases of drug development. For transformations demanding high regio- and stereoselectivity, enzymatic oxidations at the late-stage have demonstrated their efficacy, often outperforming chemical methodologies. An example of this is the enzyme-mediated synthesis of the phosphodiesterase inhibitor **5**, where the addition of a hydroxyl group improved the overall compound parameters compared to the parent molecule, leading to its selection as a potential clinical candidate based on its positive toxicological profile in rat and dog studies. [70] Another team highlighted the use of cytochrome P450 monooxygenases for the chemo-, regio-, and stereoselective oxidation of β -cembrenediol, a tobacco cembranoid with multiple potential oxidation sites, achieving high regio- and diastereoselectivity through first-sphere active site mutagenesis and minimal library screening of P450 BM3 variants. [89] Chemical oxidation methods were published as well, including an iron-catalyzed, undirected arene C-H hydroxylation using hydrogen peroxide encompassing broad substrate scope, high selectivity, functional group compatibility and good yields, [90] or an aerobic oxidation of benzylic C(sp³)-H bonds to ketones under continuous-flow conditions was achieved using *N*-hydroxyphthalimide and *tert*-butyl nitrite catalysts, with high catalyst and solvent recyclability [91].

Trifluoromethylation facilitates the synthesis of compounds with enhanced metabolic stability, augmented permeability, or superior potency. The MacMillan group demonstrated this through the photocatalytic trifluoromethylation of Prilocaine (**6**) and a curated assortment of other structurally intricate molecules. [92] Another method used a bench-stable copper(III) complex, bpyCu(CF₃), in a mild, operationally straightforward C(sp³)-H trifluoromethylation

of unactivated alkanes, involving the visible-light photoinduced generation of a CF₃ radical and anion. [93] An iron(II)-catalyzed trifluoromethylation of enamides under mild conditions employing cost-effective Togni's reagent for electrophilic CF₃ introduction, achieved good regioselectivity, broad substrate scope, and showed functional group tolerance. [94]

Despite the limited role of boron in pharmaceuticals, its regioselective incorporation into advanced drug-like entities as a synthetic handle can be a beneficial strategy to enable late-stage diversification, offering extensive synthetic options for post-borylation modifications and, thereby offering the exploration of SARs through the addition of various functional groups. The significance of this methodology is underscored by the high citation average (46 citations) of such studies. One of them is the efficient borylation of c-Met kinase inhibitor **7** published by the Hartwig lab. [95] Other examples include the C7 borylation of indole units within *Aspidosperma* alkaloids for late-stage synthesis facilitating the conversion of a polycyclic lactam to Vallesine [96] and undirected C-H borylation that generated five distinct borylated analogues of the kinase inhibitor Staurosporine, which were separated upon oxidation to phenols [97]. Outside the iridium-catalyzed borylation reaction space, Kim *et al.* developed an azine method employing stable, cost-effective amine-borane reagents and photocatalysis to generate boryl radicals for Minisci-style radical addition, enabling predictable, site-selective carbon–boron bond formation. [98]

α -Ketones have been identified as valuable intermediates for drug discovery, a concept recently exemplified by Huan *et al.* through their development of an asymmetric benzylic acylation process for the generation of α -aryl ketones from carboxylic acids. [99] This methodology was successfully applied to the functionalization of the therapeutically significant molecule Flurbiprofen (**8**), as illustrated in Figure 1.6. Segundo and Correa reported a Pd-catalyzed C–H acylation of Tyrosin-containing peptides with aldehydes. [100] The water-compatible, site-specific and scalable tagging method with full tolerance of sensitive functional groups offers direct routes to oligopeptides with varied side-chain topologies, including endomorphin-2 and neuromedin N mimetics. Another example describes the merger of transition metal and photoredox catalysis to enable the direct enantioselective acylation of α -amino C(sp³)–H bonds with carboxylic acids to offer a novel approach to stereocontrol in metal-photoredox catalysis. [101]

Even though some LSF examples on selected large, drug-like molecules can be found in the literature, broader application scopes have not been established. Consequently, despite the

plethora of LSF methodologies reported, numerous obstacles persist that must be surmounted to enable the routine and effective integration of this synthetic technology into various drug discovery campaigns.

1.2.4 Challenges and opportunities

Expectations are high that LSF will imminently establish itself as a standard technology for the efficient generation of bioactive compounds without requiring *de novo* synthesis. However, the vast array of complex pharmaceutical compounds, encompassing small molecules, proteolysis-targeting chimaeras (PROTACs), peptides, nucleic acid-based therapeutics, antibody-drug conjugates, and monoclonal antibodies, represents a substantial structural complexity. These different modalities frequently possess sterically and electronically similar C–H bonds, requiring a nuanced comprehension of the reactivity of and accessibility to specific reactive centers within diverse reaction contexts. [61, 66, 69] Enhancing the compatibility of LSF with common functional groups found in bioactive molecules, such as polar and basic moieties or heterocycles, is essential for its widespread adoption. This will involve the refinement of reaction conditions to circumvent harsh temperatures, strong acids, and potent oxidizing agents that may trigger deleterious side reactions. In general, LSF suffers from a lack of predictability and selectivity of the reaction sites mainly driven by similar C-H bond dissociation energies in complex molecules.

The resurgence of photoredox, radical, and electrochemical methodologies promises to augment current C–H activation strategies with innovative approaches. [61, 64] Furthermore, the strategic use of biocatalysis, mirroring nature-like selective modification of multifunctional molecules, holds promise for addressing these synthetic challenges. [102] In the realm of diversity-oriented synthesis, the exploration of new reactions that introduce transient functional groups, such as those containing boron or phosphorus, will enhance the repertoire of LSF techniques. [60, 103]

Low yields and the complexity of LSF transformations have restricted their application in process chemistry and scale-up operations. To overcome this barrier, catalytic systems must be refined to decrease catalyst loadings, enhance turnover frequencies, and ensure compatibility with aqueous media, while also replacing toxic transition metal complexes to foster more sustainable and cost-effective processes. [54, 104]. The integration of flow chemistry is poised to play a pivotal role, particularly in the scale-up of photochemical and electrochemical

reactions. [66] Emerging technologies that streamline the purification and characterization of reaction products, as well as the assessment of pharmacological activity, will significantly bolster the impact of the field.

Importantly, LSF can potentially be enabled by the latest developments in laboratory automation, specifically high-throughput experimentation (HTE), which enables the rapid optimization of reaction conditions with minimal material use. [61, 66] Using the generated data from such HTE campaigns and combining them with big data analysis and ML methods is anticipated to yield predictive models for the reactivity and selectivity of C–H bond transformations, facilitating the synthesis of target molecules in an environmentally conscious and material-efficient manner. [66]

1.3 High-throughput experimentation

1.3.1 Background

Established during the 1980s to enhance the efficacy of drug discovery and optimization processes, high-throughput methodologies incorporated an array of experimental approaches that enabled the swift parallel synthesis of thousands of compounds, utilizing broad reaction conditions to generate libraries of structurally related organic molecules. [105–107] In contrast, HTE emerged as an instrumental approach to expedite the identification and application of novel, efficient methodologies.

Whereas parallel synthesis applies existing reaction protocols to generate novel compounds, HTE focuses on the quick determination of optimal conditions for specific chemical transformations, particularly with substrates that are unique or pose significant challenges. In brief, HTE allows researchers to (i) swiftly refine existing methodologies, (ii) aid in the innovation of new reaction conditions, and (iii) reveal unexpected lead compounds that might be overlooked with conventional methods due to the impracticality of conducting numerous individual experiments. [108–112]

The initially developed laboratory automation systems for combinatorial work, which were able to robotically handle liquid chemicals, use multi-well plates or micro vials as experimental containers to rapidly analyze samples and were first applied for HTE approaches in the mid-1990s to early 2000s. [113] Mainly materials science, solid-supported synthesis, as well as, both heterogeneous and homogeneous catalysis benefited from these advances at that time. [114–119]

Pioneers at Merck Sharp & Dohme Corporation (MSD) started utilizing HTE for targeted applications in the pharmaceutical industry at the beginning of the 2000s. Their initial work focused on the optimization of asymmetric hydrogenation by assessing a variety of chiral phosphine ligands with noble-metal precursors using a parallel screening set-up. [120] Due to the success of the methodology, which involved treating pre-dispensed 96-vial ligand libraries in glass vials with stock solutions of catalysts, the MSD team aimed at expanding the scope to more reaction types, even though this necessitated overcoming several new engineering challenges. [110]

Transitioning to Pd- and Cu-catalyzed cross-coupling reactions required workflows for handling heterogeneous reaction mixtures with magnetic tumble stirrers, retaining volatile solvents during heating through sealing mats, and dosing bases as solids with a custom-made robot rather than using slurries with a consequent tedious evaporation step. [111] After numerous iterations validating the system with literature reactions, the first successful applications of the platform on projects included the optimization of a Kumada coupling for the synthesis of Vaniprevir, and the development of an efficient, regioselective copper-catalyzed indazole coupling for the synthesis of Niraparib. [121, 122] Despite the available automation equipment for ligand and base handling, back then manual pipetting of catalyst precursors and substrates was preferred due to its flexibility and low cost. [111]

Over time, MSD and others expanded these semi-automated HTE workflows to additional reaction types, *e.g.*, chiral phase transfer catalysis (PTC) or photoredox bond forming transformations to solve difficult synthetic problems in process development (More details and selected examples on HTE applications are described in Chapter 1.3.3). [123–130] For specific transformations, where the set-up and execution protocols did not require extensive engineering, fully automated solutions with robotic platforms were developed. [111, 131] Importantly, with increased experimental throughput, the development of innovative, automated analytical technologies, especially improved high-performance liquid chromatography (HPLC) and

supercritical fluid chromatography (SFC) methods, was required for rapid and reliable assessment of reaction outcomes. [132–134] In parallel, obtained analytical raw data processing and analyzing, either manually or automated, to identify reaction outcomes, became another important cornerstone of HTE operations. [135]

While originally started in process chemistry, HTE soon also expanded into medicinal chemistry, where it serves the purpose of generating diverse chemical entities for biological profiling. [136] Parallel screening became of high value in discovery chemistry as it helped to overcome synthetic biases by enabling the efficient synthesis of complex target molecules, which are often structurally different from simple model substrates typically used in catalytic method development. [137, 138] To align with the fast-paced and material-conserving nature of drug discovery, MSD developed reaction-specific HTE kits, which are pre-assembled 24-vial arrays containing the best catalysts and conditions derived from literature and internal data, thus facilitating rapid, resource-efficient experimentation. A move that was followed by vendors, including Sigma-Aldrich, which commercialized the solution for some re-occurring transformations, including Suzuki–Miyaura and Buchwald–Hartwig couplings. [111] Over the years, the HTE scope was further expanded to the area of C–H functionalization chemistry, which received increasing attention in academia and industry. [61, 139, 140] As a consequence, the adoption of HTE has started to foster a community that makes use of parallel screening to successfully solve synthetic bottlenecks in process research and medicinal chemistry, thereby accelerating the discovery and synthesis of molecules. [141]

This is underlined by a recent survey, in which several large pharmaceutical companies complemented by two academic institutes participated to assess the current application scope of HTE. [112] Due to perceived low engineering requirements, *e.g.*, water as the solvent, room temperature reactions, biocatalysis, in particular transaminations, keto-reductions, and hydrolysis, remains the most important application for HTE. Unsurprisingly, these enzymatic transformations are closely followed by two frequently used reaction types, namely, the Suzuki–Miyaura cross-coupling and the Buchwald–Hartwig amination. Both are key bond connecting transformations, possessing a broad substrate scope and working under rather mild reaction conditions, making them commonly applied reactions in the research and development process of active pharmaceutical ingredients (API). [71, 138, 142]

Heterogeneous catalysis, including protecting group removal and reductions, is also prevalent but requires specialized high-pressure equipment. Hence, the utilization of such transforma-

tions strongly depends on the technical capabilities available at each company. Non-catalytic reactions applied on HTE systems include chiral salt resolution, scavenger, solvent or base screenings are frequently conducted to meet regulatory specifications for metal impurities. Despite the strong academic interest and first applications (see above), C–H activation and non-Suzuki–Miyaura cross-coupling reactions are still less common in HTE due to various chemistry challenges (see Chapter 1.3.4).

The following chapters will give a more detailed overview of the HTE concept (Chapter 1.3.2), applications (Chapter 1.3.3), and remaining challenges as well as resulting opportunities for future directions of the technology (Chapter 1.3.4).

1.3.2 Concept, requirements and advantages

Concept

Today, HTE is known as a systematic approach that enables the parallel execution of multiple reactions to optimize chemical transformations, expanding the scope of known reactions and exploring mechanisms. [112]. The workflow is facilitated by the use of standardized 24- or 96-, sometimes even 384- or 1536-well plates, allowing chemists to rapidly assess a large number of variables and optimize reactions more effectively than traditional one-factor-at-a-time (OFAT) methods. [110–112, 134, 143] HTE is particularly advantageous when numerous parameters such as solvents, bases, catalysts, or additives are involved, as it allows for a comprehensive screening across a diverse set of conditions. [108, 144] Thereby, HTE enables the rational design of large arrays of experiments to test hypotheses, systematically encompassing a wide range of conditions referenced in literature and further expanded by scientific intuition. [123, 145]

The integration of automation in HTE has enhanced productivity, decreased errors and improved safety by reducing human intervention in handling hazardous materials and generating high-quality, consistent data sets. Automated workflows, including the use of liquid and solid handling robots, have been developed to efficiently generate knowledge for robust and scalable chemical processes. [48] Many systems are now designed to be robust and compatible with a wide range of reaction conditions, ensuring broad applicability. [108] Efforts to miniaturize reactions in HTE formats have enabled the use of minimal starting material amounts, thus enhancing resource efficiency and sustainability. [109, 141]

Analytical techniques such as reverse-phase HPLC or UPLC, equipped with well-plate autosamplers, are employed for rapid analysis of reaction outcomes, with UV detection and MS analysis providing insights into conversion rates, compound identification, and byproduct formation. [135, 146] For chiral compounds, fast SFC analysis with chiral columns is used to determine enantiomeric excess. [147] Recent development of advanced techniques like sample pooling and multiple injections in a single experimental run (MISER) can further expedite analysis when necessary. [109, 148]

Overall, HTE has become a cornerstone in drug discovery and process optimization, enabling the exploration of reaction optimization, catalyst design, reaction discovery, and LSF, among other applications. The adoption of HTE has expanded across all research areas in organic synthesis, leveraging the power of automation and high-throughput data generation where complexity renders first principles and rational design challenging. [48, 109, 112, 114, 123]

Requirements

For an HTE system to be highly effective, it should fulfill stringent requirements that align with the diverse and complex nature of chemical synthesis. [108] While certainly not all of the prerequisites are always or straight-away achievable, and they also depend on the purpose and application of the system, the criteria can be assembled into two main groups (Figure 1.7). The first category contains mandatory characteristics to ensure that the system is producing trustworthy results.

A paramount requirement for an HTE system is guaranteeing high accuracy throughout all operations of the screening process. This includes both, manual and automated tasks and might involve routine overhauls including potential re-calibration of equipment or the introduction of checklists to avoid human errors. [149] Especially, the analytical instrumentation is of utmost importance as these machines deliver qualitative and quantitative outputs of the HTE campaigns. [135] An equally important characteristic concerns the fidelity from vial to vial on the same plate, which guarantees that the observed outcomes of reactions are attributable to the experimental conditions rather than to experimental errors, such as incorrect dosing or temperature inconsistencies. This level of consistency is vital for the accurate interpretation of results and subsequent use of those to make informed synthesis decisions. [110]

In addition, the HTE system must be designed to be economical concerning the starting material and reagent use. It should enable the execution of a high number of reactions, ranging from

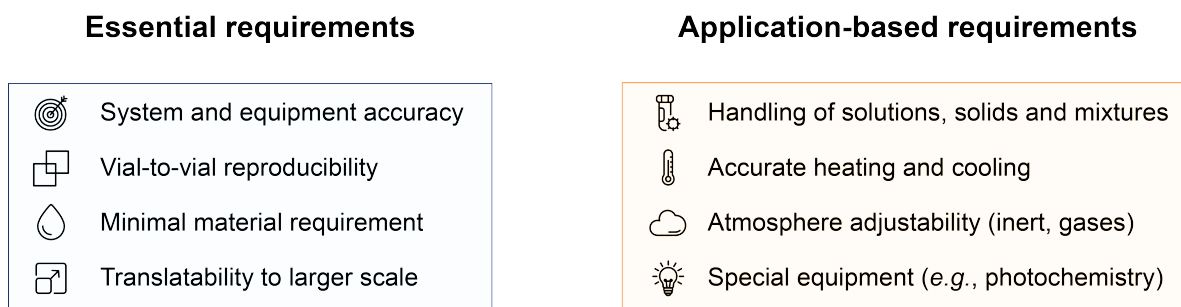


Figure 1.7: Requirements to run an HTE system. Essential criteria (l.) and application-based criteria (r.). The essential criteria need to be fulfilled to allow the smooth operation of the system and deliver accurate results. Those include the accuracy of the system including all equipment components, reproducibility of reactions between vials, very low material requirements and the translatability of small-scale screening reactions to a larger scale. Application-based requirements are closely associated with the actual reactions that run on the system. Depending on the application scope of the system, chemicals in different aggregate states need to be handled. Further temperature control could be desired if a reaction specifically necessitates heating or cooling. Since many chemical transformations only deliver the desired reaction outcome if they are conducted under an inert or gaseous atmosphere, changes thereof need to be accommodated if needed. If the HTE system should be able to run special reaction types, such as photo- or electrochemistry, the components and workflows need to be adjusted accordingly, potentially requiring a significant amount of engineering and validation.

24 to several hundred wells, requiring only minimal amounts of chemicals per reaction. This efficiency in reagent consumption not only reduces costs but also minimizes waste, aligning with sustainable laboratory practices. [109] Lastly, the system should facilitate the scaling of results from small-scale experiments, in the range of μmol or nmol , to larger scales that are of interest for medicinal or process chemistry. This scalability is crucial for the transition from experimental to practical application, ensuring that discoveries made at the microscale can be translated into tangible chemical processes. [150]

The second category of requirements depends more strongly on the desired application scope of the HTE system, *i.e.*, which reaction types and technologies are planned to be screened using the instrumentation. [108] In an ideal setting, the HTE set-up is capable of running a wide range of reactions that are used by the bench chemists within the organization. Such an approach presumably requires system compatibility with a diverse array of solvents, from non-polar solvents such as hexane to highly polar solvents like hexamethylphosphoramide (HMPA). [151]

Further, the HTE system must be able to operate effectively across the desired temperature spectrum. This could encompass the ability to conduct reactions at cryogenic temperatures as well as at temperatures that exceed the boiling points of solvents, thereby facilitating a wide range of reaction kinetics. [112] The versatility of the platform is further underscored by its requirement to handle homogeneous solutions, more complex heterogeneous or even

biphasic reaction mixtures, which are often encountered in synthetic chemistry. [151] Many chemical transformations also require anhydrous or anoxygenic conditions, which are crucial for reaction components that are sensitive to moisture or air. Therefore, the HTE system should be able to offer dosing and reaction environments in inert atmospheres. [110] Specific reactions could also demand the use of reactive gas, *e.g.*, H₂ for hydrogenation reactions, potentially requiring the installation of the necessary environment and instrumentation. [120]

The outlined requirements, especially the second group that is dependent on the type of reactions that are conducted, highlight the challenges faced in the engineering of HTE systems for chemistry applications. As opposed to biology and biochemistry, where experiments generally take place in water and ambient temperatures, chemical reactions demand solutions to avoid material decomposition, prevent solvent evaporation, and ensure heating and stirring material compatibility. [110] To circumvent these, specialized equipment has been developed over the last decade that supported the uptake and success of HTE campaigns in industry and academia. A short overview of the most used instrumentation and tools based on a recently conducted survey across HTE labs is given below. [112]

Manual tools such as single and multichannel pipettors, and a variety of well plates, remain ubiquitous due to their ease of use and low entry barrier. [110] Nitrogen-filled gloveboxes are commonly used for setting up screens, while nitrogen purge boxes are employed for compound storage and workflows requiring a less pristine inert atmosphere. [48] Solvent removal is typically performed using centrifugal evaporators within gloveboxes or nitrogen blow-down tools for more accessible HTE approaches. [110] Reaction agitation and heating are achieved through a range of equipment including tumble stirrers, hot plate stirrers, incubators, heater/cooler shakers, and custom-designed shaker/heaters. [112] Reactive gas delivery platforms are also widespread, enabling experiments under atmospheric and elevated pressures for reactions such as hydrogenations and carbonylations. [120] Automation enhances HTE by providing accurate and efficient screen setup through liquid and solid handlers, with a notable presence of Unchained Laboratories/Freeslate/Symyx systems across many teams. [112] Solid handling platforms from various vendors address the challenges of manually weighing reagents at milligram scales. [152] While the majority of automated platforms operate under an inert nitrogen atmosphere, a fraction is used on the bench for non-sensitive applications. [111]

Analytical tools are crucial for processing HTE screens, with ultra-high-performance liquid chromatography (UPLC) systems measurably reducing analysis time and maintaining high-quality data. Mass spectrometers are coupled to some UPLC systems, while HPLC and SFC systems are used to a lesser extent. Gas chromatography (GC) with mass detector sees limited use. The integration of manual, automated, and analytical tools is key to maximizing the impact of HTE, streamlining workflows, and enhancing research productivity.

Advantages

HTE, if designed and executed successfully, can offer a multitude of advantages for chemical reaction development and optimization, particularly in terms of efficiency and sustainability.

The HTE approach is distinguished by its resource-efficient experimental framework, which is particularly advantageous when only small quantities of starting materials and reactants are available. [111] In stark contrast to traditional methods that may use hundreds of milligrams of starting material, HTE operates on the micromolar scale, reducing the amount of material needed for each reaction [66, 153, 154]. This reduction in material requirements not only conserves valuable resources but also lowers the threshold for initiating experiments when the outcomes are uncertain. [112] Consequently, HTE enables the use of less material to conduct a higher number of experiments, thereby optimizing the use of resources and facilitating a more exploratory approach to chemical research. [155]

Moreover, HTE revolutionizes the optimization of chemical processes by allowing the simultaneous exploration of numerous reaction parameters, a stark contrast to the OFAT method prevalent in classic single-batch reactions. [112, 154, 156] The rapid and efficient experimental setup enables the assessment of a diverse class of variables within days, thereby expediting the development of robust processes. [120] Further, not only the reaction conditions themselves can be optimized, but HTE also unlocks the assessment of a broader chemical space by screening structurally diverse sets of compounds. [110] Consequently, HTE is an invaluable tool for tackling complex challenges in chemical synthesis through rapid multi-parameter screening, thereby contributing to a more comprehensive understanding of reaction landscapes. [112]

HTE also automates repetitive manual tasks, which are still commonplace in traditional laboratory settings. The integration of automation technology within HTE systems diminishes the need for labour-intensive activities such as the weighing of solids or the dispensing of stock solutions. [109, 110, 149] The automation of tasks substantially reduces the time required for

synthetic experimentation and as a result, affords opportunities for scientists to reallocate their time from routine tasks to more intellectually demanding and creative problem-solving endeavours.

Finally, HTE contributes to the generation of high-quality, consistent data sets, which are essential for in-depth analysis and the application of ML algorithms. The automation of sophisticated analytical instrumentation within HTE workflows ensures the acquisition of reliable experimental endpoints. [155, 157] This is specifically of high value, as data sets derived from literature often miss failed or low-yielding reactions, which can detrimentally affect the analysis and the predictive capabilities of ML models. [154, 158, 159] Additionally, the substrate scope reported in the literature for catalytic reactions is frequently limited to simple model substrates, which may not accurately represent the complexities encountered in LSF within drug discovery and development. [111] Therefore, HTE is considered an invaluable tool for generating high-quality reaction data for broad chemical space and diverse reaction space to enable successful big data analysis and reactivity prediction with ML algorithms. [160]

1.3.3 Application scope

HTE has emerged as a transformative approach in chemical research, offering a multitude of advantages that streamline the development and optimization of chemical reactions. [110–112, 143]

Development of novel reaction methodologies

Firstly, HTE can be utilized in the development of novel reaction methodologies. Through sequential screening iterations, scientists can rapidly evaluate the effect of a wide array of catalysts, ligands, reagents, additives and solvents on the success of the reaction.

MSD demonstrated the discovery of a ligand that enhanced the reactivity of a palladium-catalyzed cross-coupling reaction to aid the synthesis of a diverse set of benzophenones. [161] In another application of HTE at MSD, a late-stage direct alkylation of heterocycles was developed, utilizing iridium-based excited-state reductants to generate alkyl radicals from peracetates under mild conditions, marking the first instance of the room-temperature introduction of methyl groups into complex heterocycles. [124] The MSD HTE group facilitated the development of photoredox-catalyzed processes for late-stage pharmaceutical development in collaboration with the Britton group. This partnership yielded one-step direct fluorination

of leucine using sodium decatungstate, producing λ -fluoroleucine, a key intermediate in the synthesis of Odanacatib. [126]

The use of HTE enabled another rapid optimization of a fluorination reaction, mediated by D-proline in trifluoroethanol, that afforded a congested quaternary stereocenter with high yield and improved diastereoselectivity. The team at Lilly could also demonstrate the successful scaling to kilogram quantities, with the added benefit of trifluoroethanol recovery and reuse. Critical to the process was the optimization of α -fluorination conditions, which altered the diastereoselectivity from 1:7 to 7:1. [162]

Furthermore, AstraZeneca disclosed a cobalt-catalyzed late-stage C–H methylation strategy for complex drug molecules. HTE was instrumental in addressing this synthetic challenge through selective multiparameter optimization, culminating in a broadly applicable methodology that leverages functional groups to direct C–H activation, transforming C–H bonds into methyl groups using a boron-based methyl source. [74] Additional work from AstraZeneca reported an iridium-catalyzed directed C–H amination methodology derived from investigating numerous directing groups and substrate scope using an HTE-based strategy. [163]

Expansion of known reactions

Secondly, in addition to identifying novel methodologies, HTE also facilitates the logical extension of already-known reactions through the systematic exploration of both, substrate scope and reaction parameters. The scope of the above-described late-stage alkylation methodology by Di Rocco and colleagues [124] was expanded to hydroxymethyl groups. The enhanced protocol introduced the new scaffold into a variety of heterocycles using benzoyl peroxide as the oxidant and methanol as the hydroxymethyl radical source. [125] In another example, the MSD HTE team explored the cross-coupling of nitromethane with aryl halides, which traditionally yielded low product yields and multiple side products. The systematic screening of various ligands, bases, and solvents with a range of aryl nitromethanes identified reagent combinations that yielded high conversions and minimized side product formation. [164]

A team of Pfizer scientists noted that an existing Ni-catalyzed reductive cross-coupling protocol for electrophiles [165] was not amendable for their envisaged application. Using the broad library of nitrogen-containing molecules at Pfizer and screening those substances as nitrogen-donor ligands in an efficient HTE setting, the scope of the reaction could be extended successfully. [166] The scope of the Buchwald-Hartwig amination reaction was thoroughly

investigated through the analysis of 48 electrophiles, revealing specific challenges with amidation side reactions and the incompatibility of certain substrates, highlighting the strength of HTE to explore uncharted chemical space for a well-know reaction. [167] The assessment of coupling DNA-conjugated aryl iodides in a Ullmann-type transformation provides another example of how systematic screening can expand the scope of a frequently used chemical reaction for a specific application, in this case, DNA-encoded library technology (DELTA). [168]

Enhancement of existing chemical transformations

Thirdly, HTE also proves invaluable in the refinement of existing transformations, especially frequently used C-C cross-coupling reactions, such as the Suzuki–Miyaura or Buchwald–Hartwig reactions. MSD showed that by conducting a multidimensional HTE screen that simultaneously optimized multiple variables, including ligands, bases, and solvents, across different temperatures, a library campaign in discovery was accelerated. The comprehensive approach identified a set of conditions that yielded high conversions, which were then applied to synthesize a variety of heterodiarylmethanes with high efficiency. [169]

GlaxoSmithKline (GSK) highlighted the advances of multivariate screening through HTE compared to OFAT optimization from a paper for a single palladium-catalysed carbonylative esterification reaction. An academic group approached the problem with the well-known optimization table. 16 different phosphines were screened against a single solvent or base system of acetonitrile (MeCN) and *N,N*-diisopropylamine delivering a certain reactivity range and leading to disclosure of one condition as the ideal combination. [170] GSK approached the same transformation with HTE, screening all combinations of the 16 ligands, three bases and two solvents. Those 96 reactions revealed that the 21 reactions, which were carried out in the OFAT protocol, missed out on the ideal set of conditions. [143] This highlights the power of HTE, which overcomes the issue that limiting a reagent or catalyst assessment to a single set of parameters will only deliver the best reagent or catalyst under the tested parameters, likely missing the overall optimal set of conditions.

Further evidence can be found in the literature, *e.g.*, the HTE team from MSD demonstrated the optimization of conditions for various Pd-catalyzed C-O, C-N, and C-C cross-coupling reactions that elevated success rates. [109] An academic group also showed that a difficult esterification coupling was solved using an HTE approach, where a selection of ligands, catalysts and additives could be assessed efficiently to deliver optimal transformation parameters. [160] One publication from academia described an approach to teaching HTE to undergraduate

students using the Suzuki-Miyaura coupling as an example. [171]

Thus, HTE has become an invaluable tool for optimizing in a streamlined and efficient way contributing to significantly reducing material consumption and improving transformations from small to large scale.

Serendipity

Serendipitous discoveries have been pivotal in advancing the chemical sciences, particularly in the realm of bond-forming reactions. [172] Notably, seminal synthetic transformations including Friedel-Crafts, Wittig olefination, and Brown hydroboration transformations emerged when experimental outcomes diverged from the original objectives. [173–175] HTE can lead to directed serendipity, where the rapid screening of diverse reaction conditions can yield novel and unexpected, yet potentially efficient reaction pathways.

This was demonstrated in the improvement of a thermal cyclization protocol for the synthesis of pyrimidinone heterocycles, important intermediates in the development of HIV Integrase inhibitors. By utilizing a reaction discovery platform with pre-dosed compounds, researchers at MSD were able to conduct and analyze hundreds of experiments in a single day, ultimately identifying catalytic systems that enhanced the reaction outcome. [145] Another serendipity discovery using HTE was observed at MSD when optimizing the synthesis of Letemovir. A key step of the synthesis was the establishment of a single stereocenter through an aza-Michael reaction. It was observed that bis-quaternized PTC impurities led to more active and selective catalysts. These results provided an impetus for the development of subsequent bis-quaternized PTC libraries that found unique utility for the construction of the chiral cyclic urea moiety in Letemovir. [130] Through the deployment of an automated HTE workflow, the McMillan lab assessed numerous reactions indiscriminately, which led to a novel photoredox-catalyzed C–H arylation reaction. This transformation facilitates the synthesis of benzylic amines, a crucial structural motif in pharmaceutical compounds, from simple substrates under mild and straightforward conditions. [123]

HTE has emerged as a powerful strategy that accelerates chemical research by enabling the rapid and efficient evaluation of reaction conditions serving multiple purposes from new methodology development, through extending and refining known reactions to the discovery of unexpected outcomes, underlining its indispensable role in modern synthetic chemistry. However, limitations and challenges remain that curtail the full potential of the technology (Figure 1.8).

1.3.4 Challenges and opportunities

Capital investment

The substantial capital investment in acquiring automated, robotic synthesis and related analytical instruments has generally restricted the widespread adoption of comprehensive HTE workflows to primarily industrial environments. [135, 176] Unfortunately, equipment purchase remains a major obstacle, especially for academic institutions and smaller entities, limiting the wide adoption of HTE considerably. While there have been reports from academia where HTE techniques are applied for reaction optimization or platform development, the technology has not been broadly established so far. [110, 177–184] Therefore, it is important to address the urgent need for more affordable and accessible HTE solutions, which would widen their use and facilitate application in academia to support, *e.g.*, methodology research campaigns or total synthesis projects.

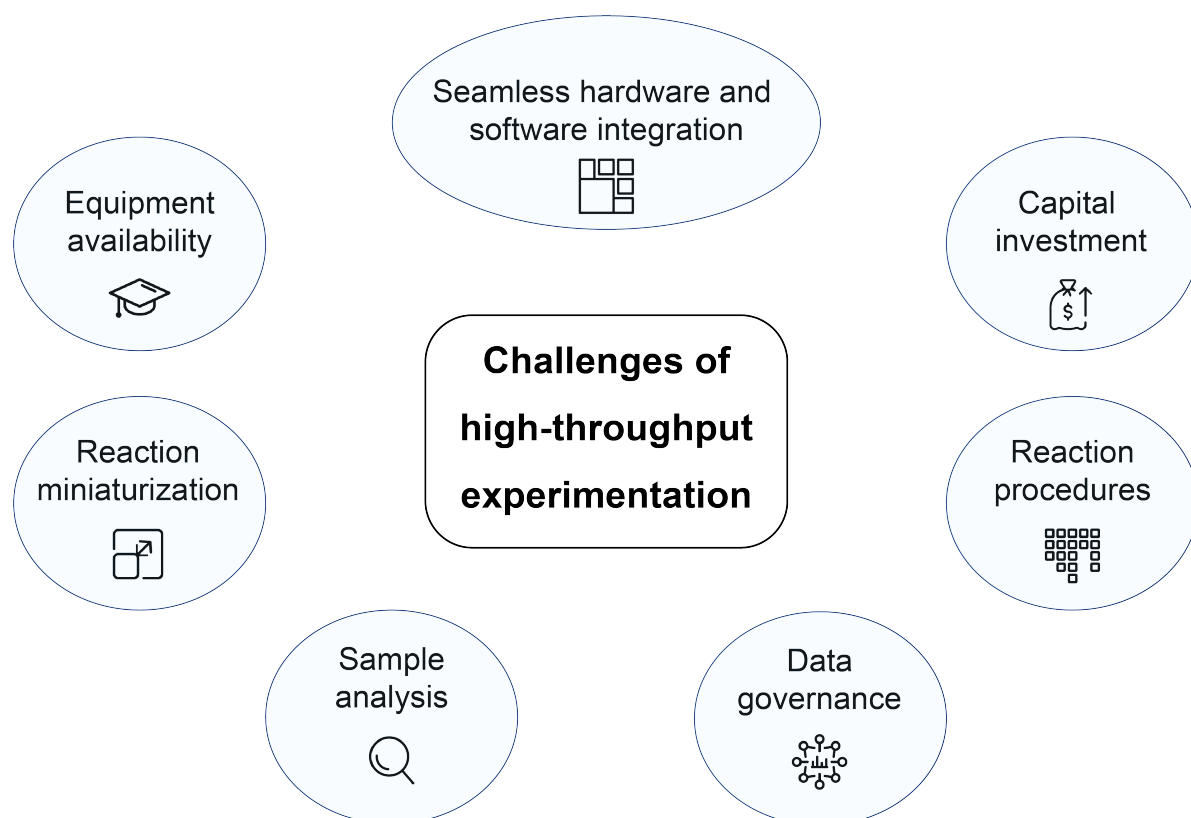


Figure 1.8: Overview of the main challenges currently observed in high-throughput experimentation (HTE). The seamless connection between equipment components (hardware) and application interfaces (software) remains a significant bottleneck leading to missing information and inefficient workflows. Further, the monetary investments to buy and set up an HTE are high and, therefore, act as an entrance barrier. This directly influences the education of new scientists on HTE systems, since equipment is hardly available outside of an industry setting. Miniaturizing reactions still belong to an area of improvement as the accurate weighing of small amounts of material (<1 mg) through robots has not been achieved so far. In addition, HTE faces a broad range of reaction procedures and needs for specialized equipment to accommodate the breadth of chemical transformations. With increasing throughput, analytical measurements and their accuracy need to be further enhanced, which requires innovative approaches and technological developments. Lastly, managing and systematically storing the generated data of the HTE system to make it accessible for machine learning (ML) applications still appears to be a bottleneck.

Equipment availability

Without HTE equipment available at the university, be it in teaching or research labs, future chemists can hardly be educated and familiarized with state-of-the-art technologies, which they might encounter later in their careers. [179] While there are certain exceptions, *e.g.*, the Cernak lab at the University of Michigan [185], the Swiss CAT+ initiative at ETH Zurich and EPFL Lausanne [186], or a rare report on undergraduate courses focusing on optimizing Suzuki–Miyaura reactions in parallel [171], most university students will not be in touch with automated synthesis equipment or parallel experimental design throughout their educational journey. While it is close to impossible to establish HTE facilities at all universities across the globe, strengthening the exchange between industry and academia could help to familiarize talent with the technology from early on. Several examples of such collaborations have shown to be fruitful for both parties, many leading to the discovery of novel chemical transformations. [163, 187–190]

Chemical reaction space

Furthermore, HTE requires a fair amount of engineering to accommodate the vast diversity of reaction types and chemical reagents. Whereas running HTE plates under standard conditions already requires a thought-through set-up to ensure reproducibility, the two main reaction types that add additional complexity to reaction protocols are photo- and electrochemical methods. [191, 192]

Employing illumination to catalyze a chemical reaction presents certain obstacles in guaranteeing the consistent reproducibility of the photochemical process. It needs to be ensured that the irradiation across all wells is identical, that the heat generated by the shining light bulbs does not influence the reaction, and that different wavelengths as well as the light intensity can be controlled. [192–195] Despite these challenges, the literature describes successful applications of photochemical HTE campaigns for different reaction types. [196–200] With the growing importance of photo-induced chemistries, further improvements and standardized set-ups are required to allow broad application of photochemical methodologies in an HTE setting that consistently deliver reliable results.

Similarly, in electrochemistry, several parameters need to be controlled, and potentially modulated, to allow the screening towards optimal reaction conditions. Those include the material of the electrodes, the voltage, the current density, type and concentration of assisting elec-

trolyte or the surface shaping during the transformation. [201] Consequently, the technical difficulties in setting up a suitable, reproducible HTE platform for electrochemistry are manifold. Examples include the independent electrical regulation within each well, the design of dependable reference electrodes and the management of potential pressure changes due to gas formation. [191] Despite these issues, over the last couple of years, progress in electrochemical HTE can be witnessed. A small number of platforms were disclosed, which helped to improve methoxylations, radical–radical cross-couplings, azidooxygenations, silylations or aminations. [201–204] These starting points need to be improved further to simplify optimization and discovery of electrochemical transformation using HTE equipment.

Despite recent progress, the accurate and efficient dispensing of a diverse range of chemical reagents on the submilligram scale remains a challenge as well. [110, 205] To work in a streamlined high-throughput fashion, the dosing of reagents needs to be carried out through automated powder dispensing or liquid handling of stock solutions. [112] The initial transformative advancements were centred around automated liquid handling to enable high-throughput nanomolar scale reaction screening using bioassay equipment in the 2010 decade. [109] MSD and Pfizer, the first movers in the field, established an array of systems that were capable of accurately dosing micro- or nanoliter amounts, *e.g.*, the TTP LabTech Mosquito HTS, which are still main parts of the industry today. [109, 206] Still, these applications are limited to chemicals that are soluble in a solvent compatible with the reaction solvent. Some workarounds, including solvent evaporation after dosing or the generation of homogeneous slurries through acoustic mixing to enable, *e.g.*, the dosing of solid inorganic bases solids, were established to overcome these limitations. [196] Yet, automated solid dosing remains an important alternative to and also circumvents the production of stock solutions, which are time-intensive to prepare and can be prone to chemical degradation during storage. [207]

When working with solid components, the HTE operations heavily rely on the ability to dispense diverse powders accurately and precisely. [206, 208] The automation of solid handling is complex due to the heterogeneity of compound properties, such as particle size distribution, powder type, density and flowability. Current methodologies rely on gravimetric distribution, primarily employing either hopper/feeder or positive displacement technology. [151] Employing rotary valves and tapping actions for flow regulation, hopper/feeder modules like the Mettler-Toledo Auto Chem Quantos provide gravimetric solid dispensing, particularly effective for free-flowing substances in quantities ranging from milligrams to grams. Positive displacement modules, such as the Chemspeed Technologies GDU-S SWILE, utilize piston-driven

capillaries for gravimetric dispensing, excelling in sub-milligram to low-milligram quantities and accommodating solids with diverse physical characteristics, including adhesiveness. [152, 206]

Recently, AbbVie pioneered an approach to allow the precise dosing of sub-milligram quantities of solid reagents using the technology of chemical-coated glass beads, often referred to as ChemBeads. [205] ChemBeads are generally produced using resonant acoustic mixing, which ensures the uniform adhesion of solid reagent "guests" onto the surfaces of inert glass bead "hosts". The application of solid reagents onto glass beads merges the intrinsic characteristics of the solid reagents with the advantageous attributes of the glass beads. [209] Importantly, the coating effectively increases the bulk of the solid reagents, thereby allowing the accurate dispensation, including with the use of automated equipment, of sub-milligram quantities directly into reaction vessels. This obviates the necessity for the preparation of reagent stock solutions, thereby conserving time, resources, potential solubility issues, and undesired co-solvent mixtures. [210, 211]

The AbbVie team also expanded their approach to enzymes to unlock the class of biocatalytic transformations. [212] Based on the reports from AbbVie, the technology has been successfully implemented and applied to a broad range of transformations, including Suzuki–Miyaura couplings, Buchwald–Hartwig aminations, nickel-catalyzed cross-electrophile couplings, C(sp²)–C(sp³) decarboxylative couplings, and nitrene transfer reactions. [205, 209, 213–215] However, apart from the AbbVie publications, only very recently reports using ChemBeads have surfaced. [216–218] To further evaluate the potential of the technology, uptake from the wider research community is required to prove that the technology is reproducible in other laboratories, applicable to a wide range of chemistries and the dosing can be carried out accurately utilizing automated equipment.

Miniaturization

With an increasing focus on sustainability, green chemistry and the limited availability of high-value intermediates in discovery chemistry campaigns, it is of utmost importance to miniaturize the reaction scale to use only milligram starting material quantities when running arrays of reaction conditions. [48, 111, 141] Main challenges of reaction miniaturization include the need to handle very small amounts of heterogeneous materials, sufficiently stir reaction mixtures and avoid the evaporation of volatile solvents. [109]

Volatility has been reported as a reoccurring issue due to the use of low solvent volumes compared to the rather large surface area of the well, which resulted in the utilization of high-boiling solvents with low flash points, *e.g.*, dimethylsulfoxide (DMSO) or *N*-methyl-2-pyrrolidone (NMP), limiting the diversity of the usable chemical reaction space. [150, 219] Current HTE equipment, predominately the Analytical Sales plates in 24-, 48- or 96-well format with glass vials in different volume sizes (1, 2, 4, and 8 mL), partially address the issue based on their plate sealing technique with rubber mats and tightening using screws. [111, 112] Further enhancements to this set-up could allow higher solvent flexibility and open up new chemical reaction space.

The above-described ChemBeads have the potential of helping to overcome the sub-milligram dispensing of compounds while also contributing to improved mixing of the reaction. [209] While miniaturization might be ideal for identifying the most suitable conditions with limited material consumption, ultimately, in most cases, the reaction will be carried out with greater material quantities. As a consequence, the translatability from small to large scale, which might vary from reaction to reaction, needs to be understood. [150] Even though there are examples in the literature where microscale screening results were seamlessly scaled to gram or kilogram quantities, [109, 144, 220–223] it is not a given for every transformation due to *e.g.*, heat and mass transfer or kinetic challenges. [224, 225]

The collection and analysis of data sets containing information on miniaturized reactions and their large-scale counterparts, further enriched through the development of suitable ML algorithms that can predict the correlations for new examples could increase synthetic success. Finally, the monitoring of miniaturized reactions also poses challenges, as sample drawing becomes technically challenging and the analytical resolution is interconnected with material quantities. [135]

Sample analysis

To determine the reaction outcome for the large number of samples generated by HTE, fast sample analysis and processing of analytical data are essential to the workflow. Often, analytics constitute a large time investment in HTE, dwarfing the time required for reaction setup. [134, 135, 151] In one of their early HTE publications, MSD reported the following time splits for their campaign with 1536 reactions: 30 minutes spent on reagent dosing, 1 hour on sampling, and 52 hours on UPLC analysis [109] However, they also highlighted the use of the MISER technology.

This method does not carry out any chromatographic separation of the sample and utilizes flow injection analysis (FIA) to sequentially introduce multiple samples in a single, continuous analytical sequence. Single samples are tracked by selected ion monitoring (SIM) across various chromatographic techniques. [135] As a consequence, MISER is capable of requiring only 10 seconds per sample analysis, [146] potentially reducing the analysis time of a 1536 well plate to roughly 4 hours. Since MISER can be used on standard liquid chromatography-mass spectrometry (LCMS) hardware if the chromatographic data system is capable of working in FIA mode, the barrier to introducing the technology is low. [135]

In addition, the automation of other analytical methods for potential application in HTE to allow rapid sample analysis was explored over the last decade. Those approaches include matrix-assisted laser desorption/ionization (MALDI), [226, 227] desorption electrospray ionization (DESI), [177, 228] acoustic ejection mass spectrometry (AE-MS), [229–231] and nuclear magnetic resonance (NMR) spectroscopy. [153, 232–234]

Ultimately, the chosen analytical method needs to be in line with the desired output from the HTE system, ranging from binary (*i.e.*, the reaction works/does not work) through quantitative information (*i.e.*, conversion, yield) to full characterization (*i.e.*, confirmed molecular structure) information depth, and the available analysis time. Advances in analytical methods will contribute to further increasing the efficiency of sample analysis in the future. [134] Importantly, the interplay between the analytical instruments and the rest of the HTE system needs to be intact to allow seamless data flow to enable accurate analysis. [135]

Hardware and software integration

The various devices that are needed to run semi-automated HTE campaigns range from solid and liquid handling through stirring and temperature control to the analysis of the reaction samples. Advanced automated solutions also include robots for the translocation of samples and materials between devices. This aggregation of heterogeneous instrumentation is of no practical use without the integration of user-friendly, yet pragmatic software solutions. In most cases, researchers need to navigate the integration of individual software for each instrument and between devices themselves as fully consolidated software frameworks are only offered for integrated robotic systems that originate from one vendor and, hence, often limit the application scope of the system. [151] While seldom emphasized in automation-related publications, error handling and data governance in automated processes critically influence the functionality and robustness of the HTE workflow. [235]

Despite the increased use of automation solutions in industry and academia, there are only a few examples in literature that contain a detailed description of the actual hardware and software integration. [236–241] The applications include the set up of a customized robotic system, limited to certain chemistries, [236] are tailored around small to medium throughput [239, 240] or focus only on one part of the HTE workflow, *e.g.*, analysis by LCMS [241]. A promising approach was very recently published by the Cernak lab, which developed *phactor*TM, a software that supports scientists throughout their HTE workflow, from set-up through execution to analysis. [237, 238]

Apart from the interplay between the systems, data governance of in- and output data remains a tremendous bottleneck, which has a direct impact on the ability to carry out rapid data analysis, detection of patterns, highlighting of chemical insights, design of experiments and reactivity prediction using ML. [155, 242] Currently, laboratory notebooks capable of systematically cataloging HTE information for straightforward access are not available. Most electronic lab notebooks support the creation of custom experiments but do not feature an intuitive interface for extracting data sets in a standardized format and conclusions from multiple experiments in aggregate. [151]

Data governance

Since HTE produces extensive and reliable data sets that capture both successful and, notably, also unsuccessful outcomes of chemical reactions across various chemical domains, it is of

utmost importance to capitalize on the information. [158, 160, 243, 244] Given the volume of data generated, it is essential to tackle the issue of data sharing by adopting standardized, human- and machine-readable formats (see Chapter 4). Initiatives like the Open Reaction Database (ORD)[245] and the Unified Data Model (UDM)[246] have contributed to enhancing the accessibility of reaction data sets, even though these mainly contain frequently used transformations. As a result, HTE remains a critical source for generating high-quality data sets to expand the use of ML for reactivity prediction and reaction condition screening. [159, 247]

Summary and outlook

As this chapter showed, HTE is a powerful technology that still requires continuous innovation to cope with the manifold chemical reaction diversity. Many challenges that remain are interconnected with one another and require thought-through innovations to be solved. Nevertheless, by addressing some of the outlined aspects, specifically reaction miniaturization, soft-/hardware integration and data governance (Figure 1.8), HTE can be used as an invaluable tool to enable LSF for drug discovery. Running multiple different methodologies in a broad chemical space and systematically collecting all data points, including all in and outputs of the reactions, can facilitate reactivity prediction and retrosynthesis among others, with ML.

1.4 Reactivity prediction

1.4.1 Background

ML falls under the umbrella of artificial intelligence (AI) and involves applying sophisticated algorithms to substantial data pools. The goal is to build systems that simulate the human learning experience. Through this process, ML algorithms progressively achieve greater accuracy, allowing for the discovery of core relationships and patterns in the data. [248] Today, ML has become integral to various technologies, [249–254] including those aiding in the acceleration of drug discovery and the exploration of chemical reaction space. [1, 44, 248, 255] Chemical reactions, which detail the transformation of reactants into products, are central to this exploration, and machine intelligence can play a pivotal role in enhancing the success rate of these chemical reactions. [256–258] The groundwork for computer-assisted synthesis planning (CASP) was laid by Corey, [259] who codified retrosynthetic rules, which was further

advanced by the development of knowledge bases and classification schemes by Hendrickson and others. [260–263] These efforts facilitated the use of ML models to recommend similar transformations in chemical reaction planning.

The encoding of chemical reactions in terms of bond-electron matrices by Dugundji and Ugi [264] marked an important milestone, inspiring subsequent developed expert systems based on formal reaction logic. [265, 266] Approaches, such as Sophia, [267] and Chematica/Synthia, [268, 269] which amassed over 100'000 rules, have expanded the capabilities of synthesis planning. Over the last decade, ML in combination with access to large reaction data sets, *e.g.*, extracted information from the US patent space or the ORD initiative, [245, 270] have propelled the field, specifically CASP, forward. [157, 271–276] More recently, the field of digital chemistry is undergoing rapid evolution, having delivered many novel developments beyond traditional CASP applications. [277–279]

A set of recent, comprehensive reviews on reactivity prediction by Coley et al., [248] Stocker et al., [255] Jorner et al., [277] Meuwly, [278], and Ertl et al., [279] can be found in the literature. In the following, only a brief overview of the topic with an emphasis on fundamental ML concepts (Chapters 1.4.2 and 1.4.3), and the forecasting of binary reaction outcome (Chapter 1.4.4), reaction yield (Chapter 1.4.5) and regioselectivity (Chapter 1.4.6) will be given as those topics are of main relevance for the desired application of reactivity prediction by training graph neural networks (GNNs) (Chapter 1.4.7) with HTE data originating from LSF screenings.

1.4.2 Molecular representation of chemical reactions

The reliance on molecular descriptors has been a cornerstone in the evolution of cheminformatics tools, *i.e.*, implementation of computer-based techniques to explore chemical phenomena, applied in drug discovery over multiple decades. [280, 281] Chemical equations are utilized by chemists to abstract the transformation of starting materials, often also referred to as reactants, into products using a defined set of reagents, catalysts, solvents and - in some cases - additives under specific physical conditions (*e.g.*, temperature, time, atmosphere). [282, 283] The molecular structures of components in chemical reactions can be represented in various computer-readable formats, including line notations, MOLfiles, and structure data (SD) files. [284–288]

Graphs are commonly used to depict molecules with nodes and edges that symbolize atoms and bonds, respectively. These descriptors emerge from the application of logical and mathe-

mathematical operations that translate the chemical information contained in a symbolic representation of a molecule into either a vector or scalar. [289] Researchers have developed a variety of molecular descriptors to represent different aspects of molecular features. Notably, the descriptors that have gained widespread prominence for use in predictive and explanatory tasks are those that are based on the two-dimensional (2D) structural attributes of molecules. [290–292] The introduction of molecular descriptors that capture pharmacophore properties in 2D has facilitated their use in the discovery of novel compounds through scaffold hopping and virtual screening. [30, 293–298]

However, 2D graphs have limitations in describing stereocenters of molecules, leading to the use of Cartesian coordinates for atomistic modelling tasks. [283] Initially developed to encapsulate the shape and structural characteristics of ligands, these three-dimensional (3D) shape descriptors have expanded to include both ligand-centric shape attributes and pharmacophore features, as well as the 2D and 3D aspects of protein-ligand complexes. [299–305] The application of 3D descriptors across various tasks relies on the *a priori* encoding of feature vectors through rule-based algorithms, with the success of such applications hinging on the premise that structurally similar features imply similar molecular properties. [26]

The Simplified Molecular-Input Line-Entry System (SMILES) notation is the prevalent format for chemoinformatic tasks, capable of encoding stereochemistry to an extent, acting as an identifier and serving as a versatile representation for database and search queries. [285, 286] Advances in chemical language models (CLMs), *i.e.*, ML models that handle molecular sequences as in- and/or outputs (Chapter 1.4.3), have popularized SMILES as molecule representation. [274, 306, 307] Reaction SMILES use characters to separate molecules and stages within reactions, and atom-mapped reactions can be represented using various formats such as Condensed Graph of Reaction (CGR), reaction SMILES arbitrary target specification (SMARTS), or ReactionCode. [308–312] Self-referencing embedded strings (SELFIES) is an alternative string-based representation ensuring syntactically valid molecules in generative tasks. [313, 314]

1.4.3 Chemical language models

CLMs are ML frameworks designed to process molecular sequences, such as SMILES, using neural networks (NNs). [307, 315, 316] NNs, often also referred to as artificial neural networks (ANNs), are computational models composed of interconnected nodes (artificial neurons)

arranged in layers, including an input layer, one or more hidden layers, and an output layer, which process information by simulating the signalling behaviour of biological neurons in the human brain. Each node is linked to others and possesses a specific weight and threshold, transmitting data to the subsequent layer only when its output exceeds this threshold. NNs depend on training data to improve their performance through iterative adjustments to the connections and thresholds. [317–319] ML that utilizes NNs with multiple layers of processing is commonly known as deep learning. [320, 321]

CLMs mostly employ recurrent neural networks (RNNs) and Transformers for sequence data handling. [322–324] RNNs are a class of NNs that process sequential data by maintaining a dynamic hidden state influenced by both the current input and the previous state, capable of handling sequences of variable lengths and often employed in an auto-regressive manner to predict subsequent elements in a sequence. [322] Transformers, the common architecture of a large language model, process sequences through graph-based structures, utilizing attention mechanisms to dynamically weigh the relevance of different tokens for predictive tasks, and are particularly effective in sequence-to-sequence applications like language translation. [325, 326]

RNNs are capable of handling variable-length sequences and predicting subsequent elements in a sequence. To address the limitations of basic RNNs, such as gradient vanishing or exploding, advanced architectures like long short-term memory (LSTM) [327] and gated recurrent units [328] have been developed. RNNs have been extensively applied to generate novel molecules with desired properties, learning both the syntax of SMILES notation and capturing molecular semantics. [307, 329–331] Techniques like data augmentation and bidirectional learning have enhanced the quality of chemical language learned by RNNs. [332, 333] RNNs have also been used for feature extraction, outperforming traditional descriptors in tasks like virtual screening and property prediction. [334] Compared to other deep learning approaches, like generative adversarial networks and variational autoencoders, RNNs have shown superior or comparable ability in learning SMILES syntax for *de novo* molecule design. [335]

Transformers have been specifically adapted for tasks such as predicting chemical reaction outcomes, multi-step syntheses, and molecular properties [274, 275, 336–338]. They were also combined with equivariant layers to predict 3D protein structures from amino acid sequences, achieving state-of-the-art results. [339, 340]

1.4.4 Binary reaction outcome assessment

The objective in binary reaction outcome prediction is to ascertain *in silico* if a reaction will deliver desirable products based on given starting materials and conditions. [341] Specifically focusing on the most recent deep learning methods applied for this task, three different NN-based methodologies are prevalent in the literature: Template-based, graph-edit-based, and sequence-based strategies. [157, 272–274, 342, 343]

Template-based methods operate by matching reactants to predefined reaction templates extracted from databases like Reaxys, which encapsulate the transformation rules and the reaction center. [283] Early attempts constructed a NN that predicts reactions by identifying electron flow within an in-house dataset of elementary reactions. [344–346] This approach was further advanced by training models to classify reactions based on molecular fingerprints and ranking the likelihood of different reaction rules from extensive template libraries. [290, 342, 347] More recently, Coley *et al.* introduced a ranking system for the multiple products that can arise from template matches, addressing the issue of template multiplicity. [157] Template-based NNs are inherently limited by the diversity and specificity of the templates in the dataset. The balance between the granularity of these templates, which may include the effects of distant functional groups, and the manageability of the template set size is a critical trade-off. [283]

Graph-based methods for reaction outcome prediction represent chemical structures as graphs and forecast alterations in molecular bonds using NN architectures. [283] Pioneered by Jin *et al.*, a first approach with a graph convolutional NN that infers bond changes in reactants without the need for predefined reaction center sizes was developed. [348] Subsequent developments that included a gated GNN, [343] a graph transformation policy network, [349] and a relational graph convolution NN [350] have further refined the prediction of bond changes. The graph-based model was, again by Coley *et al.*, further enhanced to predict a broader range of bond changes, demonstrating improved performance on the USPTO_MIT dataset which lacks stereochemical information. [273]

Sequence-based methods treat the reactants and products as textual sequences, typically employing SMILES notation. These methods adapt models from natural language processing to translate precursor sequences into product sequences, a technique initially described by Nam and Kim. [351] The efficacy of atom-wise tokenization for reactants and molecule-wise for reagents demonstrated that sequence-based models can effectively predict reactions and

handle stereochemical information when encoded in the sequence. [272] A key advantage of this approach is its ability to train on diverse data sets without the need for atom mapping, as exemplified by the Molecular Transformer, which remains the top-performing model on a benchmark data set, including those with stereochemical details. [274] This representation also eliminates the need to distinguish between reactants and reagents, a step that presupposes knowledge of the product and is not always feasible in all scenarios. Based on the Molecular Transformer, transfer learning (*i.e.*, re-using of knowledge learned from a task to boost performance on a related task) has been explored to enhance model applicability to specific reaction types. [352–354]

1.4.5 Reaction yield prediction

Reaction yield estimation has become an important tool for chemical engineering and chemistry, particularly in industrial processes where efficiency and cost-effectiveness are paramount. [355] While binary reaction outcome only covers whether the reaction takes place or not, yield prediction adds a quantitative measure of how well the reaction performs. [283] Successful estimations build on a deep understanding of the complex relationship between various reaction participants, including their stoichiometric amounts, the reaction concentration, temperature and time as well as the resulting output. [356, 357] As a consequence, reaction yield estimation can be approached as a regression problem, aiming to quantify the functional dependencies that dictate the efficiency of the reaction. [283] Given the specificity required for accurate predictions, most models are often tailored to individual reactions or closely related reaction families. [358] These predictive tools range from simple linear equations derived from principles of physical organic chemistry to sophisticated multi-variate and non-linear models that have emerged in recent years, boosting improved accuracy and broader applicability. [337, 358–361]

Recent advancements in ML have led to increased research activities in the field of reaction yield prediction. A collaboration between MSD and Princeton University pioneered these efforts in 2018 by creating a yield database from over 4000 C-N cross-coupling reactions, and training ML models with 120 descriptors, including molecular, atomic, and vibrational characteristics. [359] However, the featurization methods used sparked a debate regarding their transferability and effectiveness compared to non-chemical fingerprints. [362] Subsequent studies have sought to refine these models with Sandfort *et al.* introducing automated molecular fingerprints, [360] while Schwaller *et al.* utilized reaction SMILES with the molecular

bidirectional encoder representation Transformer (BERT) architecture, demonstrating superior performance and robustness, especially in data-scarce scenarios. [337] In addition, a performance comparison of computed descriptors and molecular fingerprints was conducted, favouring the latter for yield prediction. [361]

Despite these advances, challenges persist when data is limited, as evidenced by work on a small database of deoxyfluorination reactions, which yielded less accurate predictions but remained valuable for optimization. [363] The growing interest and capabilities in HTE promise an influx of data to further refine ML models. First explorations have applied active learning and NNs to enhance data efficiency and discover new reactivity patterns using fewer training instances using the examples of Pd-catalyzed Suzuki–Miyaura cross-coupling reactions. [364, 365] Very recently, Fitzner *et al.* evaluated the potential of ML to predict the yield of Pd-catalyzed C–N coupling reactions, where the training data was derived from chemical reaction databases. [358] The study revealed that the models are effective within the chemical space of the training data but struggle with generalization to new reactions, highlighting the need for more diverse data to improve yield predictions. Another study focusing on the same reaction type expedited the identification of substrate-adaptive conditions using NN models and relying on an experimental dataset that encompassed a wide array of reactant combinations and reaction conditions, demonstrated high efficacy in experimental validation. [366]

Yet, there remains a high need for standardized data recording to mitigate issues with noisy and incomplete data sets to make models capable of generalizing across the chemical diversity of reactants. [358, 367] The difficulty in yield prediction stems not only from computational limitations but also from a fundamental need to better understand chemical principles. This could be achieved by the use of classification models to better handle data variability. [356] Through enhanced data sets from synthesis automation, uncertainty-based predictions, and the development of chemically relevant, reaction-specific descriptors through detailed mechanistic studies, future improvements in the field are anticipated.

1.4.6 Regioselectivity forecasting

Selectivity, including regio-, site-, diastereo-, and enantioselectivity, is a crucial property in chemical reactions, often more challenging to achieve than binary outcome or yield due to the need for an understanding of competing reaction pathways. [283]

Advancements in predicting the regioselectivity of chemical reactions have been marked by the development of various ML models, each leveraging unique representations and descriptors. The RegioSQM protocol was a pioneering semi-empirical method for predicting electrophilic aromatic substitutions. [368] This method was later complemented by a GNN architecture using SMILES graphs and RDKit descriptors, offering faster performance. [369] The GNN approach was expanded to address broader regioselectivity challenges through a multitasking framework that combines atomic descriptors and quantum chemical data. [370]

In radical C-H functionalization, Li *et al.* demonstrated the efficacy of random forest models, particularly when using selected physical organic descriptors, for site selectivity prediction. [371] Banerjee *et al.* achieved over 90% accuracy in predicting difluorination outcomes on alkenes with a small dataset and expert-crafted descriptors. [372] Beker *et al.* utilized a large Diels–Alder reaction database to construct ML models, with a random forest model showing superior performance in predicting regio-, site-, and diastereoselectivity, particularly when employing physically meaningful descriptors. [373]

While underscoring the advances of recent developments in regioselectivity prediction, the limited amount of studies highlights the need for more sophisticated, generalized ML models that can be applied to a broader reaction scope. Lately, GNNs that have been partially applied for initial attempts in reaction outcome, yield and regioselectivity prediction have gained increased interest across drug discovery.

1.4.7 Graph neural networks for reactivity prediction

GNNs have emerged as a powerful class of deep learning methods tailored for graph-structured data, with significant implications for chemistry and drug discovery. [374] While chemists have utilized GNN-like algorithms for some time, [375–377] it is the recent advancements in NN design and the availability of large data sets that have propelled GNNs to the forefront, achieving state-of-the-art results in various chemical applications. [378, 379] These networks, particularly message-passing NNs, iteratively update node features through graph convolutional operations, effectively capturing the intricate relationships between atoms within molecules. [378]

In quantum chemistry, GNNs have been adept at predicting molecular properties by incorporating 3D spatial information, such as radial and angular data, into the graph's edge features. [380–382] In drug discovery, GNNs have surpassed traditional human-engineered molecular descriptors in predicting biologically relevant properties, with their performance being relatively unaffected by the inclusion of single or multiple molecular conformers during network training. [383–385] Their inherent compatibility with molecular structures makes GNNs particularly advantageous for explainable AI applications, aiding in the interpretation of models predicting molecular properties of pre-clinical and quantum chemical significance. [386, 387] Moreover, GNNs have been instrumental in *de novo* molecule generation, simulating the stepwise construction of molecules through node and edge additions. [388–390]

In parallel, GNNs have also been established for chemical reaction planning, including retrosynthesis planning, regioselectivity- and reaction product prediction, originating from small substrates and culminating in the synthesis of complex drug molecules. [207, 248, 271, 348, 370, 385, 391] Literature indicates that models trained on activation energies derived from transition-state geometries can accurately forecast competing reaction pathways. [392–394] Incorporating graph-based features with properties calculated at the density functional theory (DFT) level has been shown to enhance regioselectivity predictions in electronically driven reactions. [395] Furthermore, the integration of graph-based ML with HTE data has facilitated the refinement of reaction conditions for C-H activation in organic molecules. [396] Nevertheless, the applicability of these models is currently constrained by their focus on smaller molecular structures, presenting a challenge for their use with complex, drug-like molecules. [278] A significant knowledge gap persists regarding the impact of steric and electronic influences on model accuracy for C-H activation, particularly in the context of regioselectivity in compounds with multiple aromatic rings. This could be approached by identifying or generating reaction data sets that contain transformations of large, drug-like or drug molecules with a vast set of functional groups.

1.4.8 Reaction data availability

The availability of curated, complete and trustworthy reaction data sets containing successful and failed transformations is crucial for reaction prediction and retrosynthesis tasks. [158, 243, 244, 358, 367] Data sets extracted from US patents such as USPTO_MIT, USPTO_STEREO, and USPTO_full provide a range of reaction records with varying levels of stereochemistry and reagent information. [157, 270, 272, 352, 397, 398] Specialized data sets include NameRXN-generated reactions and reaction super classes with high-quality atom mapping. [399–401] Schwaller *et al.* introduced the USPTO 1 k TPL for reaction classification, [338] and commercial databases like Pistachio, Reaxys, SciFinderⁿ, and Science of Synthesis offer partially curated reaction data. [402–405] However, access to these databases is often restricted, and literature-extracted data sets face challenges with structural representation, missing reaction conditions and biases toward successful reactions. [337, 367, 406]

In 2021, a novel ML-based method for removing incorrect entries from chemical reaction data sets, enhanced the predictive quality of deep learning models in organic chemistry, as demonstrated by improved metrics in retrosynthetic models trained on the cleaned data. [407] Just recently, HTE analyser, a versatile and statistically sound framework that reveals interpretable correlations within HTE data sets, validated by uncovering significant hidden relationships in over 39'000 proprietary reactions, including cross-couplings and chiral salt resolutions, was presented. [242] In general, reaction data sets derived from HTE platforms offer an alternative, focusing intensively on the impact of different conditions on the yield or selectivity, but often only focus on a specific reaction or transformation type. [283, 358, 359, 363] In addition, the ORD [245] and the UDM [246] are two recent initiatives designed to standardize and centralize chemical reaction data in machine-readable formats. Both initiatives aim to enhance the accessibility and utility of reaction data across the scientific community.

UDM and ORD initiatives are pivotal in advancing the standardization of reaction data, yet they introduce complexities that can impede data entry and utilization in laboratory and data science settings due to their extensive documentation requirements and the need for IT proficiency. These systems also present obstacles to interdisciplinary data sharing, as their specialized formats are not inherently accessible to non-experts, and their intricate data structures can complicate the direct exchange of information between researchers. Hence, the integration of accessible data methodologies in chemistry is vital for the progressive enhancement of ML implementations in the field. [244]

Don't let fear of failure hold you back; let the anticipation of success propel you forward.

- Jan Frodeno

2

AIMS OF THE THESIS

LSF is a promising methodology to alter complex molecular structures in the LO stage of drug discovery delivering novel chemical matter that supports the understanding of SARs. An increasing amount of new LSF transformations are being disclosed year by year. HTE has become an established technology to efficiently assess an array of reaction conditions in plate format leveraging automation equipment to streamline synthesis. Reactivity prediction using machine learning (ML) methods receives increased interest from the chemistry community as it allows the estimation of reaction outcomes before carrying out resource and time-intensive wet lab experimentation. Interestingly, a connection between all three disciplines has not been made yet, despite the potential of making LSF a more efficient methodology to enable fast drug diversification and, consequently, speed up the development of novel medicines.

To interconnect the three research fields, LSF, HTE and reactivity prediction, the seamless integration of automation, digitalization and ML is needed. Thus, this thesis focuses on: (a) Setting up an LSF-tailored semi-automated HTE system that is capable of systematically assessing relevant methodologies in drug-like chemical space by automating and digitalizing typical procedures in the laboratory to efficiently design and execute reaction screening plates; (b) developing a simple, human- and machine-readable format that captures literature and

experimental reaction data to enable data analysis, systematic experiment designs and ML applications; (c) applying the platform (a) and reaction format (b) to two relevant LSF reaction types for drug discovery within the drug-like chemical space to generate high-quality reaction data that enable the assessment of the potential of ML tools for reactivity prediction.

(a) Set up of a digital LSF-tailored semi-automated HTE system:

- Conceptualizing, developing and implementing an end-to-end data-orchestrated semi-automated laboratory platform that allows the systematic and efficient assessment of LSF methodologies on complex drug-like substances.
- Integrating an automated liquid handler into the workflow capable of automatically setting up reactions in parallel and preparing process controls to ensure high sample quality for accurate analytical measurements.
- Establishing data workflows that serve as the backbone of the platform to allow seamless literature analysis, plate design, reaction execution, sample analysis, reaction outcome analysis, reaction data management, as well as the planning, execution and analysis of scale-up reactions.

(b) Development of the simple, user-friendly reaction format (SURF):

- Analyzing current reaction data sharing practices and assessing the reaction data format landscape to develop an understanding of the existing bottlenecks and of the needs that a new format should fulfill.
- Designing and implementing a new human- and machine-readable reaction data format that overcomes the identified gaps to feed ML algorithms without the need for pre-cleaning seamlessly.
- Curating reaction data from selected publications into the new reaction data format to enable systematic reaction screening plate design to provide high-quality HTE reaction data sets in the drug-like space using (a).

(c) Combination of (a) and (b) on selected LSF reaction types to enable reactivity prediction:

- Assessing the applicability of C-H borylation for late-stage drug diversification by carrying out hundreds of HTE reactions on systematically selected commercial drugs with a broad range of reaction conditions to enable the development of an ML algorithm that predicts binary reaction outcomes, yields and regioselectivity for novel substrates.

- Running a library-type screening approach aimed to explore the substrate landscape for late-stage C-H alkylations to facilitate the *in silico* prediction of suitable substrates that can be coupled with a diverse set of sp³-rich building blocks using Minisci-type chemistry.

Through approaches (a), (b), and (c), the overall aim of this thesis is to contribute to improving compound synthesis efficiency in early drug discovery through the systematic application of laboratory automation and artificial intelligence.

I am made of all the days you don't see, not just the ones you do.

- Jan Frodeno

3

SEMI-AUTOMATED LSF SCREENING PLATFORM

This chapter describes the set-up of a semi-automated and data-driven high-throughput experimentation (HTE) screening platform to systematically evaluate late-stage functionalization (LSF) reactions on complex drug-like molecules to increase synthesis efficiency. The platform was coined DOLPHIN (**D**ata-**o**rchestrated **l**aboratory **p**latform **h**arnessing **i**nnovative **n**eural networks). The following sections will describe the design, development, and implementation process of DOLPHIN, highlighting all important features of the system.

3.1 Approach and concept

As highlighted in Chapter 1.2, LSF can capitalize on the usually abundant presence of C–H bonds within complex molecular frameworks to streamline the derivatization process, thus eliminating the need for *de novo* synthesis, the introduction of functional handles or the development of a protecting group strategy. [54, 55, 66] However, the presence of functional groups and the array of C–H bonds, varying in bond strength and influenced by electronic and steric environments, complicates the direct application of LSF. [55] Hence, the application of general reactivity and selectivity principles for LSF must be approached cautiously, [61] and

potential wrong judgements of the envisaged reactivity can clash with the limited time and resources available in medicinal chemistry projects. As a consequence, an integrated approach combining automation, digitalization and artificial intelligence (AI) to efficiently apply LSF in drug discovery was developed.

The concept of DOLPHIN is centered around semi-automated miniaturized HTE screening to evaluate the applicability of LSF methodologies on a lead structure before conducting resource-intensive single reactions (Figure 3.1). In the first step, all important information on the molecule of interest is collected and, importantly, captured in a digital, machine-readable format (Chapter 3.2.3). Desired chemical transformations, *e.g.*, borylation or fluorination, are determined and screening plate layouts designed. Desirably, the choice of methodology and plate designs will be guided by machine learning (ML) models that are capable of assessing the reactivity *in silico*, thereby further reducing reaction failures. Initial successful case studies covering the application of reactivity predictions are described in subsequent chapters of this thesis (Chapters 5 and 6).

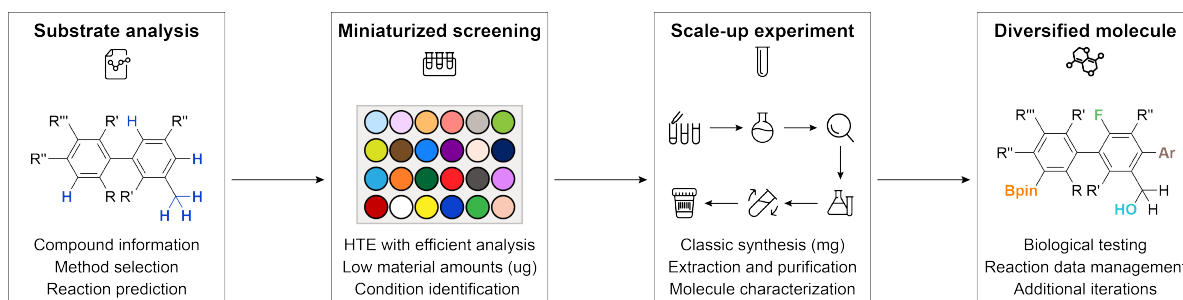


Figure 3.1: Overall approach of DOLPHIN to increase the efficiency of LSF campaigns. Potential LSF substrates are analyzed and the compound information is captured digitally. With increasing amounts of experimental data from the two consequent steps, miniaturized reaction screening and scale-up, reactivity prediction can be carried out to guide methodology and screening plate selection. The chosen transformations are then tested in small-scale parallel screening experiments with minimal material consumption to determine potential LSF products rather than running single experiments that provide little information on reactivity and require more amount of the compound. All data is reported in a structured format to aid the decisions for scale-up experiments through data analysis and to feed ML algorithms. Scale-up experiments are carried out on a standard medicinal chemistry scale to isolate sufficient material for full analytical characterization. Those analyzed molecules can then undergo biological testing, delivering SAR information that can initiate additional iterations or further refinement of LSF reaction conditions.

Cornerstones for reactivity prediction are the next two parts of the approach, which cover the actual experimental work to create new chemical matter in the laboratory. The already mentioned miniaturized HTE screening only utilizes small amounts of complex drug-like molecules to aid the identification of suitable LSF methodologies and conditions. Importantly, an efficient and data-driven workflow is needed to rapidly identify and analyze reaction out-

comes. Those results also need to be made directly available in a machine-readable format for ML applications (Chapter 4). Once, screening hits are identified, the scale-up experiments are conducted on a standard medicinal chemistry scale, typically mg amounts, to obtain and fully characterize the products of the LSF transformations. It is critical that these experiments are well connected with the previous screening workflow to avoid the use of unfavourable conditions and also report back the exact structure of the new chemical matter obtained (Chapter 3.3). Finally, the novel-obtained analogs can undergo biological testing to determine pharmacokinetic and physicochemical properties. Depending on these outcomes, new iterations of the process can be initiated, which might involve the assessment of different LSF methodologies or a more granular screening of reaction conditions to establish a highly efficient synthesis of the target molecule.

Breaking down the comprehensive four-step concept (Figure 3.1) into its essential steps and components led to the development of the DOLPHIN workflow depicted in Figure 3.2. In the initial step, the literature needs to be searched systematically for LSF reactions to avoid missing out on interesting transformations. This initial (1) *Method assessment* step is usually carried out by chemists through reading a review, which generally delivers a comprehensive overview of a specific reaction type. Yet, generally, no clear methodology for selecting the publications in the review is described. [64, 66] Therefore, DOLPHIN aimed at establishing a method that allows for a systematic literature analysis (Chapter 3.2.1, Figure 3.4), which is already conducted in many other scientific disciplines, *e.g.*, medicine or management, [408–413] to date. This overall process was termed the systematic assessment of chemical transformations (SACT), which also includes the extraction and curation of the reaction data from publications of interest (Chapter 3.2.1, Figure 3.5) to obtain high-quality data sets in a simple, user-friendly reaction format (SURF, Chapter 4) that can undergo data analysis aiding the next step of the workflow, the (2) *Plate design*.

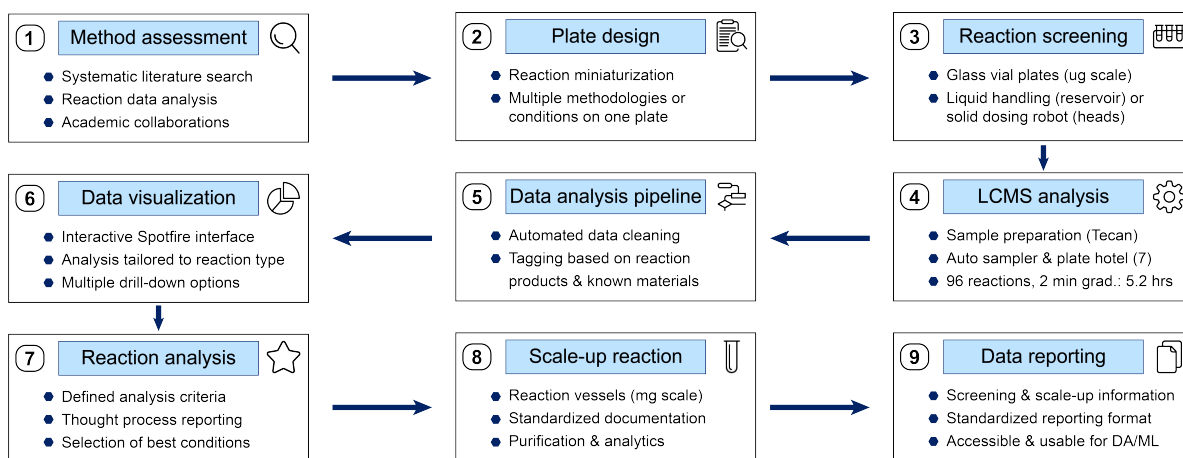


Figure 3.2: Overview of the DOLPHIN workflow. Systematic methodology assessment (1) allows efficient literature search and reaction data analysis to aid the design of screening plates (2) on a miniaturized scale covering a wide range of reaction conditions. Reaction screening (3) of the plates is executed using automated liquid and solid handling technology. At defined time points, samples are prepared and analyzed by liquid chromatography-mass spectrometry (LCMS) using automated equipment (4). The LCMS plate hotel can carry up to seven plates. The analysis of the LCMS raw data is achieved by a data analysis workflow (5) that accesses recorded and generated information on reaction input (starting materials) and potential outcomes (products). The output of the enriched data in SURF allows visualization of reaction outcomes (6) and aids reaction analysis (7). Promising conditions are scaled up in standard reaction vessels on a typical synthesis scale (mg) to obtain and characterize products (8). Standardized documentation practices (9) allow for data analysis (DA) and machine learning (ML) applications (9).

With the structured reaction data at hand, efficient analysis of employed reaction components, *e.g.*, catalysts, reagents or solvents, including their quantities and parameters, *e.g.*, temperature, time and atmosphere, can be carried out. With structural information on all components, especially those of starting materials and products, and quantitative reaction outcomes available, conclusions on reactivity relationships can be drawn as well. In terms of plate design, the analysis of such data sets reveals the influence of specific reaction components and also pin-points parameters that have not been optimized or evaluated so far (Chapter 3.2.1, Figure 3.3). Further, the information on scale and molarity can be used to guide reaction miniaturization (Chapter 3.2.2, Figure 3.6).

Once the plate is designed and the layout documented (Chapter 3.2.4, Figures 3.8-3.9), reactions are miniaturized and the (3) *Reaction screening* is carried out on an ug-mg scale in glass vials on 24- or 96-well aluminium plates with stirrers to ensure sufficient mixing of all components. The set-up of the reactions is supported through automated liquid (Tecan EVO 100) and solid (CHRONECT Quantos) handling of the required chemicals, which reduces error-prone and repetitive tasks for the scientist in the lab, accelerating the overall process. At time points of interest, samples are drawn from the plate using multichannel pipettes or the liquid handling system for (4) *Liquid chromatography-mass spectrometry (LCMS) analysis*. The sample preparation for the analytical systems is executed by the liquid handler, which delivers solutions of defined concentration in deep well plates to be inserted in the plate hotel of the LCMS. Analysis of the samples using a two-minute gradient conducted by auto sampling generally requires slightly more than five hours.

To avoid manual interpretation of the enormous amounts of data generated through the high-throughput system, a (5) *Data analysis pipeline* was built. This process takes care of cleaning and curating the LCMS raw data into a structured, tabular format. This data is then interconnected with the information on the reaction, ranging from the starting material to the potential products (Chapter 3.2.5, Figure 3.10). A comprehensive database, which will be further explained in Chapter 3.2.3, is key to this part of the workflow and allows the automated assessment of reaction outcomes through mass, and where known, retention time of components observed in the reaction sample. The output of the tagged data in the SURF (Chapter 4) supports the rapid (6) *Data visualization* of the reaction outcomes including all relevant reaction parameters in an interactive TIBCO Spotfire interface (Chapter 3.2.5, Figures 3.13-3.16). The tool allows a streamlined assessment based on the reaction types and several detailed analysis options.

Using Spotfire, (7) *Reaction analysis* of the screened plates based on defined criteria is conducted and the rationale of method or condition selection is reported. The most promising conditions are then transferred to another tool, which guides the set-up and execution of the (8) *Scale-up reaction*. Scale-ups are carried out on a standard medicinal chemistry scale, usually mg scale, to obtain and fully characterize the reaction products. A standardized documentation procedure was developed to make the data connectable with the screening outcomes (9). This also enables the export of all reaction data from screening to scale-up in SURF, allowing for in-depth data analysis and ML that can guide future DOLPHIN campaigns through plate design and reactivity prediction.

In the following sections, important parts of the DOLPHIN workflow are explained in more detail. These include the design of the screening plate (Chapter 3.2.1), the reaction miniaturization (Chapter 3.2.2), the reaction screening workflow (Chapter 3.2.3), the data structure (Chapters 3.2.4 and 3.2.5), the reaction analysis pipeline (Chapter 3.2.5) and visualization (Chapter 3.2.6), and the scale-up workflow (Chapter 3.3).

3.2 Reaction screening

The miniaturized reaction screening represents the core of the DOLPHIN workflow as it allows the efficient assessment of LSF transformations before initiating resource- and time-intensive single reactions with precious lead molecules. The following chapters will describe the main operations and features of the screening process.

3.2.1 Plate design

In contrast to traditional chemical experimentation, where the literature review aims at selecting a few testable reaction conditions, followed by single-flask laboratory experiments to identify and potentially isolate reaction products, the parallel assessment of chemical transformations with HTE requires a more streamlined process. [110] The manual iterative literature search process without digital support and documentation of data can be very time-consuming, especially when facing complex optimization challenges or a large number of different methodologies. Designing a screening plate with a clear rationale backed by a broad literature search will enhance the ability to identify successful reactions, discover patterns and support potential further iterative rounds of screening. To achieve this goal, a novel plate design process was developed, which is depicted in Figure 3.3.

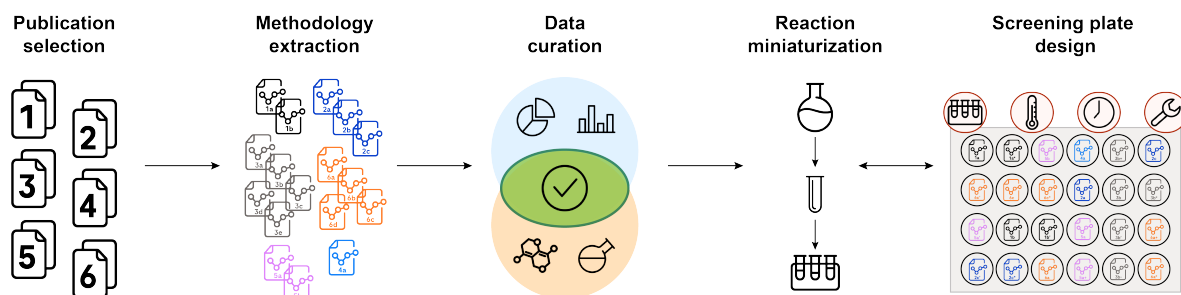


Figure 3.3: The process of designing a screening plate. At first, the literature needs to be searched for publications of interest in a streamlined fashion. From those manuscripts, the reported methodologies and reaction conditions need to be extracted and transferred to SURF. After potential manual curation of the data to avoid information gaps, the data can be analyzed to identify the most promising parameters for a screening campaign, always balancing data and chemical understanding. In the following, the reactions need to be miniaturized and extensive testing on the small scale needs to be carried out to ensure reproducibility in HTE experiments. Once, miniaturized, the plate layout can be designed, including the consideration of plate-specific characteristics, *e.g.*, all reactions on the plate will be exposed to the same temperature range, and engineering challenges (liquid handling, solid dosing, solubility issues).

In the initial step of the workflow, relevant publications covering a certain reaction type, *e.g.*, C-H borylation, or specific chemical transformation, *e.g.*, Suzuki-Miyaura, need to be identified. This approach will be further explained below (Figure 3.4). In the next step, the reaction data from those publications needs to be extracted to obtain high-quality data sets for analysis to uncover relevant screening conditions (Figure 3.5). This will consequently involve the curation of the data to close all remaining information gaps. Importantly, an understanding of the correlation between the molecular process and the corresponding reaction data is needed. This delivers the foundation for the reaction miniaturization process, which follows in the next step. Selected transformations will be tested on a smaller scale (μmol , nmol) than reported in the literature and transferred to a plate format. Further details on this step are explained in Figure 3.6. Once the reactions are validated, the design of the final screening plate can be defined, taking into account the special requirements associated with working in a parallel plate format. Those include a constant temperature for all vials and, generally, also identical time points for process controls. Further, engineering challenges need to be addressed, which can comprise the handling of a broad range of liquids and solids, the insolubility of certain components, solvent volatility and overall small amounts of the reaction mixture that need to be consistently stirred.

Figure 3.4 describes the developed systematic literature analysis for DOLPHIN. The five-step data-driven procedure aims at identifying publications containing methodologies of interest using a clearly defined rationale.

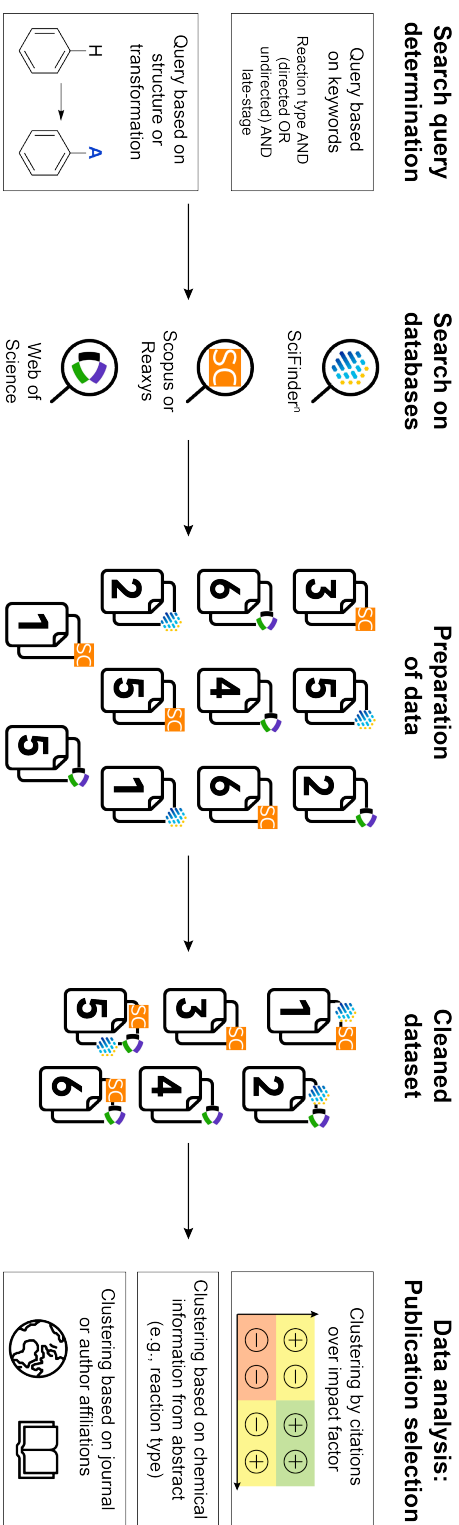


Figure 3.4: Overview of the systematic literature analysis. Upon definition of a search query, either through a string syntax or a structure-based approach, three major databases, namely, SciFinder[®], Scopus/Reaxys and Web of Science, are scanned. The obtained results are downloaded, and the data is cleaned by removing duplicate publications and enriched with additional information, e.g., citation count, and journal impact factor. The cleaned data set then undergoes data analysis to aid the selection of relevant publications using different types of clustering approaches. The figure only shows exemplary analysis options, further options are feasible.

In the first step of the literature analysis, a definition of precise search queries needs to be conducted. Generally, two types of queries can be considered for chemistry, a keyword- and a structure-based approach. Using keywords, types, names, parameters and characteristics of reactions as well as involved functional groups or reactive centers in string format can be interconnected through Boolean operators to deliver the syntax. In a structure-based approach, the query is centered around a transformation described by structural information, usually by drawing molecules in a sketch application provided by the database. To allow a comprehensive literature search, the search is executed on three different, renowned scientific databases, namely Scopus/Reaxys (Elsevier, Netherlands), [403, 414] Web of Science (Clarivate Analytics, USA), [415] and SciFinderⁿ (Chemical Abstracts Service, USA). [404] This approach helps to balance the strengths and weaknesses of each database, which include the number of records, list of titles as well as the focus on certain topics. [416]

The search results obtained from each database are downloaded and the data set is cleaned, removing duplicate entries and adding a database source tag. Consequently, the data is curated, additional information *e.g.*, the citation count or the journal impact factor, is added and calculations with the available data can be executed. The cleaned literature data set then undergoes data analysis to identify the most suitable publications for detailed inspection. Different analysis approaches can be chosen, including clustering by citations and impact factor, by chemical information in string format occurring in the abstract or by clustering based on journal and author affiliation. Of course, further means of analysis can be freely chosen as the data set contains a broad range of information for each publication. In Chapter 5.3, an example of the application of this literature analysis as part of the borylation case study using a keyword-based search query is described in detail. Further, a reference implementation of the systematic literature analysis (Figure 3.4) based on a literature search for Minisci-type reactions with the respective Alteryx Designer (Irvine, US) workflow is available at <https://doi.org/10.5282/ubm/data.469>. The information provided in this chapter also touches on the next step of the plate design process, the extraction and curation of the reaction data from the selected publications, which is displayed in Figure 3.5.

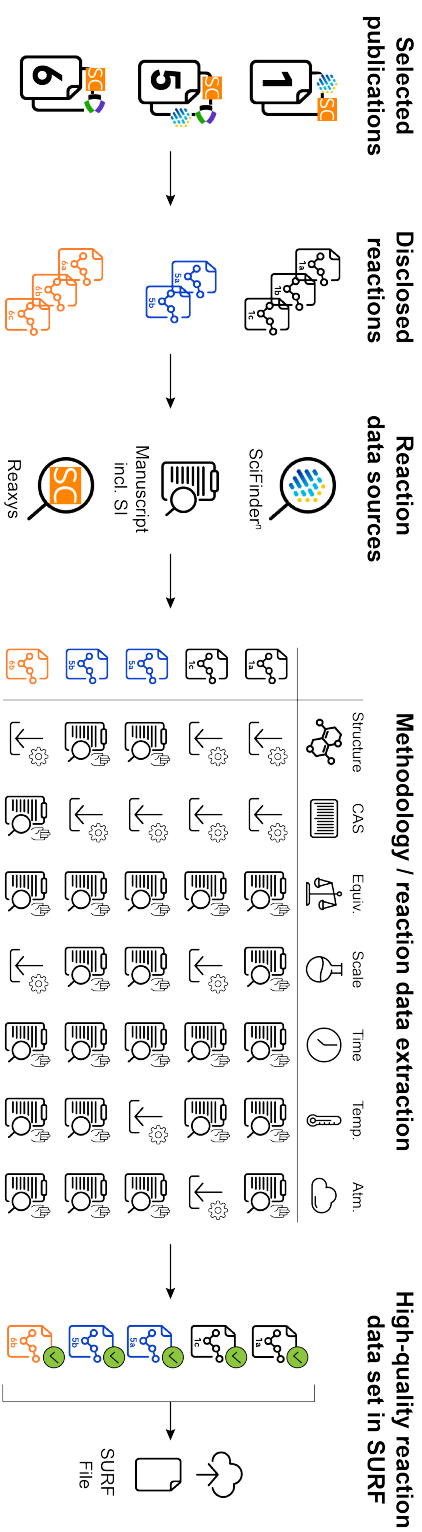


Figure 3.5: The reaction data extraction process. Each selected publication contains a certain number of reactions, of which the data is extracted using common databases and the procedures obtained from the manuscript including the supplementary information (SI). To have complete data sets available that are usable for ML applications and data analysis, a broad range of information describing the chemical reaction needs to be extracted. Those include identifiers and equivalents (equiv.) of all reaction components, scale, time, temperature (temp.) and atmosphere (atm.). The data is reported in the simple user-friendly reaction format (SURF) and once a reaction is fully curated, the data is stored in an internal database.

Automated reaction data extraction from publications has progressed from rule-based algorithms to advanced data-driven approaches using neural networks, with recent tools like Deep Learning for Chemical Image Recognition (DECIMER) achieving up to 90% accuracy in optical chemical structure recognition (OCSR). [417–421] However, the extraction from full chemical reaction schemes remains a complex challenge, with recent developments by Qian *et al.* [422] introducing an image-to-sequence translation model for parsing reaction schemes, indicating the potential for further improvement with additional annotated data. [423, 424] Despite these recent advances, the extraction of all information describing a chemical reaction, including important parameters, such as the quantities of the reaction components, into a structured format remains a challenge.

In a pragmatic approach, the DOLPHIN workflow gathers reaction information from different databases, namely SciFindern [404] and Reaxys, [403] complemented by manual analysis of the manuscript including the supplementary information (SI). To obtain all required information for SURF (Chapter 4), structural data, identifiers (CAS number) and quantities of all reaction components are extracted and curated. Further, important parameters, such as scale, time, temperature and the atmosphere need to be obtained. Unfortunately, many of these data points are not obtainable through the mentioned databases' online view, despite as a downloadable file and if they are, the information is not always in line with the one reported in the publication. Therefore, as exemplified in Figure 3.5, oftentimes, only manual extraction of parameters from the publication leads to complete data sets. In particular, the quantities of the reaction components and data on scale or atmosphere are not readily available and even require a detailed analysis of the manuscript in many cases. However, once the data of each transformation is captured in SURF, the analysis process of reaction data is accelerated, plate layouts can be designed more efficiently and the reaction information can also be directly subjected to ML models. An exemplary SURF dataset of Minisci literature conditions is available at <https://doi.org/10.5282/ubm/data.469>, additional data sets can be downloaded from <https://github.com/alexarnimueller/surf/tree/main/data>. The applications of DOLPHIN in Chapters 5 and 6 describe the design of the plate layout for C-H borylation and alkylation reactions. However, before being able to execute the screening in the laboratory, the reproducibility of the literature reaction on conditions on a small scale needs to be assessed.

3.2.2 Plate testing

The reaction miniaturization workflow is an iterative step-by-step process depicted in Figure 3.6. To avoid the generation of false positive or negative data and consequent misinterpretation of reaction outcomes, potentially leading to error-prone training sets for ML, this process requires special attention. In the initial step, the reaction conditions from the literature are reproduced as reported in the manuscript. If successful, the reaction scale can be reduced (μmol , nmol scale) subsequently. This can be conducted stepwise to avoid large jumps in scale that could have a strong impact on the reaction performance. Confirmation of positive results initiates the next step, the transfer of the single reaction to the plate format.

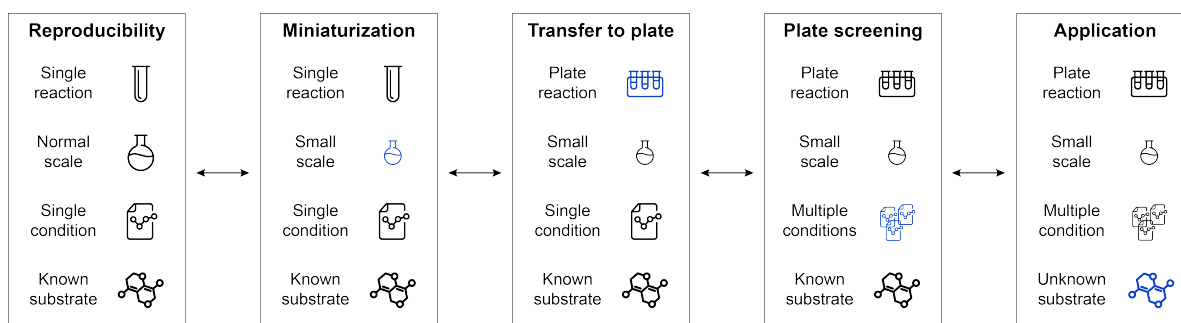


Figure 3.6: Overview of the reaction miniaturization and plate validation process. Starting from literature conditions, which are being exactly reproduced, a five-step workflow leads to the development of a robust and reproducible plate layout that can be applied to unknown starting materials. In every step of the iterative optimization, only one parameter is varied at a time to allow seamless tracing of potential errors.

To prove reproducibility across all vials of the screening plate, an experiment using a 24-well plate, in which all wells contain the same single reaction conditions tested previously, is executed. This also allows the identification of potential issues through inconsistent heating or stirring across different positions. In the case that all reactions show similar, expected product formation, reaction conditions can be varied in the subsequent iteration round. This generally involves the variation of catalysts, solvents or reagents or the addition of other methodologies to the plate that already passed the described assessment as well. Finally, if the designed plate delivers reproducible results for several experiments using a defined set of model substrates, the plate can be rolled out on an unknown starting material. If reactions fail or inconsistencies are observed throughout the whole process, the step will be repeated until reproducibility and robustness can be confirmed.

The above-described, meticulous, iterative process of designing and validating screening plates with data-driven decision-making ensures the screening of relevant, robust and reproducible reaction conditions to generate high-quality reaction data.

3.2.3 Screening workflow

The following chapter describes the experimental screening plate workflow and the interplay with the associated databases stored as interactive Google Sheets or on Google Cloud, which represent the core of DOLPHIN. An overview of the screening operations including the required data interplay (blue text), highlighting the importance of seamless software and hardware integration, is shown in Figure 3.7.

Upon the selection of the substrate for the screening campaign, all information on the compound is digitally documented in the Google Sheet of the screening platform (see "Compound data" in Figure 3.8 for details). The compound undergoes analytical quality control by liquid chromatography-mass spectrometry (LCMS) to confirm the purity of the material and capture mass pattern as well as the retention time, required for the data analysis of the screening reactions at a later point in the workflow. Next, the screening plates are selected and the consequent experimental runs are connected through unique identifiers to the substrate within the Google Sheet ("Plate definition" and "Experiment information" in Figure 3.8). Based on this information, the material preparation can be initiated as the preparation protocols for stock solutions (liquid handling) or dosing heads (solid handling) are generated. Once the reaction components are prepared, stirring bars are added to each well of the plate, and the robotic systems execute the reaction set-up using the generated files. Liquid handling with the Tecan robot is faster compared to solid dosing, where each component needs to be milled into the respective vial. This leads to set-up times of ten to 15 minutes for a 96-well plate using liquid handling and up to ten hours, depending on the amount of material and reaction participants, for solid dosing.

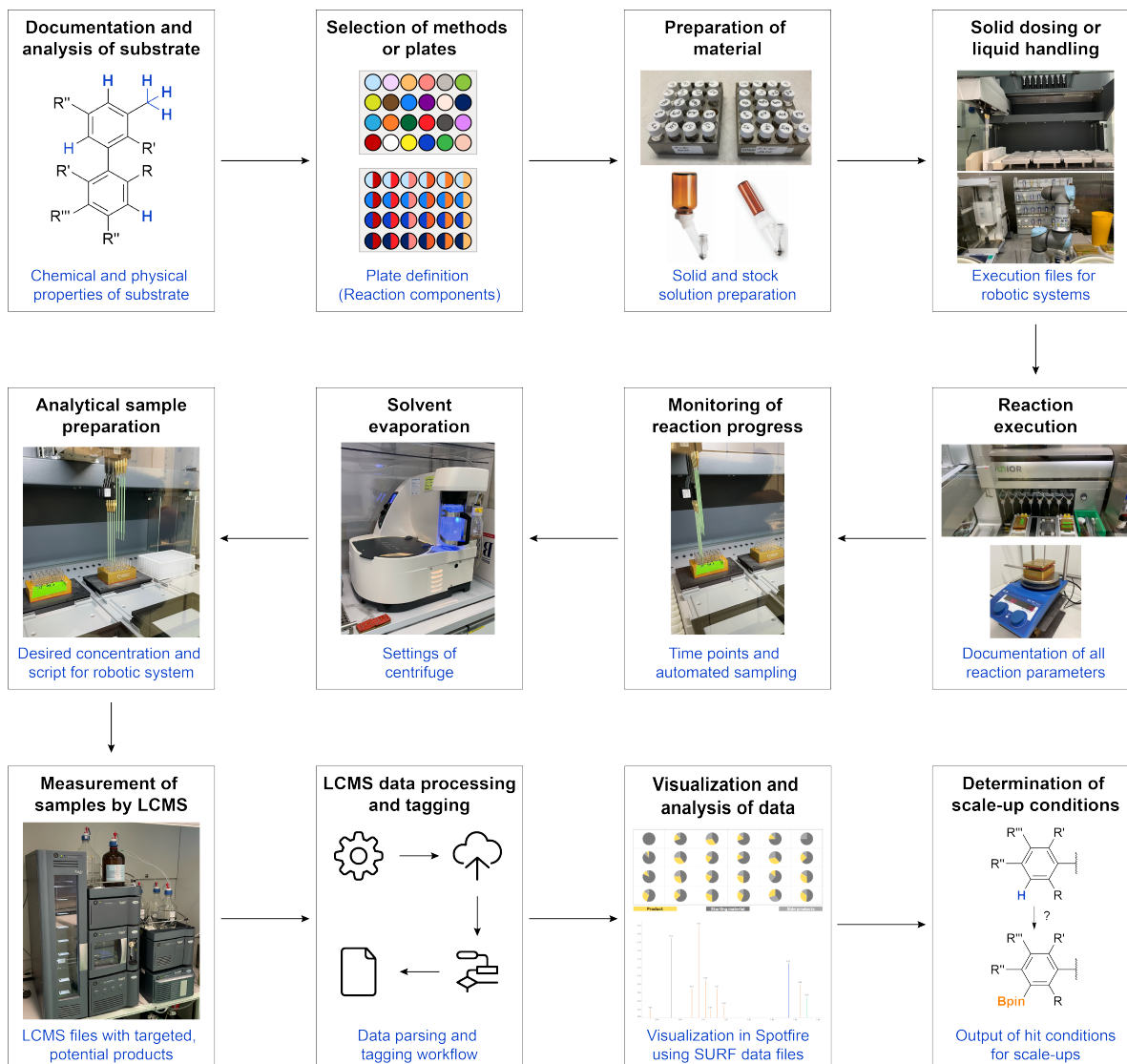


Figure 3.7: The reaction screening workflow of DOLPHIN. All steps require data operations in the background, which are highlighted in blue text. After digital documentation and analytical characterization of the starting material, screening plates of interest are selected and the required chemicals for automated liquid using a Tecan EVO 100 or solid handling with a CHRONECT Quantos are prepared. Reaction set-up is conducted using robotic systems, which are controlled through automatically generated execution files. The solid dosing of components takes place under a nitrogen atmosphere inside a glovebox. The reaction mixtures are stirred and, if required, heated on standard heating plates or a reaction bay within the glovebox. Sampling from the plate at defined time points allows the monitoring of reaction progress. The drawn samples undergo a defined preparation process, which includes the removal of the reaction solvent using centrifugal evaporators from GeneVac and automated sample re-suspension and dilution steps by the Tecan, before being subjected to liquid chromatography-mass spectrometry (LCMS) analysis. The LCMS raw data is processed, and, using the compound database, tagged based on mass patterns and retention times. Visualization of this structured and complete data provided in the simple user-friendly reaction format (SURF) allows the analysis of reaction outcomes. Those results guide the determination of suitable scale-up conditions.

Once all reaction components are dosed, the reaction plates from Analytical Sales are sealed and transferred to either a standard stirring plate with a heating function or a heating bay within the glovebox. Importantly, all reaction parameters, including the often unreported or estimated values, such as atmosphere and time, are digitally documented. The monitoring of reaction progress is achieved through sampling from the reaction plates using multichannel pipettes or the liquid handler. Time points are reported and samples are subjected to solvent evaporation in a centrifuge. Tecan workflow scripts that aid the re-dilution and mixing of the samples in a defined amount of LCMS solvent (MeCN:H₂O, 4:1) to obtain accurate sample concentrations (1 mM) across all wells are generated based on the scale of the reaction. Following transfer into deep well plates, the reactions are analyzed on the LCMS system using automatically generated files that include the chemical formula of the starting materials as well as desired products and undesired byproducts, thereby guiding the mass spectrometry (MS) search. The product formulae originate from a data workflow that uses the information of the starting material and the defined transformations on the plate to automatically construct these potential products (Chapter 3.2.5, Figures 3.9 and 3.10).

The LCMS analysis of 96 samples requires approximately 5.2 hours, after which the raw data is parsed into a tabular format and transferred to a Google Cloud. This structured data is then used in a product tagging workflow (Chapter 3.2.5, Figures 3.9 and 3.11) that compares the information in the database with the reaction data from the LCMS. Starting materials and products are tagged based on mass (pattern) and, if available, retention times. The processed data is transformed into SURF (Chapter 4), which allows visualization of the reaction outcome in an interactive Spotfire interface (Chapter 3.2.6, Figures 3.13-3.16) and the training of ML models. The selection of screening hits is documented and scale-up reactions are seamlessly determined using another customized workflow and Google Sheet database (Chapter 3.3, Figures 3.16). The following two Chapters 3.2.4 and 3.2.5 will give a more detailed insight into the data structure and two important workflows that aid the analysis of the reaction mixtures.

3.2.4 Data backbone

The data structure of the interactive Google Sheet that captures the screening operations of DOLPHIN is shown in Figure 3.8. Three main sections, which are interconnected through unique identifiers, were defined to record all relevant HTE campaign data. Exemplary data sets covering the data structure are available at <https://doi.org/10.5282/ubm/data.469>.

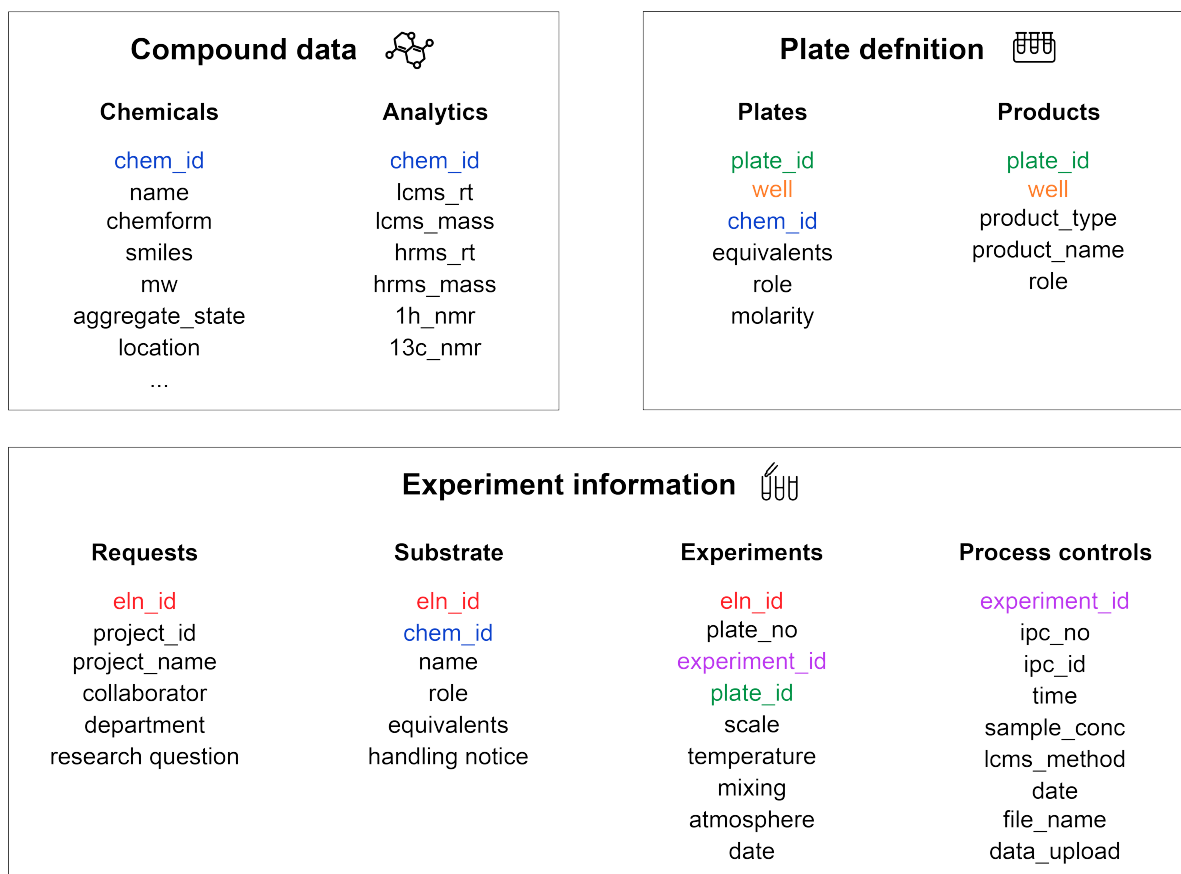


Figure 3.8: Data structure of the semi-automated screening platform. Three main components are needed to define the screening experiments. Compound data captures all important information related to a molecule from structural data and physical properties to analytical characterization and the chem_id serves as the unique identifier. The plate definition section contains all relevant information on the plate layout, describing which chemical in what quantity and role is in a certain position of the plate. Finally, the experimental section makes use of the aforementioned databases to define which chemical reactions are taking place under what type of conditions, including temperature, scale, and atmosphere among others. Further, time points for reaction monitoring are captured, allowing the precise evaluation of transformation outcomes.

The compound data section includes all information on the chemicals including the name, the CAS number, the chemical formula in Hill notation, structural information in string format (Simplified molecular-input line-entry system, SMILES), the molecular weight, the aggregate state, the solubility in certain solvents and, importantly, the location of the chemical in the laboratory. The chem_id serves as the unique identifier of the chemical, used in other tables to cover analytical information or the use of the compound as a starting material or reagent on a plate. The analytics tab, which also belongs to the compound data section contains all LCMS, high-resolution mass spectrometry (HRMS) and nuclear magnetic resonance (NMR) spectroscopy information of the compound. This allows the retrieval of analytical information for the tagging of the LCMS reaction screening data but also supports the scale-up experiments (Chapter 3.3).

The definition of the plates is captured by two sub-tabs, namely plates and products (Figure 3.8), which are connected through the identifier `plate_id`. The plates tab records the components of the plate in each well, including their role and equivalents or molarity, which are important for the plate execution and the documentation in SURF. The products tab defines the expected reaction outcomes of the plate, which depends on the selected methodology and is documented as the atomic difference between the starting material and the product. In the case of the C-H borylation (Chapter 5), the formation of a boronic acid implies the addition of a boron atom, two oxygen atoms and one hydrogen atom compared to the substrate. The associated workflow that relies on this data is described in Figure 3.10.

Finally, the experimental information section combines and adds data to define each unique screening experiment. The requests tab captures information on the screening campaign including data on the project, the collaborating colleagues or the general research question. For each experiment, an electronic lab journal (ELN) entry is generated, which serves as the overall request identifier, the `eln_id` (`elnXXXXXX-XXX`). The next part of the experimental section covers the information, on which starting material is used, captured through the corresponding `chem_id`. While the name is retrieved from the compound database to reduce tab switching in the interactive Google Sheet, other novel information pieces, including the equivalents, generally 1.0, or specific handling notices are added. In the experiments section, the plates associated with each `eln_id` are defined. Since several plates (`plate_no`, `Y`) could be run for a certain request, *e.g.*, two borylation plates and three alkylation plates, an `experiment_id` (`elnXXXXXX-XXX_Y`) is needed to keep the data structure unique. In the example case, five entries would be added to this database, generating five plate numbers (1-5), which are added to the `eln_id` using an underscore and followed by the number to generate the `experiment_ids`. The corresponding `plate_id` as well as the scale, temperature, mixing information, atmosphere and the data are captured in this section as well. Finally, for each of the plates, reaction analysis by LCMS could be carried out at different time points requiring another tab covering these process controls. Reaction controls are counted by the internal process control (IPC) number, which leads to the extension of the `experiment_id` with another underscore and the corresponding `ipc_no` (`Z`) to generate the `ipc_id` (`elnXXXXXX-XXX_Y_Z`). Within the process controls tab, the time of sampling, the concentration of the LCMS sample solution, the LCMS method and file name as well as the data upload is captured. By adding the well position (A1-D6 for 24-well or A1-H12 for 96-well plates) obtainable through the plates section, each reaction can receive a unique identifier termed `rxn_id` (`elnXXXXXX-XXX_Y_Z_AA`).

3.2.5 Data interplay

Figure 3.9 gives an overview of the required data interplay across the screening platform to aid the execution and analysis of the HTE experiments. The data interplay is key to making DOLPHIN an efficient HTE system that possesses a high degree of software-hardware integration and data governance.

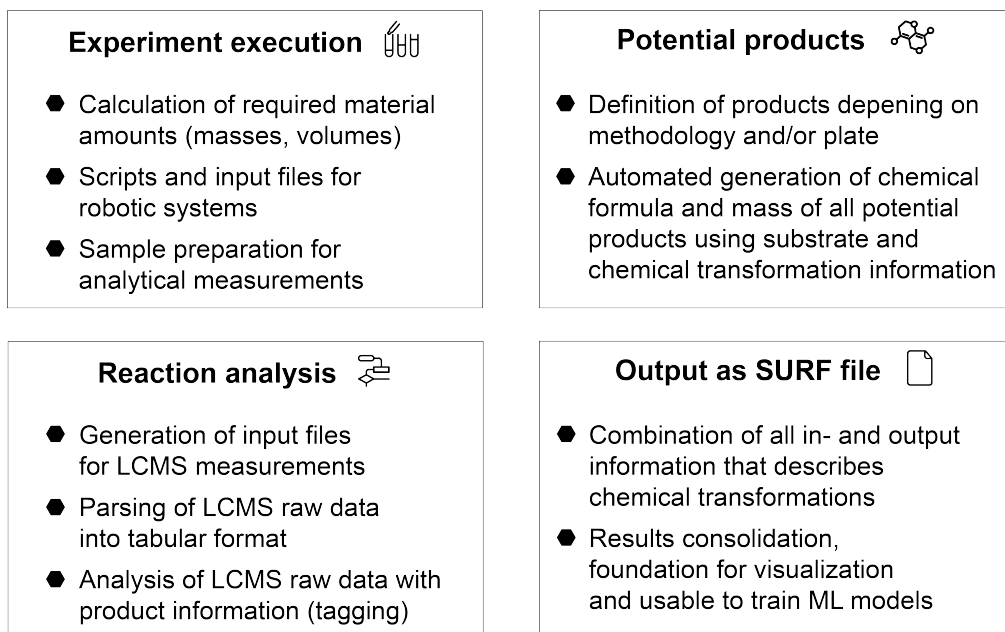


Figure 3.9: Overview of the data interplay. The data structure of the DOLPHIN screening platform allows efficient experiment execution due to the availability of all information for material calculations, script and input file generations. The automated generation of potential products is aided through an additional workflow, which uses the plate definition information and the starting material chemical formula. The reaction analysis process resembles the generation of input files for liquid chromatography-mass spectrometry (LCMS) measurements as well as the parsing and analysis of the LCMS raw data. Due to the reporting of all reaction-associated data, the output of simple user-friendly reaction format SURF files is seamlessly possible.

As highlighted in the screening workflow (Figure 3.7), the execution of experiments requires the preparation of chemicals and the generation of scripts as well as input files for the robotic systems and the analytical measurements. Combining the data from Figure 3.8 allows for conducting the needed calculations and determining the dosing of solutions and solids in the defined wells of the screening plate. Due to the capturing of all reaction information, including roles and equivalents, tabular SURF files serving as a findable, accessible, interoperable and reusable (FAIR) output of the HTE campaigns can easily be generated as well. The automated generation of the potential products using the starting material chemical formula and the analysis of the reaction data are described in the following two sections.

Potential products

Unlike a standard chemical reaction, where a certain product is targeted, LSF can yield a diverse set of products ranging from mono-, di- and tri-substituted products to various regioisomers thereof. [55, 61, 66] Since full structural elucidation in the miniaturized HTE screening was not feasible due to time, equipment and cost constraints, solely LCMS was used to analyze the reaction mixtures. Differentiation of regioisomers is possible through varying retention times but does not elucidate the exact position of the new functional group. However, to assess the general reaction outcome, including product formation within an acceptable time frame, LCMS is a feasible method. To enhance the precision of LCMS results, it is beneficial to include the chemical formulas within the input files, which the system utilizes for its search. Consequently, the desired products should be generated *in silico*. In addition, the masses of the potential products are required to enable the product tagging in the reaction analysis step (Figure 3.11). To address the inefficiency and error-proneness of manually creating several potential products for each new starting material and reaction type, a process that would adversely affect the throughput and precision of DOLPHIN, a data pipeline was developed (Figure 3.10).

To overcome the need to manually generate thousands of new potential products for each new starting material and reaction type, which would be highly inefficient and prone to errors, consequently impacting the throughput and accuracy of DOLPHIN, a data pipeline needed to be developed (Figure 3.10).

Using the established data structure (Figure 3.8), the experiment information can be linked to both, the compound data and the plate definition. While the former can provide the chemical formula of the starting material, within the latter, the element changes of the potentially observable products are defined. Hence, the data workflow needs to disassemble the chemical formula (Hill notation) of the starting material in the first step. Next, the product information data and the formula need to be combined to carry out the change in atom quantity. Subsequently, re-assembly of the chemical formula delivers the potential products, which receive a defined product tag and the associate experiment_id, before being stored in a separate tab of the database. Using the chemical formula, the molecular weight and mono-isotopic mass of the products can be calculated and added as well. Exemplified for a substrate subjected to C-H borylation reactions, the potential product generator delivers the chemical formula of the boronic acid pinacol ester (Bpin) and B(OH)₂ products in a streamlined fashion without

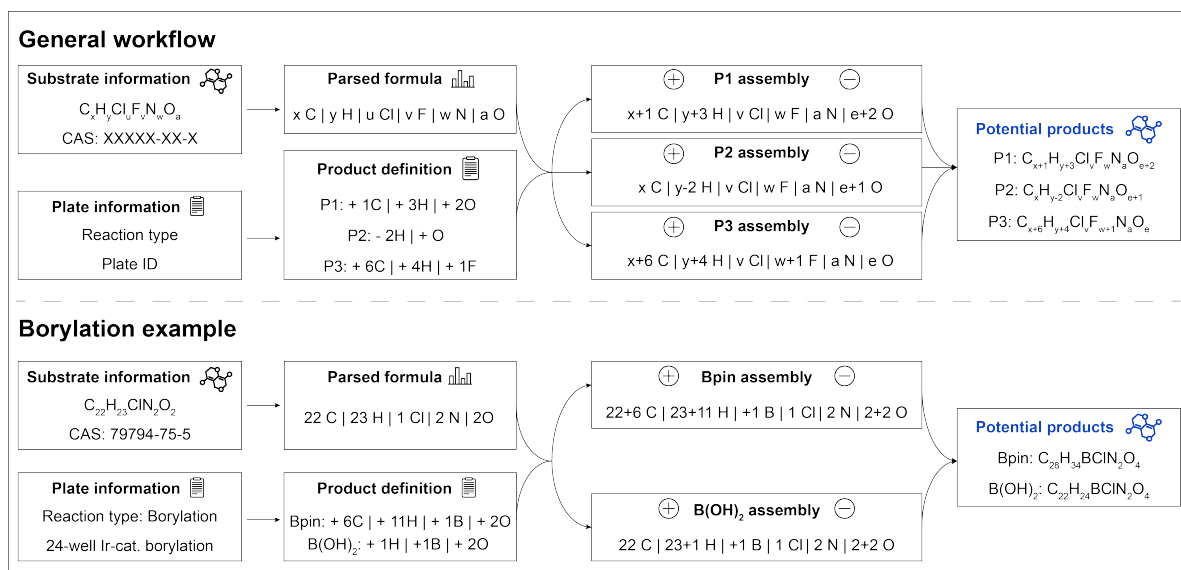


Figure 3.10: Schematic of the potential product generator including an example for the generation of products for C-H borylation reactions. Using the substrate information from the compound database, the chemical formula is parsed into its components. In parallel the plate definition is obtained, which includes information on the potential products and the atomic changes taking place. Next, both information is combined and the changes for each element are executed. In the final step, the chemical formulae are recombined following the Hill notation and tagged with an identifier (experiment_id) connecting them to the starting material and plate. The example at the bottom of the figure shows the creation of two possible products for C-H borylation, the boronic acid pinacol ester (Bpin) and B(OH)₂ derivative.

the need for manual intervention (Figure 3.11). A reference implementation of the potential product generator (Figure 3.10) based on the generation of potential products for Minisci-type reactions (case study, chapter 6.2) with the corresponding Alteryx Designer (Irvine, US) workflow is available at <https://doi.org/10.5282/ubm/data.469>.

Reaction analysis

Figure 3.11 provides an overview of the reaction data analysis pipeline, which supports the processing of LCMS raw data to determine reaction outcomes. The LCMS input files in txt only (txt) format contain the plate position of the 96-well plate, the corresponding rxn_id and the chemical formulae that the MS is intended to search for in a line-by-line structure. Upon measurement of the samples (approximately 3.2 minutes per sample), the instrument outputs the entire run in portable document format (PDF) files and a report (rpt) file. While the PDF file is not machine-readable and cannot be used for automated data analysis, the rpt file contains unstructured text. Hence, a confidential customized parsing script was developed, that transforms the data into a tabular format. These structured tables, which are stored on an internal Google Cloud contain the ultra-violet (UV) peak area and the five most abundant masses of all peaks per sample connected through the unique identifier of each reaction (rxn_id).

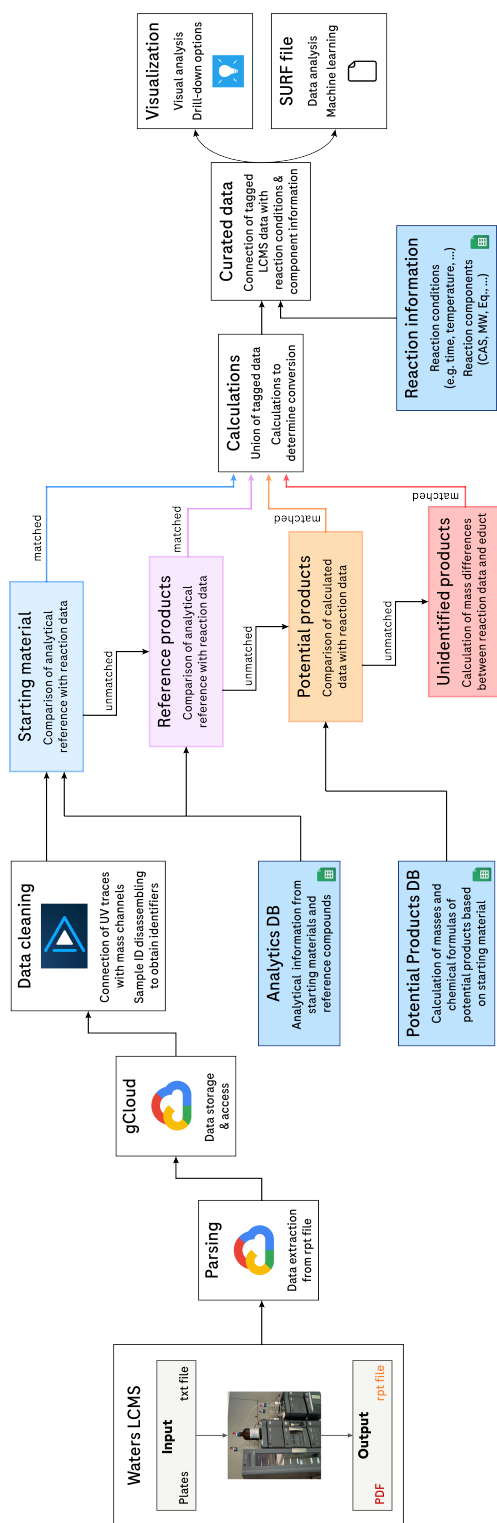


Figure 3.11: Overview of the reaction data analysis pipeline. The raw data from the liquid chromatography-mass spectrometry (LCMS) is obtained as a portable document format (PDF) and a report (rpt) file. While the PDF file is not machine-readable, the rpt file contains unstructured text, which is parsed through a customized script developed at Roche into a tabular format containing areas and masses for all peaks of each sample. The parsing script directly uploads the data to the Google Cloud (gCloud) from which the data can be accessed through the data pipeline software tool, Alteryx Designer (Irvine, US). Using Alteryx, all ultra-violet (UV) peak traces are connected with their corresponding mass information and the sample identification (ID), which equals the rxn_id is parsed to obtain required identifiers (eln_id, plate_id, etc.). In the following step, the LCMS data is compared with the analytical reference data of the starting material obtained from the analytics database (DB). If retention time and mass pattern match, the corresponding peak receives a tag. The same process takes place in the next step, where the analytical data of potential reference products is compared to the LCMS data. Then, the matching of the potential product masses with the LCMS data takes place, leading to additional tags, if correlations are observed. Finally, all other remaining peaks are tagged as unidentified products and their mass difference compared to the starting material is calculated. Subsequently, all data streams are combined to carry out the quantification of the reaction outcome. The area of each UV peak is divided by the sum of all peak areas to deliver the individual ratios. The addition of the reaction information, such as conditions (e.g., temperature, time) and the reaction components lead to the output of a SURF file usable for visualization and direct machine learning (ML) application.

Using the data pipeline software tool Alteryx Designer (Irvine, US) to connect to the Google Cloud, the UV peak traces are connected to the mass information and sample identification (ID), equivalent to the rxn_id is parsed to extract other identifiers, such as eln_id and plate_id. The latter is needed for the addition of the reaction components towards the final step of the workflow. Next, the LCMS data of each sample undergoes comparison with analytical reference data of the corresponding starting material using the eln_id and chem_id as a joint string. If retention times and mass patterns match, a peak is tagged as the starting material. The process is repeated for potential reference products, which also possess analytical information stored in the database. The potential products (Figure 3.10) are retrieved from a different database as their retention times are not known and multiple regioisomers could be observed. In the subsequent matching of potential product masses with the remaining LCMS data, the next tagging round is carried out. Peaks without any matches are designated as unidentified products, and their mass differences relative to the starting material are computed.

After the tagging process, all consolidated data streams are unioned and the area of all peaks is summed up. This allows the computing of the individual conversion ratios for each peak in the following step. Using the plate_id identifier, each sample is now enriched with information on the exact reaction conditions and parameters (*e.g.*, time, temperature). This also includes identifiers (SMILES, CAS number) and the equivalents of each component. In the end, each line contains all information of one reaction and is exported as a SURF file (Chapter 4), which can directly be used for data visualization and analysis as well as ML applications. The average run time of the data workflow takes two minutes, which ensures rapid and accurate data preparation of the screening results. A reference implementation of the LCMS reaction data analysis (Figure 3.11) leading to the generation of the visualization and SURF files exemplified on a Minisci-type reaction with the corresponding Alteryx (Irvine, US) workflow including intermediate data outputs is available at <https://doi.org/10.5282/ubm/data.469>.

3.2.6 Visualization and analysis

The automated reaction analysis pipeline reduces the time from the last measurement to the visualization of the reaction data to a minimum. Using the visualization output files (examples available at <https://doi.org/10.5282/ubm/data.469>), different visualizations were built in TIBCO Spotfire that support the interpretation and understanding of complex reaction screening campaigns. The following chapter describes the developed visualizations (Figures 3.12 to 3.15) exemplified by the analysis of a Minisci-library screening plate on Loratadine (9).

Figure 3.12 presents a general overview of the full analysis page in Spotfire. The interface is designed for intuitive navigation, allowing for the selection of experiments through the input of the electronic laboratory notebook identifier (eln_id), the screening plate layout (plate_id), the plate number (plate_no), and the process control (ipc_no). The visualization is structured to provide a comprehensive snapshot of the experimental data. The central figure displays all 24 wells of the plate, with the ratio of identified components represented in pie charts, offering an immediate visual assessment of the reaction outcomes. Adjacent to this, a scrollable tabular field lists all components, inclusive of their retention times, facilitating a detailed examination of the chromatographic data. The right side of the panel contains essential experiment information, such as the general reaction conditions, time, temperature, and atmosphere. By selecting a well, the reaction components used, are shown in the section below. The structure of the starting material, Loratadine (9), is always depicted at the bottom right of the panel. The visualization also incorporates an MS reliability score, color-coded to reflect the confidence in the results, with green indicating high reliability. A chromatogram with tagged and untagged peaks would be displayed upon the selection of a specific well, providing further analytical depth.

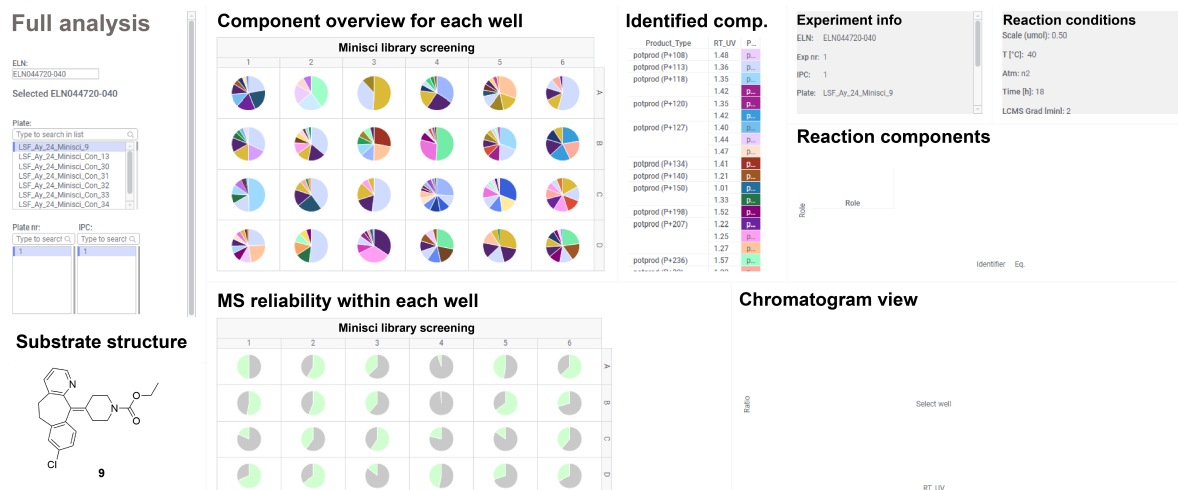


Figure 3.12: General overview of the full analysis page in Spotfire. The selection of the experiments is steered through the input of the eln_id (ELN) and subsequent selection of plate_id (Plate), plate_no (Plate nr) and ipc_no (IPC) on the left side of the panel. In this case, a Minisci-library screening plate on Loratadine (9) was selected. The top center figure shows all 24 wells of the plate and the ratio of the identified components in pie charts. Next to this chart, a scrollable tabular field contains all the components including their retention times. Further to the right, the basic experiment information, followed by the general reaction conditions, such as time, temperature and atmosphere are shown. Below this feature, the reaction components would be displayed if a well is selected. On the bottom, the structure of the starting material is depicted, in this case, Loratadine (9). The next visual of the plate contains the MS reliability score, indicating the confidence of the results (green: high reliability, grey: no reliability - unknown products). To the right, the chromatogram with all tagged and untagged peaks would appear if the user selects one specific well.

While offering a general overview of the reaction outcome, Figure 3.12 is not tailored to immediately highlight the reaction outcome of the executed reaction type, in this case, the presence of mono- or dialkylated products. Hence, a focused analysis view was developed that supports an accelerated identification of reaction outcomes (Figure 3.13). This sub-tab of the full analysis mirrors the initial experiment selection, thereby streamlining the workflow. The visualization is split into two sections: the top row illustrates the structure of Loratadine (9) and a chromatogram view, which populates data upon the selection of a well. The bottom row offers a focused analysis, with a 24-well overview on the left-hand side, highlighting the ratios of the combined ratios of the different potential product types. In this case, the visualization shows the combined ratios of all mono- (yellow) and di-alkylation (blue) products. This provides a fast overview of the carboxylic acids that have reacted well with Loratadine (9) and could be of interest for a more detailed analysis.

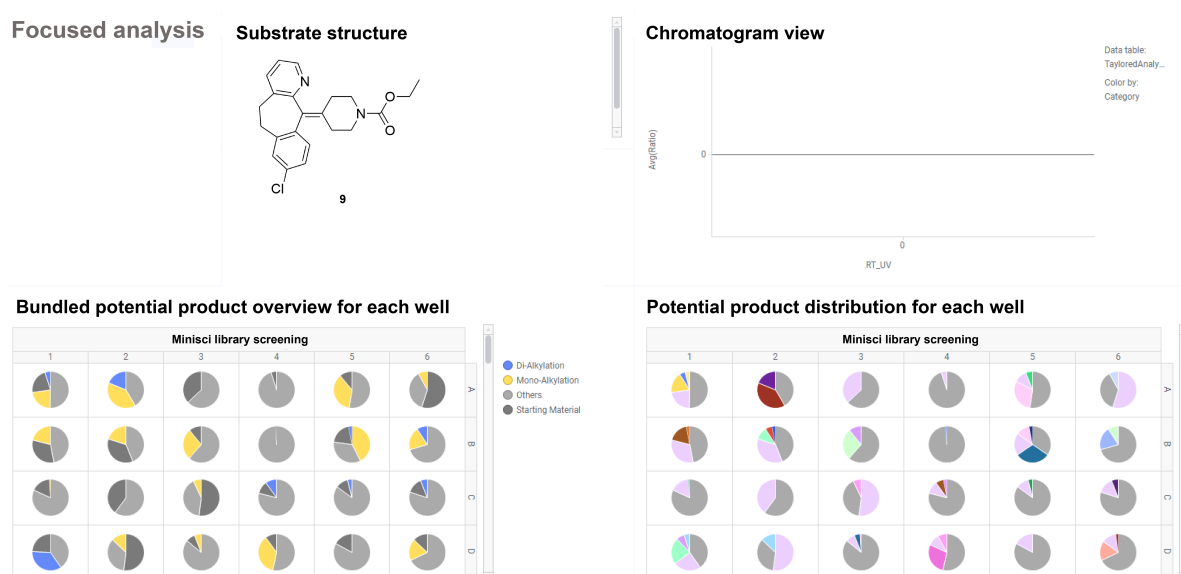


Figure 3.13: Focused analysis for the fast identification of reaction outcomes. As a sub-tab of the full analysis, no selection options for the experiment are needed as the initial choice is mirrored. The top row shows the structure of the starting material, in this case, Loratadine (9) and a chromatogram view, which only contains data if a specific well is selected. The bottom row contains the focused analysis with the 24-well overview on the right-hand side highlighting the ratios of the different potential products corresponding to the reaction type. In this example, all combined mono- and di-alkylation ratios are highlighted in yellow and blue, respectively. The right-hand side overview offers a drill-down possibility to differentiate between different mono- and di-alkylation products, helping to identify regioisomers.

Based on the viewing of the focused analysis in Figure 3.13, the reaction conditions from well B5 seemed to have delivered a good amount of mono-alkylated product. Consequently, in Figure 3.14, well B5 is selected, directing the tool to specifically highlight the potential products from this condition. The chromatogram view on the top right indicates that two mono-alkylated products were formed and the height of the peaks, which corresponds to

the ratios, shows the proportion of 2:1 of the more polar product. To reduce the amount of distraction, in the bottom row overviews, the opacity of all other wells is reduced, emphasizing the outcome of selected well B5. The right-hand side overview allows for a similar granular differentiation as the chromatogram but uses pie charts instead.

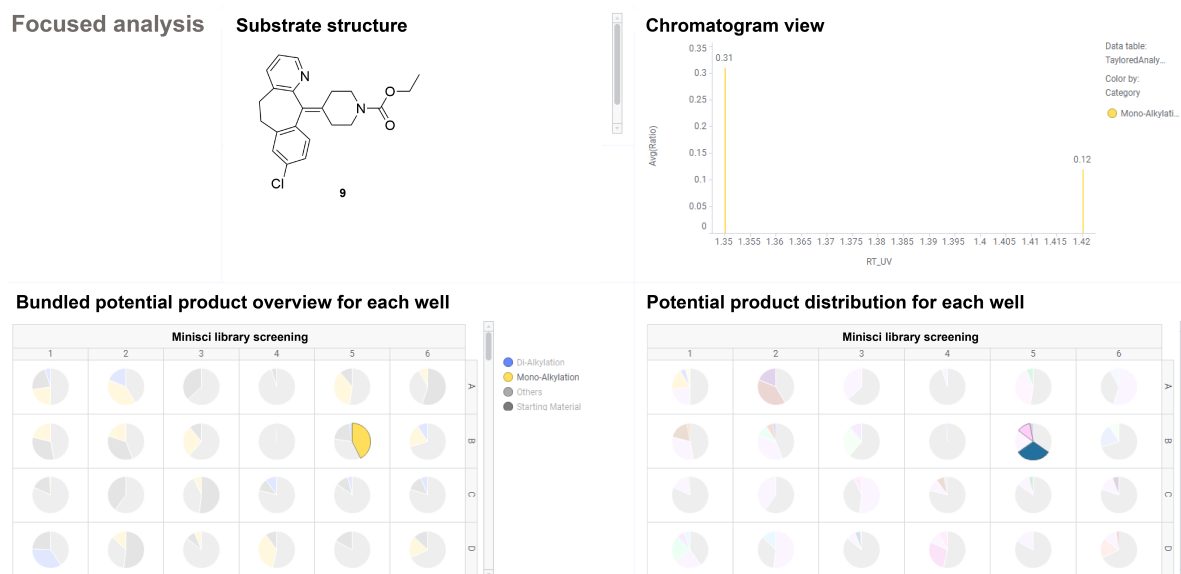


Figure 3.14: Focused analysis for the fast identification of reaction outcomes with the selection of a well. As a sub-tab of the full analysis, no selection options for the experiment are needed as the initial choice is mirrored. In this view, well B5 was selected. The top row shows the structure of the starting material, in this case, Loratadine (9) and a chromatogram view of B5 only highlighting the potential products. The height of the peaks corresponds to their ratios. In this case, a clear indication of the formation of two mono-alkylated regioisomers due to the two yellow peaks can be observed. The bottom row contains the focused analysis with the 24-well overview on the right-hand side highlighting the ratios of the different potential products corresponding to the reaction type, here the combined ratios of all mono- and di-alkylation peaks. As B5 is selected, all other wells have reduced opacity. The right-hand side overview offers a drill-down possibility to differentiate between different mono- and di-alkylation products, highlighting the two mono-alkylation isomers and their respective ratios.

With a potential scale-up condition at hand, the user can now return to the full analysis overview depicted in Figure 3.15. Due to the mirroring of the tool, B5 remains selected and all other wells are shown with reduced opacity. The tabular field now exclusively lists the components and retention times for well B5. The reaction components used for this specific reaction are detailed below the general reaction conditions (time, temperature, atmosphere) and include their role and equivalents. The MS reliability score and chromatogram now highlight the B5 results. While the MS score helps to confirm the reliability of the tagging, the chromatogram provides a better understanding of the overall composition of the reaction mixture. In this case, a good number of side products have been formed, but the potential products have a higher or similar ratio. Further, the starting material peak indicates that the reaction might be pushed further as full conversion of Loratadine (9) has not been achieved.

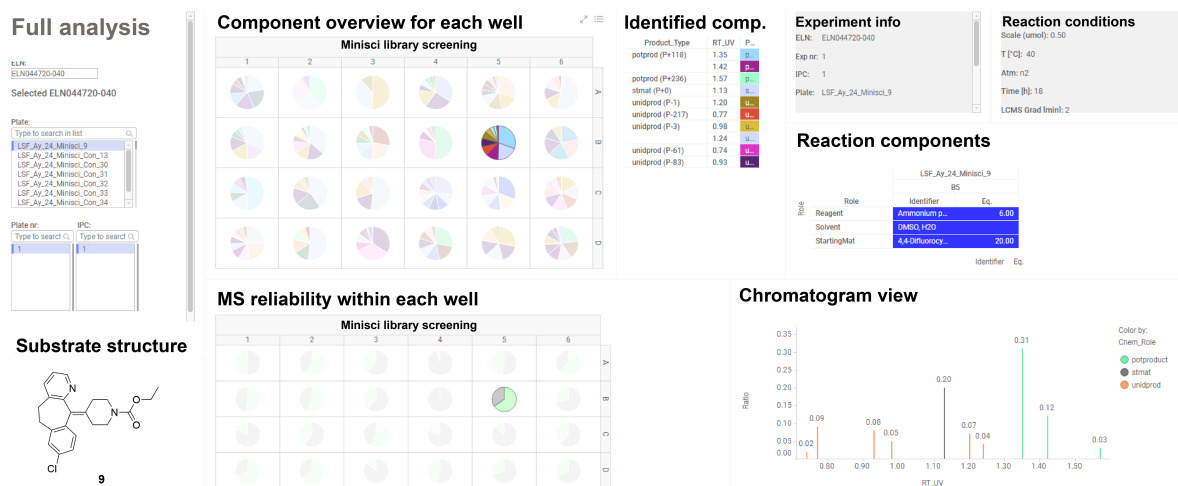


Figure 3.15: General overview of the full analysis page in Spotfire with the selection of a well. The selection of the experiments is steered through the input of the eln_id (ELN) and subsequent selection of plate_ids (Plate), plate_no (Plate nr) and ipc_no (IPC) on the left side of the panel. In this case, a Minisci-library screening plate on Loratadine (9) and well B5 was selected. The top center figure shows all 24 wells of the plate and the ratio of the identified components in pie charts. Except for well B5, all other wells have low opacity to highlight the selected well. Next to this chart, a scrollable tabular field contains all the components including their retention times of B5. Further to the right, the basic experiment information, followed by the general reaction conditions, such as time, temperature and atmosphere are shown. Below this feature, the reaction components that were used in well B5 are displayed including their role and the used equivalents. On the bottom, the structure of the starting material is depicted, in this case, Loratadine (9). The next visual of the plate contains the MS reliability score, indicating the confidence of the results (green: high reliability, grey: no reliability - unknown products). Similar to the top plate layout, the opacity for all wells except B5 is turned down. To the right, the chromatogram with all tagged and untagged peaks of well B5 is visible.

The above-described analysis workflow highlighted that the Spotfire tool provides a robust and versatile platform for the analysis of the screening data, enabling the user to navigate and interpret multiple screening plates with ease. The information depth provided through the visualization efficiently guides the scientist towards promising scale-up conditions and supports the thought process on the execution of those.

3.3 Scale-up

The scale-up of the most promising reaction conditions obtained through the miniaturized screening platform is an important step in the DOLPHIN workflow towards obtaining novel chemical matter. It requires running the identified chemical transformation on a typical medicinal chemistry scale (50 to 100 mg) to isolate and characterize material for biological testing. Executing such reactions on this scale is the daily business of discovery chemistry departments in the pharmaceutical industry and other fields, but it has received little digital innovation over the last decades. Despite the introduction of ELNs, the manual and, especially, non-standardized reporting of all involved steps is still common practice. Establishing find-

able, accessible, interoperable, and reusable (FAIR) principles to make use of the data points for data analytics and ML applications and easing the documentation in the laboratory in parallel, remains a challenge. Thus, to streamline the process and reduce repetitive and error-prone tasks in the reporting process, a streamlined, data-orchestrated workflow to support the scale-up reactions was developed (Figure 3.16).

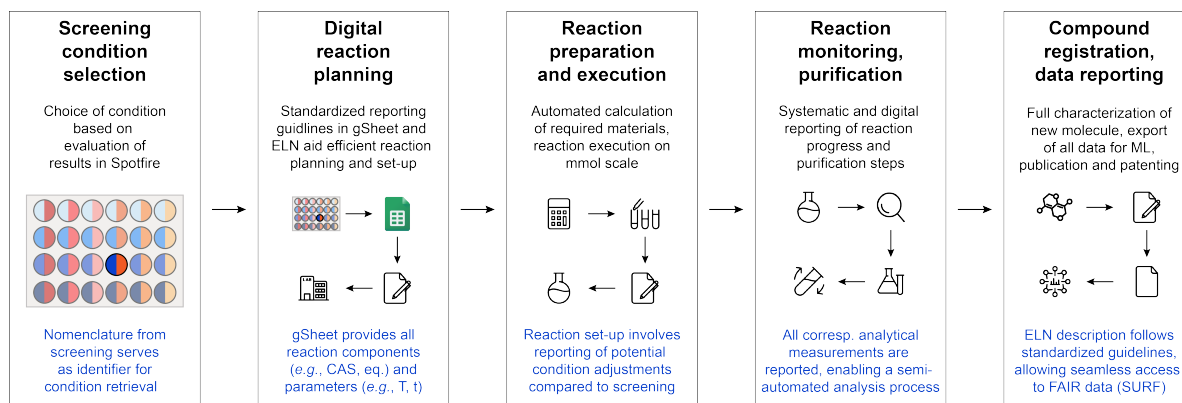


Figure 3.16: Overview of data-orchestrated scale-up workflow. The screening condition selection is aided by the Spotfire tool. Making use of the unique identifier nomenclature for each reaction (elnXXXXXX-XXX_Y_Z_AA), the reaction components from the screening are automatically provided in a Google Sheet for the seamless set-up of the scale-up experiment as a single reaction in the electronic lab journal (ELN). The output includes identifiers (CAS, Smiles) and the equivalents (eq.) as well as the reaction parameters, such as temperature (T) and time (t). With this data at hand, the reaction can be prepared digitally and rapidly executed. The standardized reporting protocol guides the scientist to document the reaction simple, yet comprehensively in a machine-readable format. Monitoring of reaction progress and purification steps are reported in the Google Sheet, which automatically generates files for the analytical instruments with unique identifiers. Upon characterization of the final product, the compound is registered and all data are exported from the ELN as a SURF file, which can be used for machine learning (ML) applications, publication and patenting, following findable, accessible, interoperable, and reusable (FAIR) principles.

Upon the identification of promising screening conditions aided through the Spotfire visualization tool (Chapter 3.2.6), the input of the unique identifier nomenclature (rxn_id, elnXXXXXX-XXX_Y_Z_AA) in an interactive Google Sheet retrieves all screening related information from the database. This helps to avoid manual look-up of the conditions in Spotfire, but rather provides the scientist digitally with all required data to plan the reaction in the ELN. In general, the structural (SMILES) and inventory identifiers (chem_id, CAS number, location) are provided together with all other parameters, such as temperature (T) and time (t). This step reduces errors when entering information, ensures completeness of all data and speeds up the overall process as the search for the conditions and chemicals is reduced.

When conducting reactions, the documentation of the reaction set-up and the execution is paramount to guarantee the reproducibility of the protocol in future experiments. Hence, a standardized reporting protocol was developed that guides the scientist in the reporting of the reaction. The protocol ensures simplicity and comprehensiveness while guaranteeing that the data is recorded in a machine-readable format, essential for data sharing. The monitoring of the reaction progress and purification steps is also captured in a dynamic Google Sheet, which automatically generates input files with unique identifiers for the analytical instruments. Hence, the scientist does not need to take any manual notes and determine a potentially inconsistent file name nomenclature but can approach the instruments with the sample directly and document the results digitally. This degree of automation and standardization minimizes the potential for human error and enhances the efficiency of executing the scale-up reaction.

Upon successful purification of the final product, the compound is fully characterized, which reveals the regioselectivity of the LSF reaction. The analytical characterization of the products by LCMS, HRMS and NMR spectroscopy is captured in the analytical database of DOLPHIN and the compound is registered in the compound database. Due to the standardized documentation procedures, all relevant reaction data, including yields and analytics, can be exported from the ELN as a SURF file. This enables comparison with the initial screening conditions and aids ML applications as well as publication and patenting activities. Hence, this workflow strictly adheres to the FAIR principles, which are key to modern data governance in chemistry and facilitate data sharing.

In summary, the developed, data-orchestrated scale-up workflow highlights that the integration of digital tools and standardized practices can reduce repetitive and error-prone tasks in the laboratory, thereby increasing efficiency and access to high-quality reaction data.

3.4 Discussion

DOLPHIN supports the assessment and execution of LSF transformations on complex drug-like molecules in an efficient and streamlined fashion based on data-driven and semi-automated workflows. The integration of automation, digitalization, and AI facilitates the systematic evaluation of LSF reactions to obtain starting points for scale-up reactions and generates high-quality reaction datasets that deliver the foundation for the development of ML models. Such an approach reduces the likelihood of reaction failures through the selection of suitable reaction conditions for single experiments, thereby optimizing material and resource utilization in the laboratory.

Central to the DOLPHIN workflow is the miniaturized HTE screening, which allows for the efficient assessment of LSF transformations on a small scale before committing to resource-intensive single reactions on a larger scale. The data-driven approach comprehensively captures all relevant reaction information in SURF to enable rapid identification and analysis of reaction outcomes. The data backbone and the interplay between software and hardware components ensure an efficient end-to-end process, from literature analysis to reaction execution and analysis. Especially, the output of the reaction data in SURF supports the data visualization and analysis enhancing the capability to guide scientists through complex screening data towards promising scale-up conditions.

The scale-up process within DOLPHIN is equally data-orchestrated, ensuring that the transition from screening to larger-scale synthesis is seamless and well-documented. The standardized reporting protocols for capturing reaction progress and purification steps minimize manual intervention and potential errors. The comprehensive documentation and data governance, adhering to FAIR principles, not only facilitate the reproducibility of experiments but also enable the use of collected data for ML applications, publications, and patenting. Overall, DOLPHIN exemplifies how digital innovation can contribute to increasing the efficiency and effectiveness of chemical synthesis in drug discovery.

However, to unlock the potential of LSF for drug discovery and further integrate digital tools into the chemical synthesis process, improvements to the platform are needed. Firstly, the development of alternatives to the mostly manual extraction of literature reaction data from publications could accelerate access to high-quality datasets. The advances in ML and the use of large language models (LLMs) have led to the development of several models. [417, 422, 423, 425] Yet, obtaining all important parameters as defined in SURF (Chapter 4), which

include scale, atmosphere and equivalents based on reaction schemes and tables as well as the procedures in the SI has not been achieved so far.

Additionally, the miniaturization of the LSF reactions utilizing the ChemBeads technology should be investigated. ChemBeads would enable the precise dosing of sub-milligram quantities of reaction components directly into reaction vessels without the need for stock solutions, thus saving time and resources while avoiding solubility and co-solvent issues. [205, 209–211] This could also contribute to further broadening the reaction scope of the platform through the implementation of C(sp²)-C(sp³) couplings. Many of these transformations, however, rely on the use of photochemistry, another current limitation of the platform that could be addressed. The integration of integration of light-emitting modules would allow access to a wide array of transformations and also contribute to more sustainable reaction execution.

Even though LCMS is a standard and commonly employed analytical method in HTE, other techniques could be tested as well. The introduction of multiple injections in a single experimental run (MISER) could reduce the time required for HTE sample processing to 15 minutes for a 96-well plate. [135, 146]. Additional methodologies of interest include matrix-assisted laser desorption/ionization (MALDI), desorption electrospray ionization (DESI), acoustic ejection mass spectrometry (AE-MS), and NMR spectroscopy. [153, 177, 226–234] Obtaining this higher volume of different analytical data would also require adjustments of the reaction analysis workflow and, consequently, the visualization tools as well. Further, different approaches to conducting the quantification of the reaction outcome could be established. Such techniques could include the use of internal standards or assay development.

Finally, to provide scientists using DOLPHIN with a seamless digital solution, all developed automation solutions and digital tools could be combined in one interface. In doing so, the current plethora of ELN, different Google Sheets, Google Cloud, different data pipelines, robots and the Spotfire visualization, would need to be integrated into a single solution. While a highly complex information technology (IT) task, this approach could generate an integrated software/hardware package, that could be deployed at different sites of a company supporting technology transfer, education and FAIR data sharing.

The greatest battles are fought within oneself, where determination meets doubt.

- Jan Frodeno

4

THE SIMPLE USER-FRIENDLY REACTION FORMAT (SURF)

This chapter describes the development of the simple, user-friendly reaction format (SURF). The following manuscript presents discusses a new human- and machine-readable reaction data format, SURF, that allows seamless feeding of machine learning algorithms without requiring data pre-cleaning. [426] SURF is a cornerstone of DOLPHIN and served as the foundation to capture experimental high-throughput experimentation (HTE) experiments for the two presented case studies (Chapter 5 and 6).

The manuscript has been uploaded on ChemRxiv and was submitted to a journal for peer-review: **Nippa, D. F.[†]**, Müller, A. T.[†], Atz, K.[†], Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Simple User-Friendly Reaction Format, *ChemRxiv* (2023), DOI: 10.26434/chemrxiv-2023-nfq7h.

The author of this thesis is the co-first author of the manuscript as he analyzed the current reaction data format landscape, conceptualised, designed and developed SURF, curated reaction data from selected publications into the new reaction data format and built a data infrastructure that generates SURF output files from HTE campaigns.

Simple User-Friendly Reaction Format

David F. Nippa^{1,2,†}, Alex T. Müller^{1†}, Kenneth Atz^{1,3,†}, David B. Konrad^{2,*},
Uwe Grether^{1,*}, Rainer E. Martin^{1,*} & Gisbert Schneider^{3,*}

¹Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, 4070 Basel, Switzerland.

²Department of Pharmacy, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, 81377 Munich, Germany.

³ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

† These authors contributed equally to this work.

* To whom correspondence should be addressed.

E-mail: david.konrad@cup.lmu.de, uwe.grether@roche.com, rainer_e.martin@roche.com, gisbert@ethz.ch

Abstract

Leveraging the increasing volume of chemical reaction data can enhance synthesis planning and improve success rates. However, machine learning applications for retrosynthesis planning and forward reaction prediction tools depend on having readily available, high-quality data in a structured format. While some public and licensed reaction databases are available, they frequently lack essential information about reaction conditions. To address this issue and promote the principles of findable, accessible, interoperable, and reusable (FAIR) data reporting and sharing, we introduce the Simple User-Friendly Reaction Format (SURF). SURF standardizes the documentation of reaction data through a structured tabular format, requiring only a basic understanding of spreadsheets. This format enables chemists to record the synthesis of molecules in a format that is both human- and machine-readable, making it easier to share and integrate directly into machine-learning pipelines. SURF files are designed to be interoperable, easily imported into relational databases, and convertible into other formats. This complements existing initiatives like the Open Reaction Database (ORD) and Unified Data Model (UDM). At Roche, SURF plays a crucial role in democratizing FAIR reaction data sharing and expediting the chemical synthesis process.

1 Introduction

The synthesis of chemical matter is often viewed as a rate-limiting step in material sciences, crop protection and drug discovery [1–4]. Crafting complex molecules typically involves multi-step syntheses, encompassing various reaction steps, each presenting multi-parameter optimization challenges [5, 6]. This high complexity makes chemical reactions time- and resource-intensive [7, 8]. Exploiting the growing volume of chemical reaction data could enhance synthesis planning and potentially boost success rates [9–11]. In recent years, machine learning has shown applications to a broad variety of challenges in chemistry [12–18]. In particular, graph neural networks, transformers, and recurrent neural networks have successfully demonstrated their value in reaction prediction and synthesis planning [19–26].

However, these tools can only achieve success when trained on high-quality data presented in a structured, machine-readable format [27]. Currently, detailed reaction data, encompassing all parameters, reagents, quantities, and roles, are often disclosed within the supplementary information of publications as unstructured text or, in some cases, substrate scope tables. These tables may also appear in the main manuscript of methodology publications but frequently include numerous footnotes highlighting exceptions, complicating human interpretation and analysis. Moreover, both documents are typically available in the challenging-to-process portable document format (PDF). Consequently, the barrier to accessing complete reaction

data sets in a time- and cost-efficient manner remains high [28]. Furthermore, data derived from scientific literature and patents often lack information regarding unsuccessful reaction outcomes. However, these negative results are of paramount importance for training machine learning models, as they play a crucial role in generating reliable predictions. [29–32].

The challenges mentioned above are evident in the state of currently accessible public and commercial databases that encompass chemical reactions. Public resources in this domain are notably limited, with examples including the dataset covering chemical reactions from US patents spanning from 1976 to 2016 [33]. Additionally, there are commercial offerings like Reaxys [34] and SciFinder [35], but these, too, face constraints in providing comprehensive and well-structured reaction data. While these databases do contain a considerable number of reactions from scientific literature and patents, they frequently fall short in terms of providing essential information regarding reaction conditions and outcomes. Moreover, there is often a noticeable bias in favor of including high-yielding reactions, potentially overlooking the valuable insights that can be gained from reactions with lower yields or unsuccessful outcomes. [36, 37]. A multitude of different file formats, in which this data is stored, further complicates access to and harmonization of reaction data. Among the most common formats are Reaction Data File (RDFFile), ChemDraw Extensible Markup Language (CDXML), Reaction International Chemical Identifier (RInChI), Reaction File (RXNFile), JavaScript Object Notation (JSON), and Chemical Markup Language

Reaction(CMLReact) [38–41]. While these formats can effectively store molecular structures and corresponding chemical reaction diagrams, they tend to lack a controlled vocabulary and detailed reaction conditions, such as equivalents. Additionally, their usability is often compromised by the specialized technical knowledge required to work with them, which can hinder accessibility and understanding. Hence, there exists a notable gap in achieving findable, accessible, interoperable, and reusable (FAIR) standards for the reporting, collection, and storage of reaction data. Addressing this gap is imperative to facilitate and advance data-driven research in the field of chemistry. [42].

Recently, two initiatives have been introduced with the aim of capturing reaction data in machine-readable and uniform formats.

1. The Unified Data Model (UDM), initially developed by Roche and Reaxys and now managed by the Pistoia Alliance, is an open, extendable, and freely available data format for exchanging experimental information on compound synthesis and testing [41]. UDM employs a controlled vocabulary, an explicit hierarchical data model, and supports various molecule and reaction representations. UDM, implemented through an Extensible Markup Language (XML) schema, provides the advantage of utilizing widely accessible, generic tools for parsing, validation, and transformation. The format also captures analytical data, literature references, and legal information, with extension points allowing the inclusion of vendor- or process-specific data.
2. The Open Reaction Database (ORD) was introduced as an open-access platform for making chemical reaction data available in a structured format [43]. The ORD schema, implemented using Protocol Buffers [44], offers nine sections to comprehensively cover all experimental details, including the integration of raw and processed analytical data, ensuring reproducibility. ORD's high flexibility accommodates varying levels of detail based on available information. Moreover, the authors of the ORD emphasize usability by enabling data submission via software programs and through a web interface. Leveraging these features, ORD data is compatible with machine learning applications and even provides descriptive fields for reaction featurization.

While UDM and ORD represent crucial steps towards improving the standardization of reaction data for information sharing and machine learning applications, they pose certain challenges in day-to-day laboratory and data science environments in both academia and industry: (i) Complexity: The availability of numerous fields and options for data entry may lead to fewer entries and missing data, as laboratory scientists have limited time for documentation. Focus and simplification, within the constraints of chemistry, are essential for capturing as many data points as possible, including

unsuccessful reactions. (ii) IT barrier: Although ORD offers the option of entering and searching reaction data through a web interface in addition to programmed input, this still necessitates multiple manual steps in an external environment. UDM provides programmed input only, requiring IT skills or dedicated specialists, which precludes most chemists from using UDM for their reaction data. (iii) Data sharing between disciplines: Efficient exchange of reaction data within and across research groups, departments, or companies can accelerate research. With UDM and ORD, direct sharing of data between scientists in the same discipline, such as chemist to chemist, or across disciplines, such as chemist to machine learning scientist, may be hindered depending on available IT skills and infrastructure, as these formats are not easily human-readable for untrained individuals. Finally, the nested data structure complicates streaming reactions from these file formats.

Implementing accessible data practices in chemistry is crucial for further enhancing machine learning applications in the field [37]. We have developed the "Simple User-Friendly Reaction Format" (SURF) at Roche. SURF addresses the limitations of UDM and ORD, complementing these existing data formats while maintaining interoperability. It structures reaction data reporting through a straightforward, yet comprehensive tabular format, requiring only a basic understanding of spreadsheets. SURF eliminates the need for coding experience, advanced IT skills, or a web interface, empowering every chemist to document and share their chemical syntheses in a human- and machine-readable format. As a result, the SURF format has the potential to further democratize reaction data. We advocate making the attachment of a SURF file to the supplementary information of manuscripts mandatory, thereby improving reaction data reporting and ultimately allowing a broad scientific community simplified access to valuable data.

2 Simple-user friendly reaction format

The development of SURF emerged from the need for efficient sharing of reaction data among laboratory chemists, data scientists, and machine learning researchers. Given the involvement of such a diverse group of stakeholders with different backgrounds in computer science and chemistry, creating a structured model interpretable by both humans and machines was of paramount importance for improving the drug discovery process. Based on these considerations, we opted to use simple spreadsheets, as they facilitate data capture in a tabular format, are widely used, and require minimal training. Using spreadsheets addresses the existing information technology barrier of other formats and democratizes FAIR reaction data documentation and sharing. Figure 1 illustrates the current role of SURF at Roche, serving as a connector between the laboratory and data world, enabling FAIR reaction data capture, storage, sharing, and application.

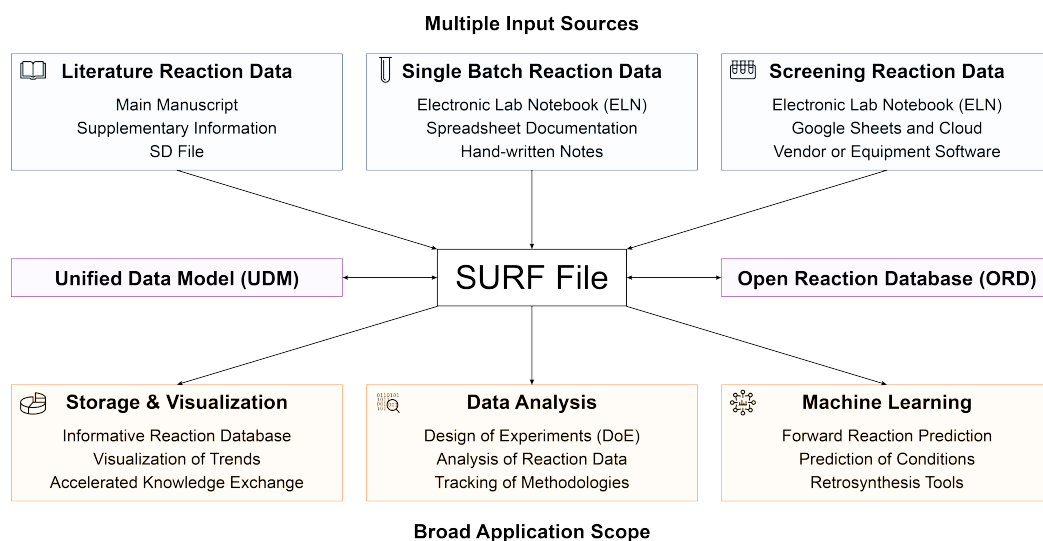


Figure 1: **The simple user-friendly reaction format (SURF) bridges the gap between the laboratory and data science worlds.** SURF files serve as a connector between multiple input sources from literature and the laboratory environment (blue, top) with a broad range of output applications in data science and machine learning (orange, bottom). The format is interoperable with the Unified Data Model (UDM) and the Open Reaction Database (ORD) (purple, middle). The data flow is demonstrated through arrows, highlighting the central role of SURF in connecting data from the laboratory with data science utilization.

Through SURF, laboratory scientists can independently report their reaction data, eliminating the need for handwritten notes, expensive software, or specialized training. Other means of single-batch reaction documentation, such as data from electronic laboratory notebooks or spreadsheets, can also be imported or transformed. Furthermore, various types of literature data can be curated into SURF. At Roche, we are funneling all high-throughput experimentation reaction data from multiple sources into SURF.

SURF enables direct data loading into machine learning models, as structural molecular features are captured through public compound identifiers, *i.e.*, Chemical Abstracts Service (CAS) numbers, simplified molecular input line entry system (SMILES) or international chemical identifier (InChI) strings. This feature enables forward reaction prediction, supports the determination of useful reaction conditions and training of retrosynthesis prediction tools. Due to its structure, reaction databases and corresponding visualization can be easily built and harnessed. Moreover, SURF files enable scientists in the laboratory to efficiently track their reaction data, directly work with their data by conducting analyses, and make data-driven decisions for designing new experiments based on previous outcomes.

3 Structure of SURF

In a SURF spreadsheet, each row stores data for one reaction. The column headers structure the data and are divided into constant (CC) and flexible (FC) columns. CCs remain unchanged and should always be present, independent of the number of reaction components. They capture the identifiers and provenance of the reaction, as well as basic characteristics (reaction type, named reaction, reaction technology) and conditions (temperature, time, atmosphere, scale, concentration, stirring/shaking). Add-ons, such as the procedure or comments, also belong to the CCs. The FCs describe the more variable part of a reaction, including different starting materials, solvents, reagents, and products. Each reaction component is represented by an identifier, such as the CAS number or molecule name, a SMILES or an InChI string storing the chemical structure. While the SMILES/InChI string is available for every compound and can serve as structural input for machine learning models, the CAS number can be useful for chemists in the laboratory to order, itemize, and find chemicals. To account for starting materials and reagents, including catalysts, ligands, and additives, a third column is incorporated to specify the stoichiometric amount, that is, equivalents. SURF's flexibility enables the capture of multiple starting materials and reagents, as these can be accommodated by adding three additional columns (CAS/name, SMILES/InChI, and equivalents). If desired, further columns for additional identifiers or lot numbers can be added. As shown in Figure 2, the headers are populated by adding ascending numbers to record all used components. The

same applies to multiple solvents or products; however, due to their role, they possess more and partly different column headers. While the CAS number/name and/or the SMILES/InChI string remain as identifiers, the solvent fraction (recorded in decimals from [0,1]) is used instead of equivalents, allowing for the exact determination of the ratio between solvents. The product category contains the largest number of headers, as the basic SURF records the reaction yield (in percent, %), complemented by the reaction yield type (*i.e.*, isolated, LCMS, GCMS, etc.), as well as the detected mass by mass spectrometry and the nuclear magnetic resonance (NMR) spectroscopy sequence(s) in addition to the common identifiers CAS and SMILES/InChI. Additional information, such as detailed product characterization (*e.g.*, enantiomeric excess (ee) or purity), can be captured by introducing respective columns with headers following the standard snake case nomenclature.

Utilizing the basic structure of SURF, all relevant data for reproducing the experiment is readily available. Laboratory chemists can order chemicals, draw structures, calculate the masses of molecules, or compare NMR data without the need to consult separate files. Since most electronic laboratory journals already record the aforementioned parameters of the basic SURF structure, enforcing documentation compliance combined with automated data extraction and cleaning pipelines has the potential to make numerous new reaction data accessible in the SURF format and available for machine learning applications.

4 File formats and interoperability

As SURF captures data in a tabular format, we recommend using universally readable file formats such as TXT, CSV, or TSV files. Since chemical data can contain delimiters such as commas or semicolons, we suggest using only TAB-delimited TXT or TSV files. These file types can be written and read with all popular spreadsheet or text editor software available on multiple operating systems. One point to consider when using SURF is that data is not validated upon capture. We acknowledge that this does not prevent users from entering false or incomplete reaction data. However, we recommend performing validation only upon reading SURF files into a database, transforming them to other formats, or using the data for machine learning purposes. SURF files are interoperable, as they can be introduced into hierarchical databases and converted into other existing reaction formats, such as the ORD Protocol Buffers format or UDM XML format. As part of this manuscript, we open-source the respective Python code enabling the transformation between different data formats (<http://reaction-surf.com>).

5 Applications

When preparing for a new series of reactions, such as in a high-throughput setting, chemists have the capability to populate a SURF file with all the necessary conditions and reagents to be tested in advance. They can link these entries to the specific vessels, tubes, or plates used for the reactions through the reaction identifier. Furthermore, having the CAS numbers available for all compounds greatly aids in locating the corresponding materials in the laboratory. Subsequently, as the reactions are executed and data on their outcomes are recorded, any potential gaps or missing data become immediately visible and accessible within the SURF format.

A frequently observed barrier to machine learning application is data pre-processing and cleaning. With SURF, reaction data is presented in a structured, both human and machine-readable format. Hence, SURF has shown to be a key enabler for several reaction prediction case studies at Roche [27, 45]. The use of SURF necessitated minimal data cleaning, mainly focusing on structural information validation and the exclusion of non-relevant columns. This approach allowed for the rapid extraction and analysis of reaction data using standard data science libraries. The SURF header convention as shown in Figure 2 ensures reproducibility and allows for easy identification of relevant columns needed for model training.

Moreover, the tabular SURF format allows users to browse and filter available reaction data directly in a spreadsheet. Simple analyses to visualize yields or find all reactions of a certain type, using a specific technology, substrate, or reagent, can be performed without loading the data into a database. Correlating individual columns like reaction characteristics with reaction outcomes becomes a straightforward task in SURF. Lastly, capturing reactions in a universally readable spreadsheet format facilitates data sharing. The snake case naming convention for headers creates human and machine-readable tables, and if CAS numbers are used as identifiers, compounds can be universally identified even without loading the SMILES/InChI.

6 Discussion and Conclusion

SURF presents a streamlined and accessible solution for chemists to document and share their chemical syntheses in a format that is both human- and machine-readable. By adopting SURF, researchers can overcome the limitations of existing data formats, promote successful data-driven chemistry research, and foster a culture of open data sharing and collaboration, thereby accelerating the pace of discovery and innovation in the field. The availability of reliable data and accompanying code provided by SURF enables other researchers to rapidly verify research findings, thereby reducing the risk of publishing irreproducible results. Importantly, the adoption of SURF facilitates efficient exchange of

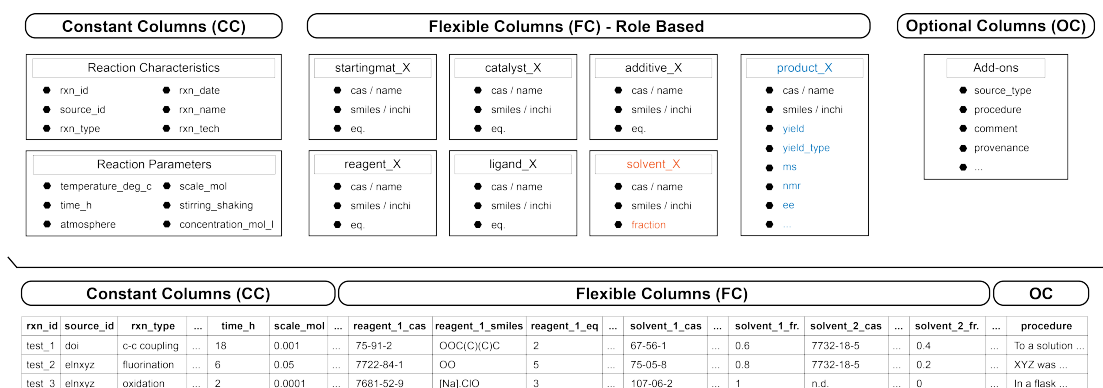


Figure 2: **The structure of the simple user-friendly reaction format (SURF).** Top: Detailed structure of a SURF file, which contains constant (CC), flexible (FC), and optional columns (OC) to comprehensively capture reaction data information. Reaction components are described with two identifiers, one of them containing structural information, *e.g.*, SMILES or InChI, and the used equivalents. For solvents, an exception applies, instead of the equivalents, the fraction is recorded (orange). In the product section, depending on the granularity required, multiple columns for product characterization can be added (blue). In the basic SURF structure, yield, yield type, nuclear magnetic resonance and mass spectroscopy information are added. Bottom: Condensed example of a SURF file that demonstrates the simple structure of the format.

reaction data within and across research groups, departments, and companies, which can accelerate research progress.

Funding agencies and journals have an opportunity to play a more prominent role in promoting open access and FAIR publication of reaction data, ensuring that the necessary incentives and support are in place for researchers to embrace these principles. By encouraging the adoption of SURF as a standard for publications and requiring its attachment to the supplementary information of manuscripts, the scientific community can facilitate reaction data sharing and ultimately advance chemistry research.

7 Acknowledgements

D.B.K. acknowledges funding from the Fonds der Chemischen Industrie (FCI) through a Liebig Fellowship and Roche Basel for funding the PhD position of D.F.N. K.A. and G.S. acknowledge support by the Swiss National Science Foundation (SNSF, grant no. 205321_182176). We thank Nadja Flückiger, Yannick Stenzhorn and Remo Hohler for their valuable contributions to the SURF curated data sets.

8 Competing interest

G.S. declares a potential financial conflict of interest as co-founder of inSili.com LLC, Zurich, and in his role as a scientific consultant to the pharmaceutical industry. D.F.N., A.T.M., K.A., U.G. and R.E.M. are full employees of F. Hoffmann-La Roche Ltd. The authors have not disclosed any additional potential conflicts of interest.

9 Data and code availability

Three SURF files containing reaction data from literature covering Minisci-type alkylations [45], C-H borylation [27] and post-borylation modification reactions as well as program code for seamless interoperability with other data formats is available at <http://reaction-surf.com>.

References

- Blakemore, D. C. *et al.* Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10**, 383–394 (2018).
- Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
- Tabor, D. P. *et al.* Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mat.* **3**, 5–20 (2018).
- Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
- Corey, E. J. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (nobel lecture). *Angew. Chem. Int. Ed.* **30**, 455–465 (1991).
- Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, 1566 (2019).
- Regalado, E. *et al.* Organic chemistry. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2014).

8. Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic synthesis: march of the machines. *Angew. Chem. Int. Ed.* **54**, 3449–3464 (2015).
9. Schneider, G. Mind and machine in drug design. *Nat. Mach. Intell.* **1**, 128–130 (2019).
10. Lowe, D. M. *Extraction of chemical structures and reactions from the literature* PhD thesis (University of Cambridge, 2012).
11. Schneider, G. & Clark, D. E. Automated de novo drug design: are we nearly there yet? *Angew. Chem. Int. Ed.* **58**, 10792–10803 (2019).
12. Von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
13. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3**, 1023–1032 (2021).
14. Unke, O. T. *et al.* Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
15. Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. *Curr. Op. Struct. Biol.* **79**, 102548 (2023).
16. Huang, B., von Rudorff, G. F. & von Lilienfeld, O. A. The central role of density functional theory in the AI age. *Science* **381**, 170–175 (2023).
17. Isert, C., Atz, K., Riniker, S. & Schneider, G. Exploring protein-ligand binding affinity prediction with electron density-based geometric deep learning. *ChemRxiv preprint 10.26434/chemrxiv-2023-585vf* (2023).
18. Atz, K. *et al.* Deep interactome learning for de novo drug design. *ChemRxiv preprint 10.26434/chemrxiv-2023-cbq9k* (2023).
19. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. *J. Neural Inf. Process.* **30** (2017).
20. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
21. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
22. Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
23. Shen, Y. *et al.* Automation and computer-assisted planning for chemical synthesis. *Nat. Rev. Methods Primers* **1**, 1–23 (2021).
24. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. *J. Neural Inf. Process.* **34**, 9405–9415 (2021).
25. Guan, Y. *et al.* Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **12**, 2198–2208 (2021).
26. Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O. & Reymond, J.-L. Retrosynthetic accessibility score (RAscore)—rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **12**, 3339–3349 (2021).
27. Nippa, D. F. *et al.* Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *ChemRxiv preprint <https://doi.org/10.26434/chemrxiv-2022-gkxm6>* (2022).
28. Guo, J. *et al.* Automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model* **62**, 2035–2045 (2021).
29. Engkvist, O. *et al.* Computational prediction of chemical reactions: current status and outlook. *Drug Discov. Today* **23**, 1203–1218 (2018).
30. Strieth-Kalthoff, F. *et al.* Machine learning for chemical reactivity: The importance of failed experiments. *Angew. Chem. Int. Ed.* **61**, e202204647 (2022).
31. King-Smith, E. *et al.* Predictive Minisci and P450 Late Stage Functionalization with Transfer Learning. *ChemRxiv preprint <https://doi.org/10.26434/chemrxiv-2022-7ddw5-v2>* (2023).
32. Caldeweyher, E. *et al.* Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation. *J. Am. Chem. Soc.* (2023).
33. Lowe, D. Chemical reactions from US patents (1976-Sep2016). *Figshare <https://doi.org/10.6084/m9.figshare.5104873>* (2017).
34. Limited, E. *Reaxys <https://reaxys.com/>* (2023).
35. Society, A. C. *Reaxys <https://scifinder.cas.org/>* (2023).
36. Fitzner, M., Wuitschik, G., Koller, R., Adam, J.-M. & Schindler, T. Machine Learning C–N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS Omega* **8**, 3017–3025 (2023).
37. Mercado, R., Kearnes, S. M. & Coley, C. W. Data Sharing in Chemistry: Lessons Learned and a Case for Mandating Structured Reaction Data. *J. Chem. Inf. Model* (2023).
38. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput.* **32**, 244–255 (1992).
39. Grethe, G., Blanke, G., Kraut, H. & Goodman, J. M. International chemical identifier for reactions (RInChI). *J. Cheminformatics* **10**, 1–9 (2018).

40. Holliday, G. L., Murray-Rust, P. & Rzepa, H. S. Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. *J. Chem. Inf. Model* **46**, 145–157 (2006).
41. Tomczak, J. *et al.* UDM (Unified Data Model) for chemical reactions - past, present and future. *Pure Appl. Chem.* (2022).
42. Jablonka, K. M., Patiny, L. & Smit, B. Making the collective knowledge of chemistry open and machine actionable. *Nat. Chem.* **14**, 365–376 (2022).
43. Kearnes, S. M. *et al.* The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
44. LLC, G. *Protocol Buffers* <https://protobuf.dev/> (2023).
45. Nippa, D. F. *et al.* Graph transformer neural network for chemical reactivity prediction. *ChemRxiv preprint* <https://doi.org/10.26434/chemrxiv-2023-8hdmv> (2023).

Success is a result of pushing your limits beyond what you thought was possible.

- Jan Frodeno

5

LATE-STAGE DRUG DIVERSIFICATION THROUGH C-H BORYLATION

This chapter describes the first application of the developed late-stage functionalization (LSF) screening platform (DOLPHIN) and reaction data format (SURF) to assess the applicability of C-H borylation reactions for late-stage drug diversification by carrying out HTE reactions on systematically selected commercial drugs with a broad range of reaction conditions. This enabled the development of a machine learning tool capable of accurately predicting binary reaction outcomes, yields and regioselectivity for novel substrates.

First, a brief introduction to C-H borylations reactions and their potential for LSF in the context of drug discovery is given (Chapter 5.1). Next, the prepared publication describing the case study in detail is reprinted with permission (Chapter 5.2). [427] The final section contains the corresponding experimental and supplementary information (Chapter 5.3).

5.1 Introduction and background

A prevalent strategy for introducing synthetic handles in organic synthesis is the incorporation of boronic acid, $B(OH)_2$, or boronic acid pinacol ester (Bpin) groups, through direct C-H borylation. [428, 429] The approach has found application in various disciplines, including materials science, drug discovery, fine chemicals, and natural product synthesis. [59, 429–431] C-H borylation has emerged as a versatile technique due to the broad utility of organoboron compounds, which can be readily converted into various functional groups. [431–433] Borylated intermediates are of particular value as they can undergo C-C cross-coupling reactions, including the Suzuki-Miyaura coupling, where they contribute to constructing complex molecular architectures. [434–436] Apart from being used to couple building blocks, the organoboron species can also be directly converted into a range of functional groups, including hydroxy (OH), [437, 438] fluorine (F), [439–442] nitrile (CN), [443–445] trifluoromethyl (CF_3), [446–448] chlorine (Cl), [449, 450] or bromine (Br) [449, 451–453].

Selective C-H functionalization via transition metal catalysis is highly dependent on steric and electronic factors, with the ligand environment around the metal center often dictating selectivity. [66] Iridium catalysts, which have been extensively studied for C-H borylation advanced the understanding of these reactions. [454] These catalytic systems facilitate the cleavage of inert C-H bonds and subsequent C-B bond formation, leveraging the electropositive nature of boron and the ability to form π -bonds with the transition metal (Figure 5.1). [428] Group 9 metals, including iridium, rhodium, and cobalt, have been explored for C-H borylation, with iridium exhibiting superior reactivity due to stronger metal-carbon and metal-hydrogen bonds. [454] However, palladium-based catalysts have seen limited use as palladium(II) species promote boronic ester decomposition and by-product formation. [430]

C-H borylation can be categorized into two main classes based on the hybridization of the carbon atom involved. In the literature, methodologies covering the aromatic $C(sp^2)$ -H borylation are more prevalent than aliphatic $C(sp^3)$ -H transformations. [59] Early aromatic C-H borylation research required harsh conditions, but the advent of iridium-based systems with nitrogen-containing heterocyclic ligands improved regioselectivity, yields, and functional group tolerance while reducing reaction temperatures and starting material excess. [59, 139, 428] In general, steric effects predominantly influence aromatic C-H borylation, whereas electronic properties are crucial for heteroaromatic compounds, where steric factors still play a role. Aromatic C-H borylations usually proceed faster than aliphatic ones, reflecting in higher

turnover rates and better functional group tolerance for C(sp²)-H centers. [428]

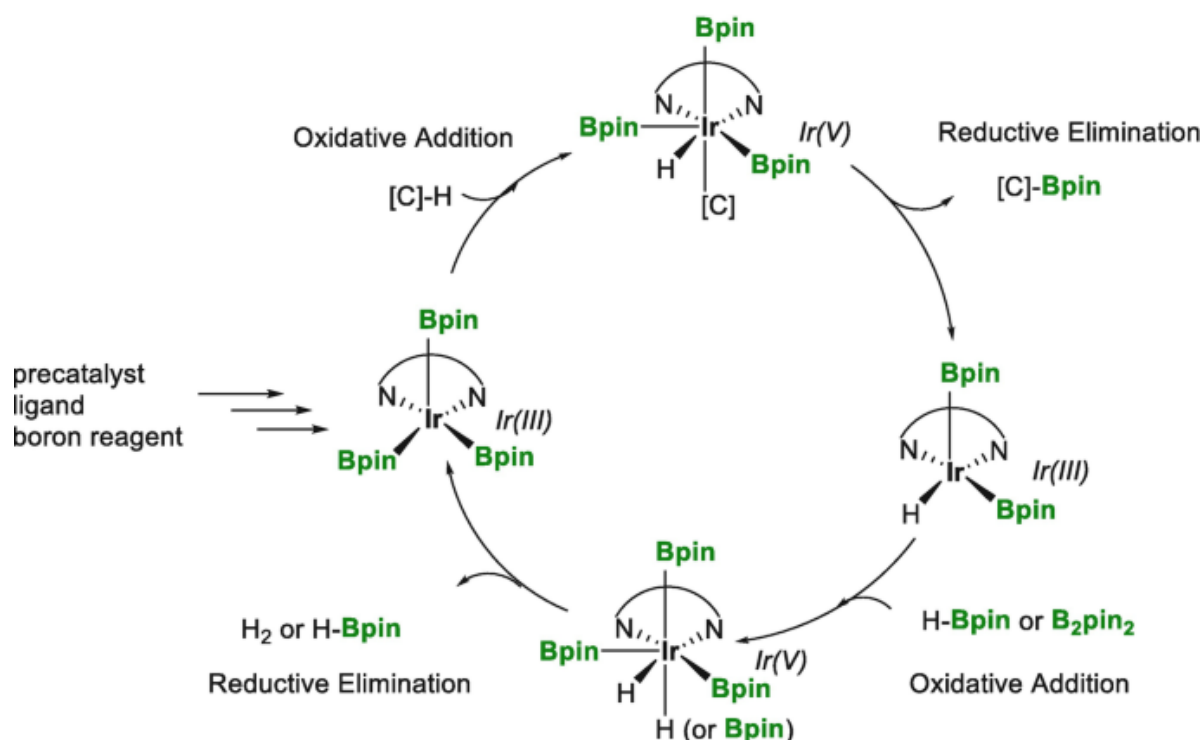


Figure 5.1: General mechanism of the C-H borylation using transition metal catalysts. Pre-catalyst, ligand and boron reagent form the activated Ir(III) complex. Through oxidative addition of the C-H bond, an Ir(V) species is formed. Reductive elimination yields the desired C-Bpin species through C-B bond formation, whereby Ir is reduced from (V) to (III). Next, oxidative addition of the boron source, B₂pin₂ or HBpin, generates an Ir(V) species. This intermediate undergoes reductive elimination of HBpin or H₂ to yield the activated Ir(III) complex, which can start the next catalytic cycle. Data and figure derived from Oro & Claver (2021). [455]

The desire of medicinal chemists to modify C(sp³)-H bonds has spurred research into C(sp³)-H borylations, which are often catalyzed by transition metals such as palladium or copper. [433] These reactions have enabled enantioselective borylations, yielding chiral molecules with high enantiomeric excess using both established and novel ligands. [454] However, stringent regulations on transition metal residues and their scarcity have limited the scalability of these methods [430, 433] Consequently, metal-free borylation techniques, including the use of BBr₃ and the concept of frustrated Lewis pairs, have gained attention. BBr₃ coupled with in-situ esterification using pinacol, has successfully borylated nitrogen-containing molecules under mild conditions with functional group tolerance. C-H borylation based on the frustrated Lewis pair approach, involves electron density transfer from the C-H bond to the Lewis acid, hydrogen abstraction by the Lewis base component of the frustrated Lewis pair, and subsequent borylation with HBpin. [430]

Running efficient C-H borylation with subsequent post-functionalization modifications would support medicinal chemistry campaigns to accelerate the make step of the design-make-test-analyze (DMTA) cycle, which represents a major bottleneck in establishing structure-activity-relationships (SARs) in the lead optimization (LO) phase. [2, 456] To evaluate the applicability and potentially increase the synthesis efficiency of borylation reactions, a case study, that connects laboratory automation with artificial intelligence was designed.

5.2 Publication

The following case study has been published as: **Nippa, D. F.[†]**, Atz, K.[†], Hohler, R., Müller, A. T., Marx, A., Bartelmus, C., Wuitschik, G., Marzuoli, I., Jost, V., Wolfard, J., Binder, M., Stepan, A. F., Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G. Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning, *Nat. Chem.*, **16**, 2, 239-248 (2024). [427] The material (DOI: 10.1038/s41557-023-01360-5) is reprinted with permission from Springer Nature Limited (Author reuse for own thesis).

The author of this thesis is the co-first author of the publication as he carried out the literature analysis, experimental work (HTE, scale-up, post-borylation), reaction data preparation for the predictive tool and the writing of the initial manuscript draft. The machine learning algorithms were designed and developed by Dr. Kenneth Atz. Further details on the contributions of all authors are stated on the last page of the publication.

A detailed description of the experiments conducted and methods used in the publication can be found in Chapter 5.3.



Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning

Received: 21 October 2022

Accepted: 3 October 2023

Published online: 23 November 2023

Check for updates

David F. Nippa^{1,2,6}, Kenneth Atz^{3,6}, Remo Hohler¹, Alex T. Müller¹, Andreas Marx¹, Christian Bartelmus¹, Georg Wuitschik¹, Irene Marzuoli⁴, Vera Jost¹, Jens Wolfard¹, Martin Binder¹, Antonia F. Stepan¹, David B. Konrad², Uwe Grether¹, Rainer E. Martin¹ & Gisbert Schneider^{3,5}

Late-stage functionalization is an economical approach to optimize the properties of drug candidates. However, the chemical complexity of drug molecules often makes late-stage diversification challenging. To address this problem, a late-stage functionalization platform based on geometric deep learning and high-throughput reaction screening was developed. Considering borylation as a critical step in late-stage functionalization, the computational model predicted reaction yields for diverse reaction conditions with a mean absolute error margin of 4–5%, while the reactivity of novel reactions with known and unknown substrates was classified with a balanced accuracy of 92% and 67%, respectively. The regioselectivity of the major products was accurately captured with a classifier *F*-score of 67%. When applied to 23 diverse commercial drug molecules, the platform successfully identified numerous opportunities for structural diversification. The influence of steric and electronic information on model performance was quantified, and a comprehensive simple user-friendly reaction format was introduced that proved to be a key enabler for seamlessly integrating deep learning and high-throughput experimentation for late-stage functionalization.

Structural novelty and complexity render the synthesis of chemical target structures challenging when aiming to establish structure–activity relationships in medicinal chemistry¹. Structure–activity relationship models guide hit-to-lead and lead optimization programmes, aiming to improve the pharmacological activity and physicochemical properties of drug candidates^{2–4}. For structure–activity relationship exploration, time-efficient synthesis is important because synthesis

represents a bottleneck of the design–make–test–analyse cycle⁵. A number of synthetic methods for the selective activation and modification of C–H bonds allow for the late-stage functionalization (LSF) of organic scaffolds, ranging from molecular building blocks to advanced drug molecules⁶. Numerous catalytic systems offer both, directed and non-directed methods, as well as chemo- and site-selective access to modified analogues. LSF methods in medicinal chemistry include

¹Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland. ²Department of Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany. ³Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland. ⁴Process Chemistry and Catalysis (PCC), F. Hoffmann-La Roche Ltd., Basel, Switzerland. ⁵ETH Singapore SEC Ltd, Singapore, Singapore. ⁶These authors contributed equally: David F. Nippa and Kenneth Atz. ✉ e-mail: david.konrad@cup.lmu.de; uwe.grether@roche.com; rainer_e.martin@roche.com; gisbert@ethz.ch

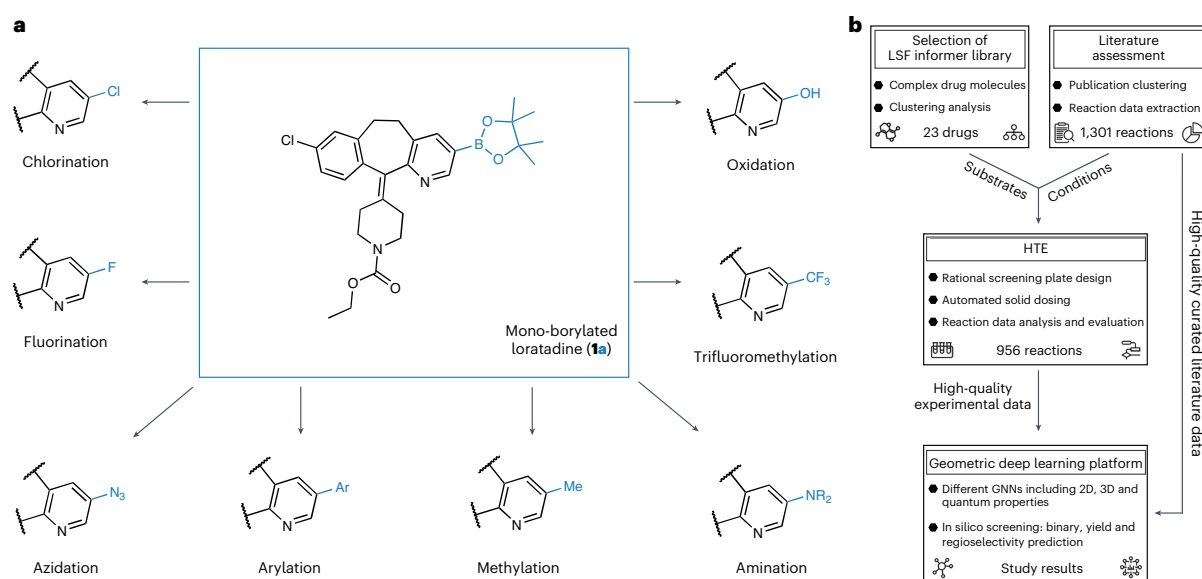


Fig. 1 | Borylation diversification opportunities and research overview of the study. **a**, Late-stage borylation of a drug molecule. The example illustrates mono-borylated Loratadine (**1a**), which can be accessed through borylation of the drug Loratadine (**1**). Borylation provides the opportunity for rapid and broad diversification, aiming to study structure–activity relationships and improve pharmacokinetic and pharmacodynamic properties. Note that the eight potential post-functionalization modifications shown are for demonstration purposes only; these transformations were not carried out in the presented research. **b**, Overview of the research study. A comprehensive literature study provided a manually curated, high-quality literature dataset containing 1,301

reactions extracted from 38 publications. The dataset was used to identify suitable borylation reaction conditions for HTE and used for machine learning. The LSF informer library resulted from a cluster analysis of 1,174 approved drug molecules. In total, 23 drugs from the LSF informer library, 12 relevant fragments and 5 simple substrates were subjected to HTE to deliver 956 experimental data points. Both experimental and literature data provided the basis for geometric deep learning using different GNNs, including 2D and 3D information and atomic partial charges. Prediction models for substrate reactivity, reaction yields and regioselectivity were developed, and the results are shared in this study.

fluorination, amination, arylation, methylation, trifluoromethylation, borylation, acylation and oxidation⁷. Among these methods, C–H borylation is considered the most versatile for rapid compound diversification. Organoboron species can be transformed into an array of functional groups and serve as a robust handle for subsequent C–C bond couplings (Fig. 1a), which enables broad structure–activity relationship studies^{8–10}.

However, only a few applications of LSF in drug discovery have been reported to date^{11,12}. Most of these rare examples focus on a single LSF reaction type^{13–15}. Multiple functional groups and various types of C–H bonds with different bond strengths, electronic properties and steric and functional group environments pose challenges for straightforward LSF; thus, generalizing guidelines for reactivity and selectivity predictions should be applied with caution¹¹. Consequently, running a successful LSF campaign often requires time-consuming and resource-intensive experimentation, which is not compatible with the tight timelines and limited assets of many medicinal chemistry projects.

High-throughput experimentation (HTE) is an established approach for reaction optimization^{16–18}, enabling semi-automated miniaturized low-volume screenings to rapidly and reproducibly perform multiple transformations in parallel with small amounts of precious building blocks and consumables^{19–21}. In combination with FAIR (Findability, Accessibility, Interoperability, Reusability)²² documentation, which generates high-quality datasets on successful and failed reactions^{23,24}, HTE provides a foundation to unlock LSF for drug discovery by enabling advanced data analysis and machine learning.

Graph neural networks (GNNs) have seen broad applications in molecular feature extraction and property prediction^{25–28}. Among the

various machine learning methods developed for chemical reaction planning^{23,29,30}, GNNs have been successfully employed for retrosynthesis planning, regioselectivity prediction and reaction product prediction^{31–34}. In addition, transformers and fingerprint-based methods were developed to tackle similar problems^{35,36}. Other studies have shown that learning the activation energies of transition-state geometries yields accurate predictions for competing reaction outcomes^{37–39}. Graph featurization with density functional theory (DFT)-level atomic partial charges improved the prediction of regioselectivity for reactions driven by electronic effects⁴⁰. The combination of graph machine learning with HTE enabled the optimization of reaction conditions for the C–H activation of organic substrates⁴¹. Recently, a GNN-based approach for predicting late-stage alkylation opportunities has been published, mainly focusing on Baran-type diversinate chemistry using alkyl sodium sulfinate salts⁴². Several studies have focused on deep learning models using transition states with the capability of predicting reaction outcomes, including, in some cases, enantioselectivity^{43–45}. However, these approaches are limited to small molecular structures and comparably small datasets, rendering the application of such models to structurally more intricate drug-like molecules challenging⁴⁶. A recent study has shown that hybrid machine learning models augmented with the quantum chemical information of transition states enable regioselectivity predictions for iridium-catalysed borylation reactions⁴⁷. Importantly, the influence of steric and electronic effects on the model performance for C–H activation reactions and their application to regioselectivity for molecules with multiple aromatic ring systems remains unexplored.

Here we introduce a geometric deep learning approach applied to automated LSF borylation screening for identifying late-stage hits

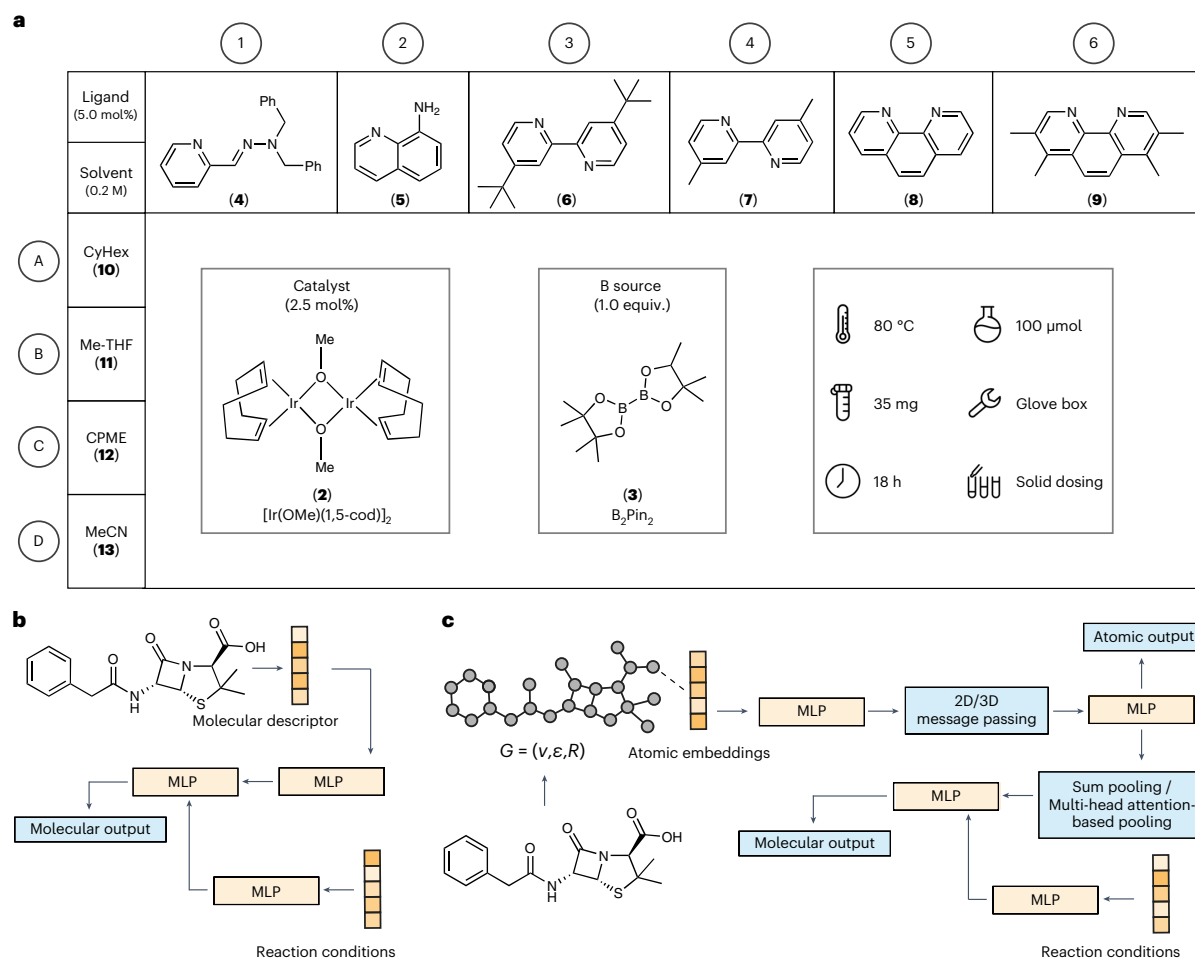


Fig. 2 | Screening plate overview and GNN architecture. a, Schematic of the 24-well borylation screening plates (columns: 1–6, rows: A–D) that were used in the experiments. One catalyst (**2**), one boron source (**3**), six ligands (**4–9**) and four solvents (**10–13**) were screened for all starting materials. B_2Pin_2 , bis(pinacolato)diboron; CyHex, Cyclohexane; $[Ir(COD)OMe]_2$, (1,5-cyclooctadiene)(methoxy)iridium(I) dimer. **b**, Baseline model composed of a feed-forward neural network, using the molecular descriptor ECFP4 and the reaction conditions as input. Multilayer perceptron (MLP) modules are highlighted in orange, and the output is in blue. This baseline model was applied for the prediction of reaction yield and binary reaction outcomes. **c**, The molecular graph is featured with 2D or

3D information, with or without atomic partial charges (Methods for details on atom featurization). After passing the atomic features through a first MLP, the atomic features are updated via three 2D or 3D message-passing layers. Subsequently, the learned atomic features are either transformed directly to the regioselectivity output, or pooled via sum pooling or multi-head attention-based pooling operations to obtain a whole-molecule feature space. This learned molecular feature space is then combined with the embedded features of the reaction conditions (Methods for details on condition featurization) and transformed to the reaction output (reaction yield, binary reaction outcome) via a final MLP.

and lead diversification opportunities (Fig. 1b). Computational deep learning was employed for predicting reaction outcomes, yields and regioselectivity for the LSF of complex drug molecules. In the first step of this study, a comprehensive analysis of the published literature was performed to provide a rationale for selecting suitable reaction conditions for HTE screening and relevant substrates reflecting the nature of late-stage lead compounds in drug discovery. Reaction conditions were chosen from manually curated literature data based on 38 selected publications (the literature dataset). LSF substrates were chosen based on a cluster analysis of 1,174 approved drugs, resulting in 23 structurally diverse drug molecules. This approach enabled us to work with relevant examples of reaction conditions and substrates in an ‘former library’ approach (that is, an approach involving a chemical space tailored to the assessment of a synthetic methodology) rather than using

idealized substrates and fragments with limited applicability to lead optimization⁴⁸. In the second step of the study, semi-automated HTE was used for data generation (the experimental dataset). The reaction data for the selected drug molecules and reaction conditions provided high-quality data for subsequent machine learning of the reaction outcomes. Finally, different GNNs were trained on two-dimensional (2D), three-dimensional (3D) and atomic-partial-charge-augmented molecular graphs, to predict binary (yes/no) reaction outcomes, reaction yields and regioselectivity.

Results

High-throughput experimentation

Using a HTE set-up and liquid chromatography–mass spectrometry (LCMS) coupled to a reaction data analysis pipeline, 23 drug compounds

Table 1 | Model performance of the GNNs

	Reaction yield <i>r</i> value	Reaction yield m.a.e. (%)	Binary reaction outcome (random split), AUC (%)	Binary reaction outcome (substrate split), AUC (%)
GTNN2D	0.896±0.006	4.53±0.09	91.8±2.1	52±2
GNN2D	0.866±0.005	5.61±0.06	87.5±1.0	51±2
GTNN3D	0.884±0.01	4.51±0.11	91.4±0.7	58±4
GNN3D	0.877±0.001	5.33±0.34	89.4±0.8	65±5
GTNN2DQM	0.898±0.003	4.41±0.17	90.9±1.5	53±5
GNN2DQM	0.876±0.01	5.41±0.10	89.0±1.1	59±5
GTNN3DQM	0.890±0.01	4.23±0.08	91.8±0.9	67±2
GNN3DQM	0.890±0.006	4.88±0.24	89.1±0.9	64±4
ECFP4NN	0.885±0.0006	4.55±0.14	89.3±1.3	52±3
	<i>F</i> -score (%)	PPV (%)	TPR (%)	Accuracy (%)
aGNN2D	38±5	56±1	30±6	88±1
aGNN2DQM	39±2	54±2	30±3	88±0.3
aGNN3D	59±3	62±2	56±4	90±1
aGNN3DQM	60±4	62±2	59±6	90±1

The top of the table shows the model performance of the nine investigated neural networks, predicting binary reaction outcomes and reaction yields. Pearson correlation coefficient (*r*) and m.a.e. values were used to quantify reaction yield predictions. Balanced accuracy (AUC) was used to quantify binary reaction outcome predictions. The bottom of the table shows the model performance of the four different aGNNs for regioselectivity prediction in terms of *F*-score, PPV, TPR and accuracy. The numbers represent mean and standard deviation for *N*=3 independent neural network runs. The numbers in bold indicate the best performance for each of the individual metrics.

(**1**, **14**, **16–36**; structures of all compounds are in the Supplementary Information (Supplementary Section 3 and Supplementary Figs. 3 and 4)) and 12 drug-like fragments (**37–48**; Supplementary Section 3 and Supplementary Fig. 5) were screened using the plate layout depicted in Fig. 2a. Herein, the ensemble of the selected 23 drug compounds and 12 drug-like fragments is referred to as the LSF informer library. The 24-well borylation screening plate was designed based on a comprehensive literature assessment that delivered 1,301 reactions for meta-analysis. A detailed description of this approach is provided in the Methods.

In addition to the LSF informer library, a small subset of five frequently occurring literature substrates (**49–53**; Supplementary Section 3 and Supplementary Fig. 5) was screened by applying the borylation conditions. In total, a dataset containing the conditions and results of 956 reactions was obtained. LCMS measurement, followed by data analysis, enabled the determination of (1) binary (yes/no) reaction outcomes, that is, whether the conditions in combination with the individual substrates resulted in the desired mono- or di-borylated products, as well as (2) reaction yields, providing information about the amount of the desired reaction product. A protocol for visualizing the reaction outcome was implemented in the data analysis pipeline, which expedited the identification of starting points for suitable scaled-up procedures. Running selected reactions on larger scales indicated that individual conditions from the miniaturized HTE screenings can be adapted to produce sufficient material for biological tests or further post-borylation modification. In addition, the scale-up reactions enabled the determination of isolated yields and elucidation of the exact structure by nuclear magnetic resonance (NMR) spectroscopy and high-resolution mass spectrometry (HRMS) of a set of selected compounds (**1**, **25**, **29**, **37–39** and **45**). These analyses generated a high-quality experimental dataset containing information on the binary reaction outcomes, reaction yields and regioselectivity, which served as the basis for the geometric deep learning platform.

Geometric deep learning

The geometric deep learning platform introduced in this study consists of a set of different GNNs tailored to learn three targets: binary reaction outcome, reaction yield and regioselectivity. Three different model architectures were investigated, and four different molecular graph representations were evaluated for each architecture (Fig. 2c).

- Architectures. For the reaction tasks (binary reaction outcome, reaction yield), two network architectures were investigated: a GNN using sum pooling and a graph transformer neural network (GTNN) using graph multiset transformer-based pooling⁴⁹. For regioselectivity, an atomistic GNN (aGNN), which learns directly from atomic features, was employed.
- Molecular graphs. To quantify the influence of steric (3D) and electronic (quantum mechanical (QM)) effects, the input molecular graph was featured using 3D- and QM-augmented information, resulting in four different molecular graphs per neural network: 2D, 3D, 2DQM and 3DQM.

The various combinations resulted in eight different GNNs for each of the reaction tasks (binary reaction outcome and reaction yield) and four for regioselectivity (Table 1). For the reaction tasks, a baseline neural network was investigated using the well-established extended connectivity fingerprint (ECFP (ref. 50); Fig. 2b).

Reaction yield and reaction outcome

Eight different GNNs and the baseline method, ECFP4NN, were optimized to predict reaction yields and binary reaction outcomes.

The performance of the reaction yield predictions was investigated on a randomly split dataset to learn reaction yields for known substrates in combination with new conditions for the experimental dataset. Figure 3a shows a scatter plot of the predictions of the best-performing neural network, GTNN3DQM, achieving a mean absolute error (m.a.e.) of 4.23 ± 0.08% and a Pearson correlation, *r*, of 0.890 ± 0.01. Figure 3d (left) shows a comparison of the nine different neural networks for this task. The four GTNNs (4.23–4.53% m.a.e.) achieved considerably higher accuracy than the ECFP4NN baseline (4.55% m.a.e.) and the four GNNs (4.88–5.61% m.a.e.). For reaction yield prediction, atomic charges as well as 3D information did not influence the performance of either the GTNNs or GNNs. GTNN models trained on the literature dataset achieve substantially higher errors with m.a.e. values of 16.15–16.73% and a correlation between *r* = 0.59 and *r* = 0.62 (Supplementary Section 9.2 for details). The observation of lower errors for reaction yield predictions for HTE data compared to literature data is in line with recent findings⁵¹.

Binary reaction outcomes were considered ‘successful’ if the reaction condition with the chosen substrate yielded a mono- or di-borylation product that could be confirmed by LCMS with a corresponding conversion of ≥1%, or ‘unsuccessful’ if the desired transformation was not traceable with LCMS. For the machine learning models trained on binary reaction outcomes, two different dataset splits were investigated: (1) a random split to investigate the performance on new conditions for known substrates; and (2) a substrate-based split for the 23 drug molecules to investigate the performance on unknown substrates with different conditions. First, the binary reaction outcome prediction was evaluated for random data splits (that is, predicting reaction outcomes for novel reaction conditions on known substrates). Figure 3d (centre left) shows a comparison of the nine different neural networks developed for this task. For the binary reaction outcome as observed for reaction yield prediction, a similar trend can be perceived; that is, GTNNs slightly outperformed (90.9–91.8% area under receiver operating characteristic curve, AUC) the ECFP4NN model (89.3% AUC) and GNN model (87.5–89.1% AUC), and the augmentation with atomic partial charges as well as 3D information did not affect the performance of the models (Table 1). Figure 3b shows a confusion matrix that is observed for predictions with a binary threshold of ≥1%. Models with

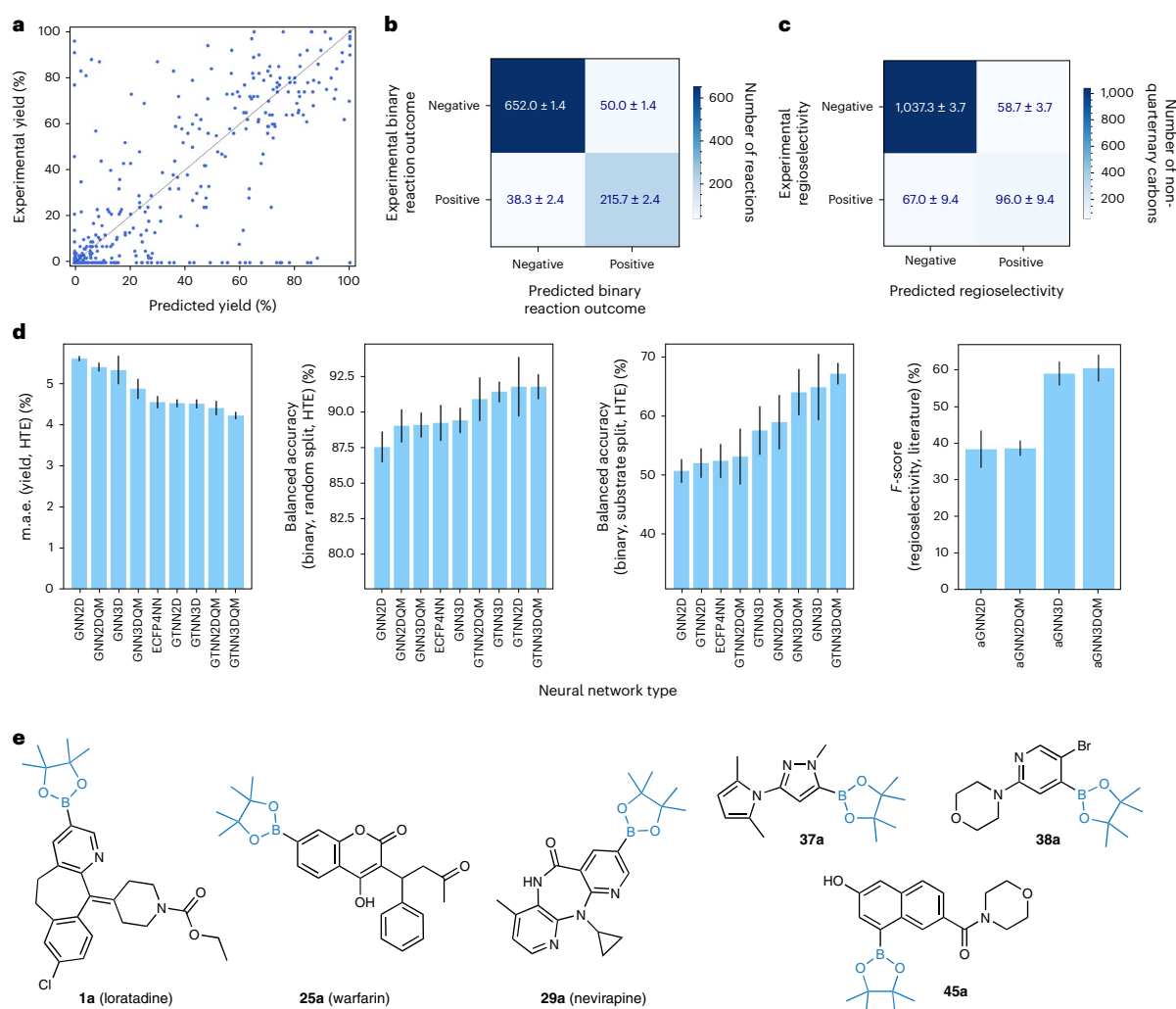


Fig. 3 | Results of binary reaction outcome, reaction yield and regioselectivity predictions.

a, Performance of reaction yield prediction on the experimental dataset. The scatter plot shows predicted reaction yields on the x axis and experimental reaction yields on the y axis for GTNN3DQM. Predictions were obtained from fourfold nested cross-validation, enabling the visualization of the whole dataset (details on dataset splitting are in Supplementary Section 1). **b**, Confusion matrix for binary reaction outcome prediction with a threshold of $\geq 1\%$ (confusion matrices with additional thresholds are in Supplementary Section 9.3). **c**, Confusion matrix for the prediction of non-quaternary carbons in the test set for aGNN3DQM. **d**, Performance of the investigated neural networks for four different tasks. Each bar plot shows the worst-performing model on the left and the best on the right. Error bars on all bar plots show the standard deviation observed on a threefold cross-validation of independent neural network

training runs on the same dataset split. The centre of the error bars denotes the mean performance observed for the threefold cross-validation. The number of predicted reaction data points in the test set (n) is annotated individually. The tasks are the m.a.e. as a percent for reaction yield prediction (left; experimental dataset, $n = 239$); balanced accuracy (centre left; AUC) as a percent on the binary reaction outcome prediction using the random dataset split (experimental dataset, $n = 239$); balanced accuracy (centre right; experimental dataset, $n = 239$); and the performance of the four aGNNs for regioselectivity prediction measured in terms of F -score (right; literature dataset, $n = 164$). **e**, Selected examples of validated borylation opportunities as predicted by the best-performing neural network (GTNN3DQM) binary reaction outcomes of unseen substrates for three drugs (**1**, **25**, **29**) and three fragments (**37**, **38**, **45**).

additional binary thresholds of $>5\%$, $>10\%$ and $>20\%$ were developed (Supplementary Section 9.3), achieving similar accuracy (AUC for 1% threshold, $94.5 \pm 0.2\%$; 5% threshold, $94.5 \pm 0.2\%$; 10% threshold, $95.6 \pm 0.3\%$; and 20% threshold, $94.4 \pm 0.2\%$).

Furthermore, the binary reaction outcome prediction was evaluated for substrate-based data splits (that is, predicting reaction outcomes for novel substrates). For 20 of the 23 unseen drugs, GTNN3DQM achieved an accuracy greater than 50%; for 16 of the 23 unseen drugs,

an accuracy greater than 80% was obtained. Overall, the GTNN3DQM model exhibited an AUC value of $67 \pm 2\%$ (Table 1). Figure 3d (centre right) shows a comparison of the nine different neural networks for this task, indicating a better performance for the GNNs trained on 3D graphs (58–67% AUC) in comparison to the ECFP4NN (52% AUC) and the GNNs and GTNNs trained on 2D graphs (51–59% AUC). Furthermore, augmentation with atomic partial charges did not show improvements for GNNs or GTNNs. Figure 3e shows three drugs (**1**, **25**, **29**) and three

Article

<https://doi.org/10.1038/s41557-023-01360-5>

fragments (**37**, **38**, **45**) that were predicted by GTNN3DQM to yield successful reaction outcomes for unseen substrates. The main reaction products of these six substrates were isolated with reaction yields ranging from 5% to 90% (Supplementary Section I1 for experimental details).

Regioselectivity

Four different aGNN models were developed for regioselectivity prediction by training the neural networks computed for all non-quaternary carbons in a given molecule to determine whether the reaction will occur. As borylation reactions regularly occur at one atom or, in rare cases, at two atoms in a molecule, the atomic labels 'reactive' and 'non-reactive' in a molecule are unbalanced (approximately 1:6). Therefore, the *F*-score (that is, the mean of positive predictive value (PPV) and true positive rate (TPR)) was used as a measure of neural network accuracy.

Figure 3d (right) shows the performance of four aGNNs trained on the literature dataset. The aGNNs trained on 3D graph structures outperformed those trained on 2D graph structures (Table 1 shows the exact numbers). The graph structures that included atomic partial charges did not appear to improve the prediction accuracy of the networks compared to their 2D and 3D equivalents. The aGNN3DQM model was the best-performing model overall, with an *F*-score of $60 \pm 4\%$. Figure 4c shows six selected predictions of the test set using aGNN3DQM; on the left side, three reactions from the top 20% are shown, and on the right side, three molecules from the bottom 20% of the test set are shown. Figure 3c features the confusion matrix of the aGNN3DQM predictions on the test set. For the 1,259 non-quaternary carbons in the test set, aGNN3DQM achieved an accuracy of $90 \pm 1\%$, a PPV of $62 \pm 2\%$ and a TPR of $59 \pm 6\%$. Table 1 lists the accuracy, PPV values, TPR values and *F*-scores of the four aGNN models. The aGNNs trained on 2D graph structures yielded a similar false positive rate (that is, similar PPV), but a much higher false negative rate (that is, lower TPR) than the aGNNs trained on 3D graph structures.

The regioselectivity prediction method aGNN3D was trained and subsequently validated on the literature dataset. Test set predictions revealed many accurate examples (Fig. 4a; **54**, **55**) but also pointed to certain limitations of the computational model (Fig. 4a; **56**, **57**). For additional testing, aGNN3D was retrospectively applied to out-of-distribution reactions containing substrates outside of the literature dataset found in Roche Medicinal Chemistry legacy projects (Fig. 4b). The model predicted three potential sites of reaction for morpholine **45**, two of which were experimentally confirmed. For carbamate **64**, the correct site of borylation and one false positive site were predicted. The aGNN3D model was then prospectively validated using six selected borylation reactions of the drugs Loratadine (**1**), warfarin (**25**) and nevirapine (**29**), and three fragments (**37**, **38**, **39**; Fig. 4c).

The prediction model achieved approximately 70% accuracy in this experiment. Five of seven experimentally observed borylation sites were correctly predicted by the model. Figure 4c illustrates the six predictions compared to the isolated and characterized products obtained through the scaled-up reactions of the best-observed

screening conditions. For fragments **37** and **38** and the drug nevirapine (**29**), the model predicted only one site of borylation. The predicted sites were experimentally confirmed, and neither false positive nor false negative predictions were observed. For Loratadine (**1**), aGNN3D predicted two potential reaction sites. The predicted mono-borylation product **1a** was isolated, and the regioselectivity prediction was confirmed. For the second predicted species, the exact position of the two pinacol esters on Loratadine (**1**) could not be directly confirmed by NMR, but the respective mass was confirmed by HRMS. Product **1b** was consequently subjected to hydrolysis to obtain the corresponding phenol **1c** (Supplementary Section I1). The analysis revealed that the second prediction was incorrect. For warfarin (**25**), aGNN3D predicted two potential reaction sites, scoring $93 \pm 5\%$ and $48 \pm 1\%$. Mono-borylation of the C–H bond with the most confident prediction (93%) was experimentally confirmed. For fragment **39** the regioselectivity model did not suggest that borylation occurs, but mono-borylation was observed during the screening, and a scale-up was conducted. This analysis revealed that **39** in fact underwent mono-borylation of the methyl group to deliver **39a**.

Finally, we investigated the influence of substitutions with different steric hindrances and electronic effects on the regioselectivity predictions. The aGNN3D model was applied to six unseen examples from the literature test set that introduce steric hindrance or directing functional groups. Figure 4d illustrates the regioselectivity predictions for four indole derivatives. Placing a directing amide functionality in position 1 yielded a prediction of $99 \pm 0\%$ at position 7 (Fig. 4d). Substituting the directing amide functionality with a bulky triisopropylsilane blocks position 7 and therefore yielded a score of $41 \pm 7\%$ for position 3 (Fig. 4d). Furthermore, blocking position 3 with a cyano group and keeping the triisopropylsilane in position 1 in place yielded a prediction score of $96 \pm 2\%$ for position 5 (Fig. 4d). For a directing keto functionality at position 3, a score of $84 \pm 3\%$ was obtained for position 4 (Fig. 4d, right). Figure 4d illustrates the regioselectivity predictions for two thiophene derivatives. Placing a directing secondary amide functionality at position 2 shows a slight preference at position 3 with a score value of $40 \pm 1\%$ (Fig. 4d). Replacing the directing secondary amide at position 2 with a bulky tertiary amide shifts the high score ($72 \pm 5\%$) to position 5 (Fig. 4d). For all of these examples, the highest prediction is in line with observed mono-borylations in the literature^{52–55}. These results conclude that the regioselectivity prediction model aGNN3D successfully considers steric and electronic substituent effects.

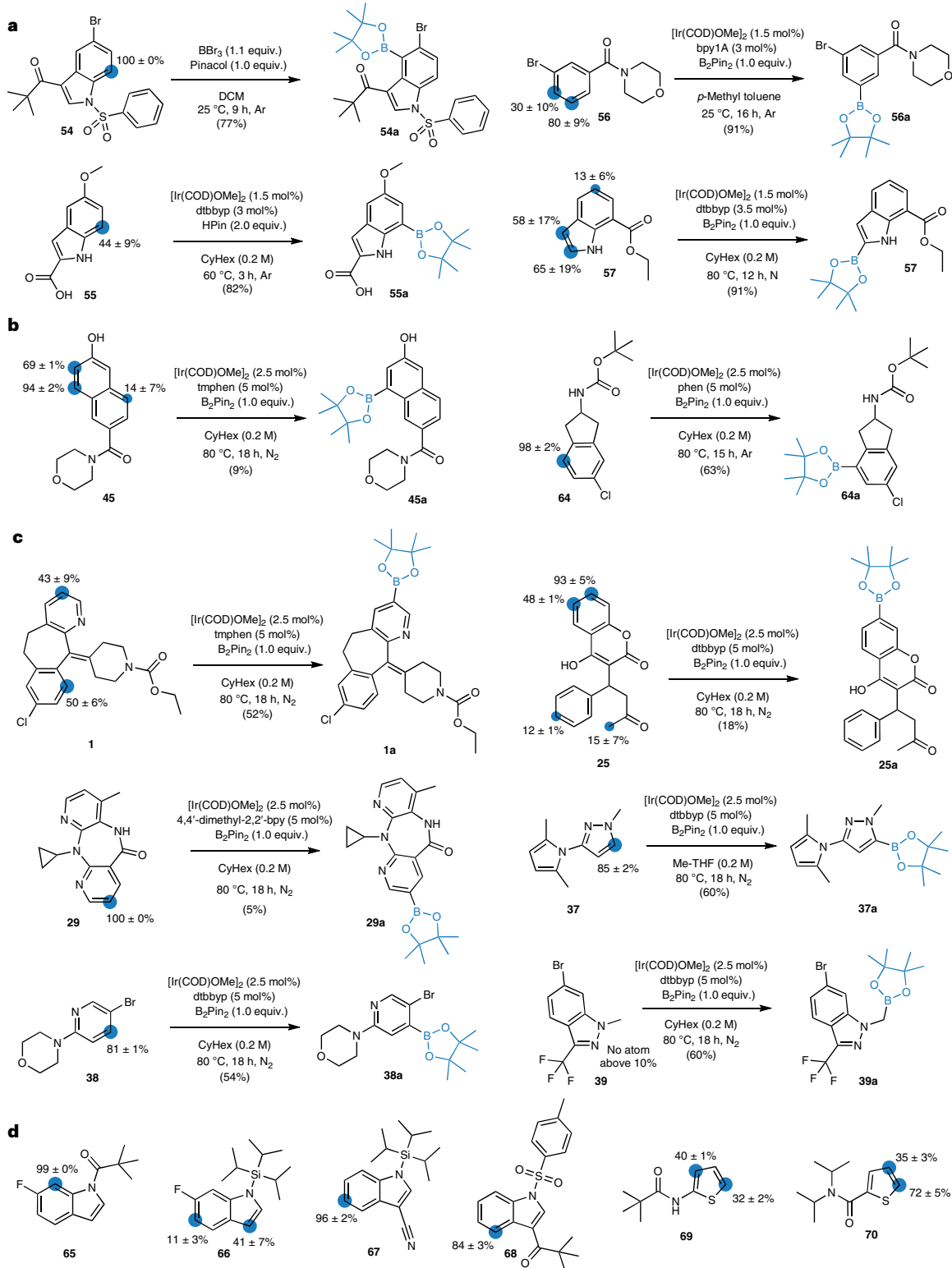
Discussion

Curated high-quality reaction data are key drivers of successful deep learning. The results of this study were obtained using two FAIR datasets (that is, literature and experimental) containing 1,301 and 956 reactions, respectively. To lower the barrier to sharing reaction data, we developed a comprehensive reaction data format (SURF, simple user-friendly reaction format) that allows for FAIR data capture. A detailed description of the SURF structure and data templates is provided in Supplementary Section 7. SURF complements similar initiatives, such as the open reaction database (ORD) and unified data model

Fig. 4 | Selected examples from the borylation regioselectivity prediction.

a–d, For each transformation, the predicted regioselectivity is shown on the left, and borylation including the reported reaction conditions and experimentally validated regioselectivity are shown on the right. The percentages for the regioselectivity predictions were generated by aGNN3D through the mean and standard deviation on ten individual conformers. Every prediction resulted in a value between zero and one, where one was set to 100%. **a**, Retrospective results obtained from the test set of the literature dataset. Results for two reactions from the top 20% (**54**, **55**) and bottom 20% (**56**, **57**) of the predictions from the literature dataset. **b**, Retrospective results obtained from out-of-distribution reactions from Roche legacy projects. Validation is shown for two molecules (**45**, **64**). **c**, Prospective experimental validation of regioselectivity prediction models that were trained on the literature dataset. Validation is shown for three

drugs, Loratadine (**1**), warfarin (**25**) and nevirapine (**29**), and three fragments, **37**, **38** and **39**. **d**, Influence of steric hindrance and directing functional groups on regioselectivity prediction for six selected examples from the test set of the literature dataset. Regioselectivity predictions of indole derivatives (**65–68**) and thiophene derivatives (**69**, **70**). The numbering of the shown indole molecule starts with 1 for the nitrogen atom and proceeds around the carbons in the ring, numbering the carbon atoms 2–7. DCM, dichloromethane; BBr₃, boron tribromide; dtbbpy, 4,4'-di-tert-butyl-2,2'-dipyridyl; byp1A, 1-(2-(1,2'-bipyridin]-5-yl)phenyl)-3-cyclohexylurea; byp, bi-pyridine; Cy, cyclohexane; HPin, pinacolborane; [Ir(COD)OMe]₂, (1,5-cyclooctadiene)(methoxy)iridium(I) dimer; phen, 1,10-phenanthroline; tmphen, 3,4,7,8-tetramethyl-1,10-phenanthroline; N₂, nitrogen.



Article

<https://doi.org/10.1038/s41557-023-01360-5>

(UDM)^{56,57}. It was developed to enable scientists to store and share reaction data in an easily editable format. High-quality literature data and newly generated experimental reaction data have enabled *in silico* estimation of reaction outcomes and reaction selectivity. The resulting geometric deep learning platform has been shown to correctly predict the reaction outcome for six substrates, and their main products were isolated (Supplementary Section 11). This approach represents a tool for identifying late-stage modifications of advanced drug-like molecules before initiating resource-intensive synthesis.

Two GNN architectures were implemented to predict the reaction tasks (binary reaction output and reaction yield). The two models, GNN and GTNN, differ only in their pooling operations. Whereas the GNN uses sum pooling, the GTNN relies on more complex graph multiset transformer-based pooling. This additional flexibility of the GTNNs slightly improved the reaction yield prediction but did not lead to increased prediction performance for binary reaction outcomes. This result suggests that greater neural network flexibility may lead to improved prediction accuracy for certain reaction prediction tasks but does not offer a general advantage.

The best-performing neural network model for reaction yield prediction (GTNN3DQM) achieved a m.a.e. of $4.23 \pm 0.08\%$ with a Pearson correlation of $r = 0.890 \pm 0.01$ on the experimental dataset (Table 1), whereas the most accurate model for literature data prediction (GTNN2DQM) achieved a m.a.e. of $16.11 \pm 0.02\%$ with $r = 0.61 \pm 0.01$ (Supplementary Section 9.2 for details). This disparity can be explained by the heterogeneity and quality of the two datasets. The experimental data were generated in the same laboratory using the same equipment for syntheses and analyses and included the same standard for determining the reaction yield in all experiments. Furthermore, the experimental dataset covers a less diverse reaction parameter space (that is, 24 versus 864 possible conditions per substrate), thereby facilitating the learning task. By contrast, the reaction outcomes in the literature dataset originate from a variety of experiments performed in different laboratories that used different methods for determining the yield (for example, isolated yield, reaction conversion assessed by NMR, LCMS). Standardized, chemically diverse, high-quality datasets will be beneficial for building accurate machine learning models that enable further optimization of reaction conditions for LSF.

Importantly, the incorporation of steric information via 3D molecular graphs led to improved neural network performance for all investigated tasks, ranging from small enhancements in reaction yield prediction (m.a.e., 4.2% versus 4.4%) and binary reaction outcomes (AUC, 67% versus 59%) to substantial improvements in regioselectivity predictions (*F*-score, 60% versus 39%). Implementing partial charges generated with DFT accuracy into neural networks did not exhibit any improvements in all investigated tasks. However, the explored borylation reactions are mainly guided by steric effects and, to a lesser extent, electronic effects^{58,59}, which could explain these observed effects. Incorporating the local 3D geometry considerably improved regioselectivity predictions from 38 ± 5 for the best-performing 2D model to 60 ± 4 for the best-performing 3D model. These observations demonstrate the relevance of the local geometries and the additional information provided by 3D graphs for reactivity prediction on the level of individual atomic environments.

Regioselectivity predictions on the literature data delivered accurate results for the majority (90%) of the cases. The four selected and validated substrates from the experimental dataset highlight the reaction biases in the literature data used for model training. Specifically, the majority of the borylations captured in the literature dataset occur at *sp*² carbons on substrates with no more than two ring systems. Substrates that fulfil these characteristics, such as fragments **37** and **38**, are predicted correctly. However, substrates outside of this scope, including the *sp*³-carbon borylation on fragment **39** or the di-borylation on the annulated pentadecanyl moiety in Loratadine (**1**), exploit the limitations of the available literature data. These results conclude that

small datasets, such as the presented 1,301 reactions from the literature in this study, are sufficient for predicting regioselectivity with GTNNs on substrates similar to the ones covered by the chemical space in the literature. However, to predict regioselectivity in a trustworthy manner for a broader chemical space including larger molecules and potentially also *sp*³ borylations, further training data will be required.

The LSF informer library containing 23 structurally diverse, approved drugs (**1**, **14–36**) complemented with 12 fragments (**37–48**) and five idealized substrates (**49–53**) yielded a dataset covering the essential chemical motifs relevant in drug discovery. A functional group analysis revealed that 33 (82.5%) of the 40 most abundant functional groups extracted from the 1,174 drug molecules are covered by the LSF informer library. Further analysis highlighted that functional groups that are known to exhibit the desired borylation reaction, such as aromatic nitrogens, aromatic alkyl-oxy groups and alcohols, are also among the functional groups in the LSF informer library that show the highest tolerance for successful reaction outcomes. On the contrary, certain functional groups such as primary amines, carbamates and carbonates, or aromatic functional groups with strong electron-withdrawing moieties (for example, nitro-aryls) were found to be less tolerated and inhibit desired reaction outcomes (Supplementary Section 8.2 for further details on the functional group analysis). Since every substrate was screened with every reaction condition, further insights about reaction conditions could be gained (Supplementary Tables 4 and 5). Whereas the best-performing ligand was **9** (33%), **6–8** (28–30%) showed similar good results, whereas **5** (22%) and especially **4** (17%) delivered fewer successful reaction outcomes. Moreover, reaction outcomes were further influenced by solvents. Cyclohexane (**10**, 50%) outperformed the other three solvents 2-methyltetrahydrofuran (Me-THF; **11**, 43%), cyclopentyl methyl ether (CPME; **12**, 38%) and acetonitrile (MeCN; **13**, 29%).

HTE and GNNs have previously been used for identifying substrates suitable for C–H activation⁴¹. This present study extends this original approach by (1) using HTE and GNNs for drug molecules, (2) introducing a literature search strategy that enables the selection of a structurally diverse set of substrates and ideal plate reaction screening conditions and (3) introducing a flexible geometric deep learning approach that considers the influence of steric and electronic effects of the substrates and allows the prediction of reaction outcome, yield and regioselectivity.

The structural and shape diversity of the compounds used for training the regioselectivity prediction model considerably exceeds the compound diversity of a recent report on regioselectivity prediction for iridium-catalysed borylation reactions⁴⁷. Compound clustering, scaffold and shape analyses of both datasets revealed greater chemical diversity of our training data. Furthermore, the neural networks were developed with more examples and broader chemical space coverage (Supplementary Section 9.1, Supplementary Figs. 14 and 15 and Supplementary Tables 6 and 7). Importantly, the estimated three dimensionality of the data is characteristic of molecules typically observed in medicinal chemistry⁶⁰. These findings positively advocate for using these computational models for drug discovery.

In conclusion, the results of this study confirm the practical applicability of the geometric deep learning platform in bioorganic and medicinal chemistry and their potential benefit for laboratory automation. The approach is routinely and successfully applied to assess binary reaction outcome, reaction yield and regioselectivity for borylation opportunities in drug discovery projects at F. Hoffmann-La Roche Ltd. Additional data points are continuously generated by standardized HTE to further enhance the predictive power of the computational models presented. For future improvements, (1) additional reaction conditions for iridium-catalysed borylation will be explored. This extended screening panel could include exchanging the catalyst or boron source as well as using a broader variety of ligands and solvents. In addition, (2) the LSF informer library can be augmented to include

Article

<https://doi.org/10.1038/s41557-023-01360-5>

more frequently occurring fragments in drug molecules to expand the relevant chemical space and potentially improve the performance of the machine learning pipeline. Finally, (3) less frequently employed transition-metal-catalysed or even metal-free synthesis methods can be investigated to enhance the coverage of the reaction conditions, addressing reactions from publications initially excluded from the analysis.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41557-023-01360-5>.

References

- Jana, R., Begam, H. M. & Dinda, E. The emergence of the C–H functionalization strategy in medicinal chemistry and drug discovery. *Chem. Commun.* **57**, 10842–10866 (2021).
- Werner, M. et al. Seamless integration of dose–response screening and flow chemistry: efficient generation of structure–activity relationship data of β -secretase (BACE1) inhibitors. *Angew. Chem. Int. Ed.* **53**, 1704–1708 (2014).
- Parry, D. M. Closing the loop: developing an integrated design, make, and test platform for discovery. *ACS Med. Chem. Lett.* **10**, 848–856 (2019).
- Sutherland, J. D. et al. An automated synthesis–purification–sample-management platform for the accelerated generation of pharmaceutical candidates. *J. Lab. Autom.* **19**, 176–182 (2014).
- Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
- Wencel-Delord, J. & Glorius, F. C–H bond activation enables the rapid construction and late-stage diversification of functional molecules. *Nat. Chem.* **5**, 369–375 (2013).
- Nippa, D. F. et al. Late-stage functionalization and its impact on modern drug discovery: medicinal chemistry and chemical biology highlights. *Chimia* **76**, 258–258 (2022).
- Hartwig, J. F. Borylation and silylation of C–H bonds: a platform for diverse C–H bond functionalizations. *Acc. Chem. Res.* **45**, 864–873 (2012).
- Wang, M. & Shi, Z. Methodologies and strategies for selective borylation of C–Het and C–C bonds. *Chem. Rev.* **120**, 7348–7398 (2020).
- Lasso, J. D., Castillo-Pazos, D. J. & Li, C.-J. Green chemistry meets medicinal chemistry: a perspective on modern metal-free late-stage functionalization reactions. *Chem. Soc. Rev.* **50**, 10955–10982 (2021).
- Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist’s toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016).
- Guillemard, L., Kaplaneris, N., Ackermann, L. & Johansson, M. J. Late-stage C–H functionalization offers new opportunities in drug discovery. *Nat. Rev. Chem.* **5**, 522–545 (2021).
- Stepan, A. F. et al. Late-stage microsomal oxidation reduces drug–drug interaction and identifies phosphodiesterase 2A inhibitor PF-06815189. *ACS Med. Chem. Lett.* **9**, 68–72 (2018).
- Halperin, S. D., Fan, H., Chang, S., Martin, R. E. & Britton, R. A convenient photocatalytic fluorination of unactivated C–H bonds. *Angew. Chem. Int. Ed.* **126**, 4778–4781 (2014).
- Friis, S. D., Johansson, M. J. & Ackermann, L. Cobalt-catalysed C–H methylation for late-stage drug diversification. *Nat. Chem.* **12**, 511–519 (2020).
- Dreher, S. D., Dormer, P. G., Sandrock, D. L. & Molander, G. A. Efficient cross-coupling of secondary alkyltrifluoroborates with aryl chlorides reaction discovery using parallel microscale experimentation. *J. Am. Chem. Soc.* **130**, 9257–9259 (2008).
- Bellomo, A. et al. Rapid catalyst identification for the synthesis of the pyrimidinone core of HIV integrase inhibitors. *Angew. Chem. Int. Ed.* **124**, 7018–7021 (2012).
- Buitrago Santanilla, A. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).
- Shevlin, M. Practical high-throughput experimentation for chemists. *ACS Med. Chem. Lett.* **8**, 601–607 (2017).
- Krska, S. W., DiRocco, D. A., Dreher, S. D. & Shevlin, M. The evolution of chemical high-throughput experimentation to address challenging problems in pharmaceutical synthesis. *Acc. Chem. Res.* **50**, 2976–2985 (2017).
- Mennen, S. M. et al. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Org. Process. Res. Dev.* **23**, 1213–1242 (2019).
- Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
- Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
- Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 1263–1272 (PMLR, 2017).
- Unke, O. T. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
- Isert, C., Kromann, J. C., Stiefl, N., Schneider, G. & Lewis, R. A. Machine learning for fast, quantum mechanics-based approximation of drug lipophilicity. *ACS Omega* **8**, 2046–2056 (2023).
- Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. *Curr. Opin. Struct. Biol.* **79**, 102548 (2023).
- Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
- Shen, Y. et al. Automation and computer-assisted planning for chemical synthesis. *Nat. Rev. Methods Prim.* **1**, 23 (2021).
- Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3**, 1023–1032 (2021).
- Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, **34**, 9405–9415, <https://proceedings.neurips.cc/paper/2021/hash/4e2a6330465c8ffcaa696a5a16639176-Abstract.html> (2021).
- Guan, Y. et al. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **12**, 2198–2208 (2021).
- Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler–Lehman network. In *Advances in Neural Information Processing Systems (NeurIPS)*, **30**, https://papers.nips.cc/paper_files/paper/2017/hash/ced556cd9f9c0c8315cfbe0744a3baf0-Abstract.html (2017).
- Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
- Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O. & Reymond, J.-L. Retrosynthetic accessibility score (RA_{score}) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **12**, 3339–3349 (2021).
- Heinen, S., von Rudorff, G. F. & von Lilienfeld, O. A. Toward the design of chemical reactions: machine learning barriers of competing mechanisms in reactant space. *J. Chem. Phys.* **155**, 064105 (2021).

Article

<https://doi.org/10.1038/s41557-023-01360-5>

38. Bragato, M., von Rudorff, G. F. & von Lilienfeld, O. A. Data enhanced Hammett-equation: reaction barriers in chemical space. *Chem. Sci.* **11**, 11859–11868 (2020).
39. von Rudorff, G. F., Heinen, S. N., Bragato, M. & von Lilienfeld, O. A. Thousands of reactants and transition states for competing E2 and S_N2 reactions. *Mach. Learn. Sci. Technol.* **1**, 045026 (2020).
40. Stuyver, T. & Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: performance, generalizability, and explainability. *J. Chem. Phys.* **156**, 084104 (2022).
41. Qiu, J. et al. Selective functionalization of hindered meta-C–H bond of o-alkylaryl ketones promoted by automation and deep learning. *Chem* **8**, 3275–3287 (2022).
42. King-Smith, E. et al. Predictive Minisci and P450 late stage functionalization with transfer learning. Preprint at *ChemRxiv* <https://doi.org/10.26434/chemrxiv-2022-7ddw5> (2022).
43. Hoque, A. & Sunoj, R. B. Deep learning for enantioselectivity predictions in catalytic asymmetric β-C–H bond activation reactions. *Digit. Discov.* **1**, 926–940 (2022).
44. Boni, Y. T., Cammarota, R. C., Liao, K., Sigman, M. S. & Davies, H. M. Leveraging regio- and stereoselective C(sp³)–H functionalization of silyl ethers to train a logistic regression classification model for predicting site-selectivity bias. *J. Am. Chem. Soc.* **144**, 15549–15561 (2022).
45. Xu, L.-C. et al. Enantioselectivity prediction of palladium-catalysed C–H activation using transition state knowledge in machine learning. *Nat. Synth.* **2**, 321–330 (2023).
46. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
47. Caldeweyher, E. et al. Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation. *J. Am. Chem. Soc.* **145**, 17367–17376 (2023).
48. Kutchukian, P. S. et al. Chemistry informer libraries: a cheminformatics enabled approach to evaluate and advance synthetic methods. *Chem. Sci.* **7**, 2604–2613 (2016).
49. Baek, J., Kang, M. & Hwang, S. J. Accurate learning of graph representations with graph multiset pooling. In *International Conference on Learning Representations (ICLR)* (2021).
50. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
51. Wiest, O. et al. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **14**, 4997–5005 (2023).
52. Yin, Q., Klare, H. F. & Oestreich, M. Catalytic Friedel-Crafts C–H borylation of electron-rich arenes: dramatic rate acceleration by added alkenes. *Angew. Chem. Int. Ed.* **56**, 3712–3717 (2017).
53. Lv, J. et al. Metal-free directed sp²-C–H borylation. *Nature* **575**, 336–340 (2019).
54. Feng, Y. et al. Total synthesis of verruculogen and fumitremorgin A enabled by ligand-controlled C–H borylation. *J. Am. Chem. Soc.* **137**, 10160–10163 (2015).
55. Bisht, R., Hoque, M. E. & Chattopadhyay, B. Amide effects in C–H activation: noncovalent interactions with L-shaped ligand for meta borylation of aromatic amides. *Angew. Chem. Int. Ed.* **57**, 15762–15766 (2018).
56. Kearnes, S. M. et al. The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
57. Tomczak, J. et al. UDM (unified data model) for chemical reactions – past, present and future. *Pure Appl. Chem.* <https://doi.org/10.1515/pac-2021-3013> (2022).
58. Hartwig, J. F. Regioselectivity of the borylation of alkanes and arenes. *Chem. Soc. Rev.* **40**, 1992–2002 (2011).
59. Wright, J. S., Scott, P. J. & Steel, P. G. Iridium-catalysed C–H borylation of heteroarenes: balancing steric and electronic regiocontrol. *Angew. Chem. Int. Ed.* **60**, 2796–2821 (2021).
60. Meyers, J., Carter, M., Mok, N. Y. & Brown, N. On the origins of three-dimensionality in drug-like molecules. *Future Med. Chem.* **8**, 1753–1767 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Article

<https://doi.org/10.1038/s41557-023-01360-5>

Methods

Literature analysis

The systematic analysis of chemical transformations (SACT) of the data retrieved from literature consisted of four steps: (1) literature search, (2) literature data curation and evaluation, (3) methodology extraction and (4) reaction data curation and analysis. All details of the literature analysis are provided in Supplementary Section 2. The literature analysis identified 38 publications describing relevant borylation methods, from which the reaction data were manually extracted to obtain a high-quality dataset containing 1,301 chemical transformations. Meta-analysis of these data provided a foundation for an informed plate design.

LSF informer library

The concept of chemical informer libraries, initially reported by Merck^{48,61}, served as the basis for developing the LSF informer library. Applying a clustering method based on structural features to a dataset containing 1,174 approved small-molecule drugs yielded eight structurally diverse groups of molecules. Details of the applied clustering and visualization of the cluster via principal component analysis are provided in Supplementary Section 3. Three molecules were selected from each cluster based on their distance from the cluster centre, price and availability and were subjected to borylation screening. To complement the model with fragments relevant to Roche's chemical space, the top 100 most popular ring assemblies found in the Roche corporate compound collection were identified. For these ring assemblies, substructure searches were performed for the entire database. The resulting compounds were retained if (1) the structures had a molecular weight below 300 g mol⁻¹ or fewer than 20 non-hydrogen atoms, (2) there was at least 1 g of powder stock available and (3) the structures were not used in any internal project or subject to legal restrictions. Out of this pool of candidates, 12 fragments were manually selected. Further details on the determination and constitution of the LSF informer library are described in Supplementary Section 3.

Screening plate design

Following the SACT approach that delivered a curated high-quality literature data set, a meta-analysis was conducted to define a clear rationale for determining the conditions for the 24-well borylation screening plate used for the LSF informer library. This analysis included the temperature (T), time (t), reaction concentration (c) and scale (n), selected based on the median values for our screening plate ($T = 80\text{ }^{\circ}\text{C}$, $t = 16\text{ h}$, $c = 0.2\text{ M}$, $n = 100\text{ mmol}$). Subsequently, the number of reaction components generally used for borylation reactions (catalyst, ligand, boron source and solvent) was determined. Owing to the limited space on the 24-well plate and the high occurrence of $[\text{Ir}(\text{COD})(\text{OMe})_2]$ (**2**), **2** was chosen as the catalyst. Analysis of the reagents used in combination with **2** provided the rationale for choosing B_2Pin_2 (**3**) as the boron source. This selection made it possible to screen a set of six ligands and four solvents. Six rather than four ligands were used because the dataset showed a greater variety of ligands than solvents. The ligands were assessed based on the chemical diversity of the converted starting materials and their commercial availability. Based on these results, six ligands from four chemical classes were selected. While the meta-analysis revealed that low-boiling solvents are the predominant solvents for borylation, their corresponding higher-boiling analogues (for example, Me-THF instead of tetrahydrofuran, THF) were selected to avoid potential solvent evaporation at 80 °C and reduce the risk of cross-contamination. The detailed meta-analysis results leading to the final plate design are described in Supplementary Section 4.

HTE borylation screening

Using a 24-well plate design (Fig. 3), all drug molecules from the LSF informer library and selected fragments (Supplementary Section 3 and Supplementary Figs. 3–5) were screened. The reaction set-up (automated

solid dosing and solvent addition) and execution (heating and stirring) in glass vials on a parallel screening plate were conducted in a glove box under a nitrogen atmosphere. Upon completion of the reaction, the solvents were removed through evaporation, followed by automated resuspension of the residues in MeCN/H₂O and dilution to a defined concentration for LCMS analysis using a liquid handler. The samples were then analysed by LCMS, and the resultant data were subjected to an automated reaction data analysis pipeline (Supplementary Figure 6) to rapidly determine all components within the mixture. Standardized reaction data output (SURF; Supplementary Section 7) allowed direct visualization of reaction outcome with the TIBCO Spotfire software as well as the direct loading into machine learning models. The general screening procedure, including detailed information on the hardware and software used, is provided in Supplementary Sections 5 and 6).

Scaled-up reactions

Selected molecules (three drugs, **1**, **25** and **29**; and four fragments, **37**, **38**, **39** and **45**) showing substantial conversion to the respective borylation products were scaled up using the most promising conditions. All reactions were conducted under a nitrogen atmosphere in a glove box using glass reaction vessels with pressure release caps and standard stirring bars. Purification was performed using flash chromatography or reversed-phase high-pressure liquid chromatography. In selected cases, where separation of the borylated species could not be achieved, the boronic ester was transformed into a hydroxyl group. Structural elucidation was performed using NMR and HRMS. The full analytical results and spectra for all compounds are shown in Supplementary Sections 11 and 12.

Deep learning

Graph neural network architecture. The following paragraphs describe the neural network architecture of the three introduced GNNs (that is, GNN, GTNN and aGNN). GNN and GTNN were trained to learn the two reaction properties (that is, binary reaction outcome and reaction yield), and aGNN was trained to learn regioselectivity. Details about dataset splitting are in Supplementary Section 1.

Molecular graph. For each of the three GNNs (that is, GNN, GTNN and aGNN), four different input molecular graph representations were investigated, which include steric (3D) and electronic (QM) features in different combinations, yielding four different molecular graphs: 2D, 2DQM, 3D and 3DQM.

E(3)-invariant message passing. The atomic features and optionally DFT-level partial charges were embedded and transformed using a MLP, resulting in atomic features \mathbf{h}_i^0 . E(3)-invariant message passing in a similar fashion as suggested by Satorras et al.⁶² was applied to l layers over all atomic representations \mathbf{h}_i^l and their edges. Edges were defined by covalent bonds for the 2D graph and all atoms within a radius of 4 Å for the 3D graph, respectively. All networks contained three message-passing layers. In each message-passing layer, the atomic representations were transformed via equation (1)

$$\mathbf{h}_i^{l+1} = \phi \left(\mathbf{h}_i^l, \sum_{j \in \mathcal{N}(i)} \psi(\mathbf{h}_i^l, \mathbf{h}_j^l) \right), \quad (1)$$

for 2D graph structures, and equation (2)

$$\mathbf{h}_i^{l+1} = \phi \left(\mathbf{h}_i^l, \sum_{j \in \mathcal{N}(i)} \psi(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{r}_{ij}) \right), \quad (2)$$

for 3D graph structures.

In equations (1) and (2), \mathbf{h}_i^l is the atomic representation \mathbf{h} of the i th atom at the l th layer; $j \in \mathcal{N}(i)$ is the set of neighbouring nodes connected via edges; \mathbf{r}_{ij} the interatomic distance features (Methods, "Atom featurization" for details); ψ is a MLP transforming node features into message features \mathbf{m}_{ij} as $\mathbf{m}_{ij} = \psi(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{r}_{ij})$ for 3D graphs and $\mathbf{m}_{ij} = \psi(\mathbf{h}_i^l, \mathbf{h}_j^l)$

Article

<https://doi.org/10.1038/s41557-023-01360-5>

for 2D graphs; Σ denotes the permutation-invariant pooling operator (that is, sum) transforming \mathbf{m}_i into \mathbf{m}_i as $\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \mathbf{m}_j$; and ϕ is a MLP transforming \mathbf{h}_i^l and \mathbf{m}_i into \mathbf{h}_i^{l+1} . The atomic features from all layers $[\mathbf{h}_i^{l=1}, \mathbf{h}_i^{l=2}, \mathbf{h}_i^{l=3}]$ were concatenated and transformed via a MLP, resulting in final atomic features \mathbf{H} . \mathbf{H} was then transformed differently by the three GNNs, using sum pooling (GNN) or multi-head attention-based pooling (GTNN) to obtain molecular outputs (that is, reaction yield and binary reaction outcome), or no pooling (aGNN) for regioselectivity prediction.

GNN. Atom features \mathbf{H} were pooled via sum pooling, transformed via an additional MLP, concatenated to a learned representation of the reaction conditions (Methods, “Condition featurization” for details) and transformed to the desired output via a final MLP.

GTNN. A graph multiset transformer⁴⁹ was incorporated into the GTNN architecture for pooling the atomic features into a molecular feature. The nodes \mathbf{H} were transformed using the Attn function: $\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{Q}\mathbf{K}^T\mathbf{V}$, where query \mathbf{Q} , key \mathbf{K} and value \mathbf{V} are learned features from the node representations \mathbf{H} . \mathbf{Q} is learned via individual embedding vectors per attention head. \mathbf{K} and \mathbf{V} are learned via individual GNNs GNN^K and GNN^V resulting in the overall graph attention head via equation (3):

$$\mathbf{o}_i = \text{Attn}(\mathbf{H}\mathbf{W}^Q, \text{GNN}^K(\mathbf{H}, \mathcal{E}), \text{GNN}^V(\mathbf{H}, \mathcal{E})) \quad (3)$$

where \mathbf{o}_i denotes the weighted pooling vector from one attention head, and \mathbf{W}^Q is a linear layer to learn the query vectors from \mathbf{H} . Herein, four attention heads are incorporated, yielding the pooling scheme graph multi-head attention block GMH: $\text{GMH}(\mathbf{Q}, \mathbf{H}, \mathcal{E}) = [\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4]\mathbf{W}^o$. This learned molecular representation was transformed via an additional MLP, concatenated to a learned representation of the reaction conditions (Methods, “Condition featurization” for details) and transformed to the desired output via a MLP network.

aGNN. No pooling of atom features was applied, and \mathbf{H} was directly transformed to the desired atomic output via a final MLP with a sigmoid activation function.

Training details. PyTorch Geometric (v.2.0.2)⁶³ and PyTorch (v.1.10.1+cu102)⁶⁴ functionalities were used for neural network training. Training was performed on a graphical processing unit, GPU (Nvidia GeForce GTX 1080 Ti) for four hours, using a batch size of 16 samples. The Adam stochastic gradient descent optimizer was employed⁶⁵ with a learning rate of 10^{-4} , a mean squared error (m.s.e.) loss on the training set, a decay factor of 0.5 applied after 100 epochs and an exponential smoothing factor of 0.9. Early stopping was applied to the model that achieved the lowest validation m.a.e. within 1,000 epochs. All the models considered in this study were trained on the Euler computing cluster at ETH Zurich, Switzerland.

Atom featurization. Atomic properties were encoded via the following atomic one-hot-encoding scheme: twelve atom types (H, C, N, O, F, P, S, Cl, Br, I, Si, Se), two ring types (true, false), two aromaticity types (true, false) and four hybridization types (sp^3 , sp^2 , sp , s). Additionally, for molecular graphs that contained electronic features, the atomic partial charges were calculated on the fly using DelFTa software^{66–68}, obtaining DFT-level (ω B97X-D/def2-SVP (refs. 69,70)) Mulliken partial charges⁷¹. For molecular graphs that contained 3D information, the interatomic distances were represented in terms of Fourier features, using a sine-based and cosine-based encoding as previously shown in ref. 66.

Condition featurization. Molecular reaction conditions, that is, solvents, ligands, catalysts and reagents, were one-hot encoded. Whereas, the experimental dataset covered six ligands and four solvent types (that is, 24 possible conditions per substrate), the literature dataset covered twelve ligands, nine solvents, two reagents and four catalyst types (that is, 864 possible conditions per substrate). Supplementary

Section 4 gives a detailed description of the structures covered by these one-hot-encodings.

Conformer generation. The 3D conformers were calculated using RDKit (AllChem.EmbedMolecule (ref. 72)) followed by energy minimization via the universal force field (UFF) method⁷³. For each molecule, ten different conformers were calculated for training and testing. A conformer was randomly selected at each training step. For testing, the final predictions were obtained by averaging the individual predictions calculated for each of the ten conformers.

Baseline model. The ECFP4NN baseline model combined three MLPs for input transformation, namely the ECFP4 fingerprint and two embedded reaction conditions (that is, solvent and ligand). The ECFP4 feature dimension was set to 256 after screening the feature dimensions in the range of 2^7 – 2^{10} . Additional baseline experiments using binary reaction fingerprints with two popular decision tree algorithms, gradient boosting and extreme gradient boosting (XGBoost), can be found in Supplementary Section 10.

Number of hyperparameters. The feature dimension of the GNN internal representation was set to 128, except for (1) the embedding dimension of the reaction and atomic properties, tr which was set to 64, and (2) the first MLP layer after the graph multiset transformer-based pooling, which was set to 256. This setting resulted in neural network sizes of ~2.0 million trainable parameters for the GNN and aGNN models and ~3.0 million trainable parameters for GTNN. The dimensions within ECFP4NN were maintained at 128 yielding a neural network size of ~2.0 million trainable parameters.

Dataset filtering and reaction yield. From the total number of 1,301 reactions in the literature dataset, 492 reactions were used for yield prediction. Two filtering criteria were applied to obtain these training data: (1) duplicate reactions were removed, that is, reactions with identical annotations for starting material, catalyst, solvent, reagent, and product, and (2) only those reactions were included that included catalysts, solvents, reagents, and that occurred at least four times in the whole dataset (in line with the one-hot encoding described in Methods, “Condition featurization”).

Dataset filtering and regioselectivity. From the total number of 1,301 reactions in the literature dataset, 656 reactions were used for regioselectivity prediction. Three filtering criteria were applied to obtain these training data: (1) duplicate products (reactions with identical products) were removed, (2) only reactions using B_2Pin_2 (that is, bis(pinacolato)diboron) as the borylation product were kept and (3) an annotated yield of $\geq 30\%$ was required.

Data availability

The SURF-formatted literature and experimental datasets containing 1,301 and 956 reactions, respectively, as well as a SURF template are available at <https://github.com/ETHmodlab/lsmfml> (<https://zenodo.org/record/8118845>).

Code availability

A reference implementation of the geometric deep learning platform based on PyTorch⁶⁴ and PyTorch Geometric⁶³ is available at <https://github.com/ETHmodlab/lsmfml> (<https://zenodo.org/record/8118845>).

References

- Dreher, S. D. & Krška, S. W. Chemistry informer libraries: conception, early experience, and role in the future of cheminformatics. *Acc. Chem. Res.* **54**, 1586–1596 (2021).
- Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)* 9323–9332 (2021).

Article

<https://doi.org/10.1038/s41557-023-01360-5>

63. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *International Conference on Learning Representations (ICLR)* (2019).
64. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, **32**, 8026–8037, https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (2019).
65. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1412.6980> (2014).
66. Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J. & Schneider, G. Δ -Quantum machine-learning for medicinal chemistry. *Phys. Chem. Chem. Phys.* **24**, 10775–10783 (2022).
67. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273 (2022).
68. Neeser, R., Isert, C., Stuyver, T., Schneider, G. & Coley, C. QMugs 1.1: quantum mechanical properties of organic compounds commonly encountered in reactivity datasets. SSRN <http://doi.org/10.2139/ssrn.4363768> (2023).
69. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
70. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
71. Mulliken, R. S. Electronic population analysis on LCAO–MO molecular wave functions. I. *J. Chem. Phys.* **23**, 1833–1840 (1955).
72. Landrum, G. *RDKit: Open-Source Cheminformatics Software*, accessed September 2020; <http://www.rdkit.org>
73. Rappé, A. K., Casewit, C. J., Colwell, K., Goddard III, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).

Acknowledgements

This research was funded by the Swiss National Science Foundation (SNSF, grant no. 205321_182176). D.B.K. acknowledges funding from the Fonds der Chemischen Industrie (FCI) through a Liebig Fellowship and from Roche Basel to fund the PhD position of D.F.N. We thank C. Isert, F. O'Hara and T. Schindler for helpful discussions.

Author contributions

D.F.N. contributed to the conceptualization, methodology, experiments, formal analysis, data curation and writing of the original

draught. K.A. contributed to the conceptualization, methodology, experiments, software development and validation, formal analysis, data curation and writing of the original draught. R.H. contributed to the experiments. A.T.M. contributed to the methodology, software validation and writing (review and editing). A.M. contributed to the experiments. C.B. contributed to the experiments. G.W. contributed to the experiments and writing (review and editing). I.M. contributed to the methodology. V.J. contributed to the experiments. J.W. contributed to the experiments. M.B. contributed to the experiments. A.F.S. contributed to the acquisition of funding and the conceptualization. D.B.K. contributed to the supervision, acquisition of funding and writing (review and editing). U.G. contributed to the supervision, acquisition of funding and writing (review and editing). R.E.M. contributed to the supervision, acquisition of funding and writing (review and editing). G.S. contributed to the supervision, conceptualization, formal analysis, investigation, methodology, acquisition of funding, project administration and writing (review and editing). All authors discussed the results and gave their approval of the final version.

Funding

Open access funding provided by Swiss Federal Institute of Technology Zurich.

Competing interests

G.S. declares a potential financial conflict of interest as cofounder of inSili.com LLC, Zurich, and in his role as a scientific consultant to the pharmaceutical industry. D.F.N., R.H., A.T.M., A.M., C.B., G.W., I.M., V.J., J.W., M.B., A.F.S., U.G. and R.E.M. are full employees of F. Hoffmann-La Roche Ltd.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41557-023-01360-5>.

Correspondence and requests for materials should be addressed to David B. Konrad, Uwe Grether, Rainer E. Martin or Gisbert Schneider.

Peer review information *Nature Chemistry* thanks Clémence Corminboeuf, Jan Jensen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

5.3 Experimental and supplementary information

The following pages contain the experimental and supplementary information supporting the results described in the publication from the previous section. Citation of the publication: **Nippa, D. F.[†]**, Atz, K.[†], Hohler, R., Müller, A. T., Marx, A., Bartelmus, C., Wuitschik, G., Marzuoli, I., Jost, V., Wolfard, J., Binder, M., Stepan, A. F., Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G. Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning, *Nat. Chem.*, **16**, 2, 239-248 (2024). [427] The material (DOI: 10.1038/s41557-023-01360-5) is reprinted with permission from Springer Nature Limited (Author reuse for own thesis).

nature chemistry



Article

<https://doi.org/10.1038/s41557-023-01360-5>

Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning

In the format provided by the authors and unedited

Supplementary Information: Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning

David F. Nippa^{1,2,†}, Kenneth Atz^{3,†}, Remo Hohler¹, Alex T. Müller¹, Andreas Marx¹, Christian Bartelmuß¹, Georg Wuitschik¹, Irene Marzuoli⁴, Vera Jost¹, Jens Wolfard¹, Martin Binder¹, Antonia F. Stepan¹, David B. Konrad^{2,*}, Uwe Grether^{1,*}, Rainer E. Martin^{1,*} & Gisbert Schneider^{3,5,*}

¹Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, 4070 Basel, Switzerland.

²Department of Pharmacy, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, 81377 Munich, Germany.

³ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

⁴Process Chemistry and Catalysis (PCC), F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, 4070 Basel, Switzerland.

⁵ETH Singapore SEC Ltd, 1 CREATE Way, #06-01 CREATE Tower, Singapore, Singapore.

† These authors contributed equally to this work.

* To whom correspondence should be addressed.

E-mail: david.konrad@cup.lmu.de, uwe.grether@roche.com, rainer_e.martin@roche.com, gisbert@ethz.ch

Contents

SI1	Data set splitting	2
SI2	Systematic literature analysis	3
SI3	LSF informer library	7
SI4	Screening plate design	11
SI4.1	Reaction conditions	11
SI4.2	Catalyst	11
SI4.3	Ligand	12
SI4.4	Reagent / boron source	13
SI4.5	Solvents	13
SI4.6	Plate design	13
SI5	HTE borylation screening protocol	14
SI6	Automated reaction data analysis pipeline	15
SI7	SURF convention	17
SI8	Further analysis of the experimental data set	18
SI8.1	Molecular property distribution	18
SI8.2	Functional group analysis	19
SI9	Further analysis of the literature data set	21
SI9.1	Diversity analysis for regioselectivity data set	22
SI9.2	Model performance on the literature data set	24
SI9.3	Different thresholds for binary reaction outcome prediction	26
SI10	Decision tree algorithms using reaction fingerprints	27
SI11	Borylation scale-ups	29
SI11.1	Reagent and purification information	29
SI11.2	Analytical information	29
SI11.3	Experimental procedures and analytical data	30
SI12	NMR spectra	42

SII Data set splitting

For the three random split tasks (yield-, binary-, and regioselectivity-prediction), the data set was randomly split into training (50%), validation (25%), and test set (25%). For two of the three tasks (yield, and binary reaction outcome prediction), three-fold cross-validation was conducted for each using the same test set, for eight different graph neural networks and one ECFP baseline, within a random split, resulting in 27 training runs for each of the two tasks. The scatter plot in Figure 3a in the main document was created by the best-performing neural network (GTNN3DQM) using a nested four-fold cross-validation with four individual test sets covering the whole data set. The regioselectivity prediction was conducted in a similar manner with the only difference that four graph neural networks and no ECFP-baseline were trained, resulting in 12 training runs. A substrate-based split was additionally conducted for the binary reaction outcome prediction, where all reactions for one substrate were placed into the test set (2.5%), and the remaining data set was randomly split into training set (65%) and validation set (32.5%). For the substrate-based split, three-fold cross-validation was conducted for eight different graph neural networks and one ECFP baseline, for one split per substrate (23), resulting in 621 training runs. See Table S1 for additional details w.r.t. data set splitting.

Table S1: Overview of the neural networks trained for the four different tasks.

Task	Folds / N	Networks / N	Splits / N	Runs / N	data set
Binary	3	9	1 (random)	27	Experimental
Yield	3	9	1 (random)	27	Experimental
Regioselectivity	3	4	1 (random)	12	Literature
Binary	3	9	23 (substrate-based)	621	Experimental
Yield	3	9	1 (random)	27	Literature

SI2 Systematic literature analysis

The systematic analysis of chemical transformations (SACT) can be split up into four major steps: (1) literature search, (2) literature data curation and evaluation, (3) methodology extraction, (4) reaction data curation and analysis.

The literature search (1) can be conducted using keyword- or structure-based queries for the desired transformation allowing a comprehensive assessment of the field. For this study, the keyword-based approach was selected, which consisted of four main query categories: Methodology (M), starting material (S), review/article (R) and catalytic system (C). The M category search aimed at identifying different types of borylations (*e.g.* directed, undirected). The S pillar focused on detecting methodologies for various starting materials (*e.g.* aromatic, aliphatic) and included the hybridisation of the reacting C atom (sp^2 , sp^3) as well. Category R is centred around the publication type (methodology or review paper). In addition to enclosing the typical borylation catalyst metals (*e.g.* Ir, Rh), metal-free methods were part of the catalytic system (C) search. Tables S2 and S3 showcase all search queries for this research paper. To balance the strengths and weaknesses, *e.g.* the number of records, of scientific databases, [1] the queries were run on three different, renowned tools, Scopus (Elsevier, Amsterdam, Netherlands), Web of Science (Clarivate Analytics, Philadelphia, USA) and SciFinder-n (Chemical Abstracts Service, Columbus, USA), on the 3rd of November 2021.

Table S2: The first four categorized queries that were carried out on SciFinder, Scopus and Web of Science. Sections indicated with (M) are modified for the other queries. Those modifications are shown in Table S3.

Query	Methodology	Starting Mate- rial	Review / Article	Catalytic System
Query Name	M1	S2	R1	C1
Title	borylation	borylation	borylation	borylation
Connector	AND	AND	AND	AND
Keyword (KW) or Abstract (ABS)	functionalization OR catalys* OR activation	functionalization OR catalys* OR activation	functionalization OR catalys* OR activation	functionalization OR catalys* OR activation
Connector	AND	AND	AND	AND
KW or ABS (M)	direct*	arene*	review	iridium OR ir
Connector (M)	AND	AND	AND	AND
KW or ABS (M)	c-h OR c h	substrate OR start- ing material	overview	ligand* OR com- plex*

The resulting publication data from Scopus was downloaded as comma-separated value files (.csv), which contained information on citation, bibliography, abstract, keywords and funding details for each record. In a similar process, extraction of full records (information density similar to Scopus) from Web of Science searches as an Excel file (.xls) took place. The download of the reference data in SciFinder required additional manual efforts as only 100 references are downloadable at once in Excel format (.xlsx). Therefore, upon completing the downloads for one search tree, all excel files were combined into one sheet.

The downloaded data was subjected to a custom-built Alteryx Designer (Irvine, US) data curation (2) workflow that removed duplicates, added information from other databases, *e.g.*, journal impact factor, and carried out further filtering as well as calculations before splitting the publications into four quadrants based on journal impact factor and citations per year (Figure S1). After the removal of duplicates, 1723 unique publication records were identified, highlighting the broad and comprehensive search, which reduces the error of not including a relevant publication. Upon additional filtering for the presence of borylation and LSF-related keywords within

Table S3: Additional search queries (M2-3, S2-7, R2-4, C2-5), only showing the two modified sections.

Query number	Methodology	Starting Material	Review	Catalytic System
2	undirect* AND c-h OR c h	aromat* AND substrate OR starting material	review AND overview	rhodium OR rh AND ligand* OR complex*
3	ligand* OR complex* AND c-h OR c h	aliphatic* AND substrate OR starting material	article OR method*	copper OR cu AND ligand* OR complex*
4	-	benzyl* AND substrate OR starting material	article AND method*	iron OR fe AND ligand* OR complex*
5	-	*sp ² * AND substrate OR starting material	-	no and metal OR metalfree OR metal AND free
6	-	*sp ³ * AND substrate OR starting material	-	-
7	-	aryl* AND substrate OR starting material	-	-

the title and the abstract, 938 publications remained in the data set. With this data, various different clustering approaches could have been carried out using a selection of the following dimensions, *e.g.*, journal and affiliation, citations, journal impact factor, technologies, catalysts, starting materials, and publication year. For this research, clustering by citations per year over journal impact factor to determine the most relevant borylation methodology publications (high citations/year, high journal impact factor) was chosen. Removal of review papers delivered 242 remaining records, which underwent manual analysis to guarantee that the papers are within the scope of the automated HTE system (*e.g.*, photochemistry not yet possible). All deselected publications received a tag containing the reason to allow the usage of these records for other purposes in future without re-initiating the manual selection process. The final set of methodology papers contained 38 records, [2–40] which were subjected to reaction data extraction (3) in the next step. Figure S1 illustrates the first two steps of SACT including the results obtained for the borylation literature methodology search campaign.

While there are multiple ongoing efforts and ideas on how to establish a FAIR, simple and standardized format for reaction data documentation, today, methodologies are still reported in a multitude of different, usually not machine-readable structures. [41, 42] Therefore, full manual extraction of the data from reaction schemes or tables was conducted and a suitable database structure that captures this relevant information of a chemical transformation was determined. Rather than recording the pure minimum, all available data was stored. In this course, the simple user-friendly reaction format (SURF) convention, a simple, yet fully comprehensive and variable format, to document and store reaction data in a tab-delimited format, was developed. More details on SURF are shared in the respective section (see Section SI7). While SciFinder and Reaxys are helpful resources to obtain certain information concerning the chemical transformation, they are missing important details, such as equivalents or reaction concentration. Therefore, those properties were sourced manually from the paper, while unique identifiers (CAS numbers or SMILES) of reaction components could mostly be obtained through SciFinder or Reaxys. In addition, yield types (*e.g.*, isolated, GC-MS or NMR) and analytical data were documented. This labour-intensive work resulted in a high-quality data set comprising 1301 borylation reactions serving as an ideal foundation for informed plate design based on data analysis and chemical understanding. Moreover, due to the flexibility of the SURF format, the data was readily available as input for machine learning pipelines.

In the final step of the SACT methodology (4), the reaction data underwent analysis on various measures. Statistical evaluations of conditions, such as temperature or reaction time, as well as equivalent ratios, were complemented by an in-depth chemical and frequency interpretation. This included *e.g.*, mapping ligands with starting materials to determine what type of functional groups can be transformed by which ligands. The main important outcomes of this analysis, *i.e.*, those used for the informed plate design, can be found in Section SI4.

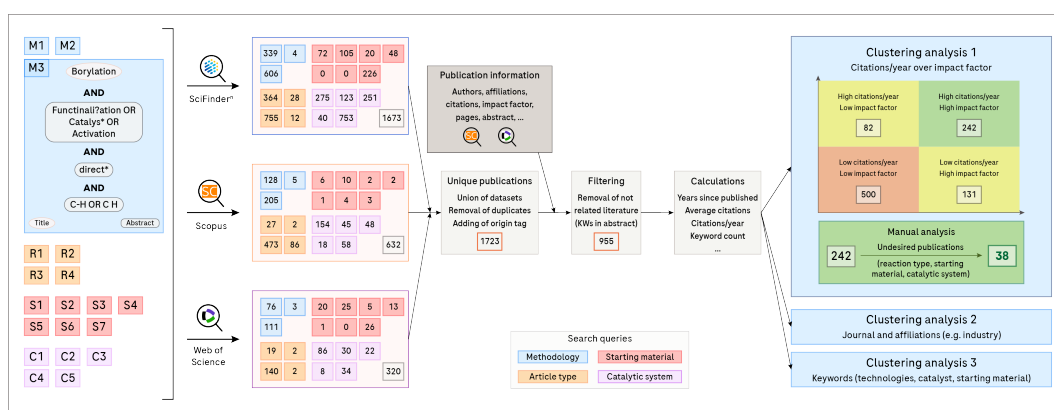


Figure S1: Literature search (SACT 1) followed by curation and evaluation of the obtained data (SACT 2).

SI3 LSF informer library

The substrates for the reaction screening and data generation were chosen through the agglomerative clustering method (a subtype of hierarchical cluster analysis), [43] of 1174 approved and accessible drugs obtained from Cortellis Drug Discovery Intelligence (Clarivate Analytics, Philadelphia, USA), and a molecular weight between 200 and 800 *g/mol*. The molecules were encoded using a similarity matrix of the Jaccard similarity of the ECFP4 [44] descriptors. Thereby, the obtained similarity matrix consisted of the dimensions NxN, where N equals the number of drugs in the similarity matrix. The similarity matrix was clustered into eight clusters from where the ten closest molecules to the cluster centre were picked using the cosine distance. 3 / 10 were then selected for the case study based on commercial availability and chemical meaningfulness. From this selection of 24 drugs (**1**, **14-36**), 23 arrived within the required time frame. Figure S2 illustrates the investigated chemical space *via* principal component analysis (PCA) and Figures S3-S4 the selected drugs.

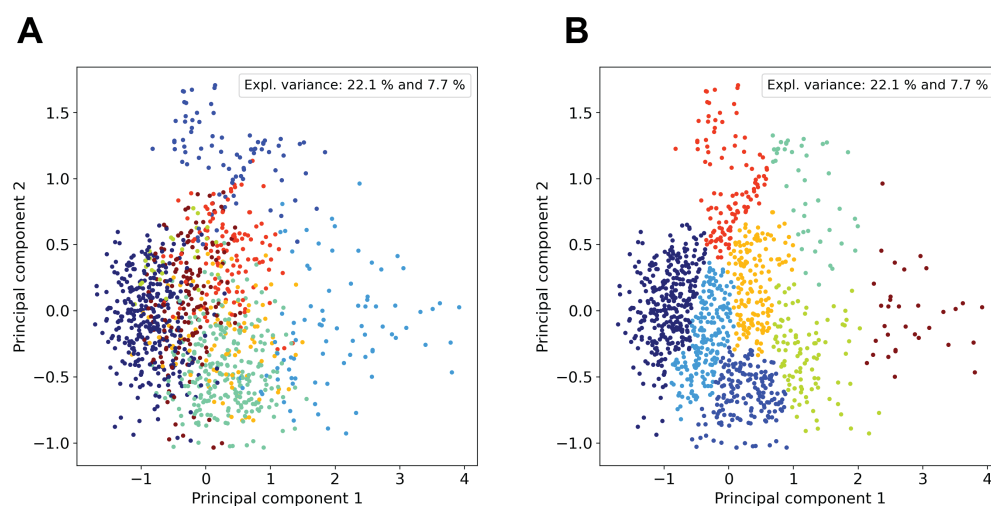


Figure S2: **Clustering.** **A** Principal component analysis (PCA) into principal component (PC) 1 and 2 of the 1174 drugs grouped into the calculated eight clusters in all dimensions. **B** PCA into PC 1 and 2 of the 1174 drugs grouped into the calculated eight clusters in the two reduced dimensions of the PCs. The explainable variance for the investigated data set in the first two PCA is 22.1% for PC1 and 7.7% for PC2.

To provide the model with fragments that are relevant to Roche's chemical space, the top 100 most popular ring assemblies in compounds of the Roche corporate compound collection were determined first. For these assemblies, substructure searches in the entire database were performed. The resulting compounds were kept if the structures had a molecular weight of less than 300 and or less than 20 heavy atoms, if there was at least 1 g of powder stock available and if the structures were not involved in any internal project or subject to legal restrictions. 268 fragments that fulfilled these criteria were identified. Out of this pool of candidates, 16 fragments (**37-48**) were manually selected by the authors. The manual selection aimed at incorporating a variety of frequently occurring functional groups and substituents in medicinal chemistry to test the broader applicability of the methodology. [45–47] Thus, fragments carrying halogen atoms (F, Br, Cl) or OH groups on the aromatic ring were chosen. Furthermore, the selection aimed to cover frequently used heterocyclic elements, such as pyridines, pyrazoles, thiazoles, morpholines, and benzimidazoles. Moreover, five idealized substrates were picked from the literature data set (**49-53**). All screened fragments and idealized substrates are depicted in Figure S5.

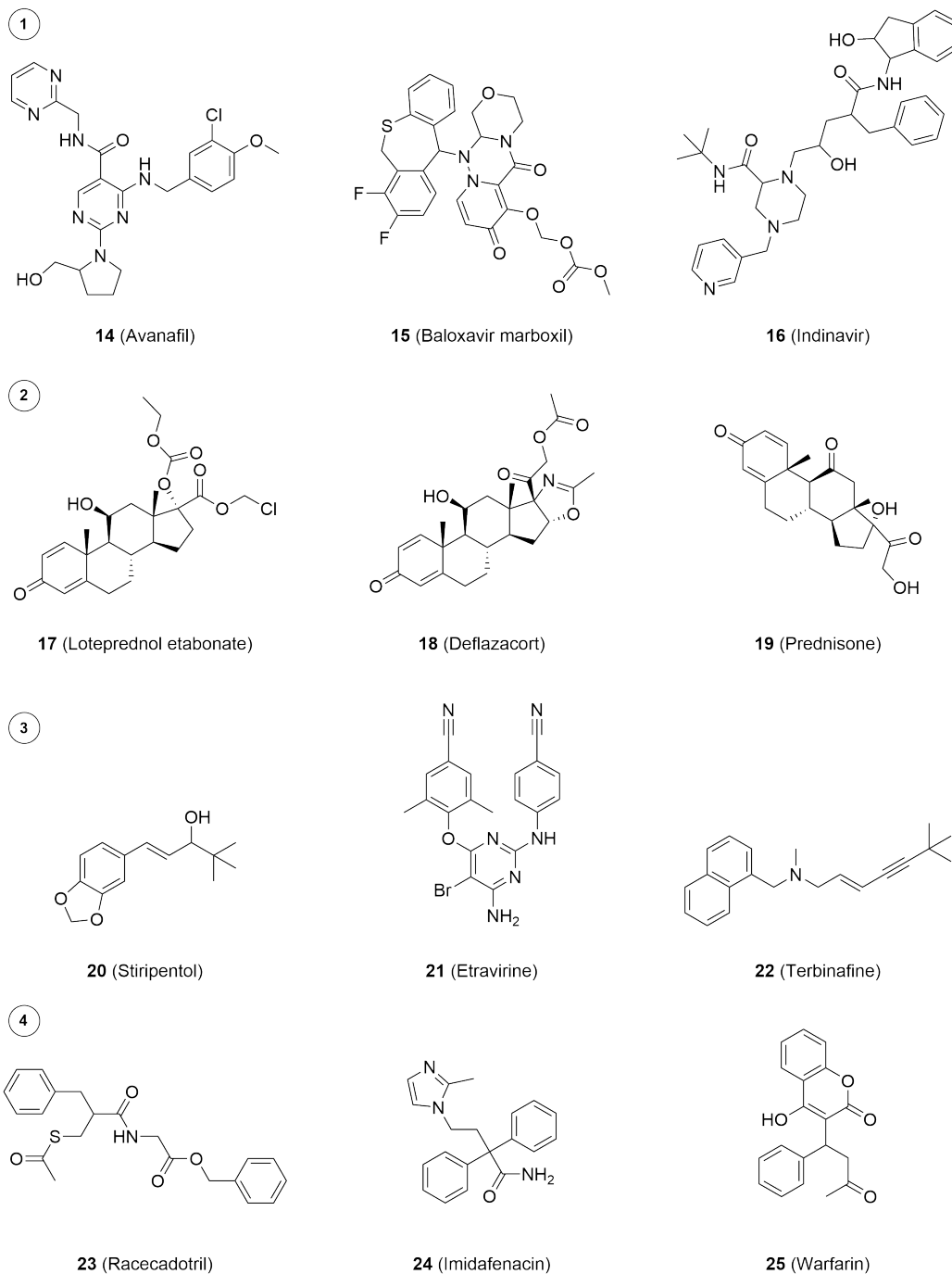


Figure S3: Selected examples from drug clusters 1-4. Note: **15** did not arrive in time and was excluded from the study.

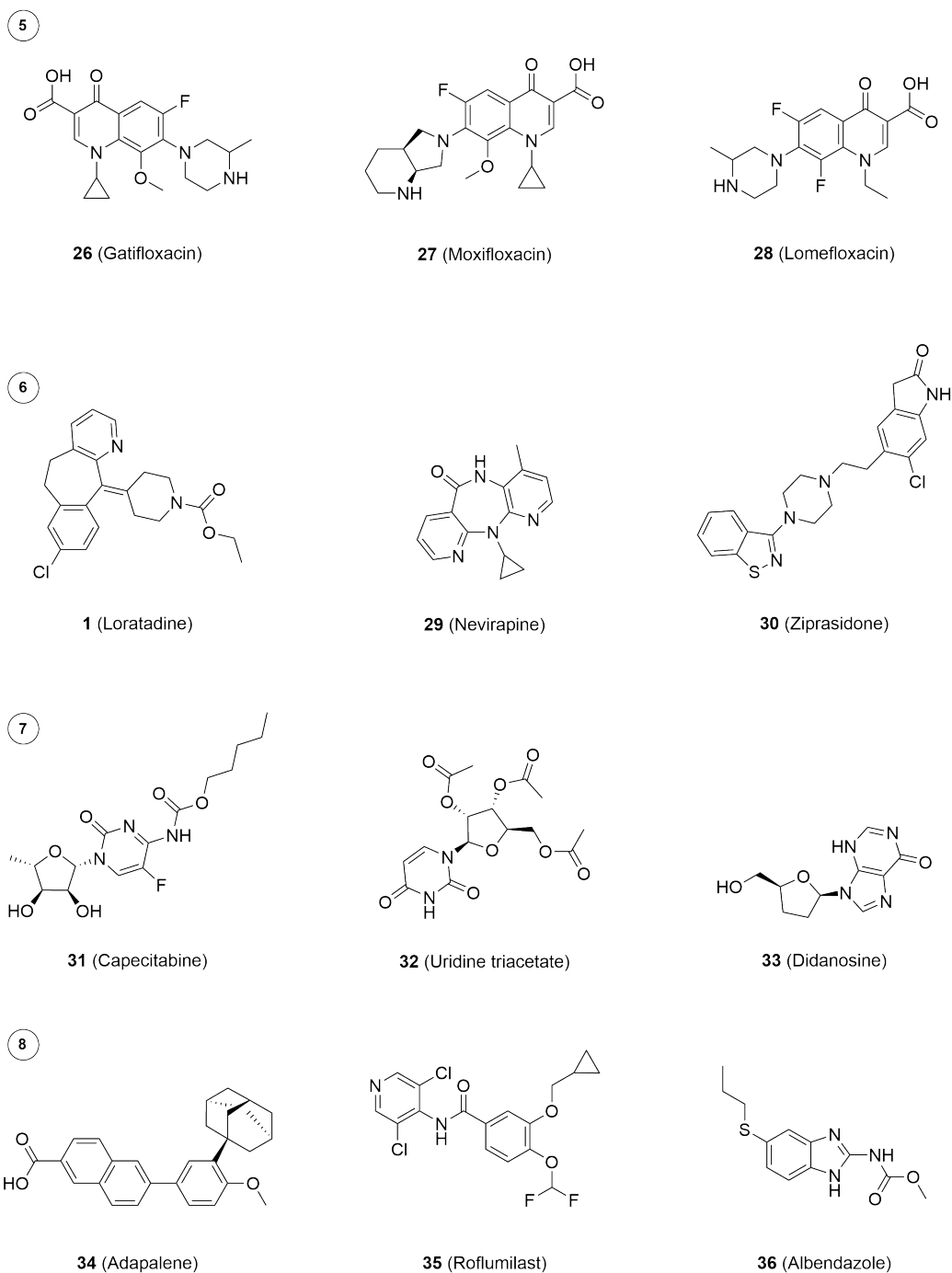
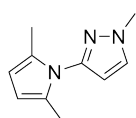
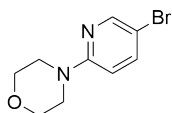


Figure S4: Selected examples from drug clusters clusters 5-8.

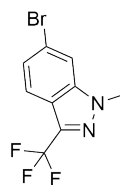
Fragments



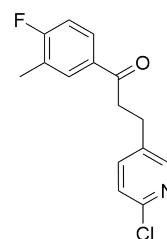
37



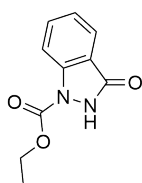
38



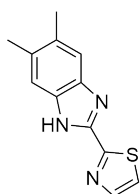
39



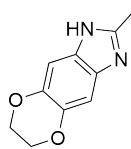
40



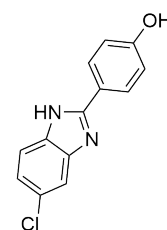
41



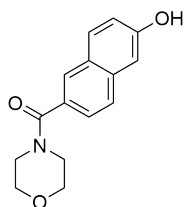
42



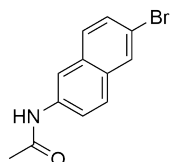
43



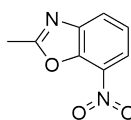
44



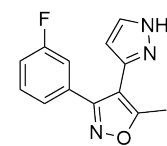
45



46

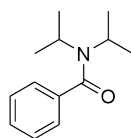


47

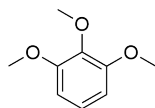


48

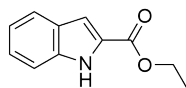
Idealized substrates



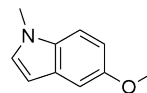
49



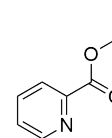
50



51



52



53

Figure S5: Screened fragments (37-48) and idealized substrates (49-53).

SI4 Screening plate design

To possess a clear rationale for the design of the screening plate, a statistical analysis, also referred to as meta-analysis, of the extracted reaction data was performed. As an initial starting point, the number of reaction components was determined. The largest number of C-H borylation reactions within the data set are constituted of four components in addition to the starting material: catalyst, ligand, reagent (boron source) and solvent. While there are examples with additives or additional reagents, initially it was aimed to reduce the complexity of a general screening plate and, therefore, only four component transformations were analyzed in detail. Further, it was opted for a 24-well plate design to reduce the time required for solid dispensing and limit the amount of required starting material (drugs, fragments) to a minimum. Future screenings to expand the data set with further reaction components are envisaged but will require reaction miniaturization and a flexible plate set-up. This could also include catalysts and ligands, which are not commercially available and were excluded from this initial study.

SI4.1 Reaction conditions

As reaction conditions are in general numerical values, a statistical analysis of the following important parameters was carried out: Reaction temperature (T) in °C, reaction time (t) in hours (h), reaction concentration (c) in mol/L and scale (n) in mmol. The plots showing the value distribution including the average and median for all four parameters are depicted in Figure S6.

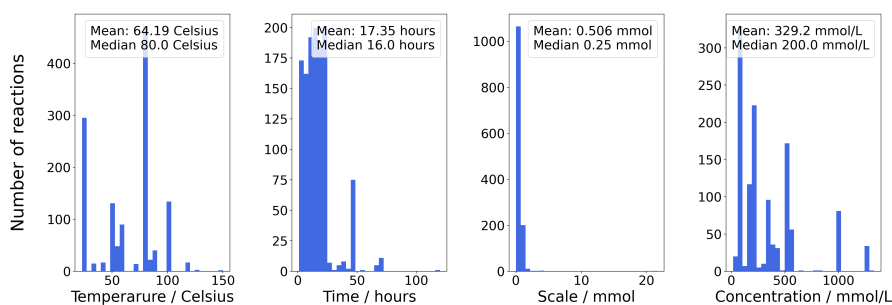


Figure S6: Core reaction parameter (T, t, c, n) distribution of literature data set.

Based on a calculated average (64.2 °C) and median temperature (80 °C) in the data set, 80 °C was selected as the reaction temperature for the 24-well plate. Determination of the average (17.4 h) and median (16.0 h) reaction time strongly indicated, running the reaction overnight for 16 hours. The reaction concentration median was found at 0.2 mol/L and used as molarity for the screening protocol. A lower scale (0.1 mmol) compared to the values (average: 0.51 mmol, median: 0.25 mmol) calculated from the data set was chosen to reduce material consumption. Moreover, the atmosphere under which the borylation reactions were performed, was analyzed. The literature data impressively showed that working under an argon or nitrogen atmosphere is preferred, which was also taken into account for the storage of the reagents. This observation can be explained due to the usage of oxygen and moisture-unstable Iridium catalysts.

SI4.2 Catalyst

Based on the data set (Figure S7), the top three catalysts used for borylation reactions have shown to be [Ir(COD)OMe]₂ (**2**, CAS: 12148-71-9), Pd(OAc)₂ (**58**, CAS: 3375-31-3) and [Ir(COD)Cl]₂ (**59**, CAS: 12112-67-3). All three catalysts are commercially available and would be suitable for the desired borylation screening. Nevertheless, **2** (N = 813) has been used 10-fold more compared to **58** (N = 74) and **59** (N = 47). Therefore, **2**

was chosen as the single catalyst for the screening plate. Based on the data set, the average catalyst loading in borylation reactions in relation to the starting material is 3 mol% with the median being 1.5 mol%. A value in the middle of both values was selected, leading to a catalyst loading of 2.5%.

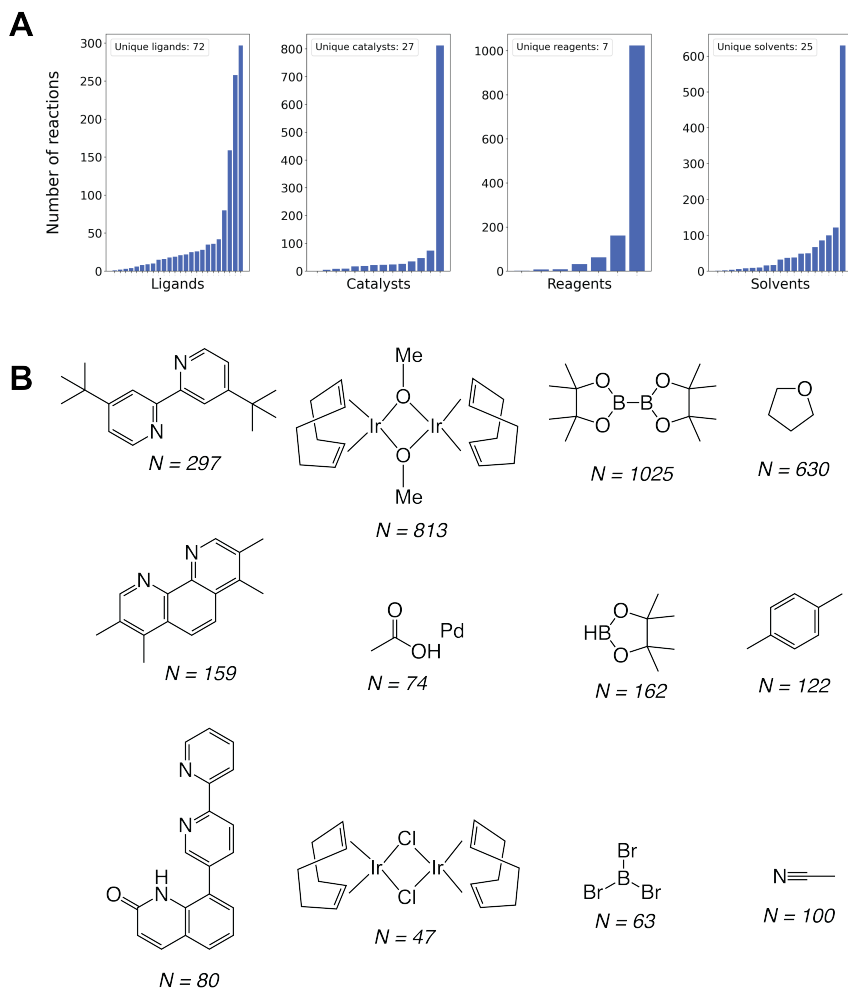


Figure S7: Analysis of the reaction components of the 1301 reactions of the literature data set. **A** Barplot illustrating the abundance of different reaction components. From left to right: ligands, catalysts, reagents, and solvents. Each bar illustrates one unique species, and the bars are sorted from least abundant (left) to most abundant (right) **B** The three most abundant species for each of the reaction components. From left to right: ligands, catalysts, reagents, and solvents. *Note: The second highest count for ligands in "none" (i.e., a ligand free reaction). Therefore, the top-3 and top-4 most abundant ligands are shown instead.*

SI4.3 Ligand

Overall, the most used ligand ($N = 297$) is dtbbpy (**6**, CAS: 72914-19-3), which was also used in combination with **2** in 256 reactions. The second most abundant ligand is tmphen (**9**, CAS: 1660-93-1), which has been used 159 times across the full data set and 127 times in combination with **2**. From the top twelve ligand combinations with **2**, only six are commercially available. Those six ligands were selected for the screening plate. In addition to the above-mentioned dtbbpy and tmphen, 2-pyridinecarboxaldehyde 2,2-bis(phenylmethyl)hydrazone (**4**, CAS:

237402-29-8), 8-aminoquinoline (**5**, CAS: 578-66-5), 4,4'-dimethyl-2,2'-bipyridine (**7**, CAS: 1134-35-6) and 1,10-phenanthroline (**8**, CAS: 66-71-7) are included into the screening plate (main paper, Figure 2a). The analysis of the ligand/catalyst ratio revealed an average of 1.59 and a median of 2, the median was used to give a 5 mol% ligand loading.

SI4.4 Reagent / boron source

B₂pin₂ (**3**, CAS: 73183-34-3, N = 1052) and HBpin (**60**, CAS: 25015-63-8, N = 162) have been the most used reagents (boron sources) across the data set, with BBr₃ (**61**, CAS: 10294-33-4, N = 63) complementing the top three. As **61** is part of a different chemistry type (metal-free borylations), only **3** and **60** were analyzed in more detail. In combination with **2**, **3** (N = 710) was used nearly nine times more often than **60** (N = 71) allowing an informed selection decision to utilize **3** as the boron source for the 24-well plate. On average, a slight excess of **3** (1.25 equivalents) was used in the analyzed reactions. The median, though, shows an equimolar in relation to the starting material (eq = 1), which was chosen for the plate design.

SI4.5 Solvents

The most used solvents have been shown to be aprotic solvents that are mainly non-polar or only slightly polar. The most used solvent by far is THF (**62**, CAS: 109-99-9, N = 630), followed by *p*-xylene (**63**, CAS: 106-42-3, N = 122) and acetonitrile (**13**, CAS: 75-05-8, N = 100). With a reaction temperature of 80 °C, **62** (boiling point: 66 °C) is not an ideal solvent if potential evaporation should be avoided. Instead, 2-methyltetrahydrofuran (**11**, CAS: 96-47-9, boiling point: 80 °C) was chosen due to the higher boiling point while maintaining key properties (*e.g.*, polarity), even though it did not appear in the data set. Due to the potential of solvent borylation, **63** was avoided, but the number three solvent **13** was included in the screening plate. Furthermore, CPME (**12**, CAS: 5614-37-9, N = 31) and cyclohexane (**10**, CAS: 110-82-7, N = 38) were selected due to their high boiling points and their regular appearance in the data set. All solvents, except **11**, were also used in combination with **2**.

SI4.6 Plate design

Based on the considerations above, the plate design was implemented and is shown in Figure 2a of the paper.

SI5 HTE borylation screening protocol

All generated screening data used the plate design depicted in the paper (Figure 2a) and the procedure below, only the starting materials (SI3) were varied. In a nitrogen-filled glovebox from mbraun (Garching, DE) that does not contain any liquids, all solid reaction components were dosed into 1 mL glass vials from Analytical Sales (Flanders, US) on a 24- or 96-well plate from Analytical Sales (Flanders, US) using a CHRONNECT Quantos from Axel Semrau GmbH & Co. KG (Spockhövel, DE) coupled with an XPE206 balance from Mettler Toledo (Greifensee, CH). The plate was sealed and discharged from the glovebox before being transferred to another glovebox from LC Technologies (Salisbury, US), where solvents were added to the vials using multichannel pipettes from Eppendorf (Hamburg, DE). The plate was heated within the glovebox (LC Technologies) on a Junior benchtop solution from Unchained Labs (Pleasanton, US) and VP 721F-1 Parylene Encapsulated Stainless Steel Stir Discs from V&P Scientific Inc. (San Diego, US) were used to stir the reaction mixture. For an intermediate internal process control (IPC), samples were drawn from the plate within the glovebox using a multichannel pipette and transferred into a new plate, which was then subjected to a Genevac centrifugal evaporator EZ3P-VVVHz-HN0 from SP Industries (Warminster, US) to remove the solvents. For a single or the final IPC, the plate was cooled and discharged from the glovebox before being placed into the Genevac centrifugal evaporator to remove the solvents. Using a Freedom EVO 100 liquid handler from Tecan (Männedorf, CH), the residues were re-suspended in MeCN/H₂O (4:1) and shook on a Teleshake 95 from Inheco (Martinsried, DE). Depending on the concentration of the suspension, further dilution steps using the Tecan liquid handler were carried out to reach an LCMS injection concentration of 1 or 0.5 mmol/L. Finally, the samples were transferred onto a 96-deep-well plate (1 mL) from Eppendorf (Hamburg, DE). The plates were analyzed on a Waters (Milford, US) UPLC-MS system equipped with a Waters Acquity sample manager with a flow-through needle, a Waters Acquity sample organizer and a Waters QDa single quadrupole mass spectrometer. The separation was achieved on a ZORBAX RRHT Eclipse Plus C18, 95 Å, 2.1 x 30 mm, 1.8 µm column (P/N 959731-902, LOT: USUXY02479) from Agilent (Santa Clara, USA) at 50 °C. A 2-minute gradient was used and the injection volume accounted for 2 µL. 2 min gradient: A: 0.1% HCOOH in H₂O; B: 0.07% HCOOH in MeCN at flow 1 mL/min. Gradient: 0 min, 3% B; 0.2 min, 3% B; 1.5 min, 97% B; 0.3 min, 97% B; 0.1 min 3% B. The raw data were processed with MassLynx V4.2 and the obtained .rpt file underwent parsing with a customized script, before being subjected to the automated reaction data analysis pipeline (SI6). Due to irrevocable data loss by the LCMS, 956 instead of 960 experimental data points (40 substrates x 24 conditions) were collected.

SI6 Automated reaction data analysis pipeline

Figure S8 illustrates the automated reaction data analysis pipeline used to rapidly identify if drugs, fragments and idealized substrates were borylated or not. Each reaction carries a unique identifier, which is reflected through the LCMS sample name and the MS searches for the sum formulas of the desired products (mono- and diborylated boronic ester and acid). The LCMS-measured data are reported into a .rpt file that needs to undergo parsing to allow a transfer into a tabular format. The obtained data is then pushed to a server from which it is accessible through multiple means. In this case, Alteryx Designer (Irvine, US) was chosen for further processing of the data. In the first step, the data stream is cleaned to remove any undesired columns that would slow down the pipeline. As the LCMS delivers a three-channel output (LC, ES+, ES-), those need to be connected for the same peak in order to allow quantitative and qualitative assessment of the peak. In addition, the Sample ID is disassembled to obtain the different identifiers required for the upcoming data curation.

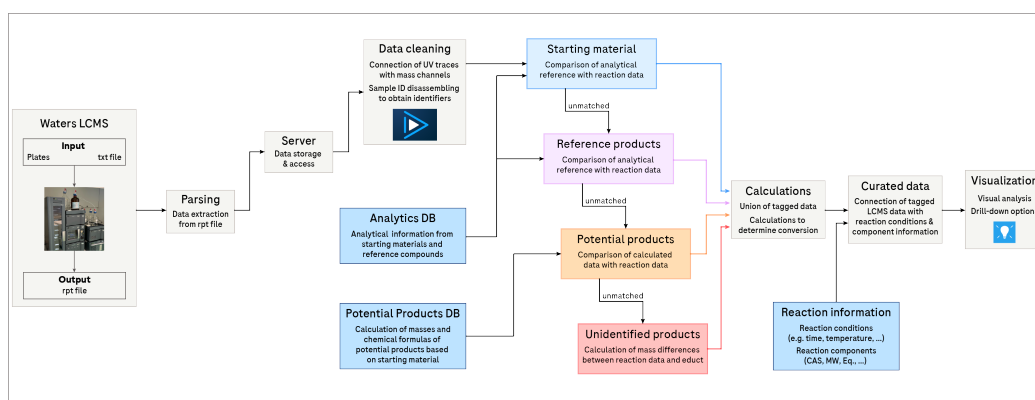


Figure S8: Simplified schematic overview of the automated reaction data analysis pipeline.

In addition to the reaction mixtures, all starting materials and, if available, reference products using the same solvent mixture (MeCN:H₂O, 4:1) are measured on the LCMS to obtain the retention time (LC) and mass pattern (MS). This data is stored in a database and needed for the initial two steps of the matching process. More relevant for LSF though, are the desired/potential products of the reaction. Those masses and chemical formulas are calculated based on the starting material information and the transformation. This Alteryx workflow allows hands-free generation of the potential products including molecular weight, mono-isotopic mass and chemical formula (Hill notation). In addition to being used for the reaction data analysis, this data is also the foundation for generating the LCMS input file.

Once the reaction data has passed through the cleaning process, it is compared to the LCMS information from the above-mentioned data sets, starting off with the identification of the starting material. If a trace from the reaction mixture matches the retention time (± 0.02 min) and the mass pattern (chemical formula detected, mass channel match with database reference), it receives the starting material tag. All unmatched traces continue through the pipeline, where reference compounds, if available, are tagged using the same criteria. The remaining data is then compared to the products that could potentially be formed and are desired (mono- or diborylated species). Since the exact position of the new functional groups is not known, no reference compounds are available. Therefore, only the five most abundant masses per peak are used for tagging and compared to information from the potential product database. Based on the abundance of the mass and if the chemical formula was found by the LCMS, the tag is complemented by an MS reliability score. The score is higher if the chemical formula was found and the correct mass of the desired product (± 0.5 Da) appears in a more abundant channel. For this study, only high MS reliability scores were subjected to the machine learning platform. Last, the unmatched

data is classified as unidentified products, and the mass differences between the peak and parent material are calculated to avoid unnecessary manual calculating of mass differences.

After the tagging is completed, the data streams are recombined and subjected to calculations in order to quantify the reaction components from starting material through reagents to products. To do so, the sum of all LC peaks (integral) is calculated and each peak is then divided by this value. This gives a quantitative measure of the product distribution within the sample, an LCMS conversion. While there are numerous approaches to using internal standards or assays, due to the nature of LSF they have not been applied. LSF reactions tap into new, unexplored chemical space and generally, multiple different components are formed. Therefore, selecting an internal standard that does not overlap with one of these unknown components, is highly difficult.

Upon completion of the calculations, using the identifiers mentioned earlier, reaction information, such as conditions and components, are added to the components that have been identified and quantified. This follows the FAIR data principle and generates a curated, high-quality LSF screening data set that can be stored and shared in the SURF convention (SI7). This allows rapid subjecting of the data to machine learning algorithms as done in this research. It also allows direct visualization of the data in known interfaces, such as TIBCO Spotfire (Somerville, USA) or Tableau (Seattle, USA). Using this workflow, the data curation of one plate usually takes less than one minute.

SI7 SURF convention

The simple user-friendly reaction format (SURF) aims at standardizing reaction data reporting through a simple, yet comprehensive and structured format that is usable with a basic understanding of a spreadsheet. SURF does not require any coding experience, advanced IT skills or a web interface. It enables every chemist within or outside the lab to document chemical synthesis in a machine-readable and shareable format. SURF allowed extraction and documentation of the borylation reactions from literature faster. The generated reaction screening data were also transformed into SURF before being directly subjected to the machine learning pipelines. Reaction documentation following SURF can be implemented in every spreadsheet as the only requirement is the existence of rows and columns.

Each row of the spreadsheet represents the information and data for one single reaction. The SURF convention contains constant (CC) and flexible (FC) categories. CCs never change and are always present, independent of the number of reaction components. They capture the origin and ids of the reaction as well as basic characteristics (reaction type, named reaction, reaction technology) and conditions (temperature, time, atmosphere, scale, concentration, stirring/shaking). Add-ons, such as the procedure or comments, belong to the CCs, too. The FCs describe the more variable part of a reaction, the starting material(s), solvent(s), reagent(s) and product(s). Two identifier options (CAS and SMILES) are available for each component. While the SMILES string is available for every compound and serves as structural input for machine learning models, the CAS number, even though not always available, can be handy for chemists in the lab to order, itemize and find chemicals. For the starting material(s) and reagent(s), *e.g.*, catalyst, ligand, additive, the number and type of columns remain the same (CAS, Smiles, equivalents). If multiple starting materials or reagents are used, additional columns are required. In that case, the three information columns are duplicated and the X is replaced by a number, starting from 1 for the first component, 2 for the second, etc. The same accounts for multiple solvents or products, however, due to their role, they possess more and partly different columns. While the CAS number and/or the SMILES string remain as an identifier, the solvent fraction (in decimals) instead of equivalents is recorded. This allows exact determination of the ratio between solvents. The product category withholds the largest amount of headers as SURF records the yield (in percent), but also the yield type (*e.g.*, isolated, lcms, gcms) as well as the detected mass by MS and the ^1H NMR sequence in addition to the common identifiers CAS and Smiles. This not only allows rapid comparison when experiments are reproduced but can also deliver important increments for machine learning models by differentiating between yield types. As most electronic lab journals already record the above-mentioned parameters, by enforcing of documentation compliance combined with simple automated data extraction and cleaning pipelines, numerous reaction data could be accessible in the SURF convention, and readily available for machine learning applications. We spent thoughts on how to further reduce complexity by introducing specific SURFs without FCs for chemical transformations where the reaction components are generally the same. An excellent example would be Suzuki-Miyaura couplings that utilize a set of six to seven components (organoboron species, halide, catalyst, ligand, base, solvents). [48, 49] However, generating different tailored templates would ultimately end up in various different formats and mismatching headers falling short of the main SURF goal to standardize reaction documentation.

The results of this paper would have not been achieved without FAIR data handling using SURF. The manually extracted reaction data (1376 reactions from 38 publications), which were used in this manuscript for data analysis and selectivity prediction, reported in SURF are attached to the SI as a tab-delimited text file. Moreover, two empty SURF templates are attached as tab-delimited text files: The first file contains the general SURF template, which can be adjusted by introducing additional columns depending on the reaction specifics. The second file is a customized SURF template that should accommodate the vast of chemical transformations: It contains columns for two starting materials, two reagents, one catalyst, one ligand, one additive, two solvents and two products.

SI8 Further analysis of the experimental data set

SI8.1 Molecular property distribution

The molecular property distribution of the 40 molecules within the LSF space library for eight different molecular properties is visualized in Figure S9. Furthermore, the reaction yield distribution of both the complete experimental data set and only the positive results of the former are visualized in a histogram in Figure S10.

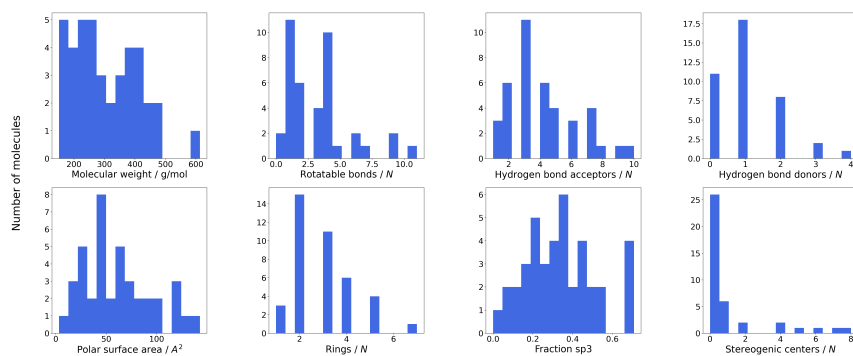


Figure S9: Molecular property distributions of the experimental data set. Top left to right: molecular weight, number of rotatable bonds, hydrogen bond acceptors, hydrogen bond donors; Bottom left to right: polar surface area, number of rings, sp^3 fraction, and number stereogenic centres.

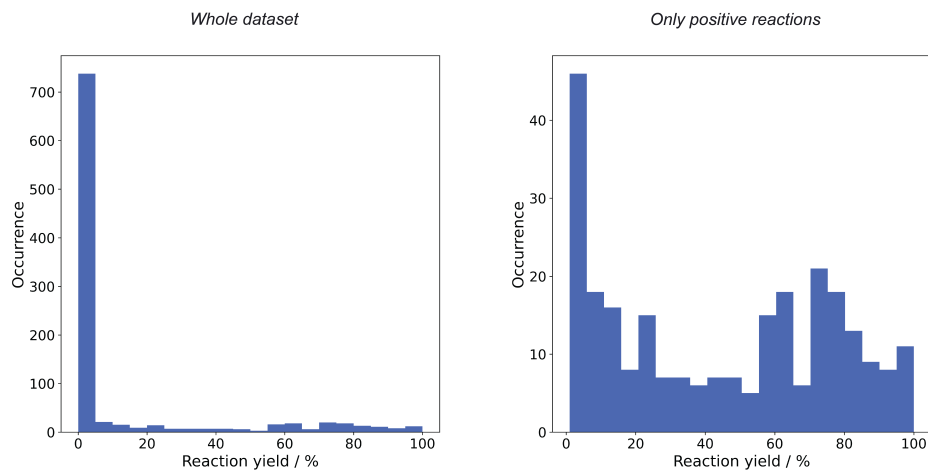


Figure S10: Histogram of reaction yield distribution of the experimental data set. Left: Reaction yield distribution on the whole dataset. Right: Histogram of reaction yield distribution of positive reactions.

SI8.2 Functional group analysis

Functional groups are known as chemical substructures in molecules that consist of atoms and bonds which are responsible for molecular properties such as reactivity or bioactivity. [50] The concept of functional groups has therefore formed a cornerstone in synthetic chemistry, medicinal chemistry and toxicology. [51] To evaluate the scope and limitations of our machine learning platform and the investigated borylation reactions, the functional groups covered by substrates in the LSF space library have been extracted and analyzed. A substructure-free algorithm has been used to extract functional groups from molecules. [52] The resulting functional groups from the LSF space library were compared to the ones from all 1174 drugs and analyzed towards their tolerance for successful borylation reactions. The 53 functional groups extracted from the LSF space library correspond to 11.6 % of the total 458 functional groups present in the 1174 drug molecules (Figure S11 A). However, of the 40 most abundant functional groups found within the drug molecules, 33 (82.5 %) are covered by the LSF space library. The top-3 most occurring functional groups in the LSF space library (40 molecules, including fragments) are aromatic nitrogens, aromatic alkyl-oxy groups and alcohols (Figure S11 B). Most abundant groups covered by the drug space but not by the LSF space (Figure S11 C) are alkyl carboxylic acids or esters (first and third orange bar from left to right, respectively), primary amines (second orange bar from left to right), and tertiary and secondary amides (fourth and fifth orange bar from left to right, respectively). Further, the functional groups which have shown to be tolerated or not tolerated were investigated. All occurring five- and six-membered aromatic heterocycles containing nitrogen, oxygen and sulfur are well tolerated or even cause the desired reaction outcomes (Figure S11 D). On the contrary, certain non-aromatic functional groups such as primary amines, carbamates and carbonates, or aromatic functional groups with strong electron-withdrawing moieties (*e.g.* nitro-aryls) are found to be less tolerated and inhibit desired reaction outcomes (Figure S11 D and E).

Further, Table S5 and S4 shows the number of successful reactions for the different solvents and ligands, respectively.

Table S4: Number of successful and failed reaction for the different ligands.

SMILES	Successful reactions / #	Failed reactions / #
<chem>N=1C=C(C(=C2C=CC3=C(N=CC(=C3C)C)C12)C)C</chem>	52	108
<chem>N=1C=CC(=CC1C=2N=CC=C(C2)C(C)(C)C(C)(C)C</chem>	48	111
<chem>N=1C=CC=C2C=CC=3C=CC=NC3C12</chem>	46	114
<chem>N=1C=CC(=CC1C=2N=CC=C(C2)C)C</chem>	46	113
<chem>N=1C=CC=CC1C=NN(CC=2C=CC=CC2)CC=3C=CC=CC3</chem>	35	124
<chem>N1=CC=CC2=CC=CC(N)=C12</chem>	27	132

Table S5: Number of successful and failed reaction for the different solvents.

SMILES	Successful reactions / #	Failed reactions / #
<chem>C1CCCCC1</chem>	80	159
<chem>O(C)C1CCCC1</chem>	68	171
<chem>O1CCCC1C</chem>	60	179
<chem>N#CC</chem>	46	193

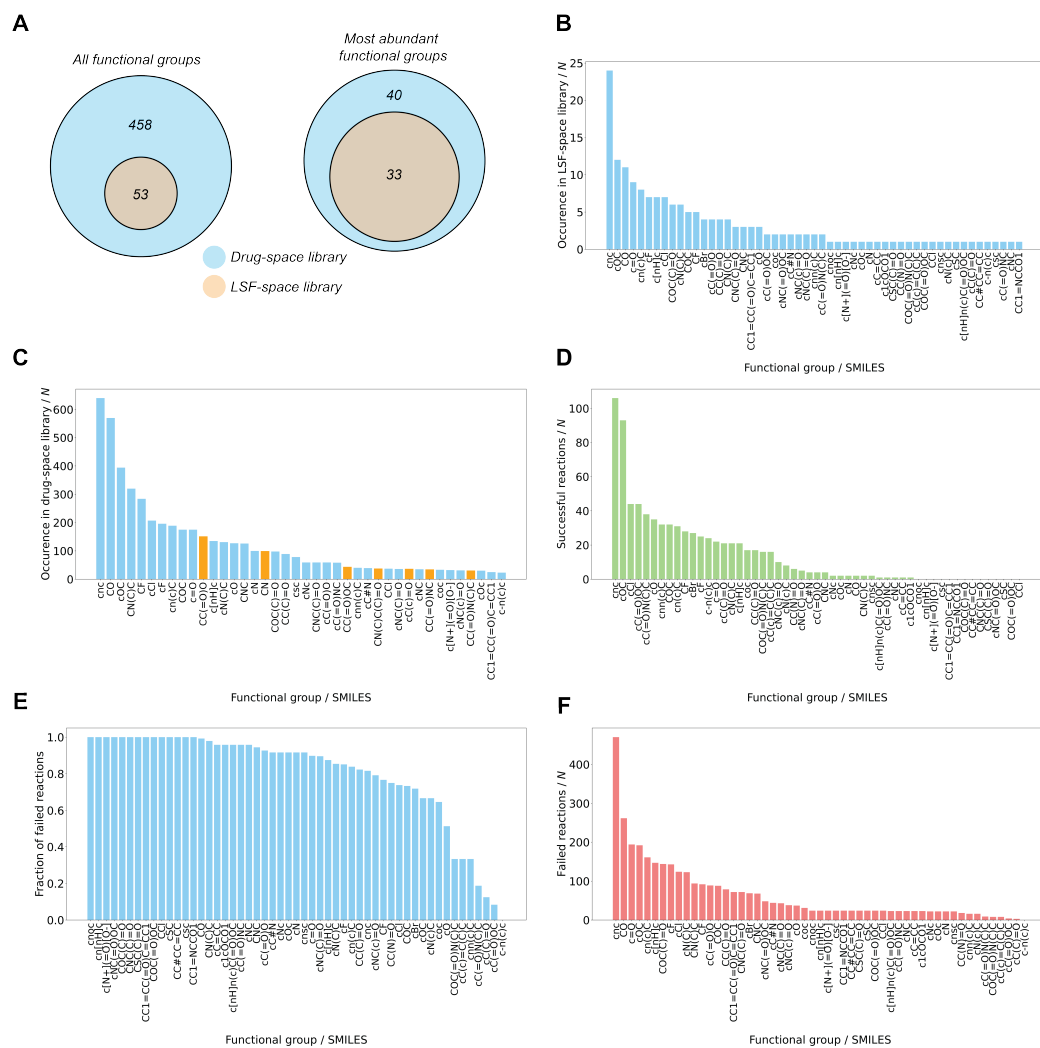


Figure S11: Functional group analysis. **A** Comparing the number of functional groups in the LSF space library to the ones in the drug space library. Left: All functional groups; right: The 40 most abundant functional groups. **B** The 53 functional groups extracted from the LSF space library are plotted by their occurrence from left to right. **C** The 40 most abundant functional groups extracted from the drug space library are plotted by their occurrence from left to right. The bars in blue (33/40) show the functional groups which are covered by the LSF space library. The bars in orange (7/40) show the functional groups which are missing in the LSF space library. **D** The 53 functional groups extracted from the LSF space library are plotted by the absolute number of successful reactions from left to right. **E** The 53 functional groups extracted from the LSF space library are plotted by the fraction of failed reactions from left to right. **F** The 53 functional groups extracted from the LSF space library are plotted by the absolute number of failed reactions from left to right.

S19 Further analysis of the literature data set

In the following, an additional analysis of the experimental data set is described. The molecular property distribution for eight different molecular properties is visualized in Figure S12. Figure S13 shows the reaction yield distribution. To learn the reaction yields the reactions have been binned into four different equally sized bins in the ranges of 0-45%, 45-65%, 65-83%, and 83-100%.

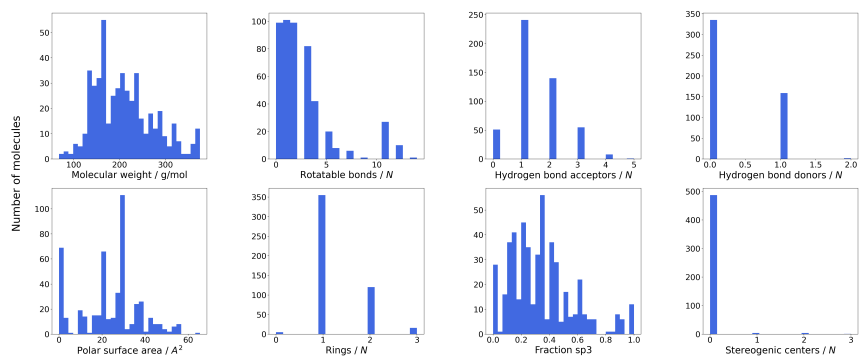


Figure S12: Molecular property distributions of literature data set showing from top left to bottom right: molecular weight, rotatable bonds, hydrogen bond acceptors, hydrogen bond donors, polar surface area, rings, sp³ fraction, and stereogenic centres.

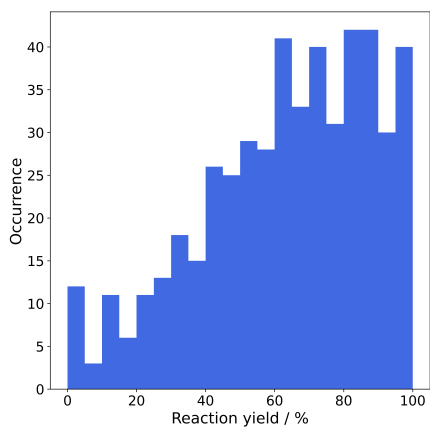


Figure S13: Reaction yield distribution of the literature data set.

SI9.1 Diversity analysis for regioselectivity data set

To further assess the diversity of the chemical space in the regioselectivity training data, the starting materials were clustered using sphere exclusion clustering on ECFP4 fingerprints using a Tanimoto threshold of 0.55 with the ChemFP toolkit [53]. To do so, starting materials were first desalted and standardized using RDKit v.2020.03.1 [54] and unique molecules were kept based on InChI keys. 656 unique starting materials remained, for which the clustering results are shown in Figure S14. Overall, 119 compound clusters and 209 Bemis-Murcko scaffolds were obtained by performing this analysis (Table S6). We argue that this is a sufficiently diverse representation for the task of interest and exceeds the chemical diversity observed in a recent pre-print [55] (Figure S7). As molecular shape potentially influences the performance of the regioselectivity prediction, principal moment of inertia plots and the fraction of sp^3 carbons were further calculated using RDKit (Figure S15). We found that the three-dimensionality of the data is in the range of structures typically observed in medicinal chemistry projects [56].

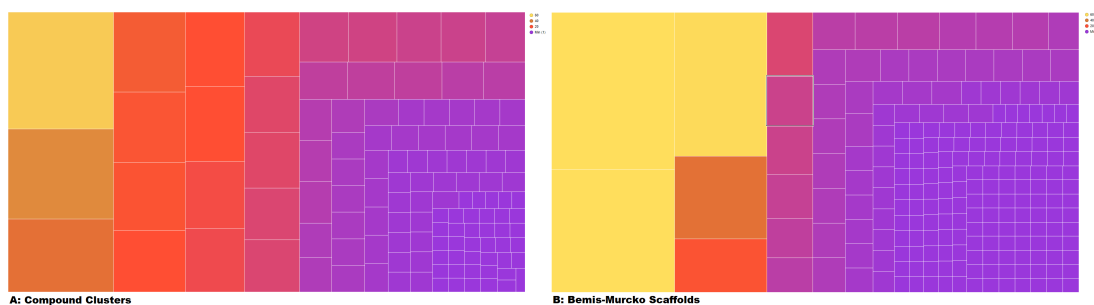


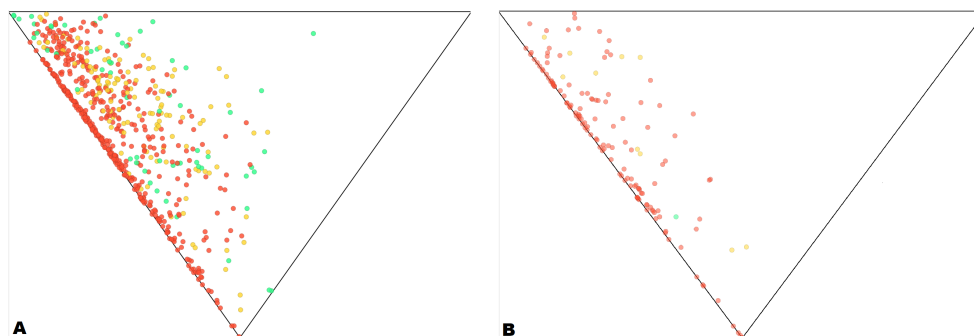
Figure S14: A / left: Tree map showing the size of the sphere exclusion clusters obtained for the regioselectivity training data when clustering on the whole molecular structure. B / right: Number of compounds per Bemis-Murcko scaffold. The size of the boxes as well as the color represents the number of compounds. For the 656 molecules, 119 clusters were obtained on the molecule level and a total of 209 scaffolds were observed. The largest molecule cluster contained 56 members and the most frequent scaffold had 86 compounds.

Table S6: Number of sphere exclusion clusters per cluster size (left) and number of compounds per Bemis-Murcko scaffold (right) observed for the regioselectivity training data.

Compounds per Cluster	Count	Compounds per Scaffold	Count
1	40	1	146
2	24	2	19
3	16	3	9
4	7	4	9
5	3	5	9
6	3	6	3
7	2	7	4
8	3	8	1
9	1	9	1
10	2	10	2
11	2	13	1
13	2	22	1
14	2	34	1
16	1	59	1
17	1	67	1
18	1	86	1
20	3		
22	1		
23	1		
26	1		
35	1		
43	1		
56	1		
Total Clusters	119	Total Scaffolds	209
Total Compounds	656		

Table S7: Number of sphere exclusion clusters per cluster size (left) and number of compounds per Bemis-Murcko scaffold (right) observed in [55].

Compounds per Cluster	Count	Compounds per Scaffold	Count
1	70	1	91
2	8	2	9
3	7	3	5
4	1	5	1
5	1	8	2
7	1		
10	1		
12	1		
Total Clusters	90	Total Scaffolds	108
Total Compounds	145		

Figure S15: Principal moments of inertia plot representing the shape of the regioselectivity training data for this publication (A/left) and a recent pre-print (B/right) [55]. Dots represent compounds and the color represents the fraction of sp^3 carbons, with red ≤ 0.3 , green ≥ 0.5 and yellow in between. Rod-shaped compounds appear in the top left, disc-shaped compounds in the bottom and sphere-shaped compounds in the top right corner.

SI9.2 Model performance on the literature data set

Table S9 and S10 show the accuracy of the investigated nine neural networks. The performance of the reaction yield predictions was investigated on a randomly split data set to learn reaction yields for known substrates in combination with new conditions for both the literature data set (Figure S16).

Table S8: Model performance of the nine investigated neural networks predicting binary reaction outcomes and reaction yields. Mean absolute errors (MAEs) were used to quantify reaction yield predictions. Area under receiver operating characteristic curve(AUC) was used to quantify binary reaction outcome predictions. The numbers represent mean and standard deviation for N=3 independent neural network runs.

	Reaction yield, PCC	Reaction yield, MAE / %
GTNN2D	0.59 (± 0.01)	4.53 (± 0.09)
GNN2D	0.61 (± 0.01)	5.61 (± 0.06)
GTNN3D	0.62 (± 0.01)	4.51 (± 0.11)
GNN3D	0.63 (± 0.01)	5.33 (± 0.34)
GTNN2DQM	0.62 (± 0.01)	4.41 (± 0.17)
GNN2DQM	0.61 (± 0.01)	5.41 (± 0.10)
GTNN3DQM	0.61 (± 0.01)	4.23 (± 0.08)
GNN3DQM	0.62 (± 0.01)	4.88 (± 0.24)
ECFP4NN	0.530(± 0.002)	4.55 (± 0.14)

Table S9: Prediction accuracy of the investigated neural networks. The numbers represent mean and standard deviation for N=3 independent neural network runs.

Prediction error	Mean absolute error / %	Accurate bin / %	1 bin off / %	2 bins off / %	3 bins off / %
GTNN2D	16.7 (± 0.13)	48.4 (± 0.7)	36.0 (± 0.9)	12.4 (± 1.3)	3.1 (± 0.5)
GNN2D	16.4 (± 0.2)	46.5 (± 1.0)	39.4 (± 0.5)	11.8 (± 0.5)	2.7 (± 0.5)
GTNN3D	16.4 (± 0.24)	49.0 (± 1.8)	37.2 (± 2.8)	11.1 (± 1.5)	2.3 (± 0.0)
GNN3D	16.2 (± 0.14)	46.3 (± 0.5)	40.4 (± 1.5)	11.6 (± 1.6)	1.9 (± 0.5)
GTNN2DQM	16.1 (± 0.02)	49.3 (± 0.5)	36.8 (± 1.1)	11.2 (± 1.0)	1.9 (± 0.5)
GNN2DQM	16.3 (± 0.04)	46.9 (± 1.4)	40.0 (± 0.5)	10.3 (± 1.7)	3.0 (± 1.1)
GTNN3DQM	16.2 (± 0.16)	47.1 (± 0.7)	38.3 (± 0.8)	11.4 (± 0.5)	2.3 (± 0.9)
GNN3DQM	16.2 (± 0.14)	46.1 (± 2.7)	39.4 (± 2.6)	12.4 (± 1.4)	1.9 (± 1.1)
ECFP4NN	18.2 (± 0.05)	46.5 (± 1.5)	36.0 (± 1.7)	13.1 (± 0.5)	1.9 (± 0.5)

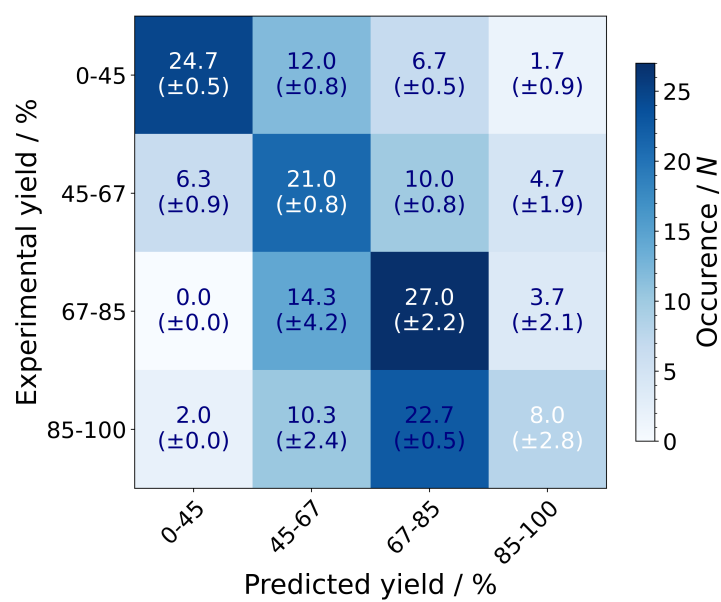


Figure S16: Performance of reaction yield prediction on the literature data set. Confusion matrix visualizing the accuracy of the best-performing neural network (GTNN3DQM) for reaction yields, divided into four equally sized bins.

SI9.3 Different thresholds for binary reaction outcome prediction

Binary reaction outcome prediction was investigated for different reaction yield thresholds (*i.e.*, >1%, >5%, >10%, and >20%) to enable tailored applications to the specific needs of different medicinal chemistry projects. Table S10 illustrates the performance of GTNN3D for the four different thresholds. Figure S17 illustrates the corresponding confusion matrices thereof.

Table S10: Model performance of GTNN3D for binary reaction outcome prediction with different thresholds at >1%, >5%, >10%, and >20%. Five metrics are shown for each of the model to quantify model performance, *i.e.*, area under receiver operating characteristic curve (AUC), *F*-score, predictive positive value (PPV), true positive rate (TPR), and absolute accuracy. The numbers represent mean and standard deviation for N=3 independent neural network runs.

Binary threshold	AUC / %	<i>F</i> -score / %	PPV / %	TPR / %	Absolute accuracy / %
>1%	94.5 (± 0.2)	82.9 (± 0.6)	80.5 (± 0.6)	85.4 (± 0.5)	91.9 (± 0.3)
>5%	94.5 (± 0.2)	84.2 (± 0.4)	82.4 (± 0.3)	86.1 (± 0.6)	93.3 (± 0.2)
>10%	95.6 (± 0.3)	81.9 (± 0.6)	80.1 (± 0.7)	83.6 (± 0.6)	92.9 (± 0.3)
>20%	94.4 (± 0.2)	82.9 (± 0.4)	81.1 (± 0.3)	84.9 (± 0.9)	90.7 (± 0.3)

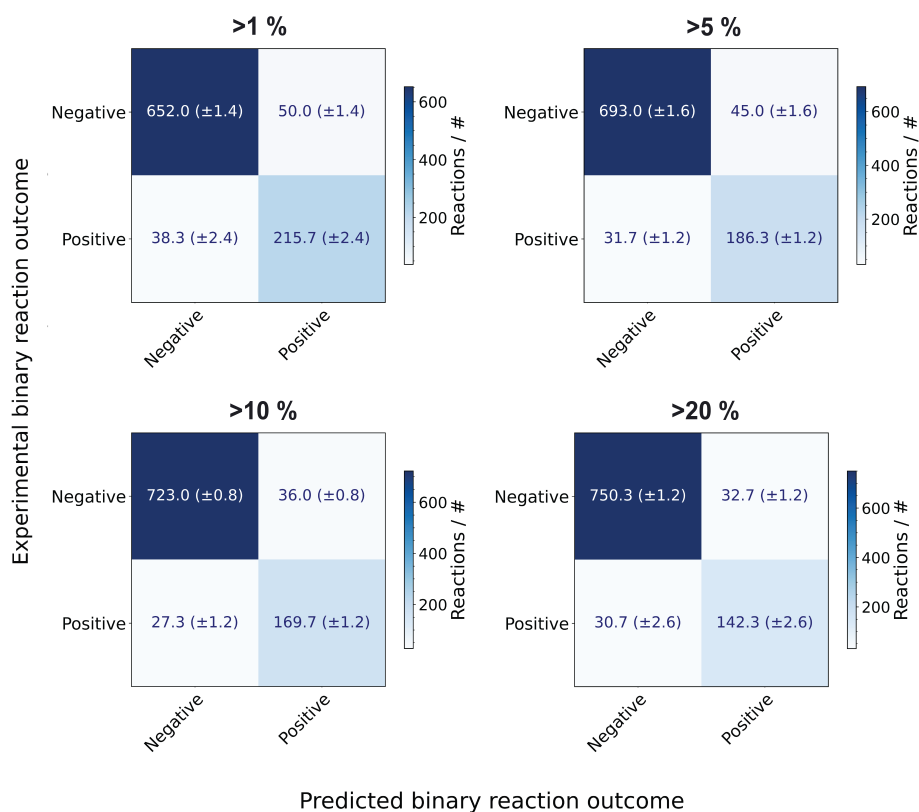


Figure S17: Model performance of GTNN3D for binary reaction outcome prediction with different thresholds at >1%, >5%, >10%, and >20%. Confusion matrix visualizing the accuracy for each thresholds.

SI10 Decision tree algorithms using reaction fingerprints

Fingerprint-based reaction representations in combination with classical machine learning algorithms (*e.g.* support vector machines, ridge regression, gradient boosting, or random forest) have shown applications in predicting reaction outcomes and reaction yields. [57] Here, we compare the results achieved through binary reactions fingerprints using two popular decision tree algorithms, gradient boosting and extreme gradient boosting (XGBoost). While both decision tree algorithms achieve comparable results for three of the four investigated tasks, they are outperformed by the best performing graph neural network (Table S11) for all the investigated reaction tasks. Binary reaction fingerprints are composed by a one-hot encoding of the reaction conditions (*i.e.* catalyst, reagent, ligand, solvent) and a structure-based fingerprint of the substrate (*e.g.* ECFP4) (Figure S18).

The two decision tree algorithms were optimized using the following hyperparameters for screening:

- **XGBoost:** The XGBoost algorithm (XGBoost Python Package version 1.6.2 [58]) was optimized by fine-tuning the following hyperparameters: `n_estimators`=[1, 2, 5, 10, 20, 50, 100, 200], `reg_lambda`=[0.01, 0.05, 0.1, 0.5, 1], `eta`=[0.01, 0.05, 0.1, 0.5, 1], `gamma`=[0.01, 0.05, 0.1, 0.5, 1], and `max_depth`=[1, 2, 4, 6, 8, 10, 12, 14, 16].
- **Gradient boosting:** The gradient boosting algorithm (GradientBoostingClassifier and GradientBoostingRegressor by Sklearn version 0.23.2 [59]) was optimized by fine-tuning the following hyperparameters: `n_estimators`=[1, 2, 5, 10, 20, 50, 100, 200], `learning_rate`=[0.01, 0.05, 0.1, 0.5, 1], and `max_depth`=[1, 2, 4, 6, 8, 10, 12, 14, 16].

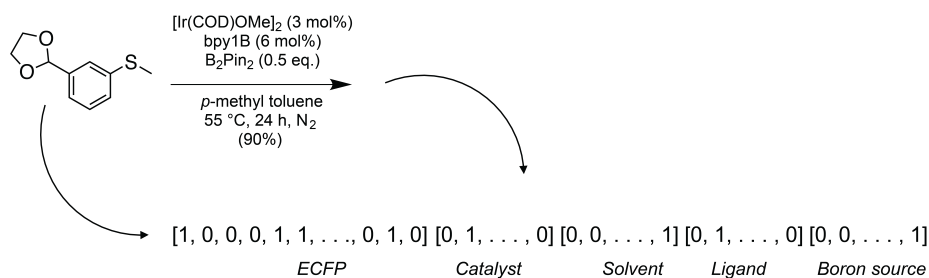


Figure S18: Illustration of binary fingerprint representations for an exemplary borylation reaction with four one-hot encoded reaction conditions (*i.e.* catalyst, reagent, ligand, solvent) and a structure-based fingerprint of the substrate.

Table S11: Model performance of the best graph neural network in comparison to the two decision tree algorithms, gradient boosting and extreme gradient boosting (XGBoost) for predicting binary reaction outcomes and reaction yields. Mean absolute errors (MAEs) were used to quantify reaction yield predictions. Balanced accuracy (AUC, area under receiver operating characteristic curve) was used to quantify binary reaction outcome predictions. The standard deviation is calculated through the results of three different hyperparameter initializations. Since the XGBoost algorithm is deterministic and uses its random state only for sub-sampling and not for initialization, the standard deviations are much lower and in all our cases even equal to zero. The numbers represent mean and standard deviation for N=3 independent neural network runs.

	Reaction yield (literature), MAE / %	Reaction yield (experimental), MAE / %	Binary reaction outcome (experimental, random split), balanced accuracy / %	Binary reaction outcome (experimental, substrate split), balanced accuracy / %
Gradient boosting	16.50 (± 0.07)	5.56 (± 0.03)	90.86 (± 0.0)	52 (± 4)
XGBoost	16.18 (± 0.0)	5.32 (± 0.0)	90.16 (± 0.0)	44 (± 0)
Best graph neural network	16.11 (± 0.02)	4.23 (± 0.08)	91.8 (± 0.9)	67 (± 2)

SI11 Borylation scale-ups

SI11.1 Reagent and purification information

Reactions were set up and conducted in nitrogen-filled gloveboxes from mbraun (Garching, DE) and LC Technologies (Salisbury, US). All chemicals were purchased from Sigma Aldrich (St. Louis, US), AstaTech (Bristol, US), Combi-Blocks (San Diego, US), TRC (Toronto, CA), Thermo Scientific (Waltham, US) or obtained from the Roche compound library and used as received. All solids were dosed using a CHRONECT Quantos from Axel Semrau GmbH & Co. KG (Spockhövel, DE) coupled with an XPE206 balance from Mettler Toledo (Greifensee, CH). Anhydrous solvents were purchased from Sigma Aldrich, stored in the glovebox and added to the reaction vials using pipettes from Eppendorf (Hamburg, DE). The vials were heated on a Junior benchtop solution from Unchained Labs (Pleasanton, US) and the reaction mixture was stirred by VP 721F-1 Parylene Encapsulated Stainless Steel Stir Discs from V&P Scientific Inc. (San Diego, US). Purification by flash column chromatography was performed using SiliaSep Premium Flash Cartridges from Silicycle (Quebec, CA) on a Combi Flash Rf from Teledyne ISCO (Nebraska, US). Eluent solvents, gradients and cartridge sizes for flash chromatography are described for each experiment.

SI11.2 Analytical information

All compounds were characterized by nuclear magnetic resonance (NMR) spectroscopy and (flow injection analysis (FIA)) high-resolution mass spectrometry (HRMS) or gas-chromatography mass spectrometry (GCMS). NMR spectra were recorded on a Bruker Avance III, 600 MHz spectrometer equipped with a 5 mm TCI, Z-gradient CryoProbe. NMR data are reported as follows: chemical shift in reference to the residual solvent peak (δ ppm), multiplicity (s = singlet, d = doublet, br d = broad doublet, dd = doublet of doublet, br dd = broad doublet of doublet, t = triplet, br t = broad triplet, m = multiplet), coupling constant (Hz), and integration. ^1H NMR residual solvent peaks in respective deuterated solvents for CHCl_3 at 7.26 ppm and DMSO at 2.50 ppm. ^{13}C NMR residual solvent peaks in respective deuterated solvents for CHCl_3 at 77.16 ppm and DMSO at 39.52 ppm.

LC-MS high-resolution spectra were recorded with an Agilent LC system consisting of Agilent 1290 high-pressure gradient system, and an Agilent 6545 QTOF. The separation was achieved on a Zorbax Eclipse Plus C18 1.7 μm 2.1 x 50 mm column (P/N 959731-902) at 55 °C; A: 0.01% HCOOH in H_2O ; B: MeCN at flow 0.8 mL/min. Gradient: 0 min 5% B, 0.3 min 5% B, 4.5 min 99% B, 5 min 99% B. The injection volume was 2 μL . Ionization was performed in an Agilent Multimode source. The mass spectrometer was run in "2 GHz extended dynamic range" mode, resulting in a resolution of about 20 000 at $m/z = 922$. Mass accuracy was ensured by internal drift correction. GC-MS spectra were recorded on an Agilent 5975B single quadrupole mass spectrometer. Separation was achieved on an Agilent 7890A using a HP-1ms column (15 m ID: 250 μm and 0.25 μm film) with He as carrier gas. Sample introduction was done via a Split injector at 270°C. After 0.5 min at a constant temperature, the temperature was ramped from 100 °C or 45 °C to 320 °C with 35 °C/min. The mass spectrometer was operated in EI (electron ionization) mode at 70 eV. FIA-HRMS spectra were recorded with an Agilent LC system consisting of an Agilent 1290 high-pressure gradient system, and an Agilent 6540 QTOF. No separation was intended and the injected sample was flushed directly into the Agilent Jetstream source. The mass spectrometer was run in "2 GHz extended dynamic range" mode, resulting in a resolution of about 20 000 at m/z 922. Mass accuracy was ensured by internal drift correction.

SI11.3 Experimental procedures and analytical data

Ethyl 4-[13-chloro-6-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene]piperidine-1-carboxylate (**1a**):

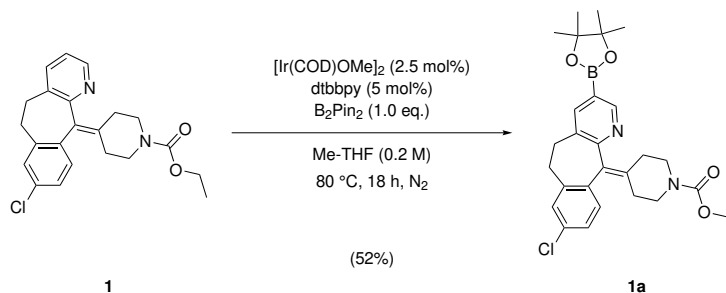


Figure S19: Monoborylation of Loratadine (**1**).

In an N_2 -filled glovebox, ethyl 4-(13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylate (**1**, 31.66 mg, 78.55 μmol , 1.00 eq.), bis(pinacolato)diboron (**3**, 19.95 mg, 78.55 μmol , 1.00 eq.), 4,4'-dimethyl-2,2'-bipyridine (**7**, 723.63 μg , 3.93 μmol , 0.05 eq.) and bis(1,5-cyclooctadiene)-dimethoxydiiridium (**2**, 1.3 mg, 1.96 μmol , 0.025 eq.) were dosed by a solid handler. Addition of 2-methyl-THF (**11**, 398 μL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 $^\circ\text{C}$ for 18 h. The crude material was purified using silica gel column chromatography (4 g) using a MeOH gradient (0%-5%) in DCM. Evaporation of solvents gave the title compound 4-[13-chloro-6-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**1a**, 23.0 mg, 52%) as a white solid.

$^1\text{H NMR}$ (600 MHz, CDCl_3) δ (ppm) 8.78 - 8.69 (m, 1H), 7.87 - 7.77 (m, 1H), 7.12 (s, 3H), 4.25 - 4.01 (m, 3H), 3.88 - 3.75 (m, 2H), 3.36 (s, 1H), 3.43 - 3.28 (m, 1H), 3.14 - 2.99 (m, 2H), 2.82 (s, 1H), 2.89 - 2.73 (m, 1H), 2.52 - 2.42 (m, 1H), 2.40 - 2.25 (m, 3H), 1.43 - 1.30 (m, 15H). $^{13}\text{C NMR}$ (151 MHz, CDCl_3) δ (ppm) 155.47, 152.51, 139.52, 137.67, 132.94, 132.57, 130.63, 129.05, 126.15, 84.17, 75.02, 61.33, 44.76, 31.78, 31.21, 24.86, 24.81, 14.68. **FIA-HRMS** $\text{C}_{28}\text{H}_{34}\text{BClN}_2\text{O}_4$; calc. for ($\text{M}+\text{H}^+$): 509.2378, found: 509.2410.

Ethyl 4-[13-chloro-6,14-bis(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene]piperidine-1-carboxylate (**1b**):

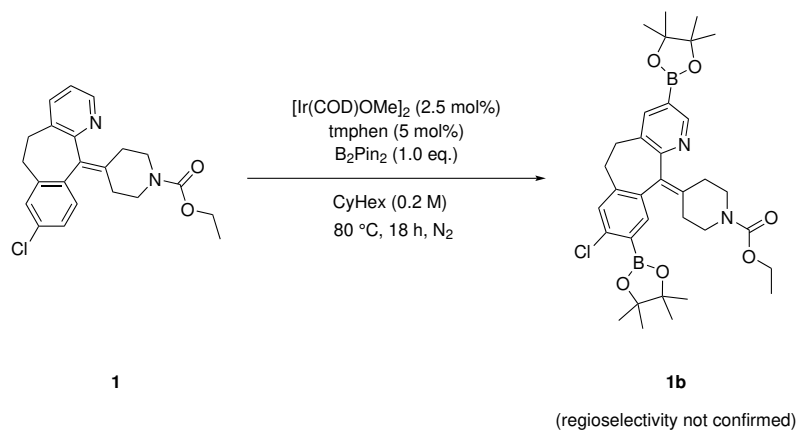


Figure S20: Diborylation of Loratadine (**1**).

In an N₂-filled glovebox, ethyl 4-(13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylate (**1**, 500 mg, 1.28 mmol, 1.00 eq.), bis(pinacolato)diboron (**3**, 325 mg, 1.28 mmol, 1.00 eq.), 3,4,7,8-tetramethyl-1,10-phenanthroline (**9**, 15.1 mg, 64.0 μmol, 0.05 eq.) and bis(1,5-cyclooctadiene)dimethoxydiiridium (**2**, 21.2 mg, 32.0 μmol, 0.025 eq.) were dosed by a solid handler. Addition of cyclohexane (**10**, 6.39 mL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 °C for 18 h. The crude material was purified using silica gel column chromatography (40 g) using an EtOAc/EtOH (3:1) gradient (10%-30%) in heptane. Evaporation of solvents gave the title compound ethyl 4-[13-chloro-6,14-bis(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene]piperidine-1-carboxylate (**1b**, 51.0 mg, 6%), which could only be characterized by HRMS. For confirmation of the regioselectivity, **1b** was transformed into **1c**.

FIA-HRMS C₃₄H₄₅B₂ClN₂O₆; calc. for (M+H⁺): 635.3231, found: 635.3458.

Ethyl 4-(13-chloro-6,14-dihydroxy-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylate (**1c**):

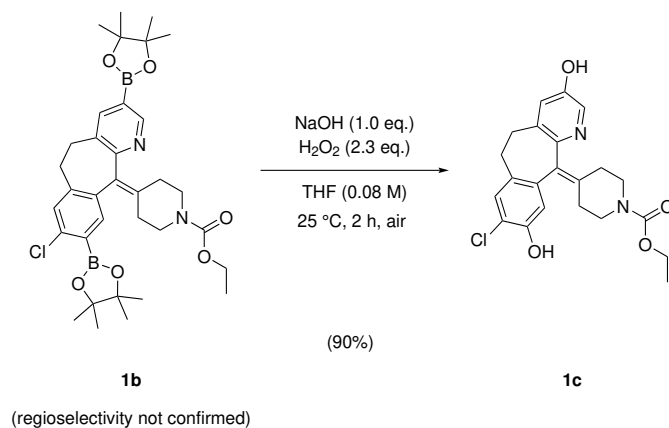


Figure S21: Hydroxylation of di-borylated Loratadine (**1b**).

Ethyl 4-[13-chloro-6,14-bis(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene]piperidine-1-carboxylate (**1b**, 51.0 mg, 0.08 mmol, 1.00 eq.) was dissolved in THF (**64**, 1.0 mL) to give a reaction concentration of 0.08 M. Next, NaOH (3.20 mg, 0.08 mmol, 1.00 eq.) and H₂O₂ (5.50 mL, 6.25 mg, 0.18 mmol, 2.30 eq.) were added to the reaction mixture, which was then stirred at 25 °C for 2 h. The reaction was worked up with H₂O₂ (10 mL) and extracted with EtOAc (3 x 10 mL). Combined organic phases were washed with brine and dried over Na₂SO₄. Evaporation of solvents gave the title compound Ethyl 4-(13-chloro-6,14-dihydroxy-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylate (**1c**, 30.0 mg, 90%) as a white solid.

¹H NMR (600 MHz, DMSO) δ (ppm) 9.93 (s, 1H), 9.73 (s, 1H), 7.88 (d, $J = 2.7$ Hz, 1H), 7.15 (s, 1H), 6.91 (d, $J = 2.7$ Hz, 1H), 6.64 (s, 1H), 4.02 - 4.05 (m, 2H), 3.56 - 3.61 (m, 2H), 3.12 - 3.18 (m, 4H), 2.68 - 2.72 (m, 1H), 2.63 - 2.67 (m, 1H), 2.25 - 2.30 (m, 1H), 2.20 - 2.24 (m, 2H), 2.13 - 2.16 (m, 1H), 1.17 (t, $J = 7.1$ Hz, 3H). ¹³C NMR (151 MHz, DMSO) δ (ppm) 155.03, 152.89, 151.08, 148.02, 139.90, 135.66, 134.78, 134.49, 134.18, 130.39, 130.02, 123.76, 118.33, 116.88, 61.15, 31.77, 30.33, 15.11. HRMS C₂₂H₂₃ClN₂O₄; calc. for (M+H⁺): 415.1424, found: 415.1420.

4-Hydroxy-3-(3-oxo-1-phenyl-butyl)-7-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)chromen-2-one (**25a**):

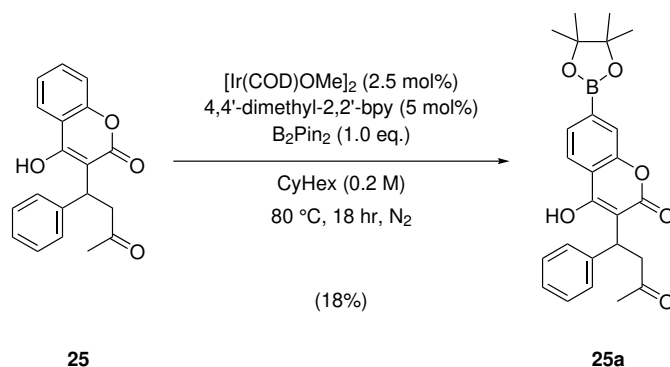


Figure S22: Borylation of Warfarin (**25**).

In an N_2 -filled glovebox, 4-Hydroxy-3-(3-oxo-1-phenylbutyl)-2H-chromen-2-one (**25**, 247 mg, 0.8 mmol, 1.00 eq.), bis(pinacolato)diboron (**3**, 203 mg, 0.8 mmol, 1.00 eq.), 4,4'-dimethyl-2,2'-bipyridine (**7**, 7.4 mg, 0.04 mmol, 0.05 eq.) and bis(1,5-cyclooctadiene)dimethoxydiiridium (**2**, 13.2 mg, 0.02 mmol, 0.025 eq.) were dosed by a solid handler. Addition of cyclohexane (**10**, 6.39 mL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 °C for 18 h. The crude material was purified using silica gel column chromatography (12 g) using an EtOAc/EtOH (3:1) gradient (0%-25%) in heptane, followed by another silica gel chromatography (4 g) using a EtOAc/EtOH (3:1) gradient (0%-10%) in heptane. Evaporation of solvents gave the title compound 4-Hydroxy-3-(3-oxo-1-phenyl-butyl)-7-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)chromen-2-one (**25a**, 48.0 mg, 18%) as a white solid.

$^1\text{H NMR}$ (600 MHz, DMSO) δ (ppm) 8.43 - 8.33 (m, 3H), 7.83 - 7.68 (m, 4H), 7.38 (s, 1H), 7.29 - 7.11 (m, 7H), 3.99 (br dd, $J = 6.7, 11.2$ Hz, 1H), 2.38 - 2.25 (m, 2H), 2.22 - 2.04 (m, 1H), 1.90 (br t, $J = 12.1$ Hz, 1H), 1.69 - 1.55 (m, 4H), 1.42 - 1.13 (m, 2H). $^{13}\text{C NMR}$ (151 MHz, DMSO) δ (ppm) 158.31, 151.58, 143.69, 128.13, 126.96, 125.83, 122.20, 121.16, 117.82, 104.47, 99.67, 84.18, 42.63, 27.07. FIA-HRMS $\text{C}_{25}\text{H}_{27}\text{BO}_6$; calc. for $(\text{M}+\text{H}^+)$: 435.1979, found: 435.1824.

2-cyclopropyl-7-methyl-13-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2,4,9,15-tetraza-tricyclo[9.4.0.03,8]-pentadeca-1(15),3(8),4,6,11,13-hexaen-10-one (**29a**):

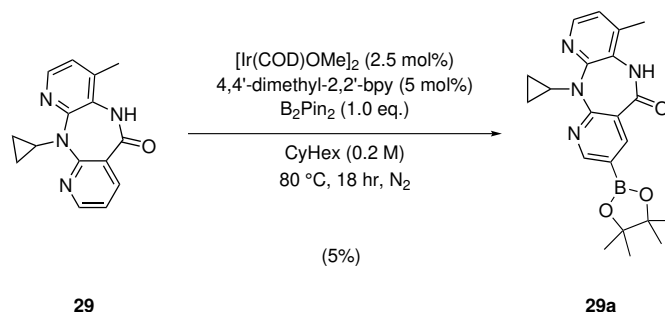


Figure S23: Borylation of Nevirapine (**29**).

In an N₂-filled glovebox, 11-cyclopropyl-4-methyl-5,11-dihydro-6H-dipyrido[3,2-b:2',3'-e][1,4]diazepin-6-one (**29**, 26.6 mg, 0.1 mmol, 1.00 eq.), bis(pinacolato)diboron (**3**, 253 mg, 1.0 mmol, 1.00 eq.), 4,4'-dimethyl-2,2'-bipyridine (**7**, 9.3 mg, 0.05 mol, 0.05 eq.) and bis(1,5-cyclooctadiene)dimethoxydiiridium (**2**, 16.5 mg, 0.025 mmol, 0.025 eq.) were dosed by a solid handler. Addition of cyclohexane (**10**, 0.5 mL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 °C for 18 h. The crude material was purified using silica gel column chromatography (4 g) using an EtOAc/EtOH (3:1) gradient (5%-25%) in heptane, followed by another silica gel chromatography (4 g) using a EtOAc/EtOH (3:1) gradient (0%-25%) in heptane. Evaporation of solvents gave the title compound 2-cyclopropyl-7-methyl-13-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2,4,9,15-tetraza-tricyclo[9.4.0.03,8]pentadeca-1(15),3(8),4,6,11,13-hexaen-10-one (**29a**, 9.0 mg, 5%) as a white solid.

¹H NMR (600 MHz, DMSO) δ (ppm) 9.89 (s, 1H), 8.66 (d, *J* = 2.0 Hz, 1H), 8.22 (d, *J* = 2.0 Hz, 1H), 8.10 (d, *J* = 4.8 Hz, 1H), 7.10 (dd, *J* = 4.8, 0.7 Hz, 1H), 3.63 (dt, *J* = 6.9, 3.3 Hz, 1H), 2.35 (s, 3H), 1.31 (d, *J* = 4.8 Hz, 10H), 1.08 (s, 1H), 0.92 (s, 2H), 0.30 - 0.44 (m, 2H). **¹³C NMR (151 MHz, CDCl₃)** δ (ppm) 168.21, 162.26, 158.12, 153.51, 147.06, 144.34, 138.52, 124.66, 121.99, 119.07, 84.18, 29.78, 24.85, 24.83, 17.68. **LCMS** C₂₁H₂₅BN₄O₃; calc. for (M+H⁺): 392.2, found: 392.2.

3-(2,5-dimethylpyrrol-1-yl)-1-methyl-5-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)pyrazole (**37a**):

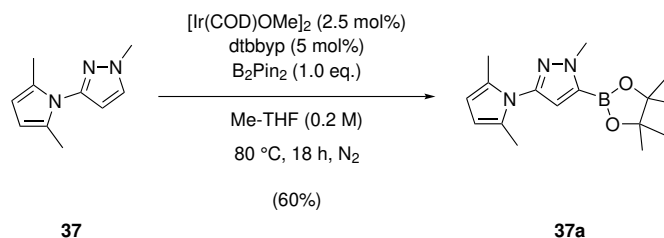


Figure S24: Monoborylation of **37**.

In an N_2 -filled glovebox, ethyl 3-(2,5-dimethylpyrrol-1-yl)-1-methyl-pyrazole (**37**, 140.18 mg, 800 μmol , 1.000 eq), bis(pinacolato)diboron (**3**, 203.15 mg, 800 μmol , 1.000 eq), dtbbpy (**6**, 10.74 mg, 40.0 μmol , 0.05 eq) and bis(1,5-cyclooctadiene)dimethoxydiiridium (**2**, 13.26 mg, 20.0 μmol , 0.025 eq) were dosed by a solid handler. Addition of Me-THF (**11**, 4.0 mL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 °C for 18 h, followed by evaporation of the solvent. The crude material was purified by silica gel column chromatography (40 g) using a MeOH gradient (0%-5%) in DCM. Evaporation of solvents gave the title compound 3-(2,5-dimethylpyrrol-1-yl)-1-methyl-5-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)pyrazole (**37a**, 146.00 mg, 60%) as an off-white solid.

^1H NMR (600 MHz, CDCl_3) δ (ppm) 6.57 (s, 1H), 5.85 (s, 2H), 4.09 (s, 3H), 2.10 (s, 6H), 1.38 (s, 12H).

^{13}C NMR (151 MHz, CDCl_3) δ (ppm) 146.14, 129.33, 1124.55, 105.86, 84.44, 39.76, 24.88, 12.92. GCMS $\text{C}_{16}\text{H}_{24}\text{BN}_3\text{O}$; calc. for (M^{*+}): 301.2, found: 301.2.

4-[5-bromo-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2-pyridyl]morpholine (**38a**):

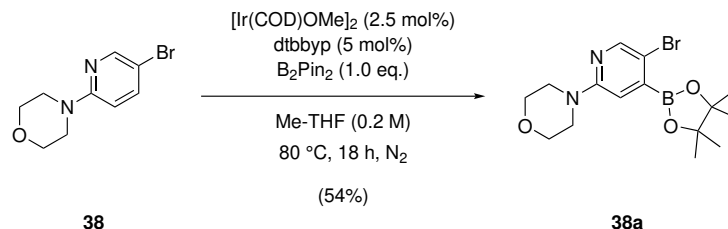


Figure S25: Monoborylation of **38**.

In an N₂-filled glovebox, 4-(5-bromo-2-pyridyl)morpholine (**38**, 194.48 mg, 800 μmol, 1.00 eq.), bis(pinacolato)-diboron (**3**, 203.15 mg, 800 μmol, 1.00 eq.), dtbbpy (**6**, 10.74 mg, 40.0 μmol, 0.05 eq.) and bis(1,5-cyclooctadiene)-dimethoxydiiridium (**2**, 13.26 mg, 20.0 μmol, 0.025 eq.) were dosed by a solid handler. Addition of 2-methyl-THF (**11**, 4.0 mL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 °C for 18 h, followed by evaporation of the solvent. The crude material was purified by silica gel column chromatography (40 g) using a MeOH gradient (0%-5%) in DCM. Evaporation of solvents gave the title compound 4-[5-bromo-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2-pyridyl]morpholine (**38a**, 161.00 mg, 54%) as a white solid.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.25 (d, *J* = 0.6 Hz, 1H), 6.84 (s, 1H), 3.80 - 3.82 (m, 4H), 3.48 - 3.50 (m, 4H), 1.38 (s, 12H). ¹³C NMR (151 MHz, CDCl₃) δ (ppm) 157.59, 149.20, 113.51, 112.82, 84.84, 66.70, 45.59, 24.83. GCMS C₁₅H₂₂BBrN₂O₃; calc. for (M⁺): 368.1, found: 368.1.

6-bromo-1-[(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)methyl]-3-(trifluoromethyl)indazole (**39a**):

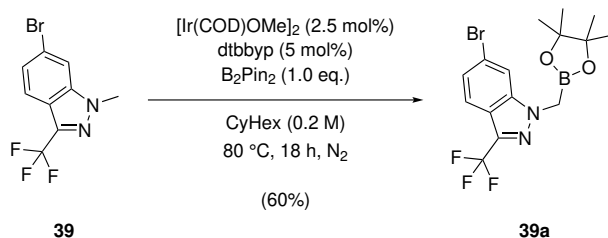


Figure S26: Monoborylation of **39**.

In an N_2 -filled glovebox, 6-bromo-1-methyl-3-(trifluoromethyl)indazole (**39**, 223.25 mg, 800 μmol , 1.00 eq), bis-(pinacolato)diboron (**3**, 203.15 mg, 800 μmol , 1.00 eq), dtbbyp (**6**, 10.74 mg, 40.0 μmol , 0.05 eq) and bis(1,5-cyclooctadiene)dimethoxydiiridium (**2**, 13.26 mg, 20.0 μmol , 0.025 eq) were dosed by a solid handler. Addition of cyclohexane (**10**, 4.0 mL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 °C for 18 h, followed by evaporation of the solvent. The crude material was purified by silica gel column chromatography (40 g) using a MeOH gradient (0%-5%) in DCM. Evaporation of solvents gave the title compound 6-bromo-1-[(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)methyl]-3-(trifluoromethyl)indazole (**39a**, 197.0 mg, 60%) as a light yellow solid.

$^1\text{H NMR}$ (300 MHz, CDCl_3) δ (ppm) 7.68 (d, $J = 8.6$ Hz, 1H), 7.61 - 7.62 (m, 1H), 7.37 (dd, $J = 1.6, 8.7$ Hz, 1H), 4.11 (s, 2H), 1.31 (s, 12H). $^{13}\text{C NMR}$ (151 MHz, CDCl_3) δ (ppm) 141.53, 125.99, 121.21, 113.11, 84.99, 24.72. GCMS $\text{C}_{15}\text{H}_{17}\text{BBrF}_3\text{N}_2\text{O}_2$; calc. for (M^{*+}): 404.1, found: 404.1.

[6-hydroxy-8-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2-naphthyl]-morpholino-methanone (**45a**) and [6-hydroxy-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2-naphthyl]-morpholino-methanone (**45b**):

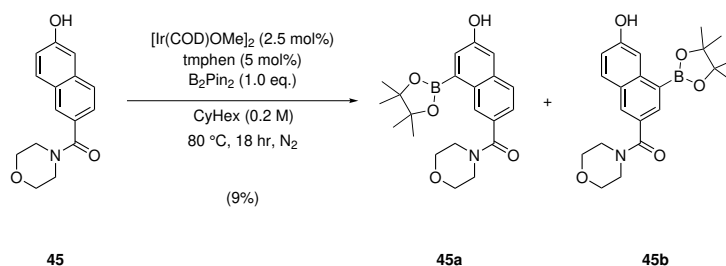


Figure S27: Monoborylation of **45**.

In an N₂-filled glovebox, (6-hydroxy-2-naphthyl)-morpholino-methanone (**45**, 25.7 mg, 0.1 mmol, 1.00 eq), bis-(pinacolato)diboron (**3**, 253 mg, 1.0 mmol, 1.00 eq.), tmphen (**6**, 11.82 mg, 0.05 mmol, 0.05 eq) and bis(1,5-cyclooctadiene)dimethoxydiiridium (**2**, 16.5 mg, 0.025 mmol, 0.025 eq.) were dosed by a solid handler. Addition of cyclohexane (**10**, 0.5 mL) dissolved all components to give a reaction concentration of 0.2 M. The reaction was stirred at 80 °C for 18 h, followed by evaporation of the solvent. The crude material was purified by silica gel column chromatography (4 g) using a MeOH gradient (0%-75%) in DCM. Evaporation of solvents gave the title compounds 6-hydroxy-8-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2-naphthyl]-morpholino-methanone (**45a**) and -[6-hydroxy-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2-naphthyl]-morpholino-methanone (**45b**) as an isomeric mixture (combined 3.4 mg, 9%).

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.72 - 8.75 (m, 1H), 8.07 - 8.10 (m, 1H), 8.05 (d, *J* = 1.9 Hz, 1H), 7.89 (d, *J* = 1.9 Hz, 1H), 7.72 (d, *J* = 2.5 Hz, 1H), 7.69 - 7.73 (m, 1H), 7.60 (s, 1 H), 7.46 - 7.50 (m, 1H), 7.15 - 7.19 (m, 1H), 7.09 - 7.14 (m, 1H), 5.54 - 6.30 (m, 1H), 3.70 - 3.89 (m, 8H), 1.40 (s, 12H). ¹³C NMR (151 MHz, CDCl₃) δ (ppm) 171.40, 153.93, 135.61, 134.55, 131.03, 130.75, 129.69, 128.26, 127.80, 127.50, 125.09, 113.17, 84.09, 83.98, 67.12, 66.96, 25.04, 25.00. HRMS C₂₁H₂₆BNO₅; calc. for (M+H⁺): 384.1982, found: 384.1979.

tert-butyl (6-chloro-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2,3-dihydro-1H-inden-2-yl)carbamate (**64a**):

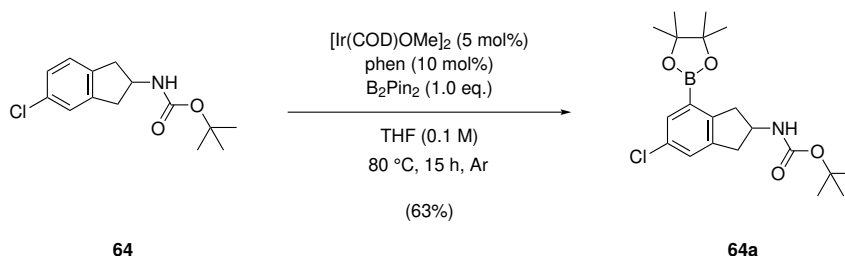


Figure S28: Monoborylation of **64**.

Under an Ar atmosphere, *tert*-butyl (5-chloro-2,3-dihydro-1H-inden-2-yl)carbamate (**64**, 2.50 mg, 9.34 mmol, 1.00 eq), bis(pinacolato)diboron (**3**, 2.42 g, 9.34 mmol, 1.00 eq), phen (**8**, 221 mg, 0.93 mmol, 0.10 eq) and bis(1,5-cyclooctadiene)dimethoxydiiridium (**2**, 309 mg, 0.47 mmol, 0.05 eq) were added to a vial. The addition of THF (**62**, 10 mL) dissolved all components to give a reaction concentration of 0.1 M. The reaction was stirred at 80 °C for 15 h, followed by evaporation of the solvent. The crude material was purified by silica gel column chromatography (20 g) using an EtOAc gradient (0%-30%) in heptane. Evaporation of solvents gave the title compound *tert*-butyl (6-chloro-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2,3-dihydro-1H-inden-2-yl)carbamate (**64a**) as an off-white solid (2.30 g, 63%).

LCMS $C_{20}H_{29}BClNO_4$; calc. for (M-Boc+H⁺): 293.1324, found: 294.2.

tert-butyl (6-chloro-4-hydroxy-2,3-dihydro-1H-inden-2-yl)carbamate (**64b**):

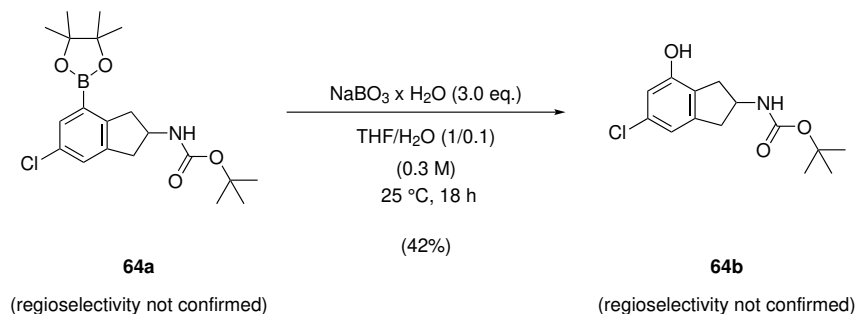


Figure S29: Conversion of **64a** to **64b**.

tert-butyl (6-chloro-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-2,3-dihydro-1H-inden-2-yl)carbamate (**64a**, 850.0 mg, 1.84 mmol, 1.00 eq.) was dissolved in THF (**62**, 5.56 mL) and H₂O (556 uL), followed by addition of sodium perborate monohydrate (549 mg, 5.51 mmol, 3.00 eq.). The reaction was stirred at 25 °C for 18 hours. The solvent evaporated and the residue was taken up in H₂O, followed by extraction with EtOAc to separate the two layers. The aqueous layer was extracted twice with EtOAc. The combined organic layers were washed with brine, dried over anhydrous sodium sulfate and evaporated to dryness. The crude material was purified by silica gel column chromatography (10 g) using an EtOAc gradient (0%-50%) in heptane. Evaporation of solvents gave the title compound *tert*-butyl (6-chloro-4-hydroxy-2,3-dihydro-1H-inden-2-yl)carbamate (**64b**) as an off-white solid (220 mg, 42%).

LCMS $C_{20}H_{29}BClNO_4$; calc. for (M-H⁺): 282.1, found: 282.2.

heptane. Evaporation of solvents gave the title compound *tert*-butyl (6-chloro-4-cyano-2,3-dihydro-1H-inden-2-yl)carbamate (**64d**) as a white solid (24.0 mg, 57%).

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.44 - 7.45 (m, 1 H), 7.42 (d, $J = 1.9$ Hz, 1 H), 4.74 (s, 1 H), 4.54 (s, 1 H), 3.43 (dd, $J = 17.1, 7.3$ Hz, 1 H), 3.34 (dd, $J = 16.6, 7.2$ Hz, 1 H), 2.97 (dd, $J = 17.1, 5.2$ Hz, 1 H), 2.88 - 2.92 (m, 1 H), 1.46 (s, 9 H). **¹³C NMR (151 MHz, CDCl₃)** δ (ppm) 155.2, 144.6, 110.3. **GCMS** C₁₅H₁₇ClN₂O₂; calc. for (M⁺): 292.1, found: 292.1.

SI12 NMR spectra

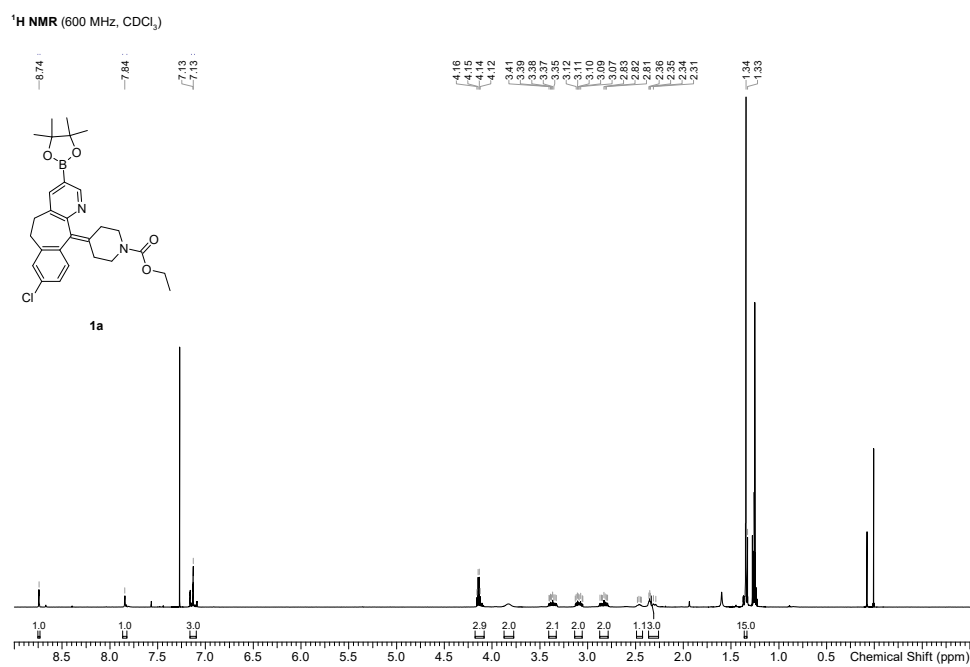


Figure S32: **1a**, ¹H-NMR spectra.

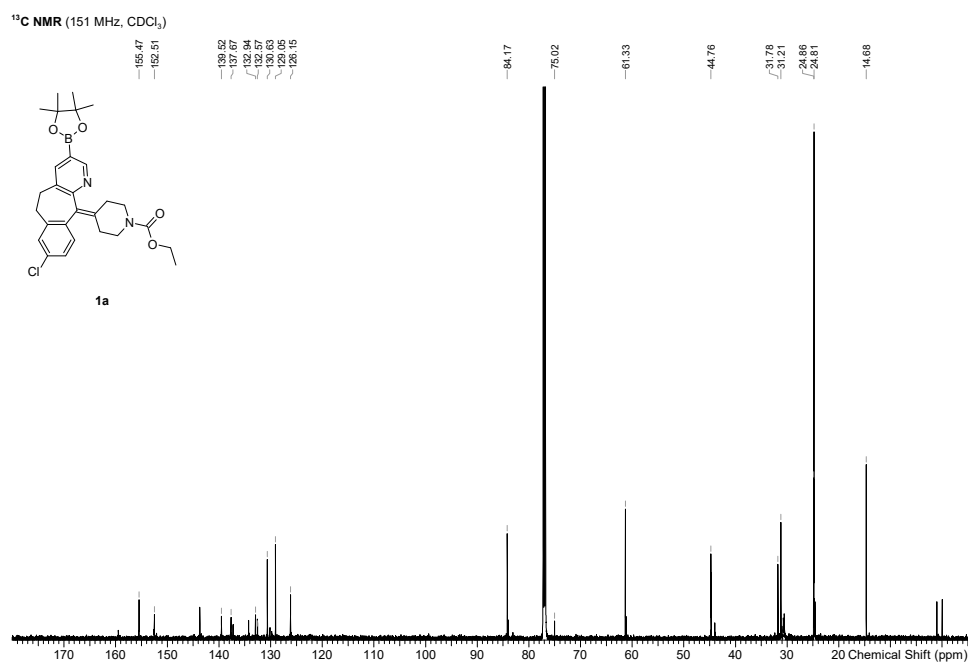


Figure S33: **1a**, ¹³C-NMR spectra.

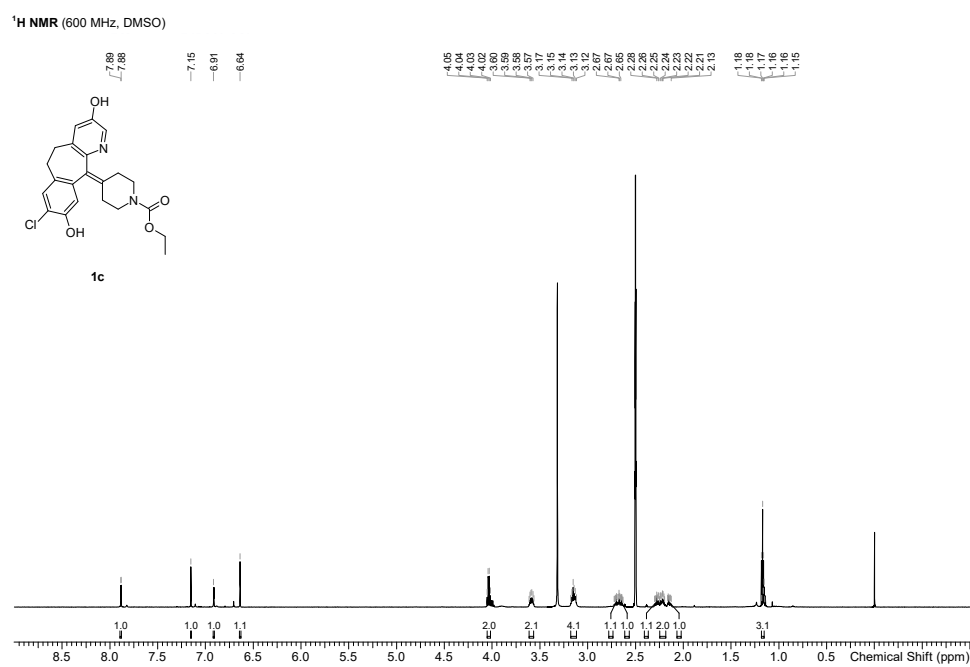


Figure S34: **1c**, ¹H-NMR spectra.

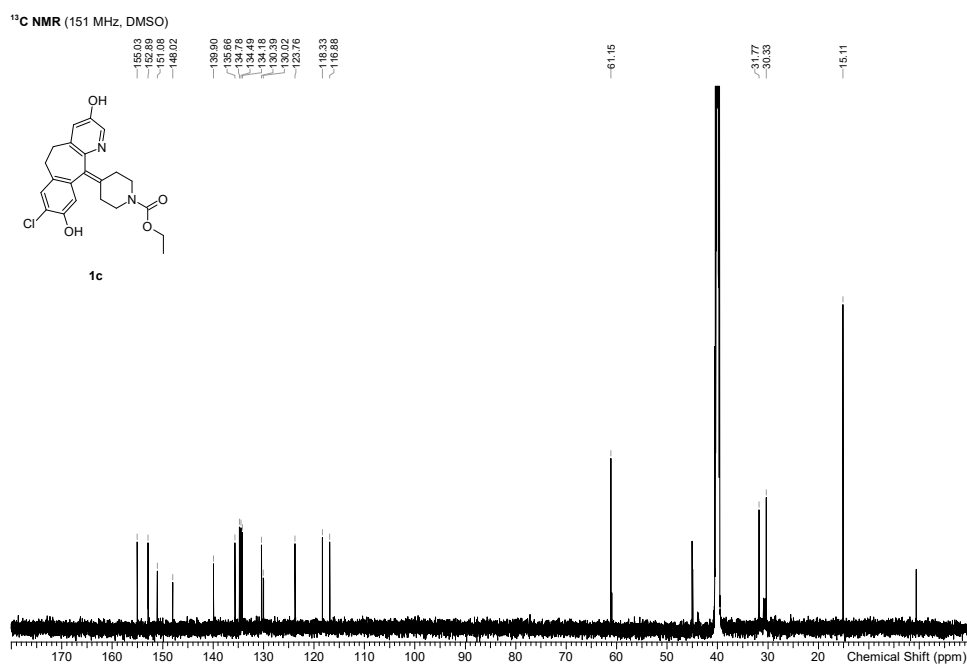


Figure S35: **1c**, ¹³C-NMR spectra.

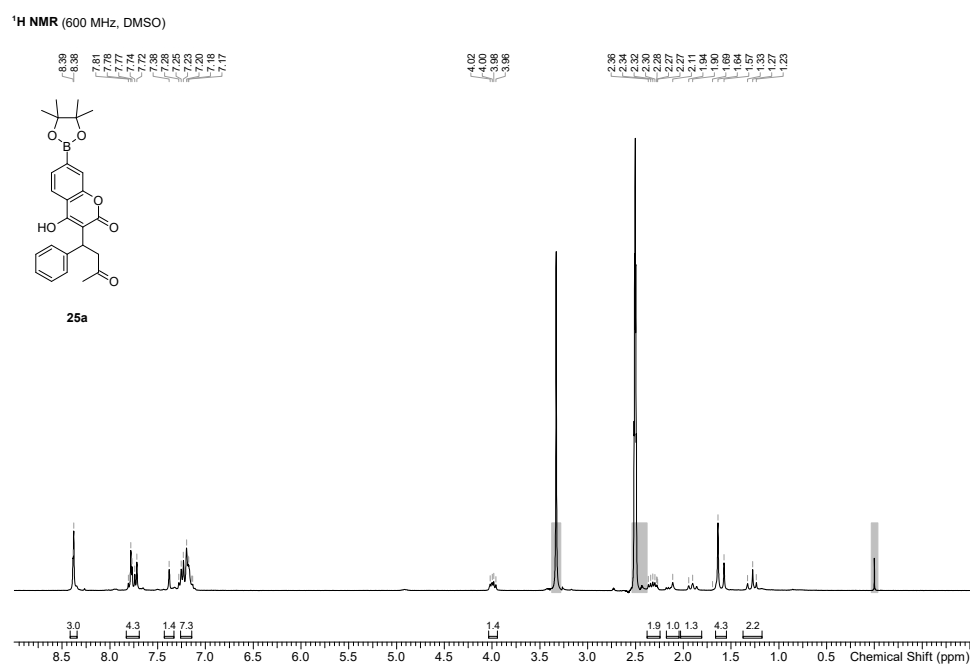


Figure S36: **25a**, ¹H-NMR spectra.

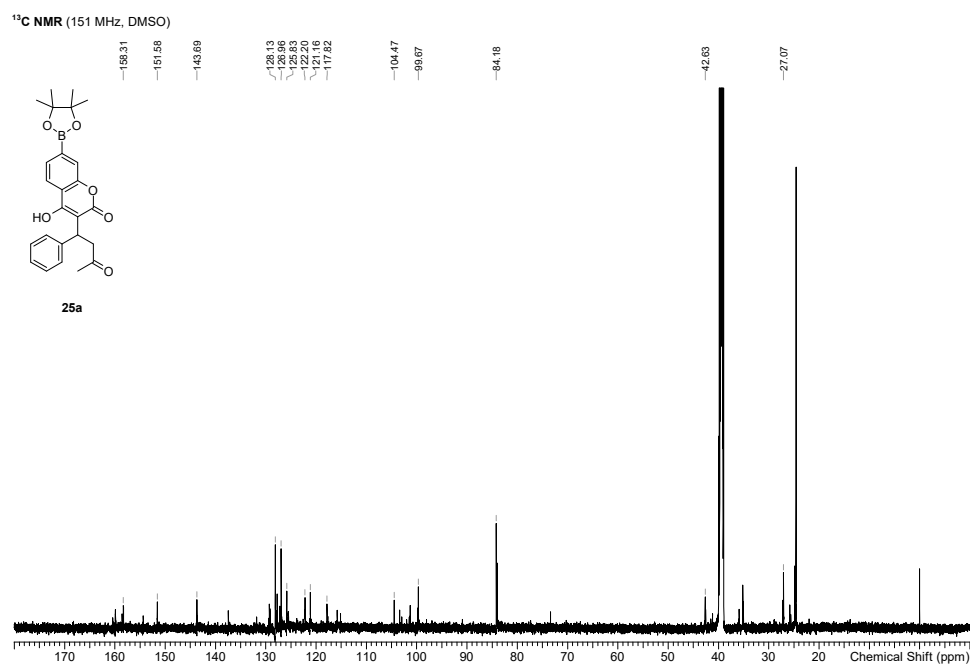
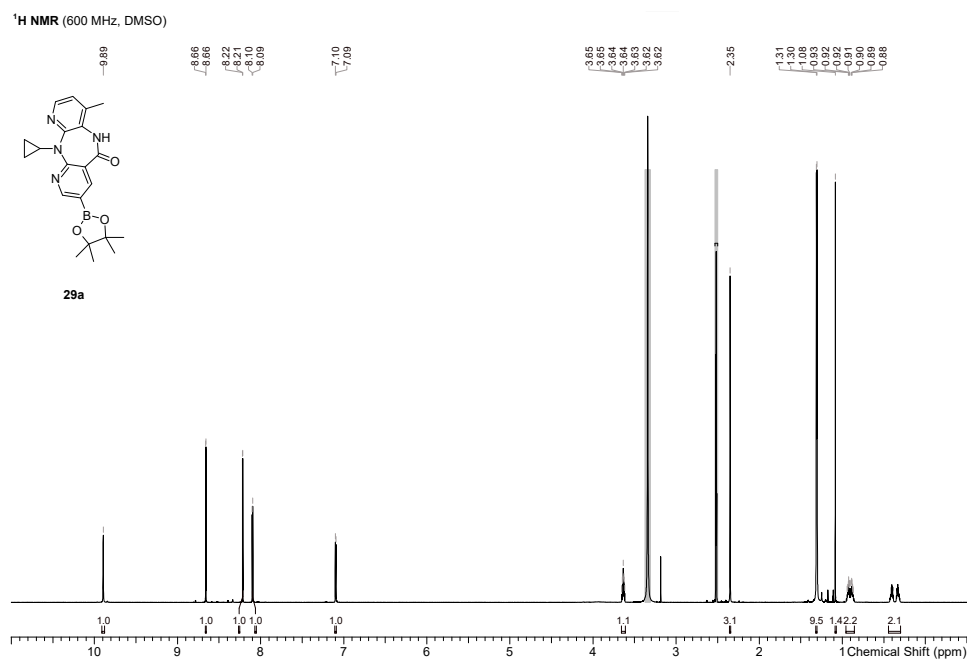


Figure S37: **25a**, ¹³C-NMR spectra.

Figure S38: **29a**, ¹H-NMR spectra.

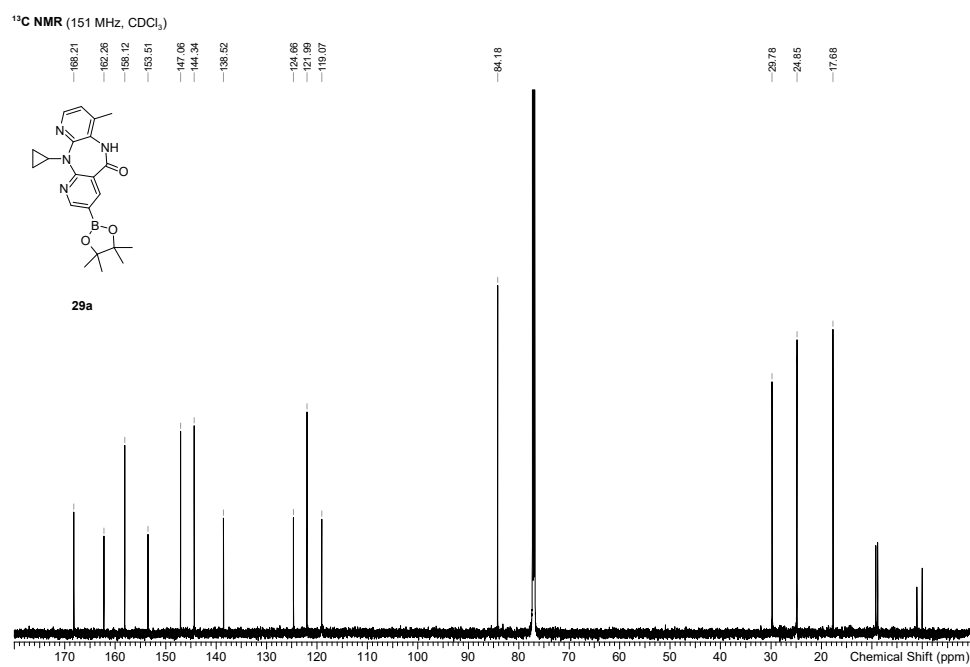


Figure S39: **29a**, ¹³C-NMR spectra.

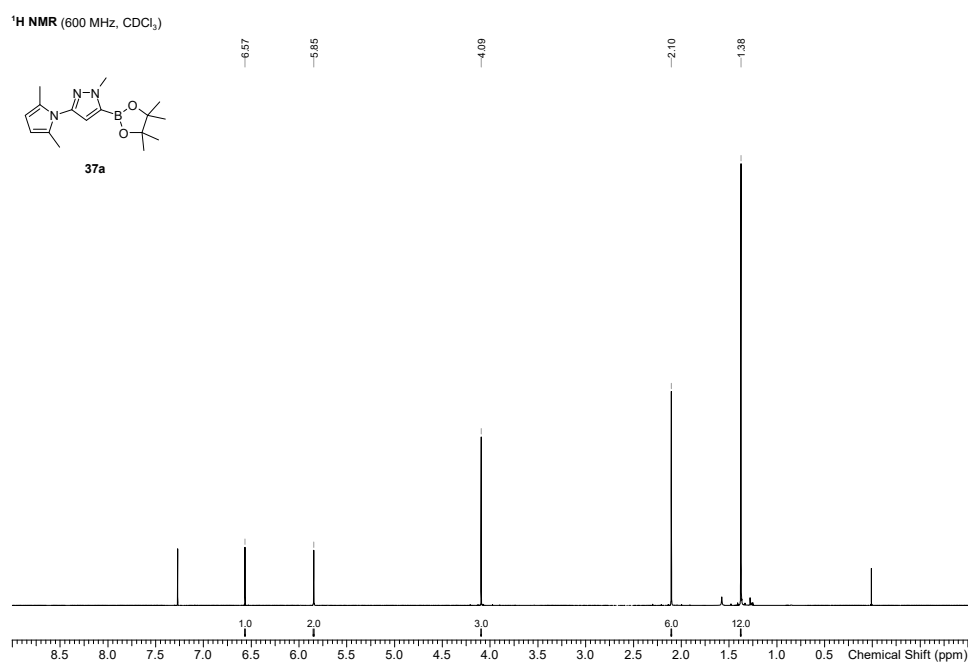


Figure S40: **37a**, ¹H-NMR spectra.

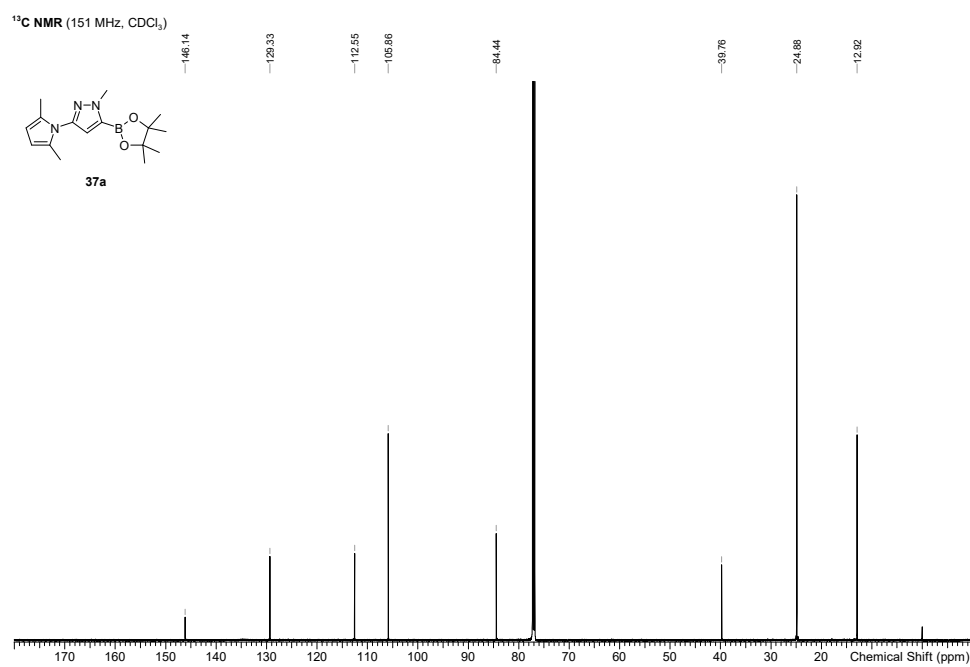


Figure S41: **37a**, ¹³C-NMR spectra.

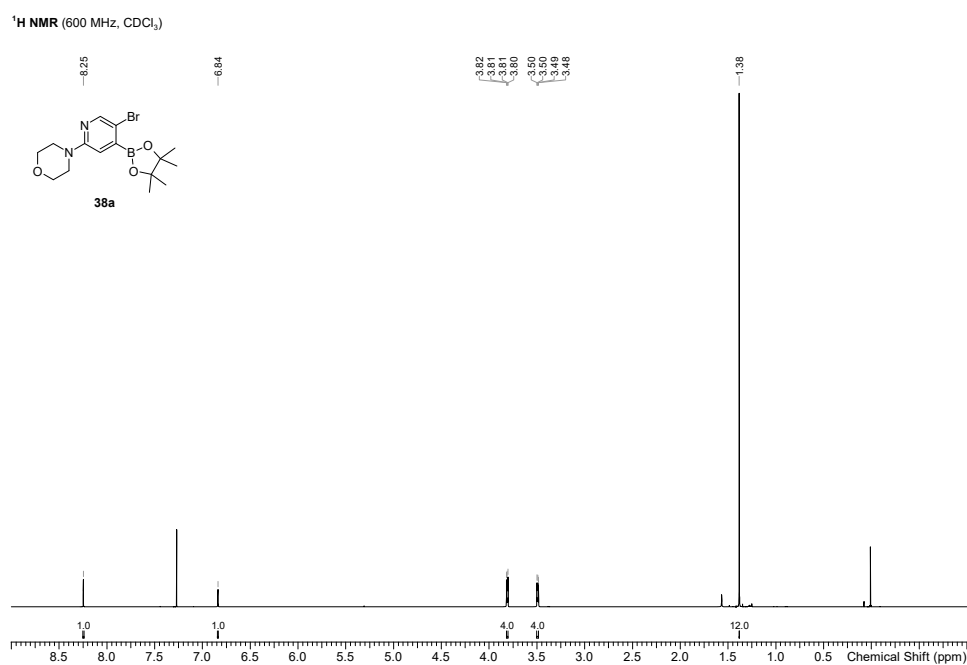


Figure S42: **38a**, ¹H-NMR spectra.

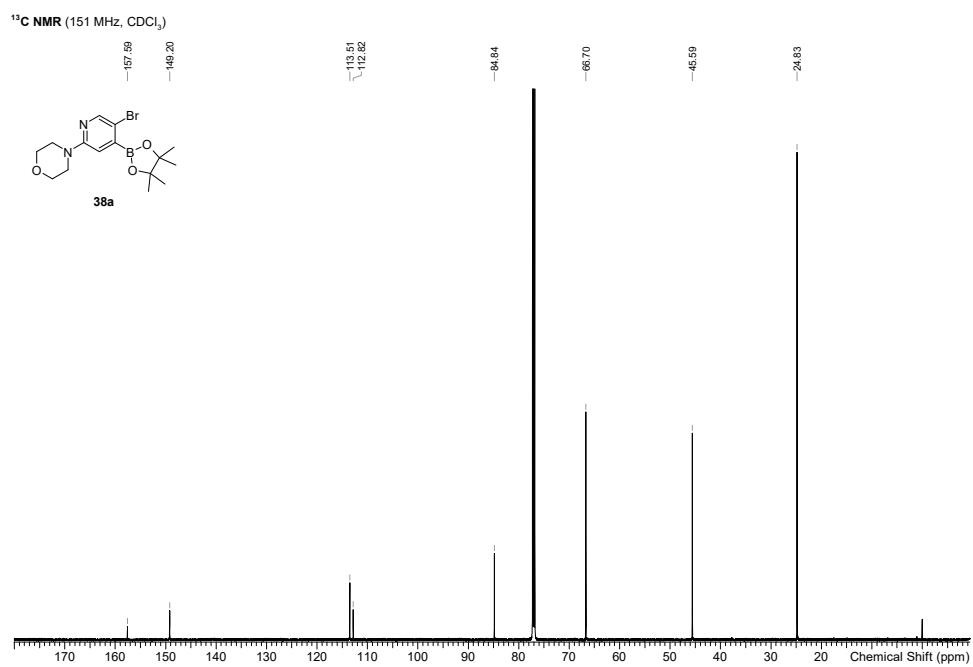
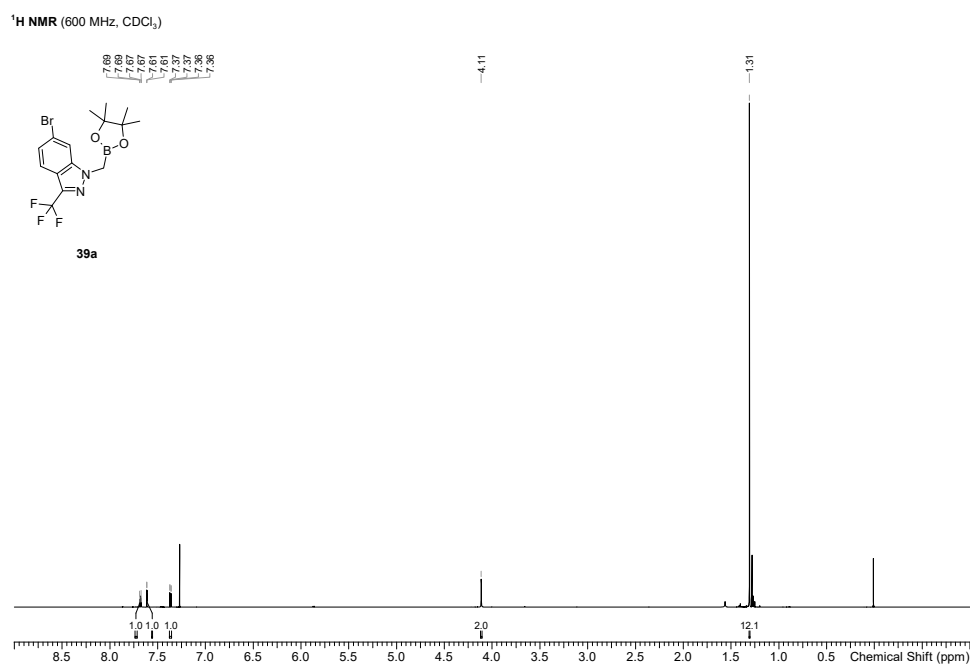


Figure S43: **38a**, ¹³C-NMR spectra.

Figure S44: **39a**, ¹H-NMR spectra.

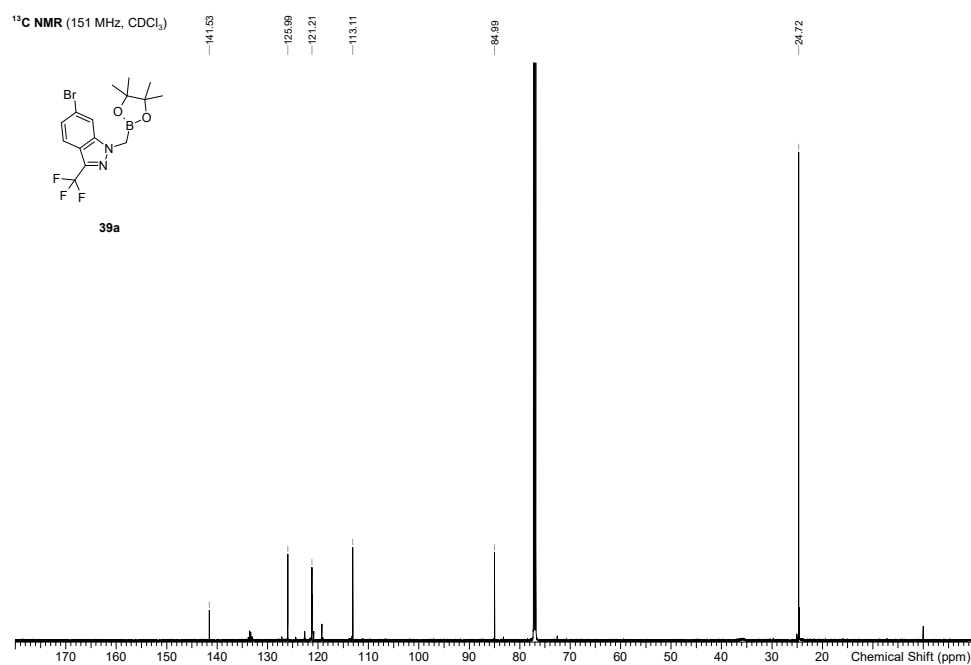


Figure S45: **39a**, ¹³C-NMR spectra.

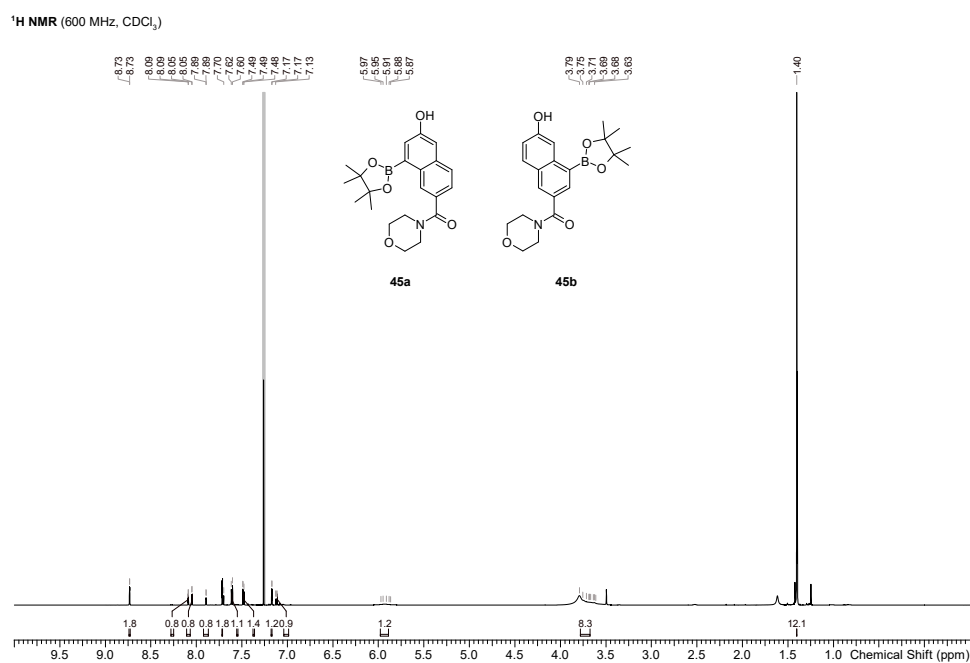


Figure S46: **45a** & **45b**, ¹H-NMR spectra.

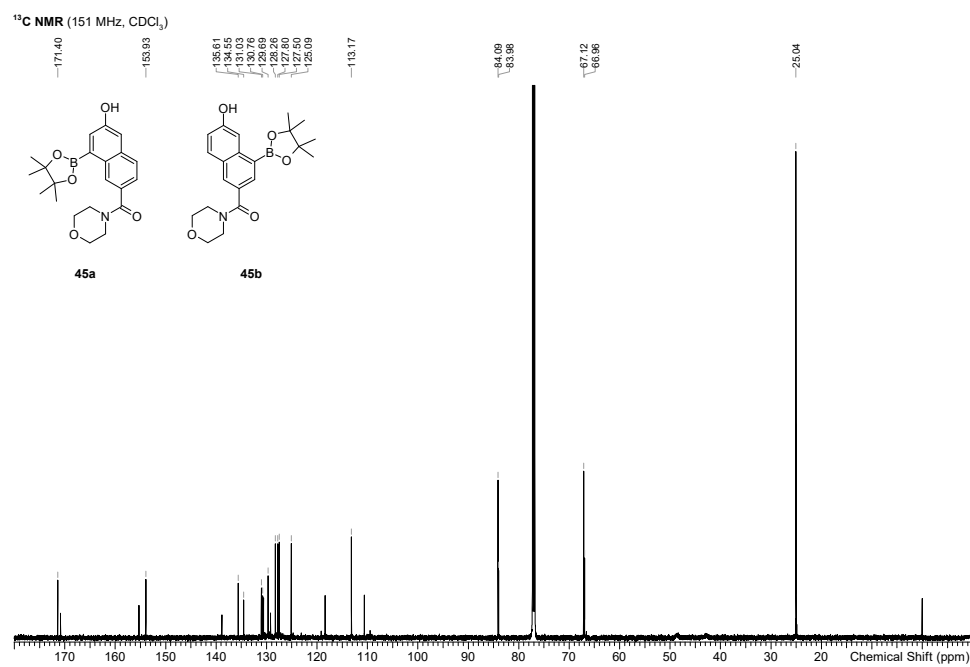


Figure S47: **45a** & **45b**, ¹³C-NMR spectra.

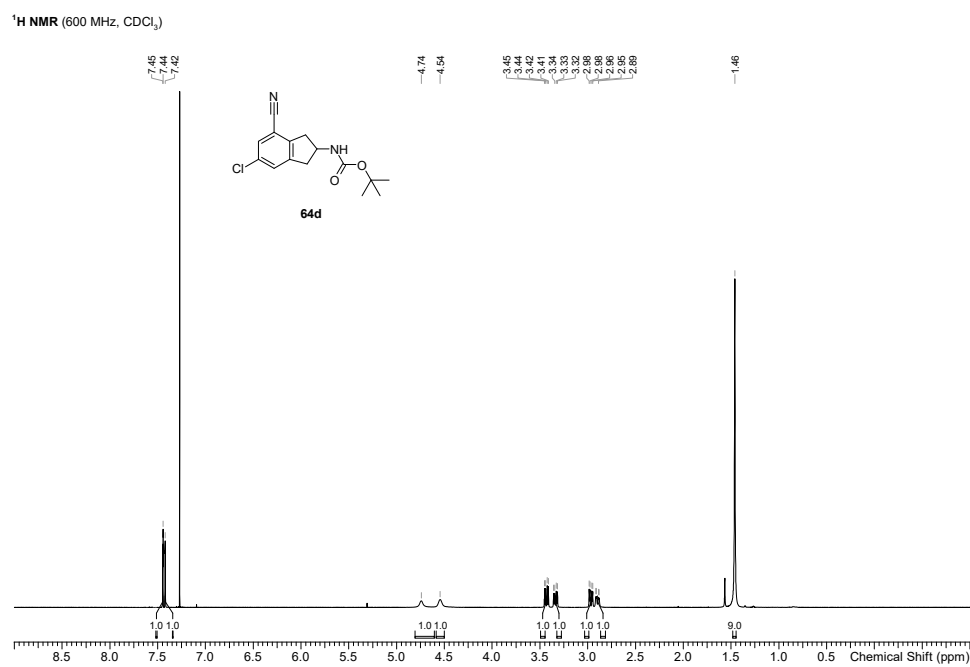


Figure S48: **64d**, ¹H-NMR spectra.

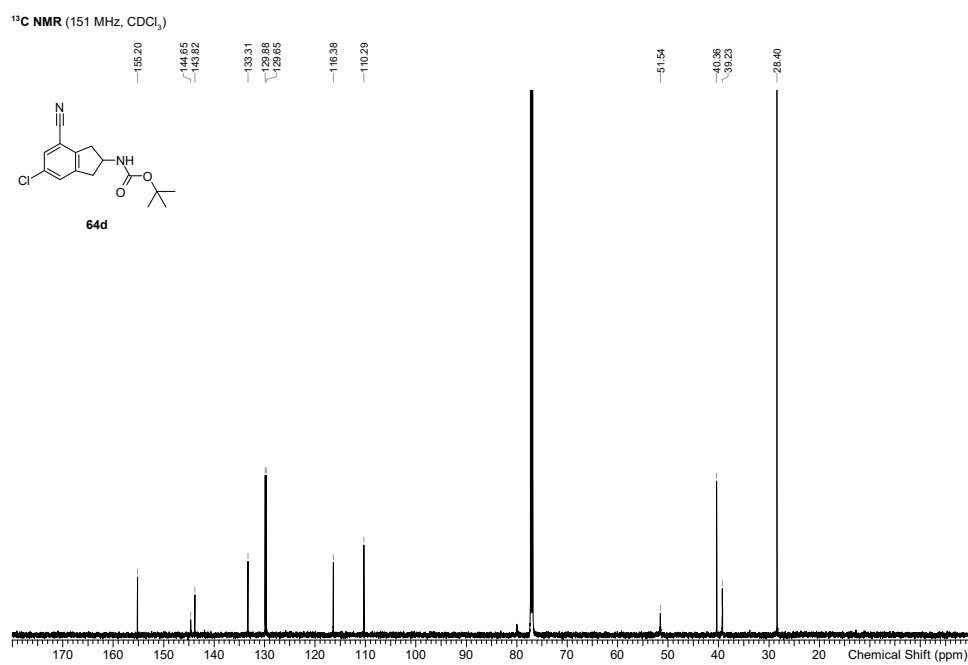


Figure S49: **64d**, ¹³C-NMR spectra.

References

1. Li, J., Burnham, J. F., Lemley, T. & Britton, R. M. Citation analysis: Comparison of web of science®, scopus™, SciFinder®, and google scholar. *J. Electron. Resour. Med.* **7**, 196–217 (2010).
2. Reyes, R. *et al.* Asymmetric remote C–H borylation of aliphatic amides and esters with a modular iridium catalyst. *Science* **369**, 970–974 (2020).
3. Tian, Y.-M. *et al.* Ni-catalyzed traceless, directed C3-selective C–H borylation of indoles. *J. Am. Chem. Soc.* **142**, 13136–13144 (2020).
4. Yu, X. *et al.* Site-selective alkene borylation enabled by synergistic hydrometallation and borometallation. *Nature Cat.* **3**, 585–592 (2020).
5. Oeschger, R. *et al.* Diverse functionalization of strong alkyl C–H bonds by undirected borylation. *Science* **368**, 736–741 (2020).
6. Larsen, M., Oeschger, R. & Hartwig, J. Effect of ligand structure on the electron density and activity of iridium catalysts for the borylation of alkanes. *ACS Catal.* **10**, 3415–3424 (2020).
7. Lv, J. *et al.* Metal-free directed sp²-C–H borylation. *Nature* **575**, 336–340 (2019).
8. Iqbal, S. *et al.* Acyl-directed ortho-borylation of anilines and C7 borylation of indoles using just BBr₃. *Angew. Chem. Int. Ed.* **58**, 15381–15385 (2019).
9. Oeschger, R., Larsen, M., Bismuto, A. & Hartwig, J. Origin of the difference in reactivity between Ir catalysts for the borylation of C–H bonds. *J. Am. Chem. Soc.* **141**, 16479–16485 (2019).
10. Bai, S.-T., Bheeter, C. & Reek, J. Hydrogen bond directed ortho-selective C–H borylation of secondary aromatic amides. *Angew. Chem. Int. Ed.* **58**, 13039–13043 (2019).
11. Bisht, R., Hoque, M. & Chattopadhyay, B. Amide effects in C–H activation: Noncovalent interactions with l-shaped ligand for meta borylation of aromatic amides. *Angew. Chem. Int. Ed.* **57**, 15762–15766 (2018).
12. Légaré Lavergne, J., Jayaraman, A., Misal Castro, L., Rochette, É. & Fontaine, F.-G. Metal-free borylation of heteroarenes using ambiphilic aminoboranes: On the importance of sterics in frustrated lewis pair C–H bond activation. *J. Am. Chem. Soc.* **139**, 14714–14723 (2017).
13. Davis, H., Genov, G. & Phipps, R. Meta-selective C–H borylation of benzylamine-, phenethylamine-, and phenylpropylamine-derived amides enabled by a single anionic ligand. *Angew. Chem. Int. Ed.* **56**, 13351–13355 (2017).
14. Chattopadhyay, B. *et al.* Ir-catalyzed ortho-borylation of phenols directed by substrate-ligand electrostatic interactions: A combined experimental/in silico strategy for optimizing weak interactions. *J. Am. Chem. Soc.* **139**, 7864–7871 (2017).
15. Hoque, M., Bisht, R., Haldar, C. & Chattopadhyay, B. Noncovalent interactions in Ir-catalyzed C–H activation: L-shaped ligand for para-selective borylation of aromatic esters. *J. Am. Chem. Soc.* **139**, 7745–7748 (2017).
16. Yin, Q., Klare, H. & Oestreich, M. Catalytic Friedel–Crafts C–H borylation of electron-rich arenes: Dramatic rate acceleration by added alkenes. *Angew. Chem. Int. Ed.* **56**, 3712–3717 (2017).
17. He, J., Shao, Q., Wu, Q. & Yu, J.-Q. Pd(II)-catalyzed enantioselective C(sp³)-H borylation. *J. Am. Chem. Soc.* **139**, 3344–3347 (2017).
18. Obligacion, J., Bezdek, M. & Chirik, P. C(sp²)-H borylation of fluorinated arenes using an air-stable cobalt precatalyst: Electronically enhanced site selectivity enables synthetic opportunities. *J. Am. Chem. Soc.* **139**, 2825–2832 (2017).
19. Li, H., Kuminobu, Y. & Kanai, M. Lewis acid–base interaction-controlled ortho-selective C–H borylation of aryl sulfides. *Angew. Chem. Int. Ed.* **56**, 1495–1499 (2017).
20. Obligacion, J., Semproni, S., Pappas, I. & Chirik, P. Cobalt-catalyzed C(sp²)-H borylation: Mechanistic insights inspire catalyst design. *J. Am. Chem. Soc.* **138**, 10645–10653 (2016).
21. Bisht, R. & Chattopadhyay, B. Formal Ir-catalyzed ligand-enabled ortho and meta borylation of aromatic aldehydes via in situ-generated imines. *J. Am. Chem. Soc.* **138**, 84–87 (2016).
22. He, J. *et al.* Ligand-promoted borylation of C(sp³)-H bonds with palladium(II) catalysts. *Angew. Chem. Int. Ed.* **55**, 785–789 (2016).
23. Furukawa, T., Tobisu, M. & Chatani, N. C–H functionalization at sterically congested positions by the platinum-catalyzed borylation of arenes. *J. Am. Chem. Soc.* **137**, 12211–12214 (2015).
24. Kuminobu, Y., Ida, H., Nishi, M. & Kanai, M. A meta-selective C–H borylation directed by a secondary interaction between ligand and substrate. *Nat. Chem.* **7**, 712–717 (2015).

25. Feng, Y. *et al.* Total synthesis of verruculogen and fumitremorgin enabled by ligand-controlled C-H borylation. *J. Am. Chem. Soc.* **137**, 10160–10163 (2015).
26. Larsen, M., Wilson, C. & Hartwig, J. Iridium-catalyzed borylation of primary benzylic C-H bonds without a directing group: Scope, mechanism, and origins of selectivity. *J. Am. Chem. Soc.* **137**, 8633–8643 (2015).
27. Wang, G., Xu, L. & Li, P. Double N,B-type bidentate boryl ligands enabling a highly active iridium catalyst for C-H borylation. *J. Am. Chem. Soc.* **137**, 8058–8061 (2015).
28. Saito, Y., Segawa, Y. & Itami, K. Para -C-H borylation of benzene derivatives by a bulky iridium catalyst. *J. Am. Chem. Soc.* **137**, 5193–5198 (2015).
29. Miyamura, S., Araki, M., Suzuki, T., Yamaguchi, J. & Itami, K. Stereodivergent synthesis of arylcyclopropylamines by sequential C-H borylation and Suzuki-Miyaura coupling. *Angew. Chem. Int. Ed.* **54**, 846–851 (2015).
30. Obligacion, J., Semproni, S. & Chirik, P. Cobalt-catalyzed C-H borylation. *J. Am. Chem. Soc.* **136**, 4133–4136 (2014).
31. Larsen, M. & Hartwig, J. Iridium-catalyzed C-H borylation of heteroarenes: Scope, regioselectivity, application to late-stage functionalization, and mechanism. *J. Am. Chem. Soc.* **136**, 4287–4299 (2014).
32. Preshlock, S. *et al.* High-throughput optimization of Ir-catalyzed C-H borylation: A tutorial for practical applications. *J. Am. Chem. Soc.* **135**, 7572–7582 (2013).
33. Liskey, C. & Hartwig, J. Iridium-catalyzed C-H borylation of cyclopropanes. *J. Am. Chem. Soc.* **135**, 3375–3378 (2013).
34. Tajuddin, H. *et al.* Iridium-catalyzed C-H borylation of quinolines and unsymmetrical 1,2-disubstituted benzenes: Insights into steric and electronic effects on selectivity. *Chem. Sci.* **3**, 3505–3515 (2012).
35. Roosen, P. *et al.* Outer-sphere direction in iridium C-H borylation. *J. Am. Chem. Soc.* **134**, 11350–11353 (2012).
36. Dai, H.-X. & Yu, J.-Q. Pd-catalyzed oxidative ortho -C-H borylation of arenes. *J. Am. Chem. Soc.* **134**, 134–137 (2012).
37. Ros, A. *et al.* Use of hemilabile N,N ligands in nitrogen-directed iridium-catalyzed borylations of arenes. *Angew. Chem. Int. Ed.* **50**, 11724–11728 (2011).
38. Robbins, D., Boebel, T. & Hartwig, J. Iridium-catalyzed, silyl-directed borylation of nitrogen-containing heterocycles. *J. Am. Chem. Soc.* **132**, 4068–4069 (2010).
39. Paul, S. *et al.* Ir-catalyzed functionalization of 2-substituted indoles at the 7-position: Nitrogen-directed aromatic borylation. *J. Am. Chem. Soc.* **128**, 15552–15553 (2006).
40. Chotana, G., Rak, M. & Smith, M. Sterically directed functionalization of aromatic C-H bonds: Selective borylation ortho to cyano groups in arenes and heterocycles. *J. Am. Chem. Soc.* **127**, 10539–10544 (2005).
41. Kearnes, S. M. *et al.* The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
42. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
43. Murtagh, F. & Legendre, P. Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion? *J. Classif.* **31**, 274–295 (2014).
44. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
45. Bhutani, P. *et al.* US FDA approved drugs from 2015–June 2020: A perspective. *J. Med. Chem.* **64**, 2339–2381 (2021).
46. Vitaku, E., Smith, D. T. & Njardarson, J. T. Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among US FDA approved pharmaceuticals: miniperspective. *J. Med. Chem.* **57**, 10257–10274 (2014).
47. Gomtsyan, A. Heterocycles in drugs and drug discovery. *Chem. Heterocyc. Compd.* **48**, 7–10 (2012).
48. Miyaura, N. & Suzuki, A. Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chem. Rev.* **95**, 2457–2483 (1995).
49. Nicolaou, K., Bulger, P. G. & Sarlah, D. Palladium-catalyzed cross-coupling reactions in total synthesis. *Angew. Chem. Int. Ed.* **44**, 4442–4489 (2005).
50. Ertl, P. In silico identification of bioisosteric functional groups. *Curr. Opin. Drug Discov. Dev.* **10**, 281–288 (2007).
51. Ertl, P., Altmann, E. & McKenna, J. M. The most common functional groups in bioactive molecules and how their popularity has evolved over time. *J. Med. Chem.* **63**, 8408–8418 (2020).

52. Ertl, P. An algorithm to identify functional groups in organic molecules. *J. Cheminformatics* **9**, 1–7 (2017).
53. Dalke, A. The chemfp project. *J. Cheminformatics* **11**, 1758–2946 (2019).
54. Landrum, G. *RDKit: Open-source cheminformatics software* May 2010. <http://www.rdkit.org/>.
55. Caldeweyher, E. *et al.* A hybrid machine-learning approach to predict the iridium-catalyzed borylation of C–H bonds. *ChemRxiv preprint* (2022).
56. Meyers, J., Carter, M., Mok, N. Y. & Brown, N. On the origins of three-dimensionality in drug-like molecules. en. *Future Med. Chem.* **8**, 1753–1767 (2016).
57. Pomberger, A. *et al.* The effect of chemical representation on active machine learning towards closed-loop optimization. *React. Chem. Eng.* (2022).
58. Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system* in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), 785–794.
59. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Stay committed to the process, and the results will follow.

- Jan Frodeno

6

LATE-STAGE MINISCI-TYPE C-H ALKYLATION CHEMISTRY

This chapter describes the first application of the developed late-stage functionalization (LSF) screening platform (DOLPHIN) and reaction data format (SURF) to explore the substrate scope of late-stage C-H alkylations. A library-type screening approach aims to facilitate the *in silico* reactivity prediction of suitable substrates coupled to a diverse set of sp³-rich building blocks using Minisci-type chemistry.

First, a short overview covering C-H alkylations, specifically Minisci-type transformations, and their potential for LSF in the context of drug discovery is given (Chapter 6.1). Next, the publication that describes the case study in detail and was published in *Communications Chemistry* is reprinted with permission (Chapter 6.2). [457] The final section contains the corresponding experimental and supplementary information (Chapter 6.3).

6.1 Introduction and background

Minisci-type reactions, first reported by Francesco Minisci with silver salts as catalysts in 1971, facilitate the introduction of alkyl groups to electron-deficient heterocycles. [458] The transformation is based on a radical mechanism that aids the substitution of a hydrogen atom on the heteroaromatic core with a nucleophilic radical. [459] Over the past decades, an expansion of the original methodology, including the development and application of a diverse array of radicals was investigated. [460, 461]

Today, a wide range of radical precursors can be employed for Minisci-type reactions, those include aldehydes, [462] alkyl halides, [463] ethers, [464] alkyl boronic acids, [465] sulfonates, [466] and even simple alkanes. [467] The generation of radicals from these precursors involves distinct mechanisms such as decarboxylation, decarbonylation, dehalogenation, hydrogen abstraction, deboronation, or desulfonylation. Additionally, the scope of functional groups that can be introduced has been extended from simple alkyl chains to aryl, [468] carbonyl, [469] and broadly functionalized methyl groups. [470, 471] The introduction of complex alkyl groups is of particular interest as it enhances molecular three-dimensionality (3D), which can lead to improved selectivity and reduced off-target effects, thereby accelerating the DMTA cycle. [472, 473]

In his foundational work, Minisci described the alkylation of *N*-heteroaromatic bases through a novel method of alkyl radical generation, catalyzed by silver ions and mediated by the decarboxylation of carboxylic acids with persulfate as the oxidant. [458] The proposed mechanism for this transformation, using cyclohexane carboxylic acid (**10**) as the radical source and 6-methoxy-2-methylquinoline (**11**) as the heterocyclic substrate, is exemplified in Figure 6.1. At first, a hydrogen atom is abstracted from the carboxylic acid **10** by the transition metal catalyst (Ag), which is then transferred to the oxidant anion through hydrogen atom transfer (HAT). This step produces the intermediate radical **I1** and regenerates the catalyst. Then, the carboxyl radical **I1** releases carbon dioxide (CO₂) to form the carbon-centered radical **I2**. Tertiary alkyl radicals are the most stable and reactive due to hyperconjugation. [474] Upon protonation of the nitrogen atom in starting material **11** to form **11a**, the alkyl radical **I2** attacks the electron-deficient heteroaromatic system to deliver intermediate **I3**. To restore the aromaticity of the heterocyclic system, intermediate **I3** loses a proton at the alkylated position by HAT to form intermediate **I4**, which is followed by deprotonation to yield the final, alkylated heterocyclic product **12**.

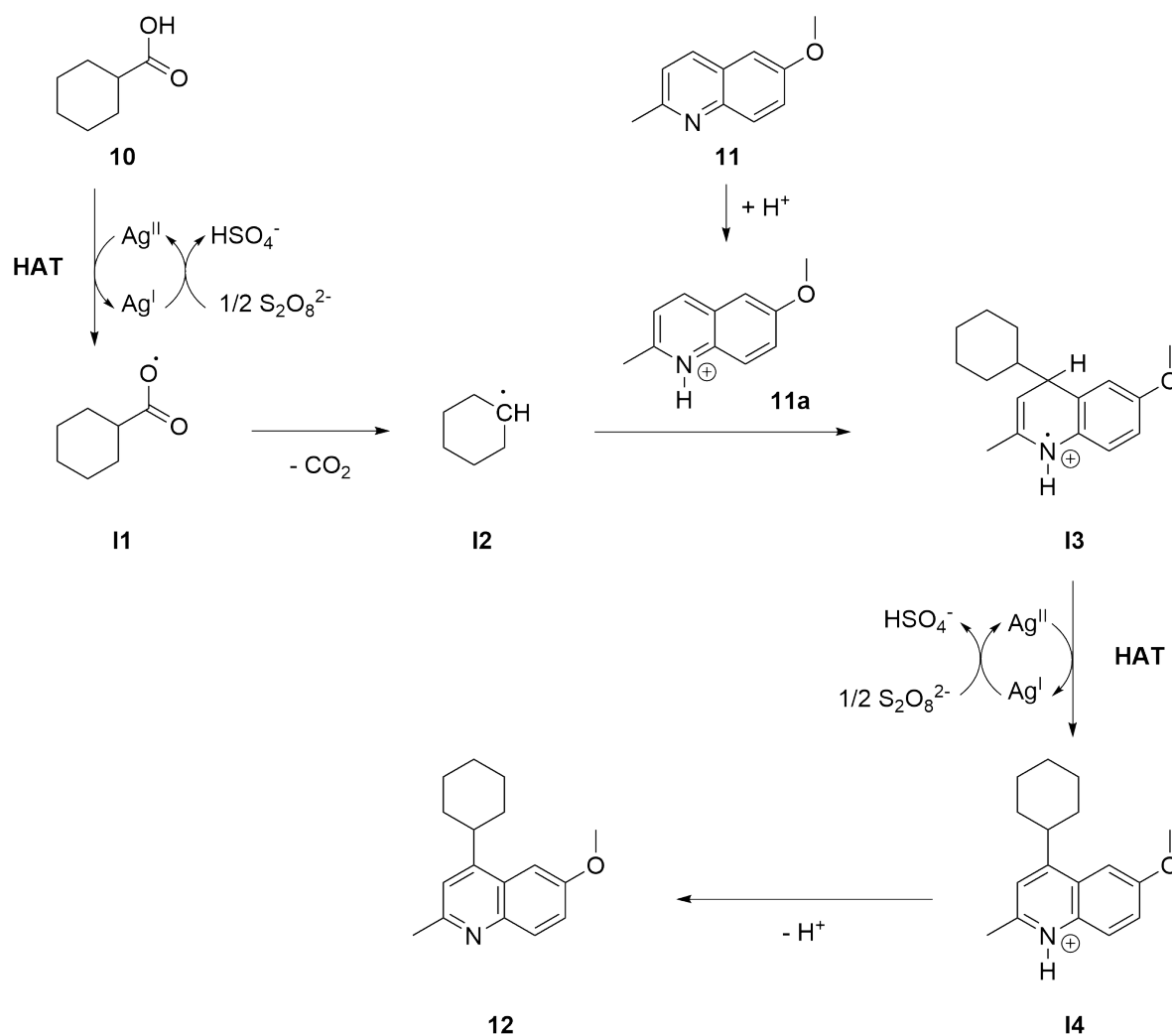


Figure 6.1: General mechanism of the Minisci-type C-H alkylation exemplified by the reaction of cyclohexane carboxylic acid (**10**) with 6-methoxy-2-methylquinoline (**11**). Hydrogen abstraction facilitated by the catalyst (Ag) on **10** generates radical intermediate **I1** and facilitates hydrogen atom transfer (HAT). Decarboxylation of **I1** leads to a carbon radical species **I2**. Heteroarene **11** is protonated and attacked by the alkyl radical **I2** to generate intermediate **I3**. Restoring of aromaticity on **I3** is facilitated through HAT to deliver **I4**, followed by deprotonation to the alkylated product **12**.

Although the original Minisci reactions were conducted at 70 °C, [458] subsequent studies have demonstrated their feasibility under milder conditions, including room temperature [475] and 40 °C. [476] The use of readily available carboxylic acids as alkylating agents, the requirement for only a few reagents, and the typically good to acceptable yields contribute to the attractiveness of Minisci-type reactions. [460] The versatility of these reactions is further supported by the ease to varying conditions, including different oxidant/catalyst systems and technologies, such as photo- and electrochemistry. [461, 477, 478]

Given these attributes, Minisci-type reactions could be a valuable methodology in the LSF toolbox for the introduction of 3D-rich fragments into advanced drug-like molecules. However,

the predictability of introducing alkyl chains and rings into structurally complex chemical matter is not always straightforward. To evaluate the applicability and possibly increase the wet-lab reaction success of alkylation reactions, a case study, that connects semi-automated high-throughput experimentation (HTE) with *in silico* reaction screening was designed and conducted.

6.2 Publication

The following case study has been published as: **Nippa, D. F.[†]**, Atz, K.[†], Müller, A. T., Wolfard, J., Isert, C., Binder, M., Scheidegger, O., Stepan, A. F., Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Identifying opportunities for late-stage C-H alkylation with in silico reaction screening and high-throughput experimentation *Comms. Chem.*, **6**, 256 (2023). [457] The material (DOI: 10.1038/s42004-023-01047-5) is reprinted with permission from Springer Nature Limited (Author reuse for own thesis).

The author of this thesis is the co-first author of the publication as he carried out the literature analysis, the experimental work (HTE, scale-up), the reaction data preparation for the predictive tool, and the writing of the first manuscript draft. The machine learning algorithms were designed and developed by Dr. Kenneth Atz. Further details on the contributions of all authors are stated on the last page of the publication.

A detailed description of the experiments conducted and methods used in the publication can be found in Chapter 6.3.









communications chemistry

ARTICLE

<https://doi.org/10.1038/s42004-023-01047-5>

OPEN

Identifying opportunities for late-stage C-H alkylation with high-throughput experimentation and in silico reaction screening

David F. Nippa ^{1,2,4}, Kenneth Atz^{3,4}, Alex T. Müller ¹, Jens Wolfard ¹, Clemens Isert ³, Martin Binder¹, Oliver Scheidegger¹, David B. Konrad ²✉, Uwe Grether ¹✉, Rainer E. Martin ¹✉ & Gisbert Schneider ³✉

Enhancing the properties of advanced drug candidates is aided by the direct incorporation of specific chemical groups, avoiding the need to construct the entire compound from the ground up. Nevertheless, their chemical intricacy often poses challenges in predicting reactivity for C-H activation reactions and planning their synthesis. We adopted a reaction screening approach that combines high-throughput experimentation (HTE) at a nanomolar scale with computational graph neural networks (GNNs). This approach aims to identify suitable substrates for late-stage C-H alkylation using Minisci-type chemistry. GNNs were trained using experimentally generated reactions derived from in-house HTE and literature data. These trained models were then used to predict, in a forward-looking manner, the coupling of 3180 advanced heterocyclic building blocks with a diverse set of sp³-rich carboxylic acids. This predictive approach aimed to explore the substrate landscape for Minisci-type alkylations. Promising candidates were chosen, their production was scaled up, and they were subsequently isolated and characterized. This process led to the creation of 30 novel, functionally modified molecules that hold potential for further refinement. These results positively advocate the application of HTE-based machine learning to virtual reaction screening.

¹Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, 4070 Basel, Switzerland. ²Department of Pharmacy, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, 81377 Munich, Germany. ³Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland. ⁴These authors contributed equally: David F. Nippa, Kenneth Atz. ✉email: david.konrad@cup.lmu.de; uwe.grether@roche.com; rainer_e.martin@roche.com; gisbert@ethz.ch

The synthesis of novel compounds represents the bottleneck in terms of time and effort for numerous small molecule drug discovery projects¹. Late-stage functionalization (LSF) is a strategy that adds extra functional groups to drug molecules, bypassing the necessity for entirely new synthesis or the requirement for specific functional handles². These subtle structural alterations simplify the process of understanding the relationships between the chemical structure and the biological activity (structure–activity relationships, SARs). Additionally, they allow for the enhancement of pharmacokinetic properties, including absorption, distribution, metabolism, and excretion, in lead compounds and drug candidates³. Importantly, these modifications can be achieved with lower synthetic costs⁴. Nonetheless, it is worth noting that not all molecules readily lend themselves to the desired functionalizations, making LSF a challenging process in experimental terms. In response to this challenge, we present a computational machine-learning framework designed for predicting the reactivity of drug molecules. This framework offers a more rational approach to LSF, potentially reducing the time and experimental costs typically associated with this endeavor.

An increasing number of experimental LSF methods have recently been published that allow medicinal chemists to fluorinate, aminate, arylate, methylate, trifluoromethylate, borylate, acylate, or oxidize structurally intricate molecules^{5,6}. Alkylation reactions have gained interest as they allow the introduction of small cyclic and acyclic alkyl groups through carbon–carbon, carbon–oxygen, or carbon–nitrogen bond formation⁷. In particular, Minisci-type alkylations^{8,9} are considered a valuable LSF methodology for incorporating alkyl building blocks into heterocyclic systems, which often form the core of drug molecules¹⁰.

Originally described in the mid-20th century, Minisci reactions have become a versatile tool in medicinal chemistry for the formation of C–C bonds¹¹. Using ammonium persulfate as the oxidant and silver nitrate as the catalyst, alkyl radicals are generated from the corresponding carboxylic acids at elevated temperatures. Upon radical addition to the heteroarene, the reaction product is formed through aromaticity-driven oxidation of the radical intermediate¹¹. The scope of both, electron-deficient heteroarenes and alkyl-donating coupling partners, has steadily been expanded^{12,13}. Various radical sources have been documented in the literature. These include alkyl carboxylic acids capable of transferring alkyl groups, boronic acids suitable for the incorporation of aryl groups, or sulfonates that were used to transfer trifluoromethyl or tert-butyl fragments^{14,15}. Employing readily accessible and cost-effective carboxylic acids, without the prerequisite for prefunctionalization, considerably broadens the applicability of this transformation for drug discovery purposes¹⁶. The growing emphasis on integrating sp³-rich building blocks into pharmaceuticals¹⁷, coupled with the ready availability of stable cyclic alkyl carboxylic acids, renders this approach particularly appealing for expanding hits into lead compounds and optimizing drugs through LSF.

It has become apparent that by decreasing the count of aromatic rings within a drug candidate, the chances of achieving clinical success can be heightened¹⁸. A higher proportion of sp³ centers allows for exploration of novel chemical territory, which can potentially improve drug selectivity¹⁹. This shift can also positively influence essential physicochemical properties, such as solubility and metabolic stability^{20–22}. While guidelines exist for predicting reactivity in Minisci-type transformations, the challenge lies in the limited range of functional groups that can be accommodated, along with the diverse array of C–H bonds and electronic effects within complex molecules. These complexities make the prediction of alkylation reactions a challenging task^{4,23}. Conducting individual reactions at the typical scale used in

medicinal chemistry (milligram scale) to enrich the reaction database with pertinent transformation examples would be a laborious and resource-intensive undertaking, yielding limited value relative to the effort invested.

High-throughput experimentation (HTE) has emerged as a valuable tool for systematically exploring and optimizing new chemical transformations in a semi-automated manner^{24,25}. To effectively accomplish the miniaturization of reactions at the nanomolar scale, it is essential to engineer the system with precision to handle extremely small quantities of materials and ensure consistent and thorough mixing of the reaction components²⁶. Advanced technologies like ultra-high-performance liquid chromatography-mass spectrometry enable the analysis and the separation of minute quantities from screening plates^{27,28}. Another crucial aspect of HTE involves the careful curation of all collected reaction data, including unsuccessful transformations, in accordance with the FAIR principles (findable, accessible, interoperable, and reusable)²⁹. This approach ensures the creation of high-quality datasets suitable for machine learning applications^{30–32}.

Graph neural networks (GNNs) that enable efficient learning on three-dimensional (3D) molecular models have found various applications in drug discovery and development^{33–35}. In addition to their prominent applications in quantum chemistry^{36,37}, GNN methods have been developed for the prediction of forward reactions, starting from small substrates and leading to the synthesis of complex drug molecules^{38–40}. Moreover, GNNs have recently found application in LSF to predict reaction yield, binary reaction outcome, and regioselectivity for borylation reactions⁴¹. A similar methodology has been introduced for predicting late-stage alkylation, with a primary emphasis on Baran-type diversinate chemistry that employs alkyl sodium sulfinate salts⁴². Additionally, a recent investigation has demonstrated that hybrid machine learning models, enriched with quantum chemical details about transition states, can achieve accurate predictions of regioselectivity for iridium-catalyzed borylation reactions, even when operating with limited data⁴³.

In this study, we showcase the application of GNNs trained on a limited set of reaction data for machine-learning-based virtual reaction screening. When combined with laboratory automation, this approach has facilitated the discovery of 276 promising alkylation possibilities with high precision (Fig. 1). This effort has resulted in the synthesis of a diverse range of novel compounds characterized by an enhanced sp³ fraction.

Results

HTE reaction screening. The Minisci-type reactions described by Sutherland et al.¹⁶ were effectively downscaled from a micromolar (150 μmol) to a nanomolar (500 nmol) level in a parallel configuration using a 24-well plate, achieving a reduction factor of 300 (Fig. 2A, B). Throughout the optimization process, it became evident that the reaction yields substantially improved when performed inside a glovebox. Conducting the reaction with 23 distinct carboxylic acids labeled as a–w (Fig. 2C) at various temperatures revealed that the highest conversions were achieved at 40 °C. Elevating the temperature beyond this point primarily resulted in the formation of di-alkylation products. To attain increased conversions, we doubled the amounts of alkyl carboxylic acids (20 equivalents instead of 10) and oxidants (6 equivalents instead of 3). This adjustment led to higher conversions, with an average improvement factor of 1.2–1.5. We included a reference reaction involving Quinoline **1** and carboxylic acid **e** in position B4 (Fig. 2C) to monitor potential performance variations and to ensure the reproducibility of the screening results. Since this reaction is anticipated to consistently

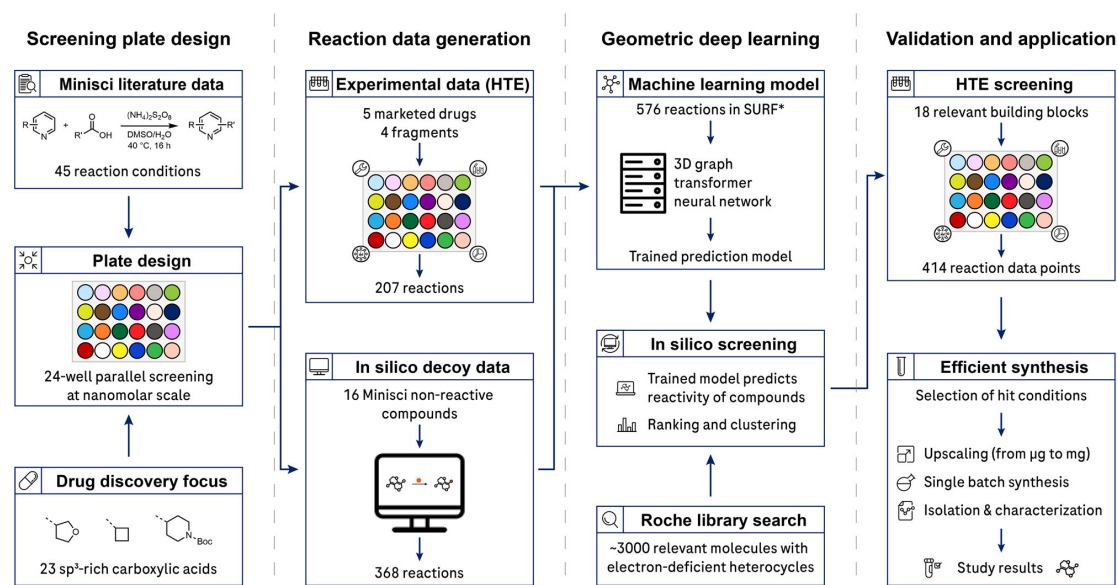


Fig. 1 Overview of the research study. *Screening plate design:* Minisci literature data containing metal-free reactions were extracted and analyzed to determine suitable reaction conditions. For parallel reaction screening, 23 sp³-rich carboxylic acids with relevance for drug discovery were included. *Reaction data generation:* Using the reaction plate design, physical experiments in high-throughput experimentation (HTE) fashion were conducted with marketed drugs and fragments from an informer library (184 reactions⁴¹) covering relevant chemical space. In addition, 16 distinctly non-reactive substrates were screened for in silico decoy data generation (368 reactions). *Geometric deep learning:* The obtained reaction data (SURF, Simple User-friendly Reaction Format)⁴¹ were subjected to geometric deep learning, incorporating 3D structural information of the chemicals. The trained model was applied to 3000 building blocks from the Roche library, with a particular focus on electron-deficient heterocycles. This in silico screening predicted the reactivity of the compounds for substrate ranking and clustering. *Validation and application:* The prediction models were experimentally validated for a diverse set of 18 building blocks. Selected scale-up reactions led to fully characterized compounds.

yield the desired outcome under the specified conditions, any unexpected outcome in this well would serve as a warning sign, indicating the potential influence of external factors or mishandling of the plates. Such deviations would prompt concerns regarding data reliability. Therefore, in the final configuration, we assessed the integration of 23 diverse alkyl groups, with a primary emphasis on compact sp³ ring systems, into electron-deficient heterocycles.

Binary reaction outcomes were labeled as “successful” when the chosen substrate, under the specified reaction conditions, produced a mono- or di-alkylation product that could be confirmed by liquid chromatography-mass spectrometry (LCMS) with a threshold of 5%. Conversely, outcomes were classified as “unsuccessful” when the intended transformation could not be detected through LCMS. In cases of di-alkylation, we consistently observed three distinct products: mono-alkylation on the two distinct carbons and di-alkylation on both. To facilitate the training of machine learning models, the yields of all three reaction products were combined together. Four fragments (1–4, Supplementary Note 5, Fig. S2) and five drug molecules (5–9, Supplementary Note 5, Fig. S2) from a chemically diverse LSF informer library⁴¹, and 18 fragments (26–43) from the Roche compound library were screened under these reaction conditions. The collected data resulted in a balanced experimental data set comprising 691 reactions, with 379 classified as successful and 312 as unsuccessful.

Machine learning-based in silico reaction screening. GNN models (Fig. 3A) were trained using an initial dataset of 621

Minisci reactions, comprising 368 generated as decoys, 45 from the literature, and 207 from the LSF informer library. These models enabled in silico reaction screening of a Roche in-house library of 3180 advanced heterocyclic building blocks. Each substrate was assigned an ensemble score, which was determined by aggregating the predictions from six independent models. Specifically, this ensemble score incorporated inputs from three models for binary reaction outcome prediction and three models for reaction yield prediction (“Graph neural network architecture”). Subsequently, the molecules were grouped into eight clusters using agglomerative compound clustering (Supplementary Note 2). Two compound clusters were excluded from consideration due to the prevalence of unsuitable structures, namely heterocycles lacking free C-H bonds, for the studied reaction. From the six remaining clusters, three molecules were chosen from each, based on their computed reactivity score, resulting in a total of 18 *N*-heteroarenes.

The selected 18 *N*-heteroarenes were subjected to automated HTE screening, generating an experimental data set of 414 reaction points. For each of the selected substrates, Minisci-type alkylation products could be identified, resulting in a total of 276 successful reactions (Fig. 3C). Among the screened *N*-heteroarenes, 10 of them facilitated between 17 and 23 successful transformations across the chosen carboxylic acids. (Fig. 3D). 7 *N*-heteroarenes allowed 10–17 successful transformations. For one substrate, specifically the meta-substituted pyridine 42 (Fig. 4), fewer than ten successful reactions were observed (Fig. 3D). Hence, for 17 out of the 18 chosen *N*-heteroarenes, a wide variety of successful Minisci-type alkylation products were identified, resulting in a 94% success rate for substrate selection.

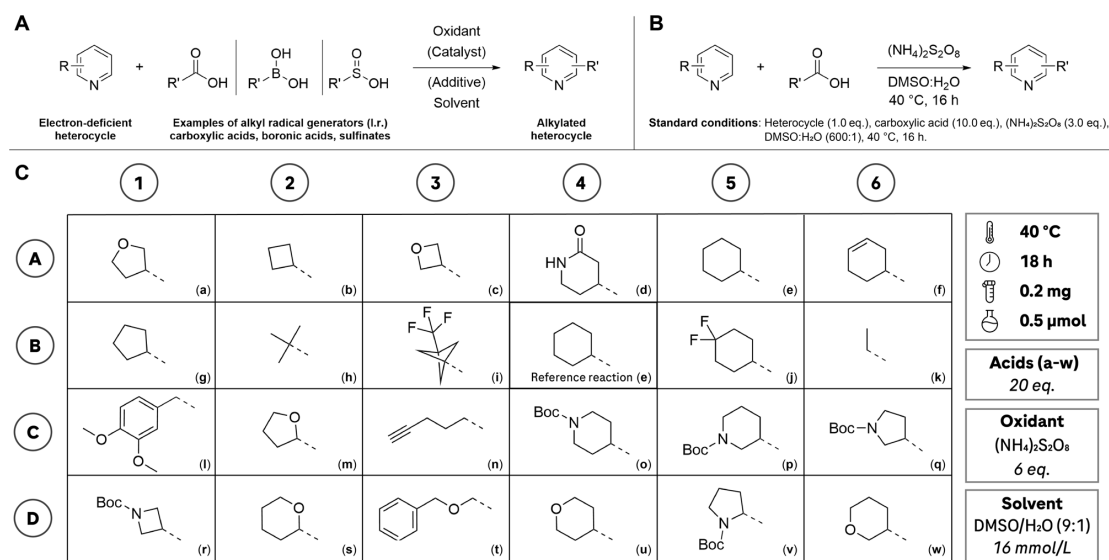


Fig. 2 Overview of Minisci-type reactions and screening plate. **A** General reaction scheme of a Minisci-type alkylation reaction. An alkyl substituent obtained from a radical generator, e.g., through decarboxylation of the carboxylic acid, is introduced to an electron-deficient heterocycle, often a pyridine. Depending on the development scope and applied technology, a variety of oxidants, catalysts, additives and solvents are used. **B** Schematic overview of the Minisci-type reaction reported by Sutherland et al.¹⁶, including the equivalents of the components. **C** Reaction screening plate used in this study. This setup allows to assess the coupling performance of a molecule of interest with 23 different alkyl carboxylic acids (**a-w**) that are relevant to medicinal chemistry applications. This configuration enables the evaluation of how well a molecule of interest couples with 23 distinct alkyl carboxylic acids (labeled as **a-w**), which are pertinent to medicinal chemistry applications. Condition B4 served as a reference reaction, ensuring consistent performance under the applied conditions. On all screening plates, B4 comprised starting material **1** and carboxylic acid **e**, providing a quality control mechanism for the generated data. If B4 had not yielded the expected outcome, the entire plate would have been reprocessed. The reaction conditions were adjusted to allow miniaturized parallel reaction screening on a nanomolar scale (0.5 μmol). Boc *tert*-Butyloxycarbonyl, DMSO dimethylsulfoxide.

However, it is worth noting that there were three five-membered *N*-heterocyclic ring systems (**2**, **4**, **9**) in the LSF informer library, for which very low reaction yields ($\leq 4\%$, averaged over 23 carboxylic acids) were observed.

To evaluate the overall performance of the GNN models that were trained on the complete experimental data set comprising 691 Minisci reactions obtained via high-throughput experimentation (207 from the LSF informer library and 414 from the virtual reaction screening), these models underwent validation for predicting reaction yield and binary reaction outcomes. This validation was conducted using a random data set split. The reaction yields were predicted with a mean absolute error (MAE) of 18.7 (± 0.2)% and a Pearson correlation coefficient (r) of 0.687 (± 0.006) (Fig. 3E). Reaction yields were categorized into four ranges: no reaction ($< 1\%$ yield), poor ($> 1-11\%$), medium ($> 11-35\%$), and high reaction yield ($> 35-100\%$). The model predicted the correct category in 55.7 (± 0.7)% of the cases. Binary reaction outcomes were predicted with an absolute accuracy of 81 (± 1), and an *F*-score of 82.7 (± 0.6)% (Fig. 3F). The failed machine learning predictions with an MAE $\geq 70\%$ (i.e., outliers) are illustrated and discussed in Supplementary Note 11 and Table S3.

Scale-up. Selected screening conditions were used for upscaling to the milligram range. LSF alkylation was carried out for the drug molecules Loratadine (**7**) and Nevirapine (**8**), and structurally complex molecular fragments. In total, 30 novel molecules were synthesized, isolated, and characterized by nuclear magnetic resonance (NMR) spectroscopy and high-resolution mass spectrometry (HRMS) (Fig. 5).

For Loratadine (**7**), a molecule from the LSF informer library, several analogs with different cyclic (**7b1**, **7b2**, **7b3**, **7j1**, **7j2**, **7e1**, **7e2**) and heterocyclic (**7s**, **7q1**, **7q2**) substituents were generated. Structurally complex scaffolds with high relevance for medicinal chemistry projects, which could serve as starting points for the development of SAR studies, also provided a variety of compelling alkylation products. Different alkyl groups, covering alkyl chains (e.g., **40h**, **33h**, **28h**), cyclic alkyls (e.g., **26e**, **41e**, **38e**) and cyclic ethers (e.g., **39u**, **35m**) could be introduced. In general, the observed regiochemistry was consistent with Minisci guidelines, with the alkyl groups being introduced in either the ortho- or para-position on the pyridine core²³. For molecule **38**, different reactivity was observed with the cyclohexyl radical reacting exclusively with the thiocarbonyl functionality affording thioether **38e**. No reaction at the pyridine core was observed.

Reactivity trends. Examination of the produced data unveiled a diverse range of observed reaction yields for both the carboxylic acids and the *N*-heteroarenes. Cyclic ethers (e.g., **u**, **s**, **a**) and alkanes (e.g., **b**, **e**, **g**) were reliably converted to the desired alkylation product, whereas cyclic boc-protected amines (e.g., **o**, **p**, **q**, **r**) and amides (**d**) resulted in low yields of the respective desired reaction products (Fig. 3B). Similarly, substituted pyridines (e.g., **30**, **31**, **36**, **39**; see Fig. 4) had lower yields compared to compounds lacking a meta-substituent (e.g., **26**, **32**, **38**, **41**; see Fig. 4). Electron-rich meta-substituted pyridines, such as **3** and **27**, had a comparably low average reaction yield compared to their less electron-rich analogs. Overall, compared to their six-membered *N*-hetero analogs, five-membered *N*-heterocyclic ring systems (e.g., **2**, **4**, **9**; see Supplementary Note 5, Fig. S2) did not show meaningful conversion to the desired alkylation product.

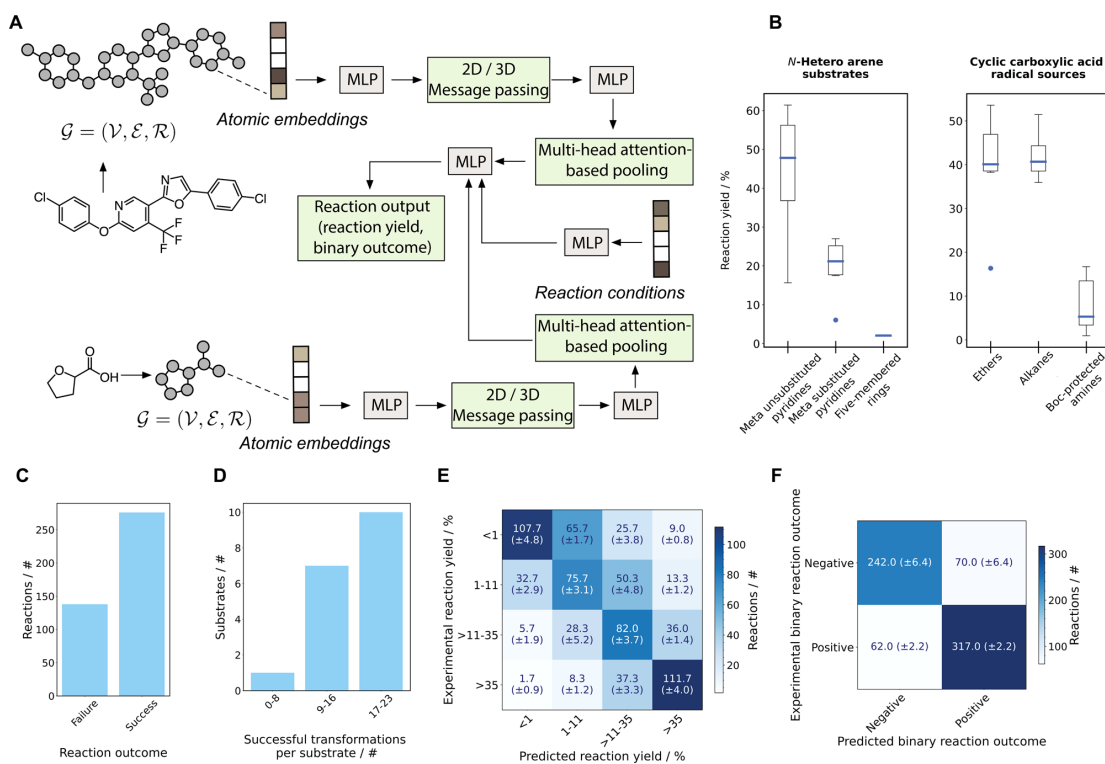


Fig. 3 Machine learning and in silico reaction screening results. **A** Schematic of the graph neural networks (GNNs) implemented within the geometric deep learning platform. Multi-layer Perceptron (MLP) modules are highlighted in gray, and the variable modules (2D/3D convolution), pooling, and outputs are highlighted in green. **B** Box plot illustrating trends observed for *N*-hetero arene (left) and carboxylic acids (right). *N*-hetero arenes: Meta-unsubstituted pyridines are observed with a reaction yield of $44 \pm 15\%$, meta-substituted pyridines with $20 \pm 6\%$ (including **27** as an outlier observed at 6%), and five-membered *N*-heterocyclic ring systems with $2 \pm 1\%$. Carboxylic acids: Cyclic ethers are observed with a reaction yield of $40 \pm 12\%$, (including **c** as an outlier observed at 16%), cyclic alkanes with $42 \pm 6\%$, and Boc-protected amines with $8 \pm 6\%$. The error bars on both box plots represent 95% confidence intervals, the bottom and top of the box are the 25th and 75th percentiles, the line inside the box is the 50th percentile (median), and any outliers are shown as open circles. **C** Bar plot illustrating the number of successful and failed reactions from HTE. The substrates selected by the model resulted in 276 successful reaction outcomes. **D** Bar plot illustrating the number of unique alkylation opportunities identified per substrate. The majority of *N*-hetero arenes (10/17) allowed for successful transformation with 17–23 carboxylic acids. **E** Confusion matrix for reaction yield prediction. Reaction yields are divided into four bins, namely, no reaction ($\leq 1\%$), poor ($>1-11\%$), medium ($>11-35\%$), and high reaction yield ($>35\%$). The model accurately predicts 54.6 ($\pm 0.9\%$)% of the reactions into the accurate bin, achieves a mean absolute error (MAE) of 18.7 ($\pm 0.2\%$)% and a Pearson correlation coefficient (r) of 0.687 (± 0.006). **F** Confusion matrix for binary reaction outcome prediction achieving an absolute accuracy of 80.8 (± 1.2) and an F -score of 82.7 (± 0.6)%.

Discussion

The Minisci reaction conditions, utilizing ammonium persulfate ($(\text{NH}_4)_2\text{S}_2\text{O}_8$) as the oxidizing agent and dimethyl sulfoxide (DMSO) as the solvent at a temperature of 40 °C, were effectively downsized and adapted into a parallel screening format. This format allowed for the efficient and resourceful execution of the reaction with a diverse range of alkyl carboxylic acids. The refined reaction protocol facilitates rapid, metal-free, and resource-efficient assessment of reaction conditions in an HTE-compatible format, aiding in informed choices for subsequent synthesis steps. Importantly, it eliminates the need for time-consuming individual reactions conducted on a milligram scale. Nonetheless, this setup has inherent limitations that merit attention in future research:

- (i) The current plate design focuses on a single set of reaction conditions for the sake of simplicity. However, examining additional oxidants or solvents, along with adjusting the equivalents of reaction components, holds the potential to deliver further enhancements in reaction yields. Moreover,

Minisci-type reactions typically involve metal catalysis, such as with silver or iron¹⁰. A systematic HTE exploration of various metal salts could lead to the discovery of even more optimized conditions.

- (ii) Instead of relying exclusively on carboxylic acids as the source of alkyl radicals, alternative radical precursors like boronic acids or sulfonates could be investigated¹³. This exploration might broaden the range of alkyl groups accessible for medicinal chemistry.
- (iii) Several photochemical Minisci-type transformations have been reported¹³. These reactions offer alternative mechanisms for radical generation that could further expand the possibilities for late-stage functionalization (LSF).

Addressing these points in future research could enhance the utility and scope of the Minisci reaction protocol.

The adoption of the user-friendly reaction data format (SURF⁴¹), facilitated the collection of reaction data from literature sources and enabled standardized reporting of results from HTE and virtual reaction screening. Sharing reaction data in a

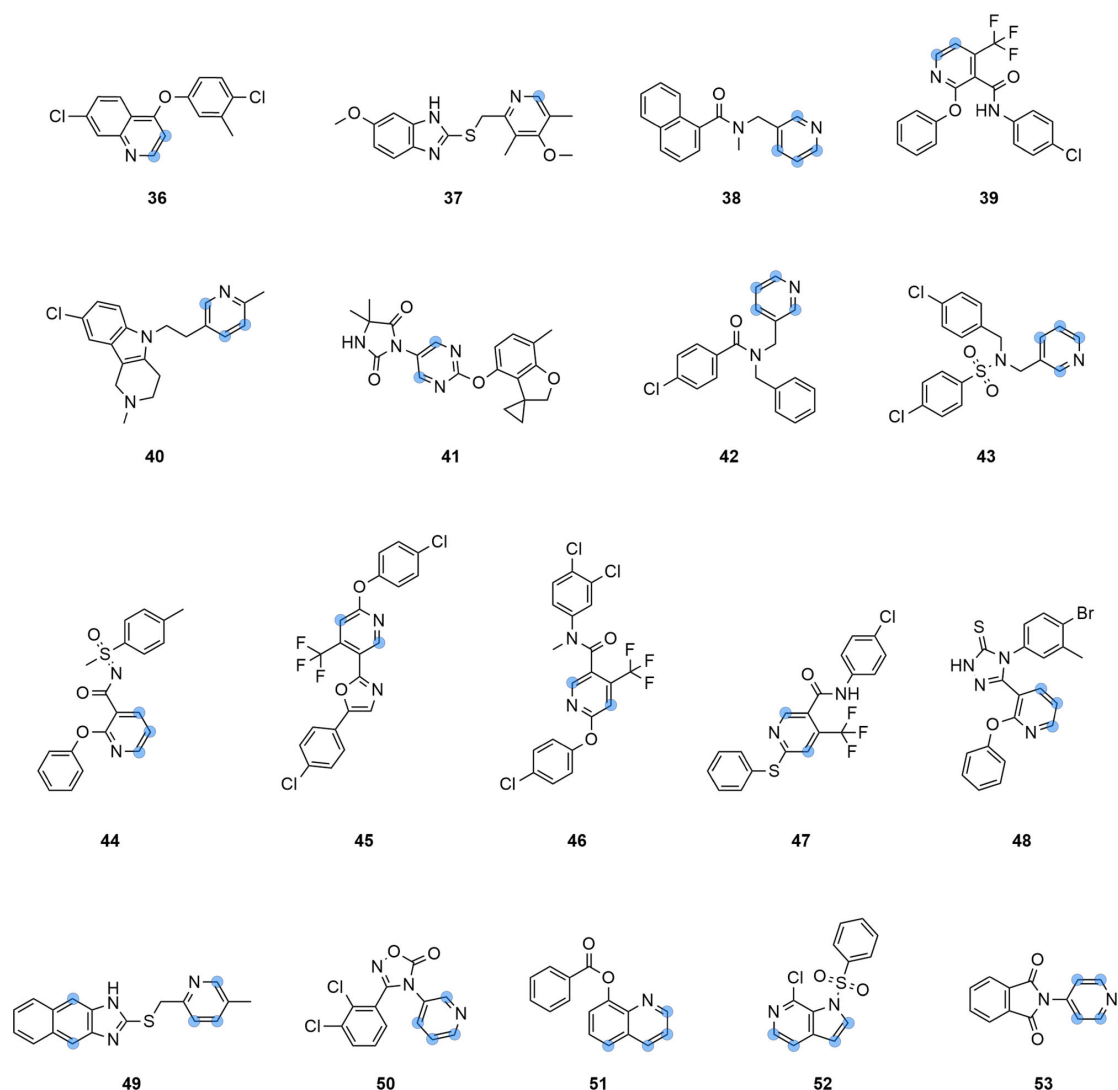


Fig. 4 Overview of selected substrates suggested by the *in silico* prediction model. Structures of the 18 selected substrates **36–53** that were suggested by the graph neural networks as suitable for Minisci-type alkylation and underwent subsequent screening to identify novel starting points. Potential, not confirmed, carbon reaction centers are marked with a blue dot.

standardized format plays a pivotal role in the effective utilization of machine learning models for predicting chemical reactivity^{44,45}. By using SURF, the initial reaction data from three distinct sources (45 from literature, 207 from experiments, and 368 decoy reactions) became readily available for machine learning, obviating the need for manual data curation. Since both the experimental and, particularly, the literature data are predominantly comprised of positive results, incorporating decoy data from unsuccessful transformations played a crucial role in constructing a dependable prediction model.

A detailed look at the experimental data revealed that cyclic Boc-protected amines (**o**, **p**, **q**, **r**, **v**), as well as amides (e.g., **d**) mainly afforded low yields (5–20%) of the desired reaction products (Supplementary Note 10, Fig. S10). This observation reflects the half-lives of the generated radical intermediates⁴⁶, e.g., with tertiary carbon radicals (e.g., **h**) having higher stability than

primary carbon radicals (e.g., **k**) and the latter thus resulting in lower product yields. Another experimental trend relates to the substitution pattern of *N*-heteroarenes. Meta-unsubstituted pyridines (e.g., **26**, **32**, **41**) consistently provided higher yields than substituted analogs, (e.g., **35**, **36**, **37**) as residues on the meta-position sterically hinder the reaction in ortho- and para-positions to the pyridine (Supplementary Note 10, Fig. S11). Finally, electron-rich meta-substituted pyridines, such as **3** and **27**, had a very low (5–10%) average reaction yield on the screening plate when compared to their less electron-rich analogs (Supplementary Note 10, Fig. S10). This low reactivity is owed to the electron-rich amine- and methoxy-substituents, respectively²³.

In contrast to a prior study⁴¹ where GNNs processed a single graph input, the GNN model outlined in this research accommodates two distinct molecular inputs, corresponding to the two

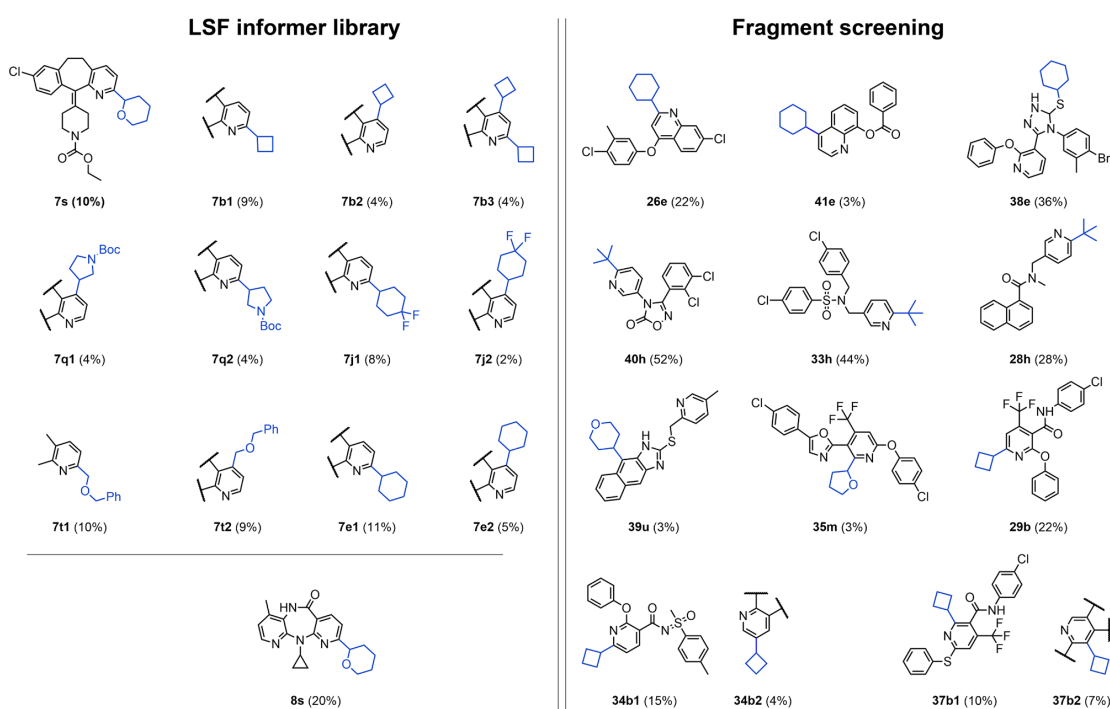


Fig. 5 Selected examples of characterized Minisci reaction products. The left panel shows examples from the LSF drug informer library and the right panel from the fragment screening. The added alkyl groups are highlighted in blue. Late-stage drug alkylation examples include derivatives of the drugs Loratadine (**7s**, **7b1**, **7b2**, **7b3**, **7q1**, **7q2**, **7j1**, **7j2**, **7t1**, **7t2**, **7e1**, **7e2**) and Nevirapine (**8s**). Fragment screening highlights the diverse range of introduced substituents, covering cyclohexanes (**26e**, **41e**, **38e**), cyclobutanes in different positions (**29b**, **34b1**, **34b2**, **37b1**, **37b2**), heterocyclic alkanes (**39u**, **35m**) and *tert*-butyl (**40h**, **33h**, **28h**). Boc *tert*-Butyloxycarbonyl, Ph Phenyl.

reactants (*N*-heteroarenes and carboxylic acids). The network architecture was tailored to the Minisci-type alkylation transformation in such a way that trained GNNs can be applied to novel *N*-heteroarenes as well as carboxylic acids. Therefore, the model can be used for in silico molecular library screening for both types of reaction inputs. It could be shown that in silico reaction screening using GNN models trained on a comparably small preliminary data set consisting of 576 Minisci reactions (i.e., 368 from decoy generation, 45 from literature, and 207 LSF from an informer library) led to the identification of 17 substrates (i.e., 94% of the 18 selected molecules). All newly identified substrates were successfully alkylated with a broad range of at least 10 different carboxylic acids. Furthermore, in total 276 successful reactions (i.e., producing alkylation products with a median yield of 26%) were identified. The low reaction yields observed for three five-membered *N*-heterocyclic ring systems (**2**, **4**, **9**) indicate that the GNN models learned to de-prioritize five-membered *N*-heteroarenes during in silico reaction screening. It was shown how a clustering approach can be combined with in silico reaction screening to assess structural diversity as well as reactivity. As previously reported⁴¹, the inclusion of partial charges did not yield improved model performance (Supplementary Note 3). This observation, in particular, led to the decision to prospectively apply GTNN models that are trained on 3D molecular graphs without electronic features. Further investigations involving more specific electronic features, such as transition state energies, could offer deeper insights into the relevance of quantum chemical attributes in machine learning for reaction prediction, as demonstrated in a recent study⁴³. Moreover, the introduced GNNs could be further advanced to facilitate regioselectivity

prediction or the prediction of multiple output properties. For instance, this could encompass predicting the proportions of mono- and di-alkylation.

With the overall goal of synthesizing novel scaffolds that are relevant to medicinal chemistry, the visualized screening data served to identify appropriate reaction conditions for upscaling to the milligram scale. Again, the SURF data format was instrumental for the laboratory chemist to set up experiments efficiently by providing the CAS number, SMILES string, equivalents, and overall reaction conditions in a comprehensive and easily accessible format. The reaction conditions were reproducible at a higher scale, underscoring the applicability of this approach to drug discovery. With the exception of compound **38e**, all reactions yielded C-C coupling products. In general, the observed regioselectivity was in agreement with the expected reaction products according to the rules reported in the literature²³.

However, when moving to more densely functionalized pyridines, these reported literature guidelines do not appear to apply. While the reaction of **34b** and **37b** primarily generated the expected *ortho*-substituted reaction products **34b1** and **37b1**, also *meta*-substituted reaction products **34b2** and **37b2** were obtained, albeit in lower amounts (Fig. 5). In the literature, amides are described as *ortho*-*para* directing groups due to their electron-withdrawing effect, and aryl ethers as *ortho*-activating moieties due to their electron-donating nature²³. The formation of regioisomer **34b2** might have been sterically hindered by the amidyl side chain, favouring the *meta*- over the *para* position. For **37b2**, an explanation of the formation could lie in the several different functional groups that are attached to the pyridine ring, which only leave the *meta* position available for substitution,

despite this position being sterically hindered by the proximity of the aryl sulfide and the CF₃ group. Lastly, **38e** showed different reactivity despite bearing a pyridine moiety. This observed reaction product can be rationalized by the greater reactivity of the lone pairs of the sulfur as compared to the C-H bonds of the pyridine side-chain. These results of the scale-up reactions underscore the importance of generating high-quality, single-batch LSF reaction data.

For the continued development of this method further exploration of Minisci-type reaction conditions is warranted, including the variation of oxidation reagents, solvents, and the incorporation of techniques like photoredox catalysis and electrochemistry⁴⁷. Also, the source of the alkyl radical precursor could be diversified, leading to an expansion of the scope for alkyl groups. Additionally, the substrate scope could be expanded to include other electron-deficient heterocyclic systems, particularly five-membered heterocycles, as they are commonly found motifs in drug-like molecules. With these possibilities in mind, the results of this study emphasize the feasibility and benefits of combining laboratory automation, parallel miniaturized screening, and machine learning to enhance the efficiency and cost-effectiveness of synthesis in drug discovery. This integrated approach is currently being effectively employed at Roche. The predictive capabilities of the computational model will be continuously enhanced by supplying the algorithm with a growing data set of newly generated LSF reaction data points that encompass the pertinent medicinal chemistry landscape.

Methods

Literature analysis. A systematic analysis of chemical transformations was carried out to determine the most feasible conditions for reaction miniaturization and parallel screening. Initially, 45 publications covering different Minisci-type alkylation reactions were selected. Most of those methods rely on photo- or electro-chemistry. Although it has been demonstrated that these approaches are amenable to HTE^{48,49}, carrying out these reaction processes requires specialized equipment that is not readily available in every laboratory. Therefore, with the goal of enabling widespread use in medicinal chemistry, publications were scrutinized for a rapid, resilient, and easily customizable procedure. Sutherland et al.¹⁶ reported a Minisci methodology that fulfilled those criteria. This transformation can be executed without the necessity for additional metals and catalysts, and it can accommodate a variety of alkyl carboxylic acids that do not demand pre-functionalization. This adaptability allows for the creation of customized templates tailored to specific project requirements. Consequently, the reaction data were manually curated and standardized in a simple user-friendly reaction format (SURF, for details, refer to Supplementary Note 9). These SURF data were used as literature data set herein. All details of the literature analysis (Supplementary Note 4) and the resulting data set in SURF are available as supplementary information (Supplementary Note 4).

Screening plate design and testing. The screening plate was designed around the literature data obtained from Sutherland et al.¹⁶, which showed good yields on average (60%) for a variety of carboxylic acid coupling partners. Aiming at assessing the reactivity of a substrate with a variety of different alkyl groups (rings and chains), a screening plate with 23 different alkyl carboxylic acids was assembled. The carboxylic acids scope from the original publication¹⁶ covering n-alkyl (e.g., **h**, **k**, depicted in Fig. 2), cyclic alkanes (e.g., **e**, **g**) and O-heterocyclic fragments (e.g., **m**, **u**) was complemented by sp³-rich N-heterocyclic carboxylic acids with relevance to drug discovery projects (**o**, **p**, **q**, **r**).

The reactions were miniaturized to 0.5 μmol scale, downsizing by a factor of 300 compared to the literature procedure¹⁶. To achieve this small reaction scale, stock solutions of all components in the reaction solvent (DMSO) were produced. Consequently, the designed screening plate only requires 4.2 mg of starting material (molar mass: 350 Da) to assess 23 different transformations. In comparison, single reactions in reference¹⁶ were carried out with 52.5 mg of starting material. Using a substrate from reference¹⁶ (Molecule **1**, structure depicted in Fig. S2 in Supplementary Note 5), different oxidant to carboxylic acid ratios (3:10, 6:10, 3:20, 6:20) were tested to identify the more favorable screening condition (higher conversion). Further, the impact of other parameters, such as the atmosphere (under air, under nitrogen in a glovebox), and the reaction concentration (2, 16 mmol/L) was investigated. Upon determining the highest-yielding reaction parameters, the best-performing condition on the plate (B4, **1** with **e**, under nitrogen, 16 mmol/L) was used as the reference reaction to monitor reproducibility across different plates. Incorporating the control experiment in position B4, which consistently remained unchanged, served the purpose of swiftly detecting potential handling errors with the plate and confirming the reliability of the generated data. The plate layout including all reaction parameters is shown in Fig. 2. Additional information on the plate testing is provided as supporting information (Supplementary Note 6, Figs. S7–S9).

LSF informer library. For the generation of the experimental reaction dataset, the previously published informer library was used as a starting point (see ref. ⁴¹ for details). From this collection, three fragments (2–4, Fig. S2 in Supplementary Note 5 for structures) and five drug molecules (5–9, Figure S2 in Supplementary Note 5) were screened. The drug molecule library in ref. ⁴¹ was assembled based on clustering of 1174 approved small molecule drugs into eight structurally diverse subsets. As three clusters did not contain any reactive functional groups required for Minisci-type reactions (e.g., electron-deficient heterocycle), only five drug molecules (5–9) were subjected to HTE alkylation screening (see “HTE alkylation screening” for details). The screening of the drugs was extended by three fragments (2–4) from ref. ⁴¹. Furthermore, a decoy data set containing 368 unsuccessful reaction examples was generated. The chemical structures of the eight N-hetero-arene substrates (2–9, Fig. S2) as well as the 16 decoy substrates (10–25, Fig. S3) used to train the machine learning are provided as supporting information (Supplementary Note 5).

To assess the performance, i.e., the prediction accuracy, of the developed machine learning model on relevant fragments for applications in medicinal chemistry, a substructure search for heteroaromatic ring systems containing at least one nitrogen atom was carried out in the Roche corporate compound collection. The resulting compounds were retained if (i) there was at least 1 g of powder stock available, and (ii) the structures were not used in any internal project or subject to legal restrictions. This pool of candidates was then clustered using sphere exclusion clustering⁵⁰ on ECFP4 fingerprints⁵¹ with a Tanimoto cutoff⁵² of 0.6. Based on the clustering results, we manually selected 18 structurally diverse fragments (26–43, Fig. 4, Supplementary Note 2, Fig. S1).

HTE alkylation screening. Using the 24-well plate design (Fig. 2, Supplementary Note 6), selected drug molecules and fragments from the LSF informer library (2–9, Supplementary Note 5, Fig. S2), a set of relevant building blocks (26–43, Fig. 4, for detailed information: Supplementary Note 5, Figs. S4, S5) and substrates from Sutherland et al.¹⁶ (44–48, Supplementary Note 5,

Fig. S6) were screened. The reaction setup (stock solution, liquid handling) and execution (heating, stirring) in glass vials on a parallel screening plate were conducted in a glovebox under nitrogen. Upon completion of the reactions, the residues were diluted in MeCN/H₂O to a defined concentration suitable for LCMS analysis, using a liquid handler. The resulting mixtures were analyzed by LCMS, and the results were subjected to automated reaction data analysis (Supplementary Note 8) for the determination of the molecular components. Standardized data output (Supplementary Note 9) allowed for direct visualization of the information in TIBCO Spotfire (Somerville, USA). The general screening procedure, including detailed information on the hardware and software utilized, is provided as Supporting Information (Supplementary Note 7).

Scale up reactions. Analysis of the screening results revealed that the drugs Loratadine (7), Nevirapine (8), and 11 fragments (26, 28, 29, 33-35, 37-41) were alkylated with different types of alkyl fragments. From this subset, conditions showing reasonable conversion (>40%, based on UV trace) were subjected to upscaling. Reactions were conducted under nitrogen in a glovebox, in glass reaction vessels with pressure release caps and standard stirring bars. Purification was performed by flash chromatography or reversed-phase high-pressure liquid chromatography (RP-HPLC). Structural elucidation was performed with NMR spectroscopy and HRMS. All comprehensive experimental details for the scale-up processes, including analytical outcomes and spectra of the purified and fully characterized compounds, can be found in the Supporting Information (Supplementary Note 12 and Supplementary Data 1, Figs. S12–S29).

Graph neural network architecture. A graph transformer neural network (GTNN) architecture was employed based on the E(3) equivariant graph neural network architecture⁵³, which has seen use in several related applications^{54,55}. The GTNN was designed using the same training procedure as in reference⁴¹ and a slightly adapted architecture that allows for two distinct and variable molecular graphs in its input, i.e., *N*-hetero arenes and carboxylic acids (Supplementary Note 1). Furthermore, the initial machine learning framework was extended to allow for prospective screening of individual substrates, carboxylic acids or single reactions. For both molecular graphs, their 3D conformers were calculated using the universal force field method⁵⁶, and the graph was constructed using nodes represented by atoms and edges defined by all neighboring atoms within a radius of 4 Å of each atom.

Atoms were featured using embeddings of four atom-level features:

- 12 atom types (H, C, N, O, F, P, S, Cl, Br, I, Si, Se);
- 2 ring types (True, False);
- 2 aromaticity types (True, False);
- 4 hybridization types (sp³, sp², sp, s).

First, the individual atomic embedding was concatenated and transformed into an initial atomic representation \mathbf{h}_i^0 via a multi-layer perceptron (MLP). Atomic representations \mathbf{h}_i^0 were subsequently transformed via three message-passing layers. In each message-passing layer, the atomic representations were transformed via Eq. (1)

$$\mathbf{h}_i^{l+1} = \phi \left(\mathbf{h}_i^l, \sum_{j \in \mathcal{N}(i)} \psi(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{r}_{i,j}, \dots) \right), \quad (1)$$

where \mathbf{h}_i^l is the atomic representation of the *i*-th atom at the *l*-th layer; $j \in \mathcal{N}(i)$ is the set of neighboring nodes connected via

edges; $\mathbf{r}_{i,j}$ the inter-atomic distance represented in terms of Fourier features, using a sine- and cosine-based encoding; ψ is an MLP transforming node features into message features \mathbf{m}_{ij} ; $\mathbf{m}_{ij} = \psi(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{r}_{i,j})$ for 3D graphs, and $\mathbf{m}_{ij} = \psi(\mathbf{h}_i^l, \mathbf{h}_j^l)$ for 2D graphs; Σ denotes the permutation-invariant pooling Operator (i.e., sum) transforming \mathbf{m}_{ij} into \mathbf{m}_i ; $\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}$; and ϕ is an MLP transforming \mathbf{h}_i^l and \mathbf{m}_i into \mathbf{h}_i^{l+1} . The resulting atomic features from all layers [$\mathbf{h}_i^{l=1}, \mathbf{h}_i^{l=2}, \mathbf{h}_i^{l=3}$] were concatenated and transformed via an MLP, resulting in final atomic features. Atomic features were then pooled via a graph multiset transformer (GMT)⁵⁷ with four attention heads yielding an overall molecular feature vector.

This procedure was conducted for both input molecular graphs, where no weights were shared between the two GNN modules except for the initial embedding layers of atom-level representations. The pooled molecular representations were then concatenated to a learned representation of the reaction conditions (Fig. 3B). This subsequent reaction representation was further transformed via a final MLP converting the latent space to the desired reaction output. Both of the examined problems, namely, reaction yield prediction and binary reaction outcome prediction, were addressed as regression tasks. The output for reaction yield was defined within the range of floating values from 0 to 1, whereas for binary reaction outcomes, it was defined as either 0 or 1.

Consistent with the results outlined in ref. 41, the performance of the models was validated for GNNs trained on molecular graphs that included atomic partial charges⁵⁸⁻⁶⁰. This evaluation revealed that there was no substantial improvement or decline in model performance. Consequently, for all the applications described, 3D graphs without electronic features were employed (Supplementary Note 3, Tables S1, S2).

Reaction condition representation. Reaction conditions were represented by one-hot-encoding for molecular entities, i.e., reagents, solvents, catalysts, additives and atmosphere, and by real numbers for scalars, i.e., equivalents for starting materials, reagents, carboxylic acids, catalysts, and additives, fractions for the solvents, temperature (°C), reaction time (h), and scale (mmol/L). The individual conditions were concatenated with each other and transformed via an MLP. This reaction condition representation was then concatenated to the learned representations of the two substrates, i.e., *N*-hetero arene and carboxylic acid.

Number of hyperparameters. The feature dimension for the internal representation of GTNN was established at 128, with the exception of the embedding dimension for the reaction and atomic properties, which was set to 64. Additionally, the first MLP layer following the graph multiset transformer-based pooling was configured to have 256 dimensions. The graph multiset transformer employed two attention heads for pooling. These parameter settings translated into neural network sizes with ~2.0 million trainable parameters for GTNN.

Metric for model validation. For model validation and optimization, mean absolute error was used for reaction yield prediction. For predicting binary reaction outcomes the models were validated using absolute accuracy and the *F*-score metric. The *F*-score (*F*₁) is used as a measure for unbalanced data sets and is calculated by the mean of precision and recall (Eq. (2)):

$$F_1 = \frac{2tp}{(2tp + fp + fn)} \quad (2)$$

where *tp* represents true positives, *fp* false positives, and *fn* false negatives.

Decoy data set. The decoy data set comprised 308 instances of unsuccessful reactions, derived from 16 substrates that lack reactivity under Minisci-type conditions due to the absence of an aromatic or heteroaromatic component in their starting materials. These selected molecules underwent thorough scrutiny by experts and were subsequently incorporated into the data set as instances of negative or unsuccessful reaction outcomes. This inclusion serves to furnish the model with knowledge about molecules that do not exhibit reactivity when subjected to Minisci conditions (Supplementary Note 5, Fig. S3).

Substrate selection. The selection of a diverse and reactive set of *N*-hetero arenes was based on a Roche-internal library of 3180 advanced heterocyclic building blocks with a molecular weight between 200 and 1000 g/mol. Aiming to check these compounds for potential reactivity in the alkylation reaction, this library was virtually screened with preliminarily trained GNN models. Each of the *N* = 3180 molecules was assigned with an average score value calculated with six independent GNNs (“Machine learning-based in silico reaction screening” for details). Subsequently, agglomerative compound clustering was performed⁶¹. The molecules were encoded as an *N* × *N* similarity matrix containing pairwise Jaccard similarity values based on ECFP4 molecular fingerprint descriptors⁵¹. Clustering resulted in eight clusters of which six were used for substrate selection. Three top-scoring compounds were selected for HTE reaction screening for each of the six clusters. This clustering approach was chosen to allow for the selection of chemically diverse reactive substrates.

In silico reaction screening. For model application, a total of six GNNs were trained. Three models were trained for predicting reaction yield, and three models were trained for binary reaction outcome prediction. These models were then utilized to predict the reaction outcomes and reaction yields for each combination of the 3180 advanced heterocyclic building blocks and the 23 carboxylic acids. The predictions yielded values for both binary reaction outcomes and reaction yields, each ranging from 0 to 1. Given that three models were employed for each of the two prediction values, mean and standard deviations were computed to provide an understanding of the model’s uncertainty. The final score was then determined as the mean of the two predictions. Subsequently, each of the six molecule clusters was ranked based on the calculated score, and molecules from the upper echelons of the list were chosen for further consideration or selection.

Data availability

The SURF-formatted literature, experimental and decoy data sets containing 45, 691 and 368 reactions, respectively, are enclosed as TSV files as Supplementary Data 2–8. Description of Supplementary Data: Supplementary Data 1: PDF file containing NMR spectra. Supplementary Data 2: TSV file containing all reactions (i.e., literature, decoy and experimental data). Supplementary Data 3: TSV file containing reactions from literature. Supplementary Data 4: TSV file containing experimental reaction data. Supplementary Data 5: TSV file containing reactions conducted to validate the literature data. These reactions were excluded in machine learning model training. Supplementary Data 6: TSV file containing decoy reactions. Supplementary Data 7: TSV file containing all investigated carboxylic acids. Supplementary Data 8: TSV file containing all investigated *N*-hetero arenes.

Code availability

A reference implementation of the geometric machine learning platform based on PyTorch⁶² and PyTorch Geometric⁶³ is available at <https://github.com/ETHmodlab/minisci> (rep. DOI: 10.5281/zenodo.8344587, <https://zenodo.org/record/8344587>).

Received: 8 June 2023; Accepted: 30 October 2023;
Published online: 20 November 2023

References

- Blakemore, D. C. et al. Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10**, 383–394 (2018).
- Wencel-Delord, J. & Glorius, F. C–H bond activation enables the rapid construction and late-stage diversification of functional molecules. *Nat. Chem.* **5**, 369–375 (2013).
- Isert, C., Kromann, J. C., Stiefl, N., Schneider, G. & Lewis, R. A. Machine learning for fast, quantum mechanics-based approximation of drug lipophilicity. *ACS Omega* **8**, 2046–2056 (2023).
- Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist’s toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016).
- Guillemand, L., Kaplaneris, N., Ackermann, L. & Johansson, M. J. Late-stage c–h functionalization offers new opportunities in drug discovery. *Nat. Rev. Chem.* **5**, 522–545 (2021).
- Nippa, D. F. et al. Late-stage functionalization and its impact on modern drug discovery: medicinal chemistry and chemical biology highlights. *Chimia* **76**, 258–258 (2022).
- Dong, Z., Ren, Z., Thompson, S. J., Xu, Y. & Dong, G. Transition-metal-catalyzed c–h alkylation using alkenes. *Chem. Rev.* **117**, 9333–9403 (2017).
- Minisci, F., Bernardi, R., Bertini, F., Galli, R., Perchinummo, M. *Tetrahedron* **27**, 3575–3579 (1971).
- Fontana, F., Minisci, F., Nogueira Barbosa, M. C. & Vismara, E. Homolytic acylation of protonated pyridines and pyrazines with α -keto acids: the problem of monoacylation. *J. Org. Chem.* **56**, 2866–2869 (1991).
- Dunston, M. A. Minisci reactions: versatile ch-functionalizations for medicinal chemists. *MedChemComm.* **2**, 1135–1161 (2011).
- Minisci, F., Bernardi, R., Bertini, F., Galli, R. & Perchinummo, M. Nucleophilic character of alkyl radicals—vi: a new convenient selective alkylation of heteroaromatic bases. *Tetrahedron* **27**, 3575–3579 (1971).
- Minisci, F., Galli, R., Cecere, M., Malatesta, V. & Caronna, T. Nucleophilic character of alkyl radicals: new syntheses by alkyl radicals generated in redox processes. *Tetrahedron Lett.* **9**, 5609–5612 (1968).
- Proctor, R. S. & Phipps, R. J. Recent advances in minisci-type reactions. *Angew. Chem. Int. Ed.* **58**, 13666–13699 (2019).
- Smith, J. M., Dixon, J. A., deGruyter, J. N. & Baran, P. S. Alkyl sulfonates: radical precursors enabling drug discovery: Miniperspective. *J. Med. Chem.* **62**, 2256–2264 (2018).
- Seiple, I. B. et al. Direct c–h arylation of electron-deficient heterocycles with arylboronic acids. *J. Am. Chem. Soc.* **132**, 13194–13196 (2010).
- Sutherland, D. R., Veguillas, M., Oates, C. L. & Lee, A.-L. Metal-, photocatalyst-, and light-free, late-stage c–h alkylation of heteroarenes and 1,4-quinones using carboxylic acids. *Org. Lett.* **20**, 6863–6867 (2018).
- Ritchie, T. J., Macdonald, S. J., Young, R. J. & Pickett, S. D. The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and-aliphatic ring types. *Drug Discov. Today* **16**, 164–171 (2011).
- Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**, 6752–6756 (2009).
- Lovering, F. Escape from flatland 2: complexity and promiscuity. *MedChemComm.* **4**, 515–519 (2013).
- Auberson, Y. P. et al. Improving nonspecific binding and solubility: bicycloalkyl groups and cubanes as para-phenyl bioisosteres. *ChemMedChem.* **12**, 590–598 (2017).
- Burkhard, J. A., Wuitschik, G., Rogers-Evans, M., Müller, K. & Carreira, E. M. Oxetanes as versatile elements in drug discovery and synthesis. *Angew. Chem. Int. Ed.* **49**, 9052–9067 (2010).
- Ishikawa, M. & Hashimoto, Y. Improvement in aqueous solubility in small molecule drug discovery programs by disruption of molecular planarity and symmetry. *J. Med. Chem.* **54**, 1539–1554 (2011).
- O’Hara, F., Blackmond, D. G. & Baran, P. S. Radical-based regioselective c–h functionalization of electron-deficient heteroarenes: scope, tunability, and predictability. *J. Am. Chem. Soc.* **135**, 12122–12134 (2013).
- Dreher, S. D., Dormer, P. G., Sandrock, D. L. & Molander, G. A. Efficient cross-coupling of secondary alkyltrifluoroborates with aryl chlorides reaction discovery using parallel microscale experimentation. *J. Am. Chem. Soc.* **130**, 9257–9259 (2008).
- Bellomo, A. et al. Rapid catalyst identification for the synthesis of the pyrimidinone core of hiv integrase inhibitors. *Angew. Chem. Int. Ed.* **124**, 7018–7021 (2012).
- Buitrago Santanilla, A. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).

27. Barhate, C. L. et al. Microscale purification in support of high-throughput medicinal chemistry. *Chem. Commun.* **57**, 11037–11040 (2021).
28. Shevlin, M. Practical high-throughput experimentation for chemists. *ACS Med. Chem. Lett.* **8**, 601–607 (2017).
29. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
30. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
31. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
32. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
33. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Process. Mag.* **34**, 18–42 (2017).
34. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3**, 1023–1032 (2021).
35. Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. *Curr. Opin. Struct. Biol.* **79**, 102548 (2023).
36. von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
37. Unke, O. T. et al. SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **12**, 7273 (2021).
38. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. *NeurIPS* **34**, 9405–9415 (2021).
39. Guan, Y. et al. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **12**, 2198–2208 (2021).
40. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. *Adv. Neural Inform. Process. Syst. (NeurIPS)* **30**, https://proceedings.neurips.cc/paper_files/paper/2017/hash/ced556cd99c08315cbe0744a3ba0f0-Abstract.html (2017).
41. Nippa, D. F. et al. Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *ChemRxiv preprint* (2022).
42. King-Smith, E. et al. Predictive minisci and p450 late-stage functionalization with transfer learning. *ChemRxiv preprint* (2022).
43. Caldeweyher, E. et al. Hybrid machine learning approach to predict the site selectivity of iridium-catalyzed arene borylation. *J. Am. Chem. Soc.* **145**, 31, 17367–17376 (2023).
44. Kearnes, S. M. et al. The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
45. Mercado, R., Kearnes, S. M. & Coley, C. W. Data sharing in chemistry: lessons learned and a case for mandating structured reaction data. *J. Chem. Inf. Model.* **63**, 4253–4265 (2023).
46. Hioe, J. & Zipse, H. Radical stability and its role in synthesis and catalysis. *Org. Biomol. Chem.* **8**, 3609–3617 (2010).
47. Bieszczad, B., Perego, L. A. & Melchiorre, P. Photochemical c-h hydroxyalkylation of quinolines and isoquinolines. *Angew. Chem. Int. Ed.* **131**, 17034–17039 (2019).
48. Buglioni, L., Raymenants, F., Slattery, A., Zondag, S. D. & Noël, T. Technological innovations in photochemistry for organic synthesis: flow chemistry, high-throughput experimentation, scale-up, and photoelectrochemistry. *Chem. Rev.* **122**, 2752–2906 (2021).
49. Wills, A. G. et al. High-throughput electrochemistry: state of the art, challenges, and perspective. *Org. Process. Res. Dev.* **25**, 2587–2600 (2021).
50. Gobbi, A., Giannetti, A. M., Chen, H. & Lee, M.-L. Atom-atom-path similarity and sphere exclusion clustering: tools for prioritizing fragment hits. *J. Cheminform.* **7**, 11 (2015).
51. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
52. Bajusz, D., Rácz, A. & Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
53. Satorras, V. G., Hoogeboom, E. & Welling, M. E (n) equivariant graph neural networks. In: *Proceedings of the International Conference on Machine Learning (ICML)* 9323–9332 (2021).
54. Isert, C., Atz, K., Riniker, S. & Schneider, G. Exploring protein-ligand binding affinity prediction with electron density-based geometric deep learning. *ChemRxiv preprint* 10.26434/chemrxiv-2023-585vf (2023).
55. Atz, K. et al. Deep interactome learning for de novo drug design. *ChemRxiv preprint* <https://doi.org/10.26434/chemrxiv-2023-cbq9k> (2023).
56. Rappé, A. K., Casewit, C. J., Colwell, K., Goddard III, W. A. & Skiff, W. M. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
57. Baek, J., Kang, M. & Hwang, S. J. Accurate learning of graph representations with graph multiset pooling. In: *Proceedings of the International Conference on Learning Representations (ICLR)* **9** (2021).
58. Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J. & Schneider, G. δ -quantum machine-learning for medicinal chemistry. *Phys. Chem. Chem. Phys.* **24**, 10775–10783 (2022).
59. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273 (2022).
60. Neeser, R., Isert, C., Stuyver, T., Schneider, G. & Coley, C. Qmugs 1.1: Quantum mechanical properties of organic compounds commonly encountered in reactivity datasets. *Chemical Data Collections*, **46**, 101040 (2023).
61. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).
62. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **32**, 8026–8037 (2019).
63. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch geometric. In: *Proceedings of the International Conference on Learning Representations (ICLR)* **7**, (2019).

Acknowledgements

This research was funded by the Swiss National Science Foundation (SNSF, grant no. 205321_182176). D.B.K. acknowledges funding from the Fonds der Chemischen Industrie (FCI) through a Liebig Fellowship and Roche Basel for funding the Ph.D. position of D.F.N. C.I. acknowledges support from the Scholarship Fund of the Swiss Chemical Industry. We thank Dr. Nicolas Zeidan for helpful discussions and proofreading.

Author contributions

To whom correspondence should be addressed: D.B.K., U.G., R.E.M. or G.S. D.F.N.: Conceptualization, methodology, experiments, formal analysis, data curation, writing—original draft. K.A.: Conceptualization, methodology, experiments, software development and validation, formal analysis, data curation, writing—original draft. A.T.M.: Methodology, software validation, writing - review and editing. J.W.: Experiments. C.I.: Software development and validation, writing—review and editing. M.B.: Experiments. O.S.: Experiments. D.B.K.: Supervision, funding acquisition, writing—review and editing. U.G.: Supervision, funding acquisition, writing—review and editing. R.E.M.: Supervision, funding acquisition, writing—review and editing. G.S.: Supervision, conceptualization, formal analysis, investigation, methodology, funding acquisition, project administration, writing—review and editing. All authors discussed the results and gave their approval of the final version.

Competing interests

G.S. declares a potential financial and non-financial conflict of interest as co-founder of inSili.com LLC, Zurich, and in his role as a scientific consultant to the pharmaceutical industry. D.F.N., K.A., A.T.M., J.W., M.B., O.S., U.G. and R.E.M. declare potential financial and non-financial conflict of interest as full employees of F. Hoffmann-La Roche Ltd. D.B.K. and C.I. declare no competing interest.

Additional information


Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42004-023-01047-5>.

Correspondence and requests for materials should be addressed to David B. Konrad, Uwe Grether, Rainer E. Martin or Gisbert Schneider.

Peer review information *Communications Chemistry* thanks the anonymous, reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

6.3 Experimental and supplementary information

The following pages contain the experimental and supplementary information supporting the results described in the publication from the previous section. Citation of the publication: **Nippa, D. F.[†]**, Atz, K.[†], Müller, A. T., Wolfard, J., Isert, C., Binder, M., Scheidegger, O., Stepan, A. F., Konrad, D. B., Grether, U., Martin, R. E., & Schneider, G., Identifying opportunities for late-stage C-H alkylation with in silico reaction screening and high-throughput experimentation *Comms. Chem.*, **6**, 256 (2023). [457] The material (DOI: 10.1038/s42004-023-01047-5) is reprinted with permission from Springer Nature Limited (Author reuse for own thesis).

Supplementary Information:
Identifying opportunities for late-stage C-H alkylation with *in silico*
reaction screening and high-throughput experimentation

David F. Nippa^{1,2,†}, Kenneth Atz^{1,†}, Alex T. Müller¹, Jens Wolfard¹, Clemens Isert³,
Martin Binder¹, Oliver Scheidegger¹, David B. Konrad^{2,*}, Uwe Grether^{1,*},
Rainer E. Martin^{1,*} & Gisbert Schneider^{3,*}

¹Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd.,
Grenzacherstrasse 124, 4070 Basel, Switzerland.

²Department of Pharmacy, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, 81377 Munich, Germany.

³ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

⁴ETH Singapore SEC Ltd, 1 CREATE Way, #06-01 CREATE Tower, Singapore, Singapore.

† These authors contributed equally to this work.

* To whom correspondence should be addressed.

E-mail: david.konrad@cup.lmu.de, uwe.grether@roche.com, rainer_e.martin@roche.com, gisbert@ethz.ch

Supplementary Note 1 Training Details

PyTorch Geometric (2.0.2) [1] and PyTorch (1.10.1+cu102) [2] functionalities were used for neural network training. Training was performed on a graphical processing unit GPU (Nvidia GeForce GTX 1080 Ti) for four hours, using a batch size of 16 samples. The Adam stochastic gradient descent optimizer was employed [3], with a learning rate of 10^{-4} , mean squared error (MSE) loss on the training set, a decay factor of 0.5 applied after 100 epochs, and an exponential smoothing factor of 0.9. The final model was stored after 1000 epochs. All the models considered in this study were trained on the Euler computing cluster at ETH Zurich, Switzerland.

Supplementary Note 2 Clustering

Figure S1 illustrates the clustered chemical space of 3180 advanced heterocyclic building blocks within principal component analysis (PCA). Exemplary chemical structures at the corners of the scatter plot are highlighted.

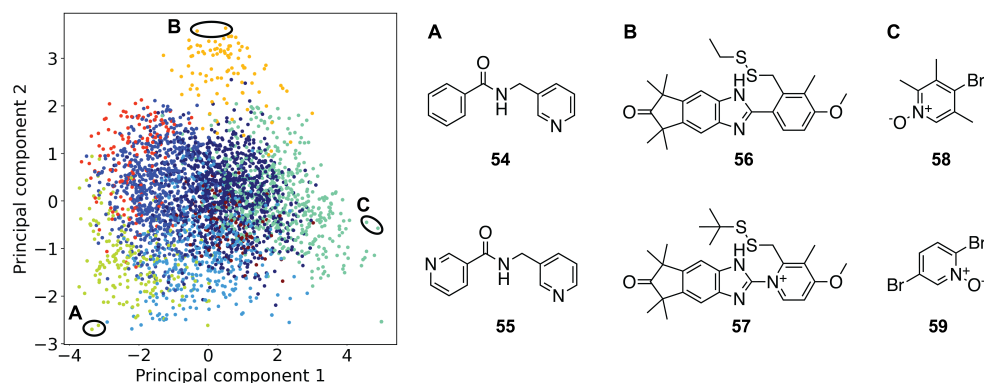


Figure S1: Compound clustering. Principal component analysis (PCA) of the 3180 advanced heterocyclic building blocks, based on ECFP4 molecular fingerprint descriptors. [4] Colors indicate the eight compound groups (clusters) obtained in the clustering process. The explainable variance for the investigated data set of the first two principal components is 22.2% and 9.6%, respectively. **A-C**: Chemical structures from three selected regions of the investigated chemical space. **A**: Amid derivatives (**54**, **55**) populating the bottom left of the PCA plot. **B**: Benzimidazole derivatives **56** and **57** populating the upper center of the PCA plot. **C**: Pyridine oxide derivatives **58** and **59** populating the center right of the PCA plot.

Supplementary Note 3 Steric and electronic graph-features

To validate the influence of electronic features in the input molecular graph on model performance as conducted in a previous study [5], two different GNNs have been trained for reaction yield and binary reaction outcome prediction (Table S1 and S2). The QM features were calculated on-the-fly using the DelFTa software package trained on the QMugs data collection [6–9]. Reaction yields were predicted with an error margin of 18 – 19 % and binary reaction outcome could be learned with an area under the receiver operating characteristic curve (AUC) of 82-83 % (Table S1). Electronic effects have shown significant improvements for the investigated tasks.

Table S1: Neural network performance for reaction yields prediction.

	PCC	MAE / %
GNN3D	0.686 (± 0.006)	18.7 (± 0.2)
GNN3DQM	0.67 (± 0.01)	18.6 (± 0.6)

Table S2: Neural network performance for binary reaction outcome prediction.

	Absolute accuracy / %	F-score
GNN3D	80.8 (± 1.2)	82.7 (± 0.6)
GNN3DQM	80.5 (± 0.7)	82.5 (± 0.2)

Supplementary Note 4 Systematic literature analysis

Following the previously reported systematic analysis of chemical transformations (SACT) concept [5], suitable Minisci-type alkylation reactions were identified. SACT comprises of (1) literature search, (2) literature data curation and evaluation, (3) methodology extraction, (4) reaction data curation and analysis.

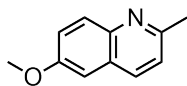
The literature search (1) was conducted using three different, renowned tools, Scopus (Elsevier, Amsterdam, Netherlands), Web of Science (Clarivate Analytics, Philadelphia, USA) and SciFinder-n (Chemical Abstracts Service, Columbus, USA) on the 30th of August 2022. On all databases a keyword search for "Minisci reaction" was carried out and the results download.

These files were subjected to a custom-built Alteryx Designer (Irvine, US) data curation (2) workflow that removed duplicates, added information from other databases, *e.g.*, journal impact factor, and carried out further filtering as well as calculations before splitting the publications into four quadrants based on journal impact factor and citations per year. After the removal of duplicates, 114 unique publication records were identified. With the available data, various different clustering approaches could have been carried out using a selection of the following dimensions, *e.g.*, journal and affiliation, citations, journal impact factor, technologies, catalysts, starting materials, and publication year. For this work, clustering by citations per year over journal impact factor to determine the most relevant Minisci methodology publications (high citations/year, high journal impact factor and high citations/year, low journal impact factor) was chosen. Removal of review papers delivered 45 remaining records [10–54], which underwent manual analysis to guarantee that the papers are within the scope of the automated HTE system (*e.g.*, photo- and electrochemistry out of scope). To allow for a broad utilization in medicinal chemistry, the publications were specially screened for a fast, robust and easily adaptable procedure. Among those, only Sutherland and co-workers [41] delivered a Minisci methodology that fulfilled those precise criteria. The metal- and catalyst-free setup as well as the ability to work with multiple non-pre-functionalized alkyl carboxylic allows for customized screening template design depending on project needs.

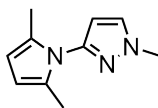
Next, using the simple user-friendly reaction format (SURF, see Section SI9 for further details), the reaction data from [41] was extracted and curated manually (3). This resulted in a high-quality data set comprising 45 borylation reactions serving as an ideal starting point for the development of the screening plate (see Section SI 6) through data analysis (4). Further, as a SURF file is machine-readable, the data was directly available as input for the machine learning pipelines.

Supplementary Note 5 LSF informer library

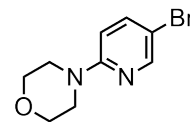
LSF informer library (drugs and fragments)



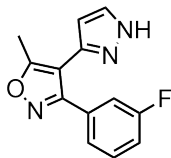
1
(1078-28-0)



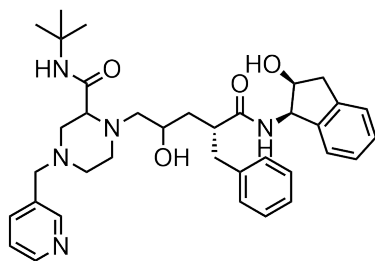
2
(34605-66-8)



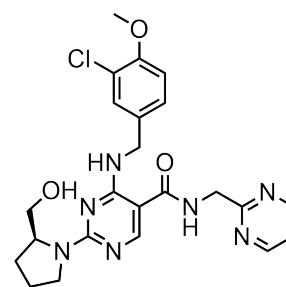
3
(200064-11-5)



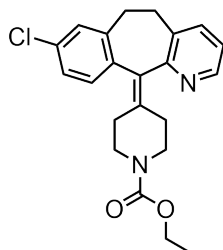
4
(n.d.)



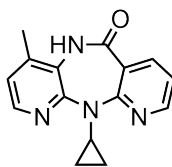
5
(150378-17-9)
Indinavir



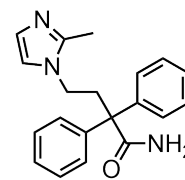
6
(330784-47-9)
Avanafil



7
(79794-75-5)
Loratadine



8
(129618-40-2)
Nevirapine



9
(170105-16-5)
Imidafenacin

Figure S2: Fragments and drugs of the LSF informer library 1-9. For compounds that have a trading name and a CAS number, the identifiers are depicted below the molecule.

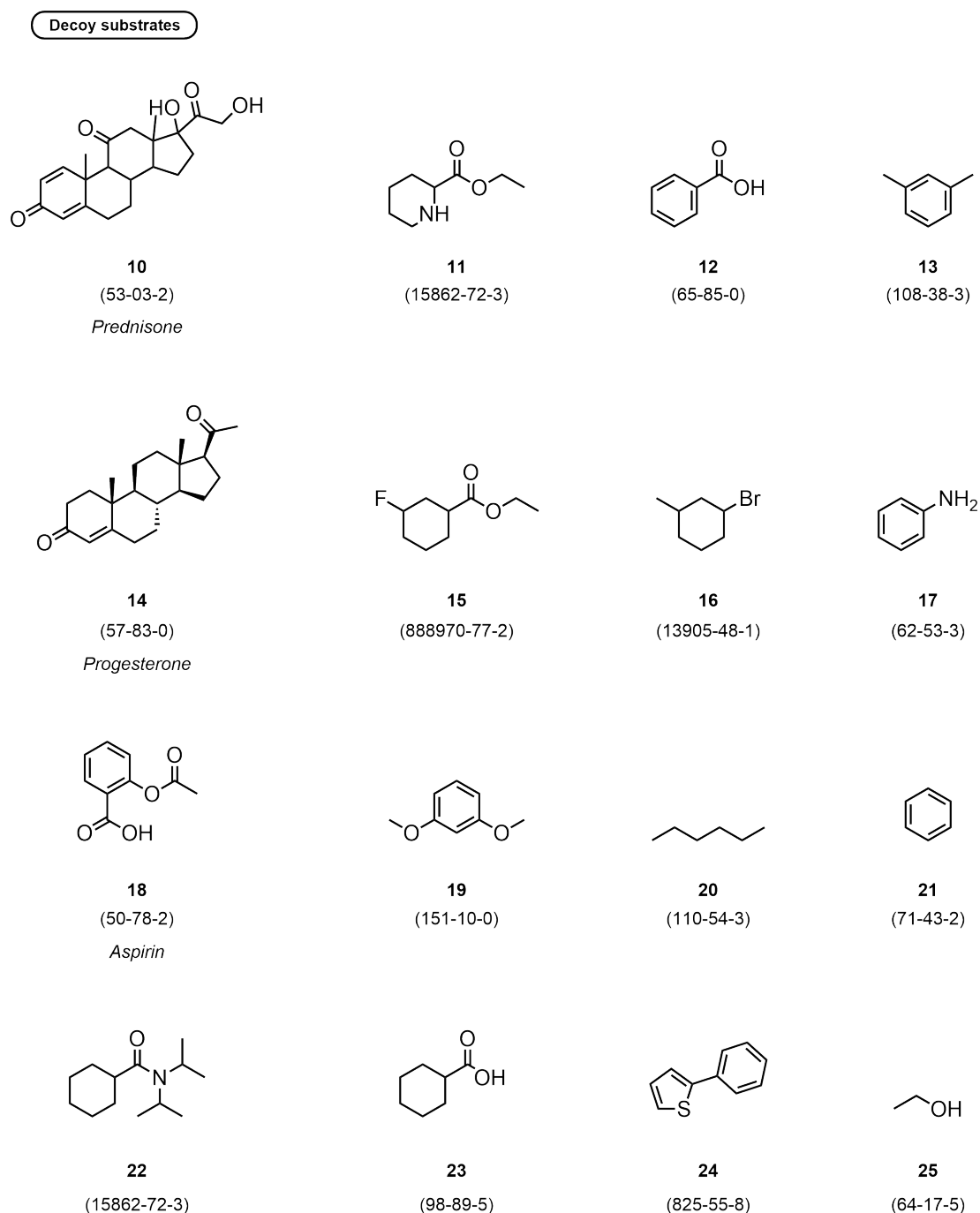
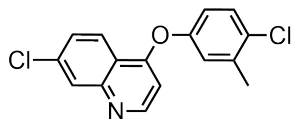
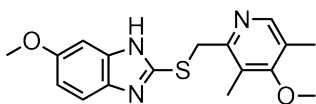


Figure S3: Selection of decoy substrates **10-25** used to generate unsuccessful reaction data to create a balanced training set. For compounds that have a trading name and a CAS number, the identifiers are depicted below the molecule

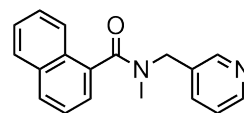
Fragments



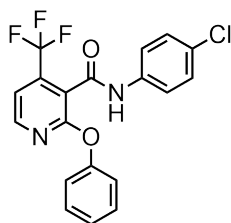
26
(124496-23-7)



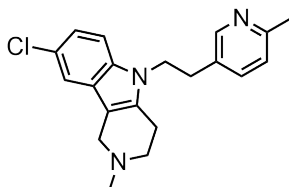
27
(73590-85-9)



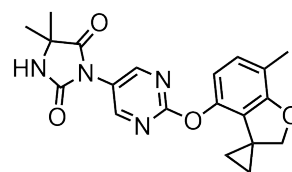
28
(1624057-23-3)



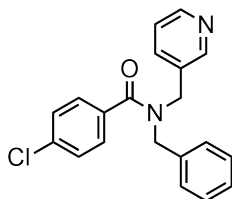
29
(685112-46-3)



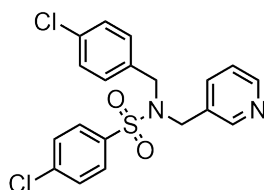
30
(21228-13-7)



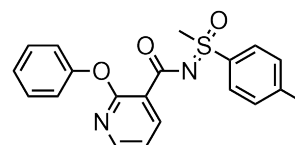
31
(1380696-64-9)



32
(287918-44-9)



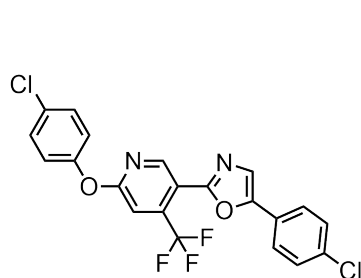
33
(685123-66-4)



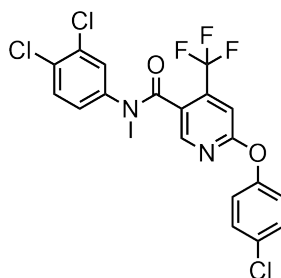
34
(882253-85-2)

Figure S4: Relevant fragments from the Roche library **26-34**. Part 1. For compounds that have a CAS number, the identifier is depicted below the molecule.

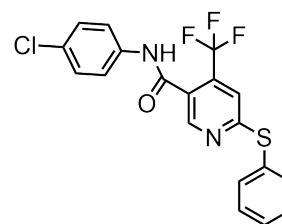
Fragments (continued)



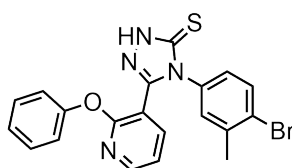
35
(883099-18-1)



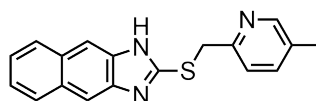
36
(287978-94-3)



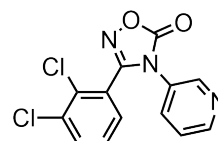
37
(685124-85-0)



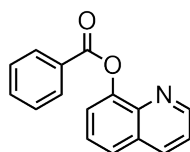
38
(218157-26-7)



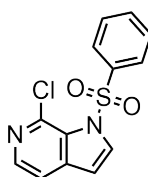
39
(71670-51-4)



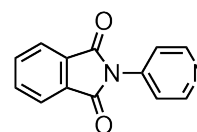
40
(288246-51-5)



41
(86-75-9)



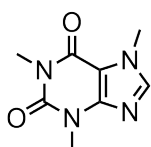
42
(1415124-76-3)



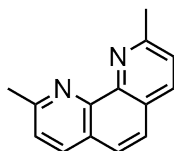
43
(69076-65-9)

Figure S5: Relevant fragments from the Roche library **35-43**. Part 2. For compounds that have a CAS number, the identifier is depicted below the molecule.

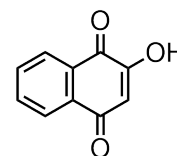
Substrates (publication)



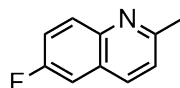
44
(58-08-2)



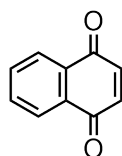
45
(484-11-7)



46
(83-72-7)



47
(1128-61-6)



48
(130-15-4)

Figure S6: Selection of substrates from [41], which also underwent screening with the designed plate and showed similar performance. For compounds that have a CAS number, the identifier is depicted below the molecule.

Supplementary Note 6 Screening plate design and testing

As indicated in the main manuscript (Section 4.2), the screening plate was designed around the literature data obtained from *Sutherland et al.* [41], which showed good yields on average (60%) for a variety of carboxylic acid coupling partners. To assess the reactivity of a substrate with a variety of different alkyl rests (rings and chains), a screening plate with 24 different alkyl carboxylic acids was assembled. The carboxylic acids scope from *Sutherland et al.* [41] was extended by sp^3 -rich *N*-heterocyclic carboxylic acids with relevance to drug discovery projects (**o**, **p**, **q**, **r**, Figure 2). The reactions were miniaturized to 0.5 μ mol scale, downsizing by a factor of 300 compared to the literature procedure. [41] To achieve this small reaction scale, stock solutions of all components in the reaction solvent (DMSO) were produced. Using simple substrate **1**, different oxidant and carboxylic acid ratios (3:10, 6:10, 3:20, 6:20) were tested to identify the more favourable screening condition (higher conversion). Further, the influence of the atmosphere (under air, under nitrogen in a glovebox), and the concentration (2, 16 mmol/L) on the reaction outcome were assessed. The results of this optimization process, which led to the final plate design (Figure 2) are disclosed below.

Air and inert atmosphere

To understand if the reaction requires an air-free atmosphere to deliver good yields, a selection of acid combinations were tested under air and in the glovebox on starting material **1**. The experimental results revealed that carrying out the reaction in the glovebox leads to significantly higher yields compared to the corresponding reaction under air (Figure S7).

	1	2	3		
A	 +19% (a)	 +54% (b)	 +30% (f)	40 °C 18 h 0.2 mg 0.5 μmol	
B	 +55% (g)	 +31% (h)	 +54% (j)		Acids (a-w) 20 eq.
C	 +34% (o)	 +29% (p)	 +15% (q)		Oxidant (NH ₄) ₂ S ₂ O ₈ 6 eq.
D	 +81% (u)	 +73% (s)	 +57% (w)		Solvent DMSO/H ₂ O (9:1) 16 mmol/L

Figure S7: Results from the plate testing with different carboxylic acids (**a**, **b**, **f**, **g**, **h**, **j**, **o**, **p**, **q**, **u**, **s**, **w**) under air and in the glovebox. The yield difference shown in the columns reflects the yield improvements when carrying out the reaction in a glovebox. *Reaction conditions*: Starting material (**1**, 5 μ mol), oxidant (NH₄)₂S₂O₈, solvent DMSO/H₂O (600:1), c = 2 mmol/L, 18 h, 40 °C.

Based on these outcomes, all further screening experiments took place in the glovebox.

Carboxylic acid and oxidant ratio

To identify the influence of the oxidant and carboxylic acid equivalents on the reaction success, four different combinations (3:10, 3:20, 6:10, 6:20) were experimentally tested. For these experiments, three different carboxylic acids (**e**, **j**, **o**) were reacted with starting material **1**. All other parameters were held constant ($c = 2$ mmol/L, $t = 18$ h, $T = 40$ °C, glovebox).


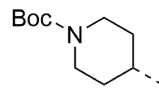
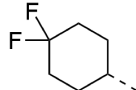





		①	②	③		
		 (e)	 (j)	 (o)	 40 °C	
		Oxidant : Acid (X : Y eq.)			 18 h	
					 0.2 mg	
					 0.5 μmol	
					 N₂	
A	3:10	15%	12%	13%	Oxidant (NH₄)₂S₂O₈	
B	6:10	39%	11%	38%	Solvent DMSO/H₂O (600:1) 2 mmol/L	
C	3:20	27%	12%	16%		
D	6:20	59%	19%	31%		

Figure S8: Results from plate testing with different oxidant and carboxylic acid equivalents on model substrate **1**. The influence on three different carboxylic acids (**e**, **j**, **o**) was tested. *Reaction conditions*: Starting material (**1**, 5 μmol), oxidant (NH₄)₂S₂O₈, solvent DMSO/H₂O (600:1), $c = 2$ mmol/L, 18 h, 40 °C, glovebox.

The results in Figure S8 show that for two of the three alkyl carboxylic acids (**e**, **j**), the 6:20 oxidant to carboxylic acid ratio delivered the best results. Even though for **o**, the 6:10 ratio showed the best results, the performance of 6:20 was in a similar range. Thus, the 6:20 ratio was incorporated into the final plate layout.

Reaction concentration

To avoid requiring additional solvent addition, *i.e.*, only dosing the stock solutions and using their solvent volume, a final comparison test was carried out comparing the influence of the solvent concentration. In the initial experiments, a reaction concentration of 0.002 mol/L was used. This baseline was compared to an increased concentration (0.016 mol/L, factor 8).

Experimental results (Figure S9) revealed that an increased reaction condition yields a higher amount of alkylated products. This experiment concluded the plate testing as all reactions nearly reached full starting material conversion. The final plate layout used for the screening experiments is depicted in Figure 2C of the main manuscript.

Application of reaction conditions to literature substrates

To verify that we did not optimize our conditions for a specific substrate, *i.e.*, compound **1**, five other starting materials (**44-48**, Section SI5, Figure S6) from Sutherland et al. [41] were screened with the optimized conditions on the whole plate (Figure 2). Similar performance as described in the paper was observed for the coupling with carboxylic acid **e**, confirming that the miniaturized plate can be used for a broader range of substrates and carboxylic acids. The detailed screening data is attached as a SURF file as one of the supplementary files of this manuscript.

	1	2	3	
A	 +72% (a)	 +27% (b)	 +47% (f)	40 °C 18 h 0.2 mg 0.5 μmol Acids (a-w) <i>20 eq.</i> Oxidant $(\text{NH}_4)_2\text{S}_2\text{O}_8$ <i>6 eq.</i> Atmosphere Glovebox
B	 +39% (g)	 +55% (h)	 +38% (j)	
C	 +31% (o)	 +48% (p)	 +55% (q)	
D	 +19% (u)	 +24% (s)	 +43% (w)	

Figure S9: Results from plate testing with two different reaction concentrations on model substrate **1**. The influence of using a slightly higher reaction concentration (0.016 compared to 0.002 mol/L) with different carboxylic acids (**a, b, f, g, h, j, o, p, q, u, s, w**) was tested. The yield difference shown in the columns reflects the yield improvements when carrying out the reaction with higher concentration. *Reaction conditions*: Starting material (**1**, 5 μmol), oxidant $(\text{NH}_4)_2\text{S}_2\text{O}_8$, solvent DMSO/H₂O (600:1), *c* = varied, 18 h, 40 °C, glovebox.

Supplementary Note 7 HTE screening protocol

All generated screening data used the plate design depicted in the paper (Figure 2) and the procedure below, only the starting materials (**1-9** and **26-43**) were varied. In a nitrogen-filled glovebox from mbraun (Garching, DE) that does not contain any liquids, all solid reaction components were dosed into 4 mL glass vials from Analytical Sales (Flanders, US) using a CHRONECT Quantos from Axel Semrau GmbH & Co. KG (Spockhövel, DE) coupled with an XPE206 balance from Mettler Toledo (Greifensee, CH). The vials were sealed and discharged from the glovebox before being transferred to another glovebox from LC Technologies (Salisbury, US), where solvents were added to the vials to generate the stock solutions. The remaining stock solutions with liquid components were prepared in a similar manner. Then according to the plate design, the stock solutions were transferred into 1 mL glass vials from Analytical Sales (Flanders, US) on a 24- or 96-well plate from Analytical Sales (Flanders, US) using multichannel pipettes from Eppendorf (Hamburg, DE). The plate was heated within the glovebox (LC Technologies) on a Junior benchtop solution from Unchained Labs (Pleasanton, US) and VP 721F-1 Parylene Encapsulated Stainless Steel Stir Discs from V&P Scientific Inc. (San Diego, US) were used to stir the reaction mixture. Only one internal process control (IPC) was taken after the overnight reaction by diluting the reaction mixture using a Freedom EVO 100 liquid handler from Tecan (Männedorf, CH) with MeCN/H₂O (4:1) to a defined concentration (1 mmol/L). After shaking on a Teleshake 95 from Inheco (Martinsried, DE), the samples were transferred onto a 96-deep-well plate (1 mL) from Eppendorf (Hamburg, DE). The plates were analyzed on a Waters (Milford, US) UPLC-MS system equipped with a Waters Acquity sample manager with a flow-through needle, a Waters Acquity sample organizer and a Waters QDa single quadrupole mass spectrometer. The separation was achieved on a ZORBAX RRHT Eclipse Plus C18, 95 Å, 2.1 x 30 mm, 1.8 µm column (P/N 959731-902, LOT: USUXY02479) from Agilent (Santa Clara, USA) at 50 °C. A 2-minute gradient was used and the injection volume accounted for 2 µL. 2 min gradient: A: 0.1% HCOOH in H₂O; B: 0.07% HCOOH in MeCN at flow 1 mL/min. Gradient: 0 min, 3% B; 0.2 min, 3% B; 1.5 min, 97% B; 0.3 min, 97% B; 0.1 min 3% B. The raw data were processed with MassLynx V4.2 and the obtained .rpt file underwent parsing with a customized script, before being subjected to the automated reaction data analysis pipeline (Section S18).

Supplementary Note 8 Automated reaction data analysis pipeline

In general, the same automated reaction data analysis pipeline as described in the SI (Section SI6) in a previous manuscript [5] was used. In this case, the framework was applied to rapidly identify if drugs or fragments were alkylated or not. Therefore, the MS searched for the sum formulas of the desired products (mono-, di- and tri-alkylated products), which were generated hands-free using a customized script based on the substrate chemical formula.

In addition to the reaction mixtures, all starting materials and, if available, reference products using the same solvent mixture (MeCN:H₂O, 4:1) are measured on the LCMS to obtain the retention time (LC) and mass pattern (MS). This data is stored in a database and needed for the initial two steps of the matching process. More relevant for LSF though, are the desired/potential products of the reaction. Those masses and chemical formulas are calculated based on the starting material information and the transformation. This Alteryx workflow allows hands-free generation of the potential products including molecular weight, mono-isotopic mass and chemical formula (Hill notation). In addition to being used for the reaction data analysis, this data is also the foundation for generating the LCMS input file.

Once the reaction data has passed through the cleaning process, it is compared to the LCMS information from the above-mentioned data sets, starting off with the identification of the starting material. If a trace from the reaction mixture matches the retention time (± 0.02 min) and the mass pattern (chemical formula detected, mass channel match with database reference), it receives the starting material tag. The remaining data is then compared to the products that could potentially be formed and are desired (mono-, di- or trialkylated species). Since the exact position of the new functional groups is not known, no reference compounds are available. Therefore, only the five most abundant masses per peak are used for tagging and compared to information from the potential product database. Based on the abundance of the mass and if the chemical formula was found by the LCMS, the tag is complemented by an MS reliability score. The score is higher if the chemical formula was found and the correct mass of the desired product (± 0.5 Da) appears in a more abundant channel. For this study, only high MS reliability scores were subjected to the machine learning platform. Last, the unmatched data is classified as unidentified products, and the mass differences between the peak and parent material are calculated to avoid manual calculating of mass differences.

After the tagging is completed, the data streams are recombined and subjected to calculations in order to quantify the reaction components from starting material through reagents to products. To do so, the sum of all LC peaks (integral) is calculated and each peak is then divided by this value. This gives a quantitative measure of the product distribution within the sample, an LCMS conversion. While there are numerous approaches to using internal standards or assays, due to the nature of LSF they have not been applied. LSF reactions tap into new, unexplored chemical space and generally, multiple different components are formed. Therefore, selecting an internal standard that does not overlap with one of these unknown components, is highly difficult.

Upon completion of the calculations, using the identifiers mentioned earlier, reaction information, such as conditions and components, are added to the components that have been identified and quantified. This follows the FAIR data principle and generates a curated, high-quality LSF screening data set that can be stored and shared in the SURF convention (Section SI9). This allows rapid subjecting of the data to machine learning algorithms as done in this research. It also enables direct visualization of the data in known interfaces, such as TIBCO Spotfire (Somerville, USA) or Tableau (Seattle, USA). Using this workflow, the data curation of one plate usually takes less than one minute.

Supplementary Note 9 SURF convention

The simple user-friendly reaction format (SURF) aims at standardizing reaction data reporting through a simple, yet comprehensive and structured format that is usable with a basic understanding of a spreadsheet. SURF does not require any coding experience, advanced IT skills or a web interface. It enables every chemist within or outside the lab to document chemical synthesis in a machine-readable and shareable format. SURF allowed extraction and documentation of the alkylation reactions from literature faster. The generated reaction screening data were also transformed into SURF before being directly subjected to the machine learning pipelines. Reaction documentation following SURF can be implemented in every spreadsheet as the only requirement is the existence of rows and columns.

Each row of the spreadsheet represents the information and data for one single reaction. The SURF convention contains constant (CC) and flexible (FC) categories. CCs never change and are always present, independent of the number of reaction components. They capture the origin and ids of the reaction as well as basic characteristics (reaction type, named reaction, reaction technology) and conditions (temperature, time, atmosphere, scale, concentration, stirring/shaking). Add-ons, such as the procedure or comments, belong to the CCs, too. The FCs describe the more variable part of a reaction, the starting material(s), solvent(s), reagent(s) and product(s). Two identifier options (CAS and SMILES) are available for each component. While the SMILES string is available for every compound and serves as structural input for machine learning models, the CAS number, even though not always available, can be handy for chemists in the lab to order, itemize and find chemicals. For the starting material(s) and reagent(s), *e.g.*, catalyst, ligand, additive, the number and type of columns remain the same (CAS, Smiles, equivalents). If multiple starting materials or reagents are used, additional columns are required. In that case, the three information columns are duplicated and the X is replaced by a number, starting from 1 for the first component, 2 for the second, etc. The same accounts for multiple solvents or products, however, due to their role, they possess more and partly different columns. While the CAS number and/or the SMILES string remain as an identifier, the solvent fraction (in decimals) instead of equivalents is recorded. This allows exact determination of the ratio between solvents. The product category withholds the largest amount of headers as SURF records the yield (in percent), but also the yield type (*e.g.*, isolated, lcms, gcms) as well as the detected mass by MS and the ^1H NMR sequence in addition to the common identifiers CAS and Smiles. This not only allows rapid comparison when experiments are reproduced but can also deliver important increments for machine learning models by differentiating between yield types. As most electronic lab journals already record the above-mentioned parameters, by enforcing of documentation compliance combined with simple automated data extraction and cleaning pipelines, numerous reaction data could be accessible in the SURF convention, and readily available for machine learning applications. We spent thoughts on how to further reduce complexity by introducing specific SURFs without FCs for chemical transformations where the reaction components are generally the same. An excellent example would be Suzuki-Miyaura couplings that utilize a set of six to seven components (organoboron species, halide, catalyst, ligand, base, solvents). [55, 56] However, generating different tailored templates would ultimately end up in various different formats and mismatching headers falling short of the main SURF goal to standardize reaction documentation.

The results of this paper would have not been achieved without FAIR data handling using SURF. The manually extracted reaction data (45 reactions from one publication), which were used in this manuscript for data analysis and selectivity prediction, reported in SURF are attached to the SI as a tab-delimited text file. Further, the reaction data of the scale-up experiments in SURF is also added to the SI of this manuscript and significantly increased the efficiency when compiling the experimental part (Section SI12). Moreover, two empty SURF templates are attached as tab-delimited text files: The first file contains the general SURF template, which can be adjusted by introducing additional columns depending on the reaction specifics. The second file is a customized SURF template that should accommodate the vast of chemical transformations: It contains columns for two starting materials, two reagents, one catalyst, one ligand, one additive, two solvents and two products.

Supplementary Note 10 Screening results of acids and fragments

For the investigated carboxylic acids (Figure S10) as well as for the *N*-hetero arenes (Figure S11) different average reaction yields were observed. Figure S10 illustrates the observed reaction yields for 22 carboxylic acids (**a-w**). Carboxylic acids substituted with cyclic ethers (*e.g.*, **u**, **s**, **a**) and alkanes (*e.g.*, **b**, **e**, **g**) were observed with high reaction yields, whereas cyclic boc-protected amines (*e.g.*, **o**, **p**, **q**, **r**) and amides (**d**) resulted in low yields of the respective desired reaction products. Figure S11 illustrates the observed reaction yields for 27 *N*-hetero arenes (**1-9** and **26-43**). Substituted pyridines (*e.g.*, **30**, **31**, **36**, **39**; see Section S15) were observed with lower yields compared to compounds lacking meta-substituents (*e.g.*, **26**, **32**, **38**, **41**).

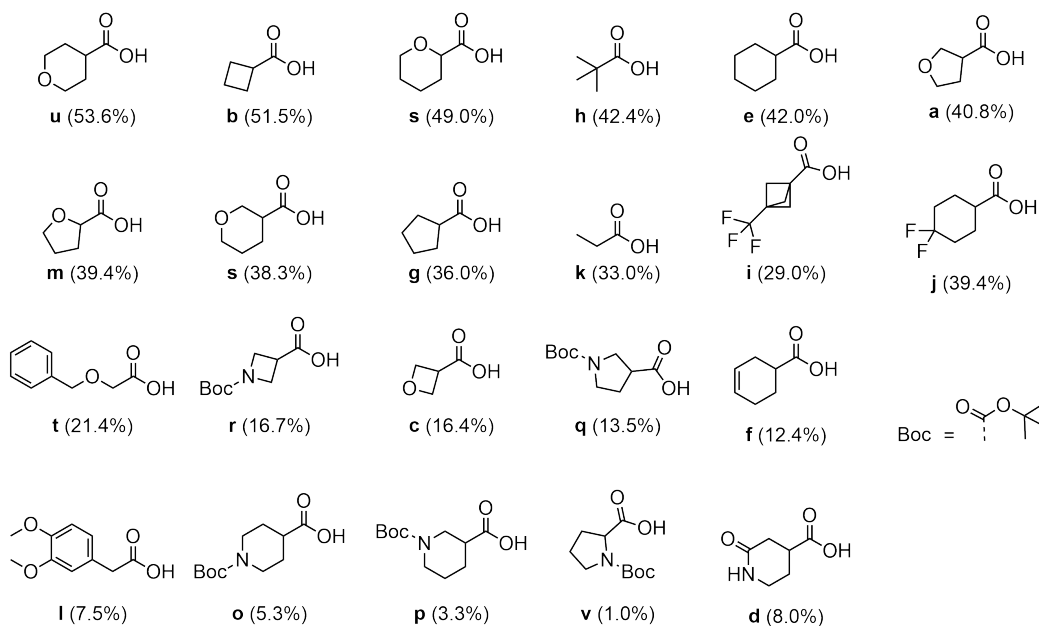


Figure S10: Analysis of screening results for acids.

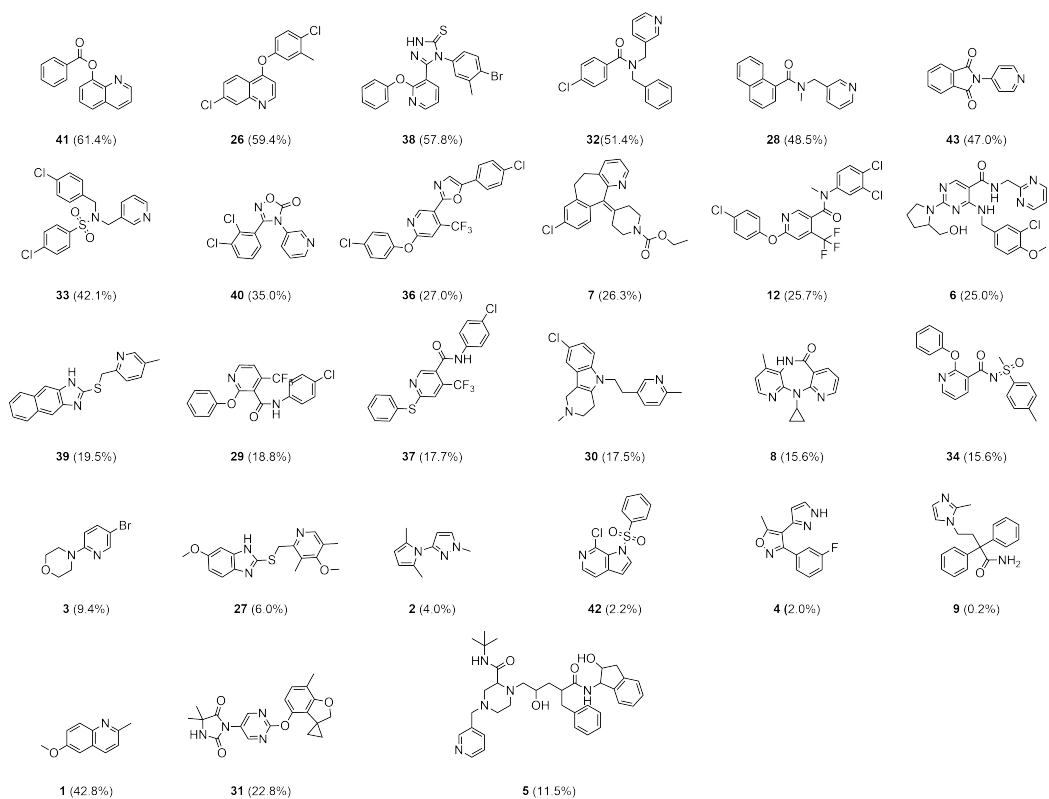


Figure S11: Analysis of screening results for substrates.

Supplementary Note 11 Machine learning outliers

The observed outliers (17/1148) for reaction yield prediction with a mean absolute error (MAE) $\geq 70\%$ are illustrated in Table S3.

Table S3: Machine learning outliers for reaction yield prediction.

Reaction ID	MAE (%)	Ground truth (%)
eln044720-053-1-1-b1	82.1	0.0
eln044720-053-1-1-b2	100.0	0.0
eln036496-183-15-1-b6	81.9	3.9
eln036496-188-2-1-d3	73.0	75.9
eln046486-037-1-1-a2	70.8	0.0
eln044720-053-1-1-a2	82.3	0.0
eln036496-183-15-1-b3	71.2	99.0
eln044720-053-1-1-a5	70.3	0.0
eln044720-045-1-1-d2	71.9	87.0
eln036496-191-1-1-d2	73.7	75.0
eln046486-035-1-1-a1	70.5	12.9
eln046486-037-1-1-a1	80.5	88.9
eln044720-053-1-1-c2	75.5	95.9
eln046486-035-1-1-a2	78.1	0.0
eln044720-041-1-1-d2	91.7	0.0
eln044720-053-1-1-a1	72.7	97.0
eln044720-041-1-1-c2	97.6	0.0

The outlier data does not indicate any clear trends, where the model did not perform well. However, selected outliers are discussed below. The two most occurring acids (3x) are A1 (**a**) and D2 (**s**). For **s**, the oxygen in proximity to the reactive site could have led to increased difficulties when predicting the outcome of the transformation as the results for this acid vary broadly across the data set, which is already visible based on the three data points in the table. While the coupling of **s** worked well with **1** and **5**, it was not reactive at all with **38**. For **a**, a conversion could always be observed experimentally, however, the values for the three substrates (**35**, **45**, **46**) differ largely as well. Discrepancies of **45** and **46** might originate due to the new chemical classes (quinones), which are not broadly represented in the data set. **35** generated difficulties for the model (six appearances in the table) as it showed varying experimental behaviour depending on the paired acid. The complexity of the structure including various functional groups with different demanding electronic and steric effects around the pyridine might have led to the observed high MAEs.

Supplementary Note 12 Scale-up reactions

Supplementary Note 12.1 Reagent and purification information

Reactions were set up and conducted in nitrogen-filled gloveboxes from mbraun (Garching, DE) and LC Technologies (Salisbury, US). All chemicals were purchased from Sigma Aldrich (St. Louis, US), AstaTech (Bristol, US), Combi-Blocks (San Diego, US), TRC (Toronto, CA), Thermo Scientific (Waltham, US) or obtained from the Roche compound library and used as received. All solids were dosed using a CHRONECT Quantos from Axel Semrau GmbH & Co. KG (Spockhövel, DE) coupled with an XPE206 balance from Mettler Toledo (Greifensee, CH). Anhydrous solvents were purchased from Sigma Aldrich, stored in the glovebox and added to the reaction vials using pipettes from Eppendorf (Hamburg, DE). The vials were heated on a Junior benchtop solution from Unchained Labs (Pleasanton, US) and the reaction mixture was stirred by VP 721F-1 Parylene Encapsulated Stainless Steel Stir Discs from V&P Scientific Inc. (San Diego, US). Purification by flash column chromatography was performed using SiliaSep Premium Flash Cartridges from Silicycle (Quebec, CA) on a Combi Flash Rf from Teledyne ISCO (Nebraska, US) or by reversed-phase high-pressure liquid chromatography (RP-HPLC) on a Gilson (Middleton, USA) GX-281 liquid handler equipped with a Shimadzu (Kyoto, JP) LC-20AP dual pump, a Thermo Fisher Scientific (Waltham, US) UV/VIS-Thermo Ultimate 300 Detector, a VWR (Radnor, US) ELSD90 ELSD detector and a Thermo Fisher Scientific (Waltham, US) Thermo MSQ Plus MS Single Quadrupole using a Phenomenex (Torrance, US) Gemini NX C18 column (12 nm, 5 μ m silica, 30 mm diameter, 100 mm length, flow rate of 40 mL/min) or YMC (Kyoto, JP) Triart C18 (12 nm, 5 μ m, 100x30 mm) column. The used eluent solvents, gradients and cartridge sizes for flash chromatography and RP-HPLC are described individually for each experiment.

Supplementary Note 12.2 Analytical information

All compounds were characterized by nuclear magnetic resonance (NMR) spectroscopy and (flow injection analysis (FIA)) high-resolution mass spectrometry (HRMS) or gas-chromatography mass spectrometry (GCMS). NMR spectra were recorded on a Bruker Avance III, 600 MHz spectrometer equipped with a 5 mm TCI, Z-gradient CryoProbe, a Bruker Avance Neo, 400 MHz spectrometer equipped with a 5 mm Z-gradient iProbe or a Bruker Avance III HD, 300 MHz spectrometer equipped with a 5 mm BBI-Probe. NMR data are reported as follows: chemical shift in reference to the residual solvent peak (δ ppm), multiplicity (s = singlet, d = doublet, dd = doublet of doublet, t = triplet, dt = doublet of triplet, td = triplet of doublet, q = quintet, m = multiplet), coupling constant (Hz), and integration. ^1H NMR residual solvent peaks in respective deuterated solvents for CHCl_3 at 7.26 ppm and DMSO at 2.50 ppm. ^{13}C NMR residual solvent peaks in respective deuterated solvents for CHCl_3 at 77.16 ppm and DMSO at 39.52 ppm.

LC-MS high-resolution spectra were recorded with an Agilent LC system consisting of Agilent 1290 high-pressure gradient system, and an Agilent 6545 QTOF. The separation was achieved on a Zorbax Eclipse Plus C18 1.7 μ m 2.1 x 50 mm column (P/N 959731-902) at 55 $^\circ\text{C}$; A: 0.01% HCOOH in H_2O ; B: MeCN at flow 0.8 mL/min. Gradient: 0 min 5% B, 0.3 min 5% B, 4.5 min 99% B, 5 min 99% B. The injection volume was 2 μL . Ionization was performed in an Agilent Multimode source. The mass spectrometer was run in “2 GHz extended dynamic range” mode, resulting in a resolution of about 20 000 at $m/z = 922$. Mass accuracy was ensured by internal drift correction. GC-MS spectra were recorded on an Agilent 5975B single quadrupole mass spectrometer. Separation was achieved on an Agilent 7890A using a HP-1ms column (15 m ID: 250 μ m and 0.25 μ m film) with He as carrier gas. Sample introduction was done via a Split injector at 270 $^\circ\text{C}$. After 0.5 min at a constant temperature, the temperature was ramped from 100 $^\circ\text{C}$ or 45 $^\circ\text{C}$ to 320 $^\circ\text{C}$ with 35 $^\circ\text{C}/\text{min}$. The mass spectrometer was operated in EI (electron ionization) mode at 70 eV. FIA-HRMS spectra were recorded with an Agilent LC system consisting of an Agilent 1290 high-pressure gradient system, and an Agilent 6540 QTOF. No separation was intended and the injected sample was flushed directly into the Agilent Jetstream source. The mass spectrometer was run in “2 GHz extended dynamic range” mode, resulting in a resolution of about 20 000 at m/z 922. Mass accuracy was ensured by internal drift correction.

Supplementary Note 12.3 Experimental procedures and analytical data

4-(13-chloro-5-cyclobutyl-4-azatricyclo[9.4.0.0.3,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7b1**),

4-(13-chloro-7-cyclobutyl-4-azatricyclo[9.4.0.0.3,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7b2**),

4-[13-chloro-5,7-di(cyclobutyl)-4-azatricyclo[9.4.0.0.3,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7b3**):

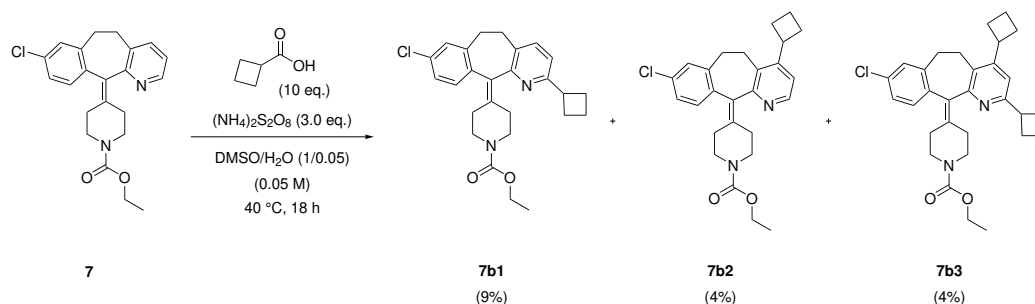


Figure S12: Alkylation of Loratadine (**7**).

To a solution of 4-(13-chloro-4-azatricyclo[9.4.0.0.3,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7**, 57.4 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μL H₂O, cyclobutanecarboxylic acid (**b**, 150.2 mg, 1.50 mmol, 10.0 eq.) and ammonium persulfate (102.7 mg, 450 μmol , 3.00 eq.) was added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μm , 100 x 30 mm) using a MeCN gradient (20-40-55%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compounds 4-(13-chloro-5-cyclobutyl-4-azatricyclo[9.4.0.0.3,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7b1**, 6.3 mg, 9%) as an off-white powder, 4-(13-chloro-7-cyclobutyl-4-azatricyclo[9.4.0.0.3,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7b2**, 2.7 mg, 4%) as an off-white powder and 4-[13-chloro-5,7-di(cyclobutyl)-4-azatricyclo[9.4.0.0.3,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7b3**, 3.8 mg, 4%) as an off-white powder.

7b1:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.34 (d, $J = 7.9$ Hz, 1H), 7.12 - 7.18 (m, 3H), 7.01 (d, $J = 7.9$ Hz, 1H), 4.15 (q, $J = 7.1$ Hz, 2H), 3.76 - 3.81 (m, 1H), 3.60 - 3.66 (m, 1H), 3.35 - 3.40 (m, 1H), 3.26 - 3.31 (m, 1H), 3.19 - 3.24 (m, 2H), 2.75 - 2.85 (m, 2H), 2.53 - 2.56 (m, 1H), 2.30 - 2.38 (m, 6H), 2.20 - 2.30 (m, 2H), 2.00 - 2.06 (m, 1H), 1.85 - 1.89 (m, 1H), 1.27 (t, $J = 7.1$ Hz, 3H). ¹³C NMR (151 MHz, CDCl₃) δ (ppm) 161.96, 155.75, 140.18, 138.04, 137.48, 134.69, 132.91, 130.76, 130.28, 128.92, 126.19, 119.22, 61.51, 45.17, 42.16, 31.97, 31.68, 31.15, 30.87, 29.04, 28.73, 18.48, 14.92, 14.89. HRMS C₂₆H₂₉ClN₂O₂; calc. for (M+H⁺): 437.1918, found: 437.2.

7b2:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.36 (d, $J = 5.1$ Hz, 1H), 7.12 (s, 3H), 7.08 (d, $J = 4.9$ Hz, 1H), 4.15 (q, $J = 7.2$ Hz, 2H), 3.75 - 3.85 (m, 2H), 3.61 - 3.68 (m, 1H), 3.34 - 3.40 (m, 1H), 3.11 - 3.19 (m, 3H), 2.77 - 2.87 (m, 2H), 2.35 - 2.44 (m, 5H), 2.21 - 2.24 (m, 2H), 2.07 - 2.11 (m, 2H), 1.85 - 1.90 (m, 1H), 1.26 (t, $J = 7.1$ Hz, 3H). ¹³C NMR (151 MHz, CDCl₃) δ (ppm) 158.25, 155.71, 152.61, 146.89, 139.42, 136.99, 136.81, 134.78,

133.01, 131.34, 131.15, 129.53, 126.19, 120.12, 61.52, 44.99, 44.88, 37.91, 31.91, 30.80, 30.80, 29.45, 28.30, 26.95, 18.57, 14.89. **HRMS** $C_{26}H_{29}ClN_2O_2$; calc. for $(M+H^+)$: 437.1918, found: 437.2.

7b3:

1H NMR (600 MHz, $CDCl_3$) δ (ppm) 7.17 (d, $J = 8.1$ Hz, 1H), 7.10 - 7.12 (m, 2H), 6.98 (s, 1H), 4.15 (q, $J = 7.2$ Hz, 2H), 3.75 - 3.86 (m, 2H), 3.58 - 3.66 (m, 2H), 3.33 - 3.38 (m, 1H), 3.08 - 3.21 (m, 3H), 2.74 - 2.82 (m, 2H), 2.32 - 2.38 (m, 6H), 2.24 - 2.26 (m, 3H), 1.99 - 2.12 (m, 3H), 1.85 - 1.89 (m, 2H), 1.26 (t, $J = 7.1$ Hz, 3H). **HRMS** $C_{30}H_{35}ClN_2O_2$; calc. for $(M+H^+)$: 491.2387, found: 491.2.

4-[5-(benzoxymethyl)-13-chloro-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]-piperidine-1-carboxylic acid ethyl ester (**7t1**),
 4-[7-(benzoxymethyl)-13-chloro-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]-piperidine-1-carboxylic acid ethyl ester (**7t2**)

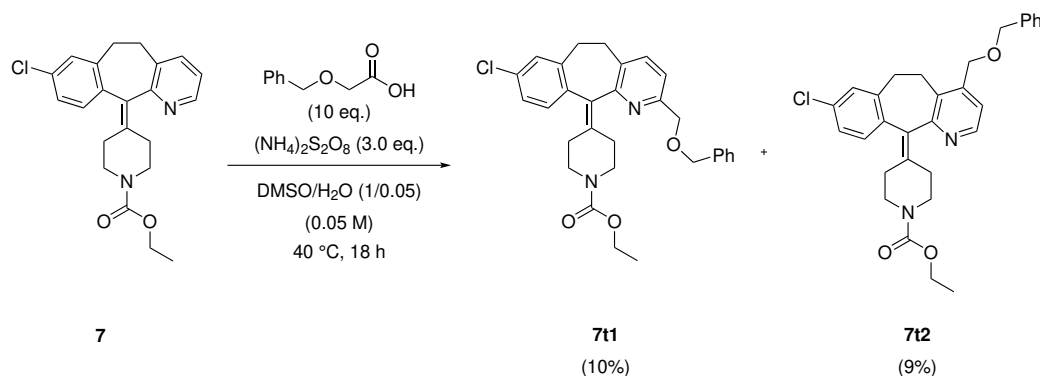


Figure S13: Alkylation of Loratadine (**7**).

To a solution of 4-(13-chloro-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7**, 57.4 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, 2-benzoxymethylacetic acid (**t**, 249.3 mg, 1.50 mmol, 10.0 eq.) and ammonium persulfate (102.7 mg, 450 μ mol, 3.00 eq.) was added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (55-75-90-100%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compounds 4-[5-(benzoxymethyl)-13-chloro-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7t1**, 8.1 mg, 10%) as an off-white powder and 4-[7-(benzoxymethyl)-13-chloro-4-azatricyclo[9.4.0.0^{3,8}]pentadeca-1(15)-3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7t2**, 6.6 mg, 9%) as an off-white powder.

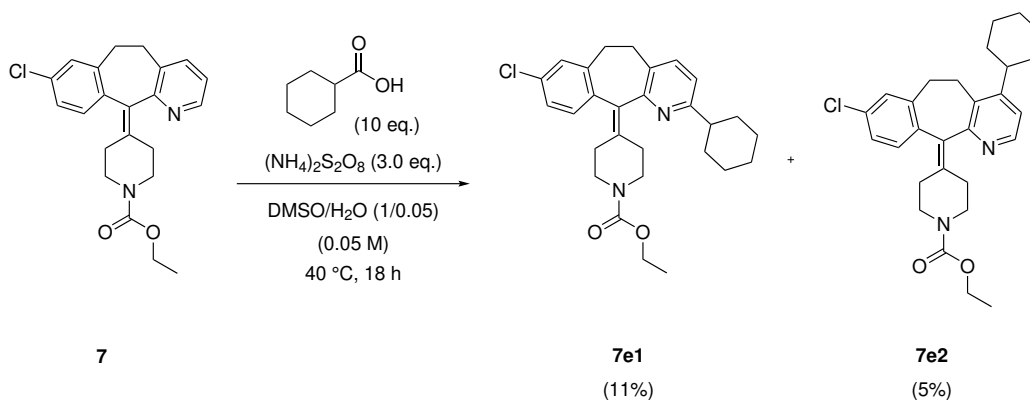
7t1:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.47 (d, J = 8.0 Hz, 1H), 7.35 - 7.38 (m, 4H), 7.33 - 7.34 (m, 1H), 7.29 - 7.31 (m, 1H), 7.16 - 7.17 (m, 1H), 7.14 - 7.15 (m, 2H), 4.66 (d, J = 2.4 Hz, 2H), 4.63 (s, 2H), 4.15 (q, J = 7.1 Hz, 2H), 3.84 - 3.86 (m, 2H), 3.37 - 3.41 (m, 1H), 3.31 - 3.36 (m, 1H), 3.03 - 3.10 (m, 2H), 2.78 - 2.87 (m, 2H), 2.42 - 2.47 (m, 1H), 2.31 - 2.36 (m, 3H), 1.26 (t, J = 7.1 Hz, 3H). HRMS C₃₀H₃₁ClN₂O₃; calc. for (M+H⁺): 503.2023, found: 503.2.

7t2:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.42 (d, J = 5.0 Hz, 1H), 7.38 - 7.40 (m, 1H), 7.35 - 7.38 (m, 3H), 7.32 - 7.35 (m, 1H), 7.30 (d, J = 5.0 Hz, 1H), 7.11 - 7.15 (m, 3H), 4.62 (s, 2H), 4.52 (s, 2H), 4.15 (q, J = 7.1 Hz, 2H), 3.81 - 3.82 (m, 2H), 3.36 - 3.31 (m, 1H), 3.12 - 3.19 (m, 3H), 2.78 - 2.86 (m, 2H), 2.44 - 2.46 (m, 1H), 2.33 - 2.41 (m, 2H), 2.26 - 2.29 (m, 1H), 1.26 (t, J = 7.1 Hz, 3H). HRMS C₃₀H₃₁ClN₂O₃; calc. for (M+H⁺): 503.2023, found: 503.2.

4-(13-chloro-7-cyclohexyl-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7e1**),
 4-(13-chloro-5-cyclohexyl-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7e2**)

Figure S14: Alkylation of Loratadine (**7**).

To a solution of 4-(13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7**, 57.4 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, cyclohexanecarboxylic acid (**e**, 192.3 mg, 186.7 μ L, 1.50 mmol, 10.0 eq.) and ammonium persulfate (102.7 mg, 450 μ mol, 3.00 eq.) was added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (YMC-Triart C18, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-98%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compounds 4-(13-chloro-7-cyclohexyl-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7e1**, 8.1 mg, 11%) as an off-white powder and 4-(13-chloro-5-cyclohexyl-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7t2**, 5.4 mg, 5%) as an off-white powder.

7e1:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.34 - 7.35 (m, 1H), 7.17 - 7.18 (m, 2H), 7.14 - 7.15 (m, 1H), 6.97 (d, $J = 7.8$ Hz, 1H), 4.15 (q, $J = 7.1$ Hz, 2H), 3.82 - 3.87 (m, 2H), 3.35 - 3.40 (m, 1H), 3.27 - 3.31 (m, 1H), 3.09 - 3.13 (m, 1H), 2.68 - 2.84 (m, 1H), 2.33 - 2.36 (m, 3H), 1.94 - 1.97 (m, 1H), 1.88 - 1.90 (m, 1H), 1.73 - 1.76 (m, 1H), 1.38 - 1.49 (m, 6H), 1.27 (t, $J = 7.1$ Hz, 3H). **HRMS** C₂₈H₃₃ClN₂O₂; calc. for (M+H⁺): 465.2231, found: 465.2.

7e2:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.32 (d, $J = 5.2$ Hz, 1H), 7.11 - 7.13 (m, 3H), 7.07 (d, $J = 5.3$ Hz, 1H), 4.15 (q, $J = 7.1$ Hz, 2H), 3.38 - 3.42 (m, 1H), 3.18 - 3.28 (m, 2H), 3.12 - 3.15 (m, 1H), 3.01 - 3.05 (m, 1H), 2.81 - 2.87 (m, 1H), 2.75 - 2.79 (m, 1H), 2.38 - 2.42 (m, 3H), 2.19 - 2.21 (m, 1H), 1.87 - 1.91 (m, 1H), 1.80 - 1.85 (m, 3H), 1.72 - 1.73 (m, 1H), 1.40 - 1.46 (m, 6H), 1.26 (t, $J = 7.1$ Hz, 3H). **HRMS** C₂₈H₃₃ClN₂O₂; calc. for (M+H⁺): 465.2231, found: 465.2.

4-[13-chloro-5-(4,4-difluorocyclohexyl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7j1**),
 4-[13-chloro-7-(4,4-difluorocyclohexyl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7j2**)

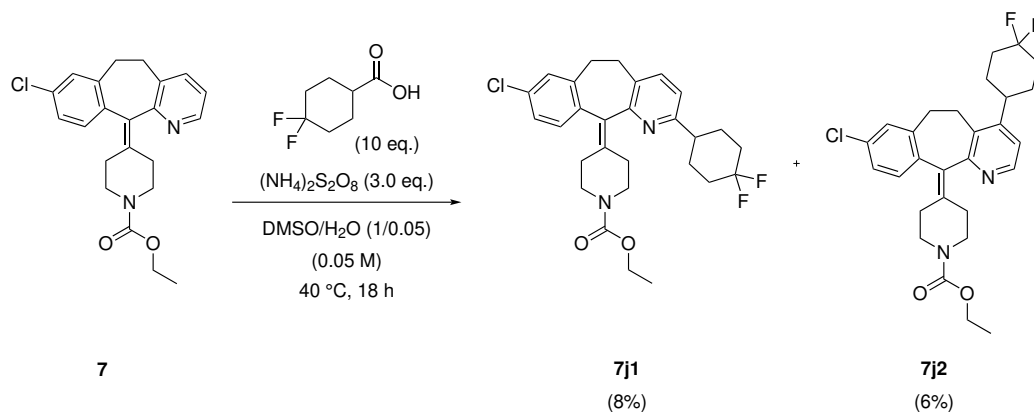


Figure S15: Alkylation of Loratadine (**7**).

To a solution of 4-(13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7**, 57.4 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, 4,4-difluorocyclohexanecarboxylic acid (**j**, 246.2 mg, 1.50 mmol, 10.0 eq.) and ammonium persulfate (102.7 mg, 450 μ mol, 3.00 eq.) was added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (YMC-Triart C18, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-98%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compounds 4-[13-chloro-5-(4,4-difluorocyclohexyl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7j1**, 7.8 mg, 8%) as an off-white powder and 4-[13-chloro-7-(4,4-difluorocyclohexyl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7j2**, 6.0 mg, 6%) as an off-white powder.

7j1:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.37 - 7.39 (m, 1H), 7.17 - 7.18 (m, 1H), 7.14 - 7.15 (m, 2H), 6.98 (d, J = 7.9 Hz, 1H), 4.15 (q, J = 7.1 Hz, 2H), 3.36 - 3.41 (m, 1H), 3.28 - 3.32 (m, 1H), 3.11 - 3.16 (m, 2H), 2.77 - 2.86 (m, 4H), 2.46 - 2.50 (m, 1H), 2.29 - 2.35 (m, 4H), 2.18 - 2.23 (m, 2H), 2.02 - 2.04 (m, 1H), 1.96 - 1.98 (m, 1H), 1.82 - 1.90 (m, 4H), 1.27 (t, J = 7.1 Hz, 3H). HRMS C₂₈H₃₁ClF₂N₂O₂; calc. for (M+H⁺): 501.2042, found: 501.2.

7j2:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.37 (d, J = 5.3 Hz, 1H), 7.12 - 7.14 (m, 3H), 7.09 (d, J = 5.3 Hz, 1H), 4.15 (q, J = 7.1 Hz, 2H), 3.41 - 3.45 (m, 1H), 3.28 - 3.33 (m, 1H), 3.14 - 3.25 (m, 2H), 2.98 - 3.02 (m, 1H), 2.80 - 2.90 (m, 3H), 2.38 - 2.44 (m, 3H), 2.24 - 2.31 (m, 2H), 2.16 - 2.19 (m, 1H), 1.78 - 1.93 (m, 6H), 1.26 (t, J = 7.1 Hz, 3H). HRMS C₂₈H₃₁ClF₂N₂O₂; calc. for (M+H⁺): 501.2042, found: 501.2.

4-(13-chloro-5-tetrahydropyran-2-yl-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7s**)

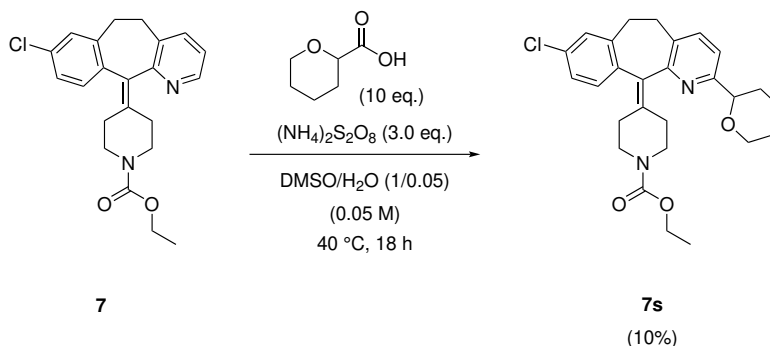


Figure S16: Alkylation of Loratadine (**7**).

To a solution of 4-(13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7**, 57.4 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 uL H₂O, tetrahydropyran-2-carboxylic acid (**s**, 195.2 mg, 1.50 mmol, 10.0 eq.) and ammonium persulfate (102.7 mg, 450 umol, 3.00 eq.) was added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 um, 100 x 30 mm) using a MeCN gradient (30-65-90%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound 4-(13-chloro-5-tetrahydropyran-2-yl-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7s**, 7.1 mg, 10%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.45 (dd, *J* = 11.9, 7.9 Hz, 1H), 7.28 - 7.31 (m, 1H), 7.13 - 7.18 (m, 3H), 4.44 (dt, *J* = 11.2, 2.4 Hz, 1H), 4.12 - 4.17 (m, 3H), 3.59 - 3.66 (m, 1H), 3.29 - 3.39 (m, 2H), 2.98 - 3.12 (m, 2H), 2.78 - 2.87 (m, 2H), 2.31 - 2.46 (m, 4H), 1.88 - 1.95 (m, 1H), 1.65 - 1.73 (m, 3H), 1.57 - 1.60 (m, 1H), 1.40 - 1.47 (m, 1H), 1.26 (td, *J* = 7.1, 3.3 Hz, 4H). HRMS C₂₇H₃₁ClN₂O₃; calc. for (M+H⁺): 467.2023, found: 467.2.

2-cyclopropyl-7-methyl-14-tetrahydropyran-2-yl-2,4,9,15-tetraza-tricyclo[9.4.0.03,8]pentadeca-1(15)-3(8),4,6,11,13-hexaen-10-one (8s)

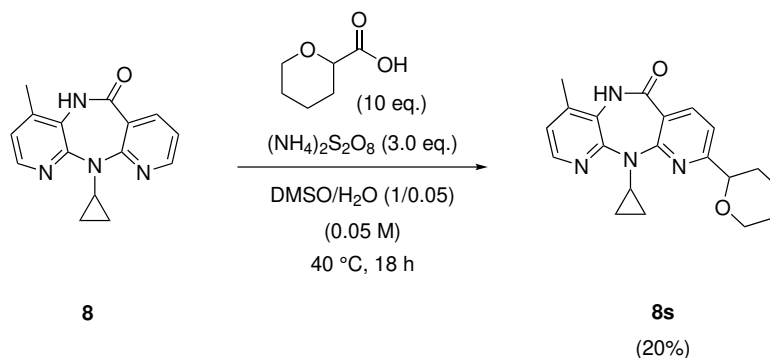


Figure S17: Alkylation of Nevirapine (**8**).

To a solution of 2-cyclopropyl-7-methyl-2,4,9,15-tetraza-tricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-10-one (**8**, 40.0 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, tetrahydropyran-2-carboxylic acid (**s**, 195.2 mg, 1.50 mmol, 10.0 eq.) and ammonium persulfate (102.7 mg, 450 μ mol, 3.00 eq.) was added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-45-65%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound 2-cyclopropyl-7-methyl-14-tetrahydropyran-2-yl-2,4,9,15-tetraza-tricyclo[9.4.0.03,8]pentadeca-1(15),3(8),4,6,11,13-hexaen-10-one (**8s**, 10.6 mg, 20%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.16 (d, J = 4.9 Hz, 1H), 8.12 (d, J = 7.9 Hz, 1H), 7.38 (s, 1H), 7.25 (m, 1H), 6.91 (dd, J = 4.9, 0.7 Hz, 1H), 4.39 - 4.46 (m, 1H), 4.14 - 4.16 (m, 1H), 3.71 - 3.74 (m, 1H), 3.61 - 3.66 (m, 1H), 2.35 (s, 3H), 2.16 - 2.23 (m, 1H), 1.93 - 1.96 (m, 1H), 1.58 - 1.73 (m, 4H), 1.43 - 1.50 (m, 1H), 0.90 - 0.96 (m, 2H), 0.41 - 0.48 (m, 2H). HRMS C₂₀H₂₂N₄O₂; calc. for (M+H⁺): 351.1743, found: 351.18.

4-[13-chloro-5-(4,4-difluorocyclohexyl)-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7q1**),

4-[5-(1-tert-butoxycarbonylpyrrolidin-3-yl)-13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7q2**)

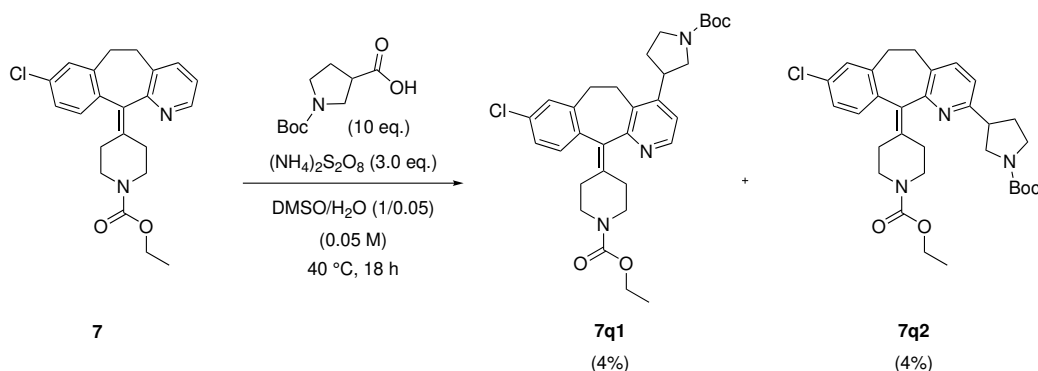


Figure S18: Alkylation of Loratadine (**7**).

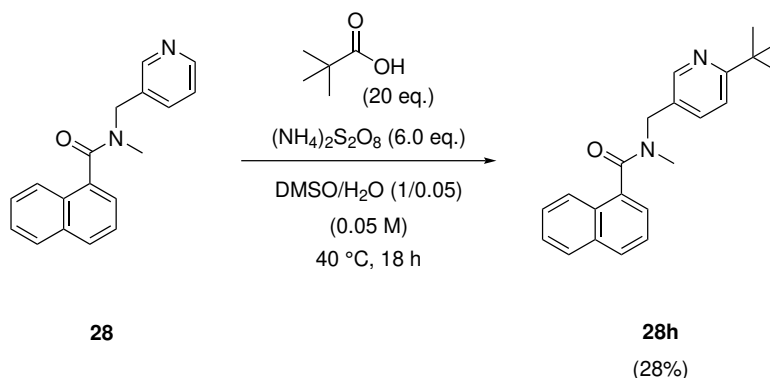
To a solution of 4-(13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(11),3(8),4,6,12,14-hexaen-2-ylidene)piperidine-1-carboxylic acid ethyl ester (**7**, 57.4 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, 1-tert-butoxycarbonylpyrrolidine-3-carboxylic acid (**q**, 322.9 mg, 1.50 mmol, 10.0 eq.) and ammonium persulfate (102.7 mg, 450 μ mol, 3.00 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (50-70-85-100%) in H₂O + 0.1% TEA. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compounds 4-[5-(1-tert-butoxycarbonylpyrrolidin-3-yl)-13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7j1**, 3.73 mg, 4%) as an off-white powder and 4-[5-(1-tert-butoxycarbonylpyrrolidin-3-yl)-13-chloro-4-azatricyclo[9.4.0.03,8]pentadeca-1(15),3,5,7,11,13-hexaen-2-ylidene]piperidine-1-carboxylic acid ethyl ester (**7q2**, 3.69 mg, 4%) as an off-white powder.

7q1:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.37 - 8.38 (m, 1H), 7.13 - 7.14 (m, 3H), 7.05 - 7.06 (m, 1H), 4.15 (d, J = 7.1 Hz, 2H), 3.77 - 3.81 (m, 2H), 3.56 - 3.63 (m, 2H), 3.41 - 3.46 (m, 2H), 3.25 - 3.34 (m, 2H), 3.14 - 3.20 (m, 2H), 3.00 - 3.05 (m, 1H), 2.82 - 2.86 (m, 1H), 2.41 - 2.43 (m, 3H), 2.18 - 2.26 (m, 2H), 1.50 (s, 9H), 1.26 (t, J = 7.1 Hz, 3H). HRMS C₂₈H₃₁ClF₂N₂O₂; calc. for (M+H⁺): 501.2042, found: 501.2.

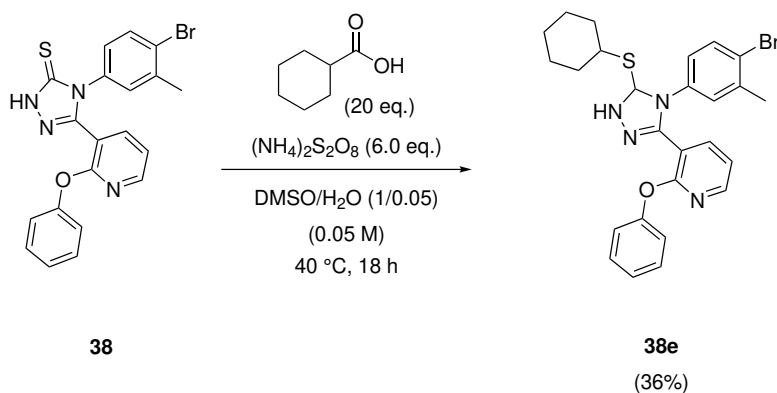
7q2:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.33 - 7.35 (m, 1H), 7.17 - 7.18 (m, 1H), 7.12 - 7.15 (m, 2H), 6.97 - 6.99 (m, 1H), 4.15 (d, J = 7.1 Hz, 2H), 3.73 - 3.81 (m, 3H), 3.55 - 3.64 (m, 1H), 3.43 - 3.49 (m, 2H), 3.36 - 3.41 (m, 2H), 3.28 - 3.32 (m, 1H), 3.14 - 3.26 (m, 2H), 2.76 - 2.86 (m, 2H), 2.49 - 2.53 (m, 1H), 2.34 - 2.38 (m, 1H), 2.28 - 2.30 (m, 2H), 2.19 - 2.23 (m, 1H), 2.05 - 2.15 (m, 1H), 1.49 (s, 9H), 1.27 (t, J = 7.1 Hz, 3H). HRMS C₂₈H₃₁ClF₂N₂O₂; calc. for (M+H⁺): 501.2042, found: 501.2.

N-[(6-tert-butyl-3-pyridyl)methyl]-N-methyl-1-naphthamide (**28h**)Figure S19: Alkylation of Fragment **28**.

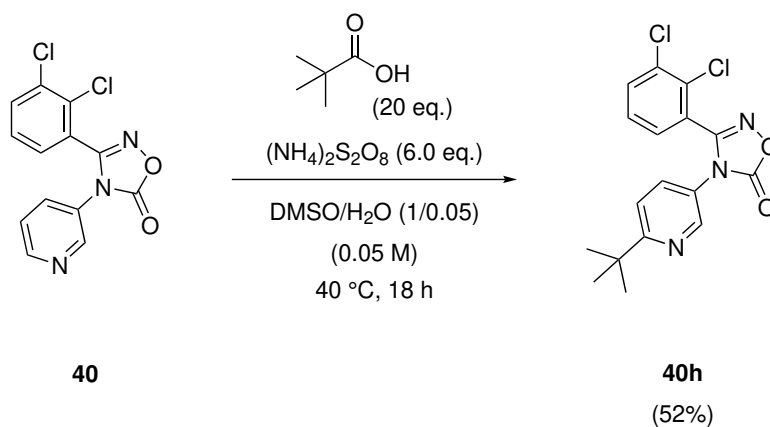
To a solution of N-methyl-N-(3-pyridylmethyl)-1-naphthamide (**28**, 41.5 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H_2O , pivalic acid (**h**, 306.4 mg, 348.2 μ L, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 $^\circ\text{C}$ for 18 hr. The reaction mixture was quenched with NaHCO_3 solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na_2SO_4 , filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (30-50-65-100%) in H_2O + 0.1% TEA. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound N-[(6-tert-butyl-3-pyridyl)methyl]-N-methyl-1-naphthamide (**28h**, 14.2 mg, 28%) as an off-white powder.

^1H NMR (600 MHz, CDCl_3) δ (ppm) 8.62 - 8.63 (m, 1H), 7.88 - 7.90 (m, 4H), 7.86 - 7.91 (m, 2H), 7.42 - 7.47 (m, 3H), 3.17 (s, 2H), 2.75 (s, 3H), 1.41 (s, 10H). **HRMS** $\text{C}_{22}\text{H}_{24}\text{N}_2\text{O}$; calc. for $(\text{M}+\text{H}^+)$: 333.1889, found: 333.19.

3-[4-(4-bromo-3-methyl-phenyl)-5-(cyclohexylthio)-1,2,4-triazol-3-yl]-2-phenoxy-pyridine (**38e**)Figure S20: Alkylation of Fragment **38**.

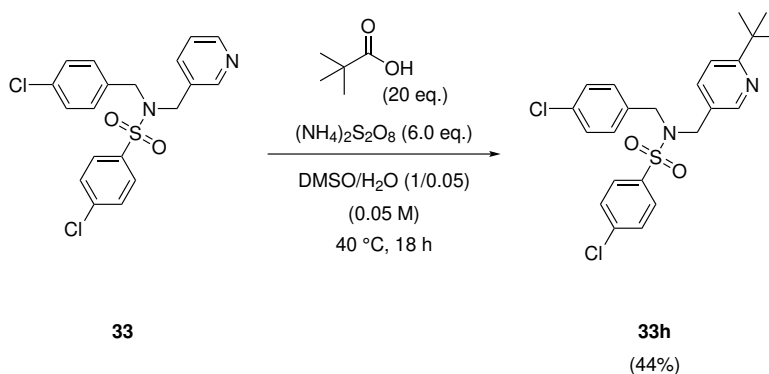
To a solution of 4-(4-bromo-3-methyl-phenyl)-3-(2-phenoxy-3-pyridyl)-1H-1,2,4-triazole-5-thione (**38**, 41.5 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, cyclohexanecarboxylic acid (**e**, 384.5 mg, 373.31 μ L, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (YMC-Triart C18, 12 mm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-98%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound 3-[4-(4-bromo-3-methyl-phenyl)-5-(cyclohexylthio)-1,2,4-triazol-3-yl]-2-phenoxy-pyridine (**38e**, 28.5 mg, 36%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.18 (dd, J = 4.9, 2.0 Hz, 1H), 8.11 - 8.15 (m, 1H), 7.48 (d, J = 8.4 Hz, 1H), 7.28 - 7.29 (m, 1H), 7.26 - 7.27 (m, 1H), 7.14 - 7.16 (m, 1H), 7.11 - 7.13 (m, 1H), 7.02 (d, J = 2.6 Hz, 1H), 6.77 (d, J = 8.4 Hz, 1H), 6.47 (d, J = 7.7 Hz, 2H), 3.86 - 3.91 (m, 1H), 2.19 (s, 3H), 2.16 - 2.18 (m, 1H), 1.73 - 1.77 (m, 2H), 1.61 - 1.64 (m, 2H), 1.49 - 1.52 (m, 2H), 1.41 - 1.49 (m, 3H), 1.25 - 1.30 (s, 1H). HRMS C₂₂H₂₄N₂O; calc. for (M+H⁺): 523.1089, found: 523.09.

4-(6-tert-butyl-3-pyridyl)-3-(2,3-dichlorophenyl)-1,2,4-oxadiazol-5-one (**40h**)Figure S21: Alkylation of Fragment **40**.

To a solution of 3-(2,3-dichlorophenyl)-4-(3-pyridyl)-1,2,4-oxadiazol-5-one (**40**, 46.2 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, pivalic acid (**h**, (306.4 mg, 348.2 μ L, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (YMC-Triart C18, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-98%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound 4-(6-tert-butyl-3-pyridyl)-3-(2,3-dichlorophenyl)-1,2,4-oxadiazol-5-one (**40h**, 28.6 mg, 52%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.24 (dd, J = 2.6, 0.7 Hz, 1H), 7.60 (dd, J = 8.7, 2.6 Hz, 1H), 7.49 (d, J = 1.6 Hz, 1H), 7.38 - 7.42 (m, 2H), 1.33 (s, 9H). HRMS C₁₇H₁₅Cl₂N₃O₂; calc. for (M+H⁺): 364.0541, found: 364.06.

N-[(6-tert-butyl-3-pyridyl)methyl]-4-chloro-N-(4-chlorobenzyl)benzenesulfonamide (**33h**)Figure S22: Alkylation of Fragment **33**.

To a solution of 4-chloro-N-(4-chlorobenzyl)-N-(3-pyridylmethyl)benzenesulfonamide (**33**, 61.1 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, pivalic acid (**h**, (306.4 mg, 348.2 μ L, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-98-100%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound N-[(6-tert-butyl-3-pyridyl)methyl]-4-chloro-N-(4-chlorobenzyl)benzenesulfonamide (**40h**, 30.8 mg, 44%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.17 - 8.19 (m, 1H), 7.77 (d, J = 8.6 Hz, 2H), 7.51 (d, J = 8.1 Hz, 2H), 7.36 - 7.38 (m, 1H), 7.16 (d, J = 8.4 Hz, 3H), 7.02 (d, J = 7.9 Hz, 2H), 4.29 (s, 2H), 4.28 (s, 2H), 1.32 (s, 9H). HRMS C₂₃H₂₄Cl₂N₂O₂S; calc. for (M+H⁺): 463.0936, found: 463.10.

N-(4-chlorophenyl)-2-cyclobutyl-6-(phenylthio)-4-(trifluoromethyl)nicotinamide (**37b1**),
 N-(4-chlorophenyl)-5-cyclobutyl-6-(phenylthio)-4-(trifluoromethyl)nicotinamide (**37b2**)

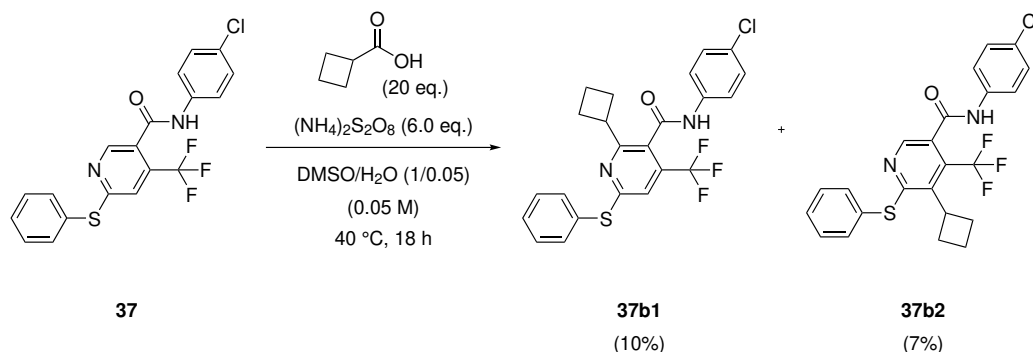
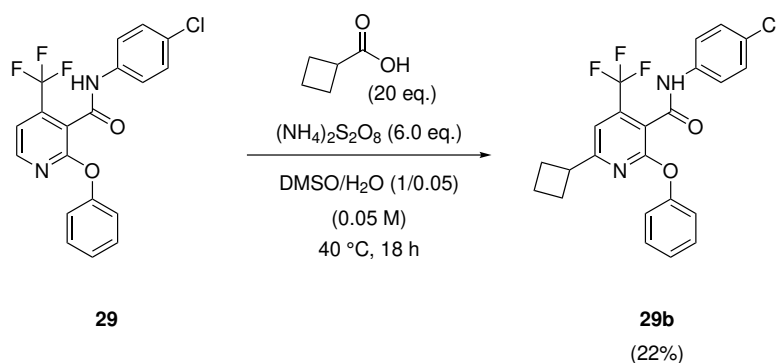


Figure S23: Alkylation of fragment **37**.

To a solution of N-(4-chlorophenyl)-6-(phenylthio)-4-(trifluoromethyl)nicotinamide (**37**, 61.3 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H_2O , cyclobutanecarboxylic acid (**b**, 300.4 mg, 3.0 mmol, 10.0 eq.) and ammonium persulfate (205.4 mg, 900 μ mol, 3.00 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 $^\circ\text{C}$ for 18 hr. The reaction mixture was quenched with NaHCO_3 solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na_2SO_4 , filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (YMC-Triart C18, 12 mm, 5 μ m, 100 x 30 mm) using a MeCN gradient (50-20-98%) in H_2O + 0.1% TEA. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compounds N-(4-chlorophenyl)-2-cyclobutyl-6-(phenylthio)-4-(trifluoromethyl)nicotinamide (**37b1**, 11.5 mg, 10%) and N-(4-chlorophenyl)-5-cyclobutyl-6-(phenylthio)-4-(trifluoromethyl)nicotinamide (**37b2**, 11.5 mg, 7%) as a mixture (60:40) in an off-white powder.

37b1 & 37b2:

^1H NMR (600 MHz, CDCl_3) δ (ppm) 8.31 - 8.35 (m, 1H), 8.11 - 8.13 (m, 1H), 7.64 - 7.67 (m, 2H), 7.51 - 7.54 (m, 2H), 7.50 - 7.55 (m, 2H), 7.50 - 7.52 (m, 3H), 7.45 (d, $J = 3.8$ Hz, 1H), 7.35 - 7.39 (m, 3H), 7.33 (s, 1H), 7.21 - 7.26 (m, 1H), 7.01 (s, 1H), 4.03 - 4.17 (m, 1H), 3.74 - 3.86 (m, 1H), 2.77 - 2.87 (m, 1H), 2.54 - 2.65 (m, 1H), 2.31 - 2.46 (m, 3H), 2.14 - 2.25 (m, 2H), 2.08 - 2.16 (m, 1H), 1.98 - 2.04 (m, 1H), 1.92 - 1.97 (m, 1H), 1.74 - 1.84 (m, 1H). HRMS $\text{C}_{23}\text{H}_{18}\text{ClF}_3\text{N}_2\text{OS}$; calc. for $(\text{M}+\text{H}^+)$: 463.0780, found: 463.08.

N-(4-chlorophenyl)-6-cyclobutyl-2-phenoxy-4-(trifluoromethyl)nicotinamide (**29b**)Figure S24: Alkylation of Fragment **29**.

To a solution of N-(4-chlorophenyl)-2-phenoxy-4-(trifluoromethyl)nicotinamide (**29**, 58.9 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, cyclobutanecarboxylic acid (**b**, 300.4 mg, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (YMC-Triart C18, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-98%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound N-(4-chlorophenyl)-6-cyclobutyl-2-phenoxy-4-(trifluoromethyl)nicotinamide (**29b**, 15.0 mg, 22%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.55 - 7.58 (m, 2H), 7.33 - 7.36 (m, 2H), 7.18 - 7.24 (m, 3H), 7.13 (s, 1H), 3.56 (t, J = 8.4 Hz, 1H), 2.22 - 2.28 (m, 2H), 2.11 - 2.16 (m, 2H), 1.92 - 2.00 (m, 1H), 1.77 - 1.82 (m, 1H).
HRMS C₂₃H₁₈ClF₃N₂O₂; calc. for (M+H⁺): 447.1009, found: 447.10.

6-cyclobutyl-N-[keto-methyl-(p-tolyl)persulfuranylidene]-2-phenoxy-nicotinamide (**34b1**),
4-cyclobutyl-N-[keto-methyl-(p-tolyl)persulfuranylidene]-2-phenoxy-nicotinamide (**34b2**)

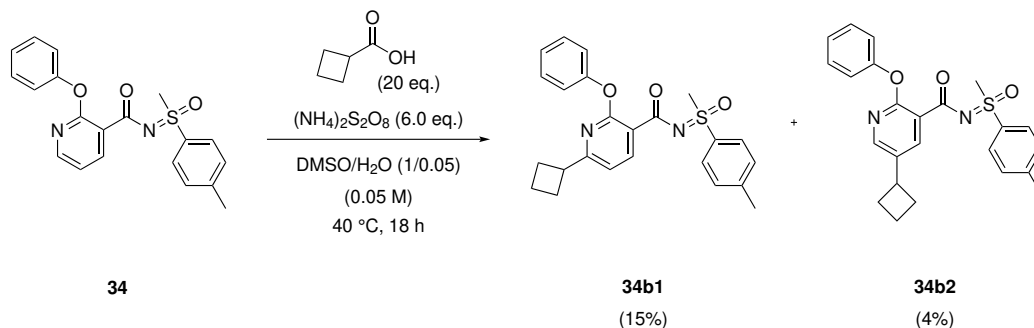


Figure S25: Alkylation of fragment **34**.

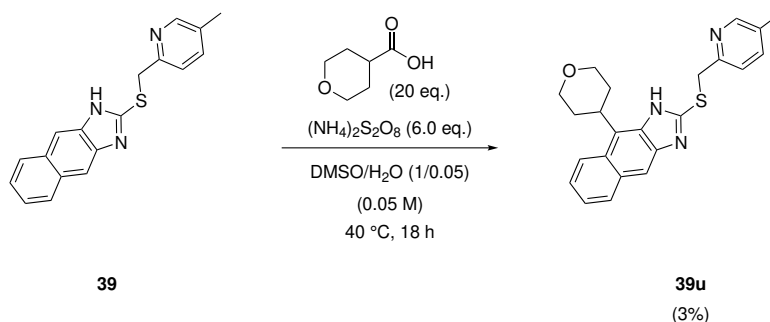
To a solution of N-[keto-methyl-(p-tolyl)persulfuranylidene]-2-phenoxy-nicotinamide (**34**, 55.0 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, cyclobutanecarboxylic acid (**b**, 300.4 mg, 3.0 mmol, 10.0 eq.) and ammonium persulfate (205.4 mg, 900 μ mol, 3.00 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (20-98-100%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compounds 6-cyclobutyl-N-[keto-methyl-(p-tolyl)persulfuranylidene]-2-phenoxy-nicotinamide (**34b1**, 9.7 mg, 15%) as an off-white powder and 4-cyclobutyl-N-[keto-methyl-(p-tolyl)persulfuranylidene]-2-phenoxy-nicotinamide (**34b2**, 2.7 mg, 4%) as an off-white powder.

34b1:

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.25 (d, J = 7.7 Hz, 1H), 7.89 (d, J = 8.4 Hz, 1H), 7.32 - 7.37 (m, 4H), 7.13 - 7.15 (m, 3H), 6.93 (d, J = 7.7 Hz, 1H), 3.49 - 3.55 (m, 1H), 3.36 (s, 3H), 2.44 (s, 3H), 2.16 - 2.23 (m, 4H), 1.89 - 1.97 (m, 1H), 1.76 - 1.81 (m, 1H). HRMS C₂₄H₂₄N₂O₃S; calc. for (M+H⁺): 421.1508, found: 421.15.

34b2:

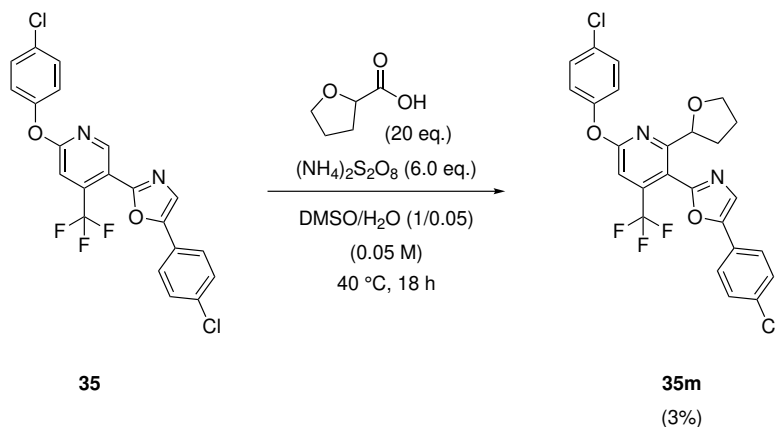
¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.09 (d, J = 5.3 Hz, 1H), 7.91 (d, J = 8.4 Hz, 2H), 7.40 - 7.44 (m, 2H), 7.28 - 7.29 (m, 2H), 7.17 - 7.21 (m, 3H), 7.02 (dd, J = 5.3, 0.7 Hz, 1H), 3.86 - 3.92 (m, 1H), 3.35 (s, 3H), 2.42 (s, 3H), 2.21 - 2.24 (m, 2H), 2.02 - 2.07 (m, 1H), 1.83 - 1.87 (m, 1H). HRMS C₂₄H₂₄N₂O₃S; calc. for (M+H⁺): 421.1508, found: 421.15.

2-[(5-methyl-2-pyridyl)methylthio]-4-tetrahydropyran-4-yl-3H-benzo[f]benzimidazole (**39u**)Figure S26: Alkylation of Fragment **39**.

To a solution of 2-[(5-methyl-2-pyridyl)methylthio]-1H-benzo[f]benzimidazole (**39**, 45.8 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μL H₂O, tetrahydropyran-4-carboxylic acid (**u**, 390.4 mg, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μm , 100 x 30 mm) using a MeCN gradient (20-50-55-100%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound 2-[(5-methyl-2-pyridyl)methylthio]-4-tetrahydropyran-4-yl-3H-benzo[f]benzimidazole (**39u**, 2.0 mg, 3%) as an off-white powder.

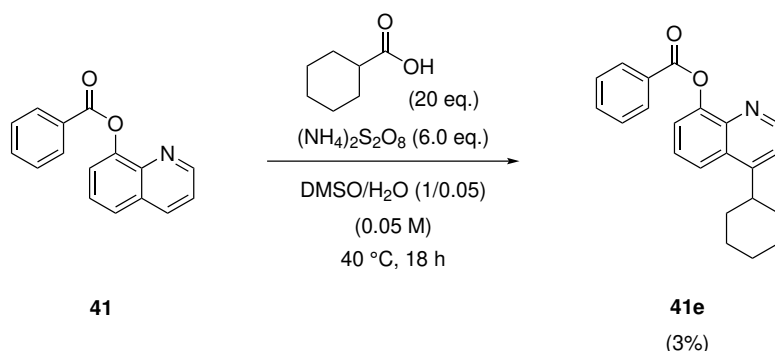
¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.83 (s, 1H), 8.30 - 8.31 (m, 1H), 7.99 - 8.03 (m, 2H), 7.61 - 7.63 (m, 1H), 7.46 - 7.48 (m, 1H), 7.39 - 7.41 (m, 1H), 7.30 - 7.31 (m, 1H), 4.40 (s, 2H), 4.32 - 4.31 (m, 2H), 4.04 - 4.08 (m, 1H), 3.81 - 3.85 (m, 2H), 2.68 - 2.74 (m, 2H), 2.41 - 2.43 (m, 3H), 1.89 - 1.91 (m, 2H). **¹³C NMR (151 MHz, CDCl₃)** δ (ppm) 138.8, 130.9, 129.5, 127.4, 124.1, 123.1, 123.0, 114.1, 60.0, 37.4, 31.0, 18.5. **HRMS** C₂₃H₂₃N₃OS; calc. for (M+H⁺): 390.1562, found: 390.16.

2-[6-(4-chlorophenoxy)-2-(trifluoromethyl)-3-pyridyl]-5-(4-chlorophenyl)oxazole (35m)

Figure S27: Alkylation of Fragment **35**.

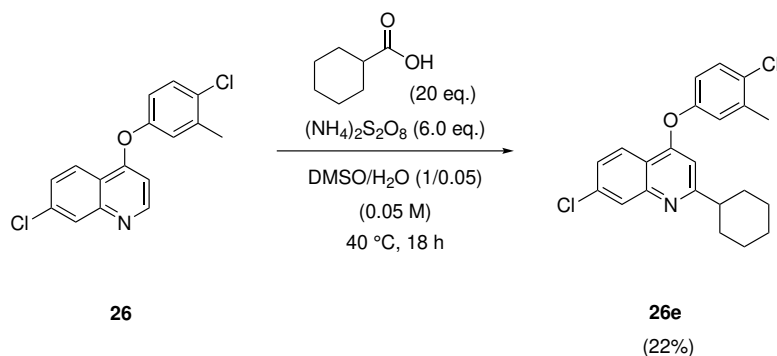
To a solution of 2-[6-(4-chlorophenoxy)-4-(trifluoromethyl)-3-pyridyl]-5-(4-chlorophenyl)oxazole (**35**, 67.7 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μL H₂O, tetrahydrofuran-2-carboxylic acid (**m**, 348.5 mg, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μm , 100 x 30 mm) using a MeCN gradient (60-80-95-100%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound 2-[6-(4-chlorophenoxy)-2-(tetrahydrofuran-2-yl)-4-(trifluoromethyl)-3-pyridyl]-5-(4-chlorophenyl)oxazole (**35m**, 2.0 mg, 3%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 7.60 (d, J = 8.1 Hz, 2H), 7.50 (s, 1H), 7.40 - 7.43 (m, 4H), 7.24 (s, 1H), 7.16 (d, J = 8.4 Hz, 2H), 4.99 (dd, J = 7.6, 5.5 Hz, 1H), 3.76 (td, J = 7.6, 5.7 Hz, 1H), 3.68 (q, J = 7.3 Hz, 1H), 2.02 - 2.08 (m, 1H), 2.01 - 2.18 (m, 1H), 1.79 - 1.87 (m, 2H). **¹³C NMR (151 MHz, CDCl₃)** δ (ppm) 163.9, 163.8, 155.3, 151.6, 151.4, 134.7, 130.7, 129.6, 129.3, 126.1, 125.6, 123.3, 123.0, 114.1, 107.6, 78.0, 69.7, 34.3, 32.3, 30.3, 29.7, 26.0. **HRMS** C₂₅H₁₇Cl₂F₃N₂O₃; calc. for (M+H⁺): 521.0568, found: 521.06.

Benzoic acid (4-cyclohexyl-8-quinolyl) ester (**41e**)Figure S28: Alkylation of Fragment **41**.

To a solution of Benzoic acid 8-quinolyl ester (**41**, 67.7 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H_2O , cyclohexanecarboxylic acid (**e**, 384.5 mg, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 $^\circ\text{C}$ for 18 hr. The reaction mixture was quenched with NaHCO_3 solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na_2SO_4 , filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μm , 100 x 30 mm) using a MeCN gradient (60-80-95-100%) in H_2O + 0.1% HCOOH . The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound Benzoic acid (4-cyclohexyl-8-quinolyl) ester (**41e**, 1.5 mg, 3%) as an off-white powder.

^1H NMR (600 MHz, CDCl_3) δ (ppm) 8.81 (d, $J = 4.5$ Hz, 1H), 8.36 (dd, $J = 8.4, 1.3$ Hz, 2H), 8.04 - 8.06 (m, 1H), 7.65 - 7.68 (m, 1H), 7.58 - 7.61 (m, 1H), 7.54 - 7.56 (m, 1H), 7.53 - 7.57 (m, 2H), 7.30 - 7.31 (m, 1H), 3.31 - 3.39 (m, 1H), 2.02 - 2.07 (m, 2H), 1.93 - 1.99 (m, 2H), 1.84 - 1.90 (m, 1H), 1.54 - 1.59 (m, 6H). **^{13}C NMR (151 MHz, CDCl_3)** δ (ppm) 165.6, 153.4, 150.6, 148.4, 141.6, 133.5, 130.6, 129.7, 128.5, 125.6, 121.4, 120.9, 118.1, 39.2, 33.7, 27.0, 26.4. **HRMS** $\text{C}_{22}\text{H}_{21}\text{NO}_2$; calc. for $(\text{M}+\text{H}^+)$: 332.1572, found: 332.16.

7-chloro-4-(4-chloro-3-methyl-phenoxy)-2-cyclohexyl-quinoline (**26e**)Figure S29: Alkylation of Fragment **26**.

To a solution of 7-chloro-4-(4-chloro-3-methyl-phenoxy)quinoline (**26**, 45.6 mg, 0.15 mmol, 1.00 eq.) in 3 mL degassed DMSO and 5 μ L H₂O, cyclohexanecarboxylic acid (**e**, 384.5 mg, 3.0 mmol, 20.0 eq.) and ammonium persulfate (205.4 mg, 0.9 mmol, 6.0 eq.) were added. The reaction mixture was degassed while bubbling nitrogen through it. The reaction mixture was stirred at 40 °C for 18 hr. The reaction mixture was quenched with NaHCO₃ solution and extracted with DCM. The combined organic layers were washed with water, and brine, dried over Na₂SO₄, filtered and concentrated to dryness. The crude material was purified by reversed-phase HPLC (Gemini NX, 12 nm, 5 μ m, 100 x 30 mm) using a MeCN gradient (60-80-95-100%) in H₂O + 0.1% HCOOH. The solvent was removed from product containing fractions. Evaporation of solvents gave the title compound 7-chloro-4-(4-chloro-3-methyl-phenoxy)-2-cyclohexyl-quinoline (**26e**, 12.9 mg, 22%) as an off-white powder.

¹H NMR (600 MHz, CDCl₃) δ (ppm) 8.18 (d, J = 8.9 Hz, 1H), 8.05 (d, J = 2.1 Hz, 1H), 7.45 (dd, J = 8.9, 2.1 Hz, 1 H), 7.44 (d, J = 8.6 Hz, 1 H), 7.07 - 7.08 (m, 1 H), 6.94 - 6.96 (m, 1 H), 6.46 (s, 1 H), 2.72 - 2.76 (m, 1 H), 2.43 (s, 3 H), 1.86 - 1.92 (m, 2 H), 1.83 - 1.85 (m, 2 H), 1.73 - 1.75 (m, 1 H), 1.46 - 1.50 (m, 2 H), 1.38 - 1.41 (m, 2 H), 1.25 - 1.31 (m, 1 H). **HRMS** C₂₂H₂₁Cl₂NO; calc. for (M+H⁺): 386.1000, found: 386.10.

Supplementary References

1. Fey, M. & Lenssen, J. E. *Fast Graph Representation Learning with PyTorch Geometric in ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).
2. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, 8026–8037 (2019).
3. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).
4. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
5. Nippa, D. F. *et al.* Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning (2022).
6. Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J. & Schneider, G. Δ -Quantum machine-learning for medicinal chemistry. *Phys. Chem. Chem. Phys.* **24**, 10775–10783 (2022).
7. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 1–11 (2022).
8. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. *QMugs: Quantum Mechanical Properties of Drug-like Molecules* en. 2021.
9. Neeser, R., Isert, C., Stuyver, T., Schneider, G. & Coley, C. Qmugs 1.1: Quantum Mechanical Properties of Organic Compounds Commonly Encountered in Reactivity Datasets. *SSRN 4363768* (2023).
10. Proctor, R., Chuentragool, P., Colgan, A. & Phipps, R. Hydrogen Atom Transfer-Driven Enantioselective Minisci Reaction of Amides. *J. Am. Chem. Soc.* **143**, 4928–4934 (2021).
11. Reid, J., Proctor, R., Sigman, M. & Phipps, R. Predictive Multivariate Linear Regression Analysis Guides Successful Catalytic Enantioselective Minisci Reactions of Diazines. *J. Am. Chem. Soc.* **141**, 19178–19185 (2019).
12. Bieszczad, B., Perego, L. & Melchiorre, P. Photochemical C-H Hydroxyalkylation of Quinolines and Isoquinolines. *Angew. Chem. Int. Ed.* **58**, 16878–16883 (2019).
13. Chen, X. *et al.* Histidine-Specific Peptide Modification via Visible-Light-Promoted C-H Alkylation. *J. Am. Chem. Soc.* **141**, 18230–18237 (2019).
14. Fu, M.-C., Shang, R., Zhao, B., Wang, B. & Fu, Y. Photocatalytic decarboxylative alkylations mediated by triphenylphosphine and sodium iodide. *Science* **363**, 1429–1434 (2019).
15. Wang, Z., Ji, X., Zhao, J. & Huang, H. Visible-light-mediated photoredox decarbonylative Minisci-type alkylation with aldehydes under ambient air conditions. *Green Chem.* **21**, 5512–5516 (2019).
16. Dong, J. *et al.* Visible-light-mediated Minisci C-H alkylation of heteroarenes with unactivated alkyl halides using O₂ as an oxidant. *Chem. Sci.* **10**, 976–982 (2019).
17. Li, G.-X., Hu, X., He, G. & Chen, G. Photoredox-Mediated Minisci-type Alkylation of N-Heteroarenes with Alkanes with High Methylene Selectivity. *ACS Catalysis* **8**, 11847–11853 (2018).
18. Proctor, R., Davis, H. & Phipps, R. Catalytic enantioselective Minisci-type addition to heteroarenes. *Science* **360**, 419–422 (2018).
19. Nuhant, P. *et al.* Visible-Light-Initiated Manganese Catalysis for C-H Alkylation of Heteroarenes: Applications and Mechanistic Studies. *Angew. Chem. Int. Ed.* **56**, 15309–15313 (2017).
20. Liu, P., Liu, W. & Li, C.-J. Catalyst-Free and Redox-Neutral Innate Trifluoromethylation and Alkylation of Aromatics Enabled by Light. *J. Am. Chem. Soc.* **139**, 14315–14321 (2017).
21. Ermanis, K. *et al.* A Computational and Experimental Investigation of the Origin of Selectivity in the Chiral Phosphoric Acid Catalyzed Enantioselective Minisci Reaction. *J. Am. Chem. Soc.* **142**, 21091–21101 (2020).
22. Matsui, J., Primer, D. & Molander, G. Metal-free C-H alkylation of heteroarenes with alkyltrifluoroborates: A general protocol for 1°, 2° and 3° alkylation. *Chem. Sci.* **8**, 3512–3522 (2017).
23. Li, G.-X. *et al.* Photoredox-mediated Minisci C-H alkylation of N-heteroarenes using boronic acids and hypervalent iodine. *Chem. Sci.* **7**, 6407–6412 (2016).
24. Jin, J. & MacMillan, D. Direct α -arylation of ethers through the combination of photoredox-mediated C-H functionalization and the minisci reaction. *Angew. Chem. Int. Ed.* **54**, 1565–1569 (2015).
25. Graham, M. *et al.* Development and Proof of Concept for a Large-Scale Photoredox Additive-Free Minisci Reaction. *Organic Process Research and Development* **25**, 57–67 (2021).
26. Dong, J., Yue, F., Song, H., Liu, Y. & Wang, Q. Visible-light-mediated photoredox minisci C-H alkylation with alkyl boronic acids using molecular oxygen as an oxidant. *Chem. Commun.* **56**, 12652–12655 (2020).

27. Rammal, F. *et al.* Visible-Light-Mediated C-H Alkylation of Pyridine Derivatives. *Org. Lett.* **22**, 7671–7675 (2020).
28. Ikarashi, G., Morofuji, T. & Kano, N. Terminal-oxidant-free photocatalytic C-H alkylations of heteroarenes with alkylsilicates as alkyl radical precursors. *Chem. Commun.* **56**, 10006–10009 (2020).
29. Dong, J., Wang, X., Song, H., Liu, Y. & Wang, Q. Photoredox-Catalyzed Redox-Neutral Minisci C-H Formylation of N-Heteroarenes. *Advanced Synthesis and Catalysis* **362**, 2155–2159 (2020).
30. Li, T. *et al.* Three-Component Minisci Reaction with 1,3-Dicarbonyl Compounds Induced by Visible Light. *Org. Lett.* **22**, 2386–2390 (2020).
31. Zidan, M., Morris, A., McCallum, T. & Barriault, L. The Alkylation and Reduction of Heteroarenes with Alcohols Using Photoredox Catalyzed Hydrogen Atom Transfer via Chlorine Atom Generation. *Eur. J. Org. Chem.* **2020**, 1453–1458 (2020).
32. Ji, X. *et al.* LiBr-promoted photoredox neutral Minisci hydroxyalkylations of quinolines with aldehydes. *Green Chem.* **22**, 8233–8237 (2020).
33. Perkins, J., Schubert, J., Streckfuss, E., Balsells, J. & ElMarrouni, A. Photoredox Catalysis for Silyl-Mediated C–H Alkylation of Heterocycles with Non-Activated Alkyl Bromides. *Eur. J. Org. Chem.* **2020**, 1515–1522 (2020).
34. Jian, Y., Chen, M., Yang, C. & Xia, W.-J. Minisci-Type C–H Cyanoalkylation of Heteroarenes Through N–O/C–C Bonds Cleavage. *Eur. J. Org. Chem.* **2020**, 1439–1442 (2020).
35. Li, X. *et al.* Complementary oxidative generation of iminyl radicals from α -imino-oxy acids: Silver-catalyzed c-h cyanoalkylation of heterocycles and quinones. *Journal of Organic Chemistry* **85**, 2504–2511 (2020).
36. Laha, J., Kaur Hunjan, M., Hegde, S. & Gupta, A. Aroylation of Electron-Rich Pyrroles under Minisci Reaction Conditions. *Org. Lett.* **22**, 1442–1447 (2020).
37. Xie, X., Zhang, Y., Hao, J. & Wan, W. Ag-Catalyzed minisci C-H difluoromethylarylation of N-heteroarenes. *Org. Biomol. Chem.* **18**, 400–404 (2020).
38. Wang, Z., Ji, X., Han, T., Deng, G.-J. & Huang, H. LiBr-Promoted Photoredox Minisci-Type Alkylations of Quinolines with Ethers. *Advanced Synthesis and Catalysis* **361**, 5643–5647 (2019).
39. Vijeta, A. & Reisner, E. Carbon nitride as a heterogeneous visible-light photocatalyst for the Minisci reaction and coupling to H₂ production. *Chem. Commun.* **55**, 14007–14010 (2019).
40. Dou, G.-Y., Jiang, Y.-Y., Xu, K. & Zeng, C.-C. Electrochemical Minisci-type trifluoromethylation of electron-deficient heterocycles mediated by bromide ions. *Organic Chemistry Frontiers* **6**, 2392–2397 (2019).
41. Sutherland, D. R., Veguillas, M., Oates, C. L. & Lee, A.-L. Metal-, photocatalyst-, and light-free, late-stage C–H alkylation of heteroarenes and 1, 4-quinones using carboxylic acids. *Org. Lett.* **20**, 6863–6867 (2018).
42. Bosset, C. *et al.* Minisci-Photoredox-Mediated α -Heteroarylation of N-Protected Secondary Amines: Remarkable Selectivity of Azetidines. *Org. Lett.* **20**, 6003–6006 (2018).
43. Fuse, H. *et al.* Photocatalytic redox-neutral hydroxyalkylation of: N -heteroaromatics with aldehydes. *Chem. Sci.* **11**, 12206–12211 (2020).
44. Sherwood, T., Li, N., Yazdani, A. & Dhar, T. Organocatalyzed, Visible-Light Photoredox-Mediated, One-Pot Minisci Reaction Using Carboxylic Acids via N-(Acyloxy)phthalimides. *Journal of Organic Chemistry* **83**, 3000–3012 (2018).
45. Zhang, L. & Liu, Z.-Q. Molecular Oxygen-Mediated Minisci-Type Radical Alkylation of Heteroarenes with Boronic Acids. *Org. Lett.* **19**, 6594–6597 (2017).
46. Galloway, J., Mai, D. & Baxter, R. Silver-Catalyzed Minisci Reactions Using Selectfluor as a Mild Oxidant. *Org. Lett.* **19**, 5772–5775 (2017).
47. Wang, Q.-Q. *et al.* Electrocatalytic Minisci Acylation Reaction of N-Heteroarenes Mediated by NH₄I. *Org. Lett.* **19**, 5517–5520 (2017).
48. Tang, R.-J., Kang, L. & Yang, L. Metal-Free Oxidative Decarbonylative Coupling of Aliphatic Aldehydes with Azaarenes: Successful Minisci-Type Alkylation of Various Heterocycles. *Advanced Synthesis and Catalysis* **357**, 2055–2060 (2015).
49. Siddaraju, Y., Lamani, M. & Prabhu, K. A transition metal-free Minisci reaction: Acylation of isoquinolines, quinolines, and quinoxaline. *Journal of Organic Chemistry* **79**, 3856–3865 (2014).
50. Stephenson, C., McClain, E., Monos, T., Mori, M. & Beatty, J. Design and implementation of a catalytic electron donor-acceptor complex platform for radical trifluoromethylation and alkylation. *ACS Catalysis* **10**, 12636–12641 (2020).

51. Wang, Q., Duan, J., Tang, P., Chen, G. & He, G. Synthesis of non-classical heteroaryl C-glycosides via Minisci-type alkylation of N-heteroarenes with 4-glycosyl-dihydropyridines. *Science China Chemistry* **63**, 1613–1618 (2020).
52. Dong, J. *et al.* Visible-light-mediated minisci C-H alkylation of heteroarenes with 4-alkyl-1,4-dihydropyridines using O₂ as an oxidant. *Green Chem.* **22**, 5599–5604 (2020).
53. Wang, Z., Liu, Q., Ji, X., Deng, G.-J. & Huang, H. Bromide-promoted visible-light-induced reductive minisci reaction with aldehydes. *ACS Catalysis* **10**, 154–159 (2020).
54. Ding, H., Xu, K. & Zeng, C.-C. Nickel-catalyzed electrochemical Minisci acylation of aromatic N-heterocycles with α -keto acids via ligand-to-metal electron transfer pathway. *Journal of Catalysis* **381**, 38–43 (2020).
55. Miyaura, N. & Suzuki, A. Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chem. Rev.* **95**, 2457–2483 (1995).
56. Nicolaou, K., Bulger, P. G. & Sarlah, D. Palladium-catalyzed cross-coupling reactions in total synthesis. *Angew. Chem. Int. Ed.* **44**, 4442–4489 (2005).

Supplementary Data:
Identifying opportunities for late-stage C-H alkylation with *in silico*
reaction screening and high-throughput experimentation

David F. Nippa^{1,2,†}, Kenneth Atz^{1,†}, Alex T. Müller¹, Jens Wolfard¹, Clemens Isert³,
Martin Binder¹, Oliver Scheidegger¹, David B. Konrad^{2,*}, Uwe Grether^{1,*},
Rainer E. Martin^{1,*} & Gisbert Schneider^{3,*}

¹Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd.,
Grenzacherstrasse 124, 4070 Basel, Switzerland.

²Department of Pharmacy, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, 81377 Munich, Germany.

³ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

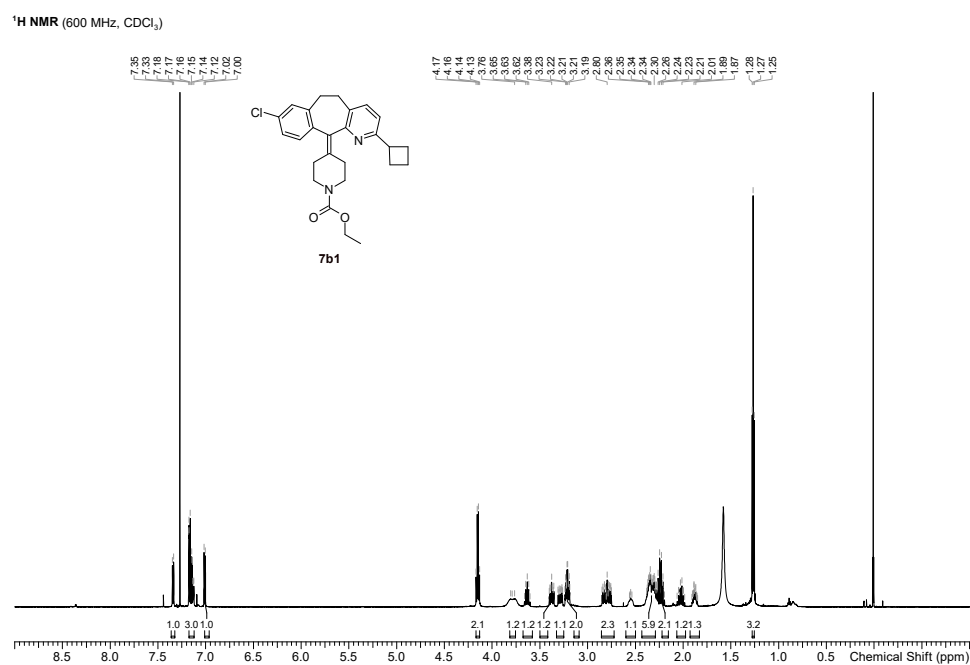
⁴ETH Singapore SEC Ltd, 1 CREATE Way, #06-01 CREATE Tower, Singapore, Singapore.

† These authors contributed equally to this work.

* To whom correspondence should be addressed.

E-mail: david.konrad@cup.lmu.de, uwe.grether@roche.com, rainer_e.martin@roche.com, gisbert@ethz.ch

1 NMR spectra

Figure 1: **7b1**, ¹H-NMR spectrum.

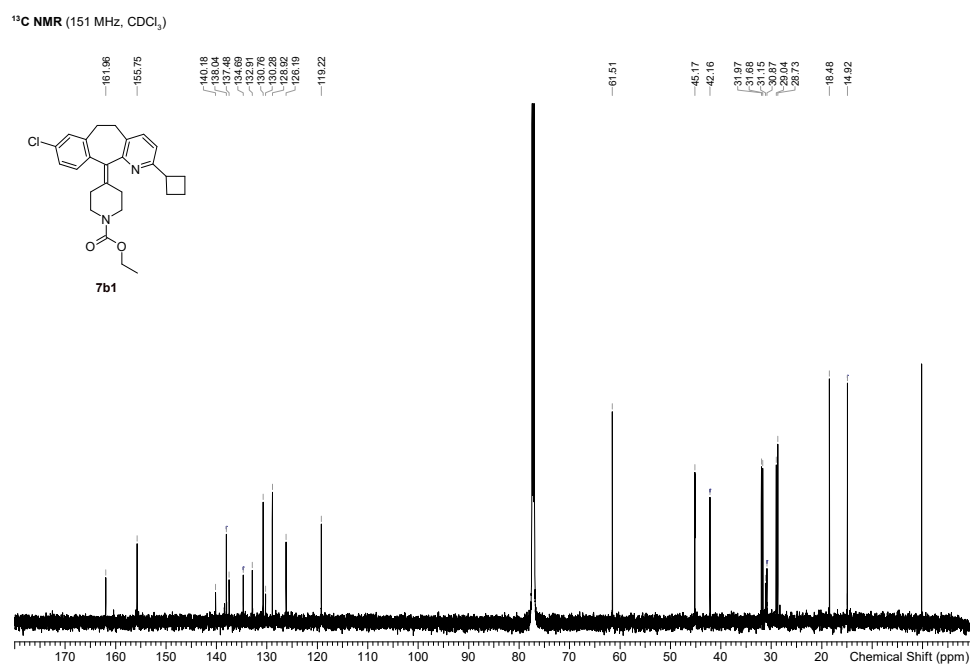


Figure 2: **7b1**, ¹³C-NMR spectrum.

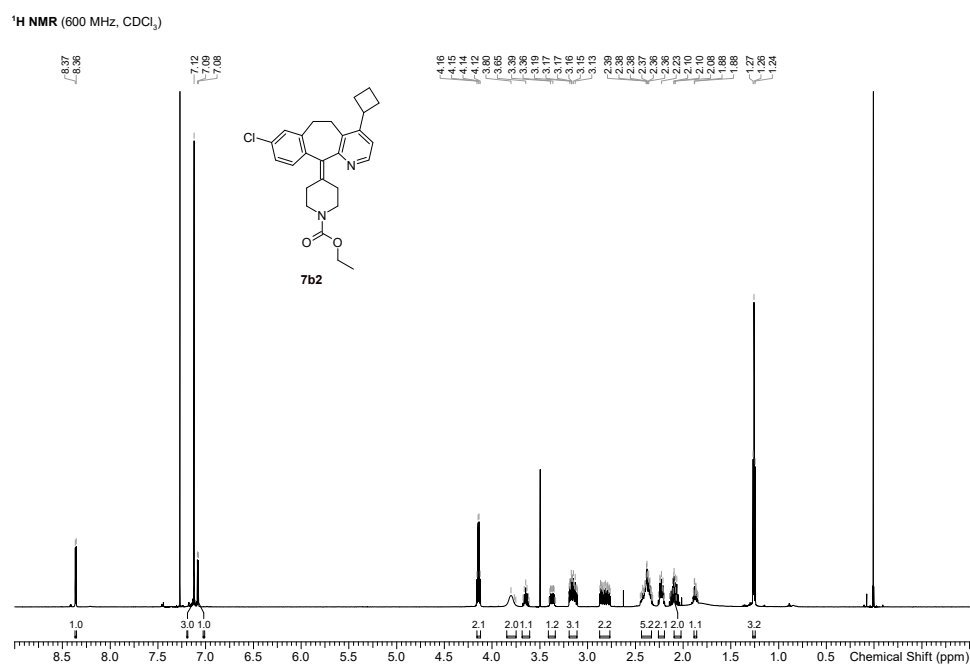


Figure 3: **7b2**, ¹H-NMR spectrum.

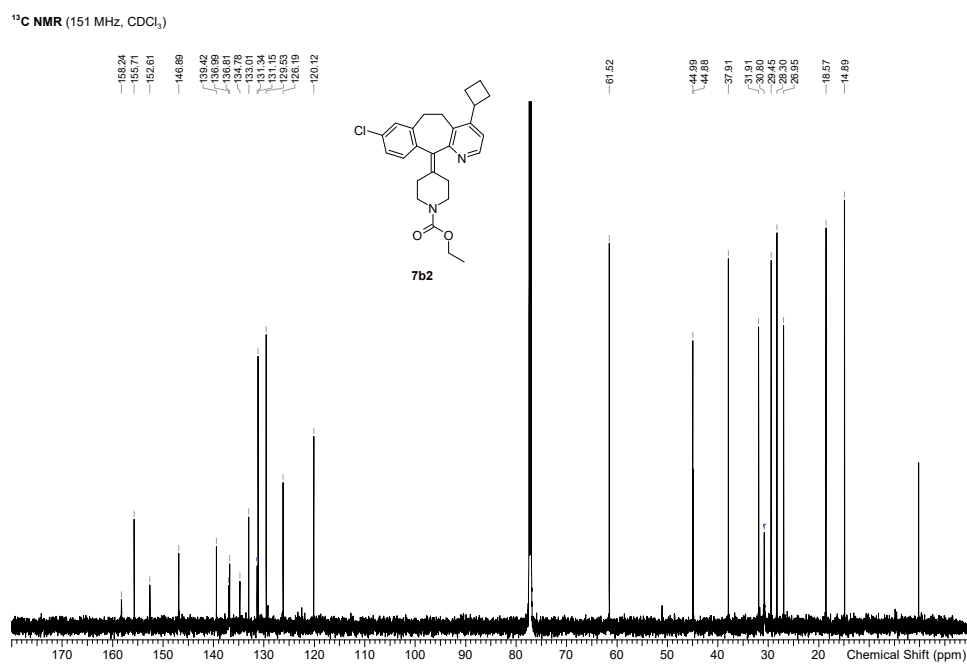
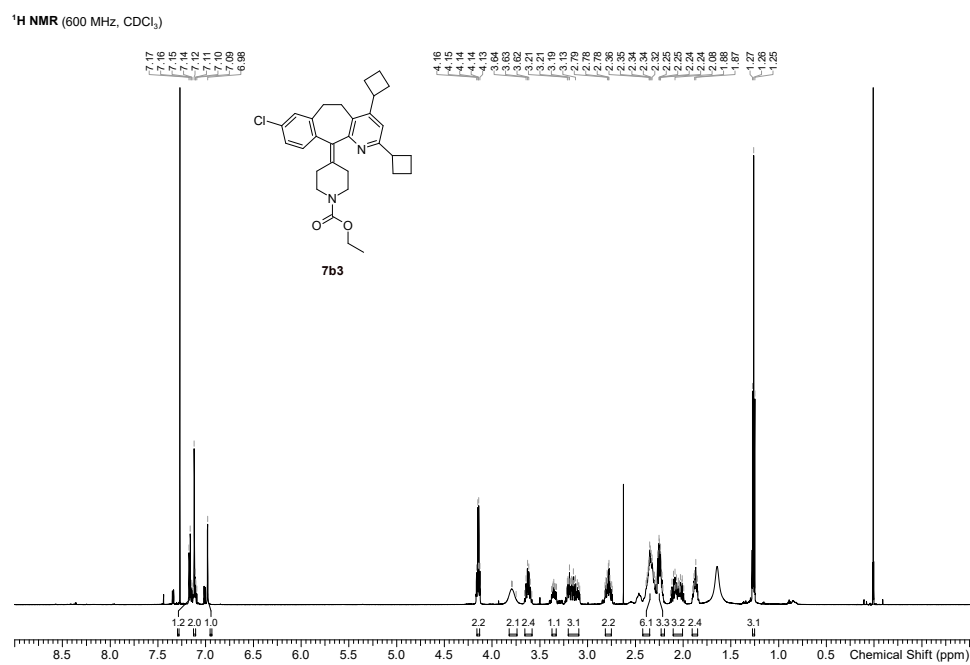


Figure 4: **7b2**, ¹³C-NMR spectrum.

Figure 5: **7b3**, ¹H-NMR spectrum.

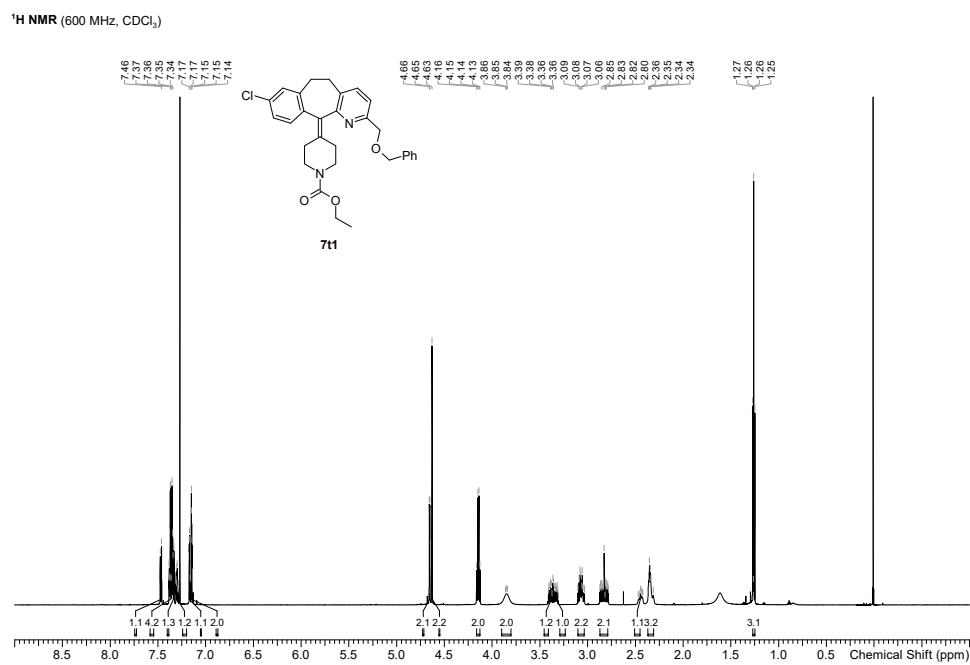
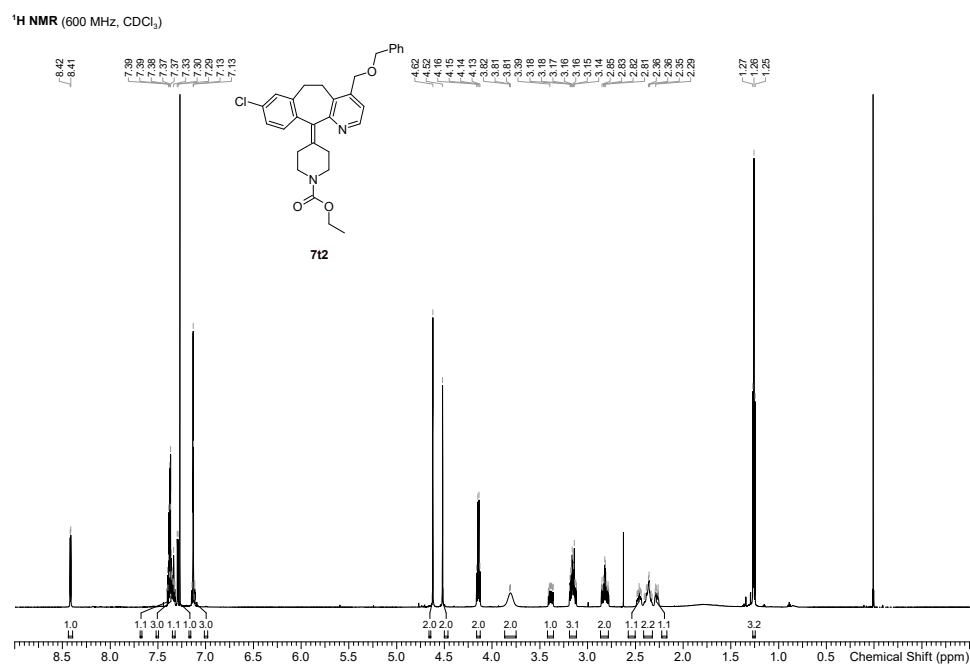


Figure 6: **7t1**, ¹H-NMR spectrum.

Figure 7: **7t2**, ¹H-NMR spectrum.

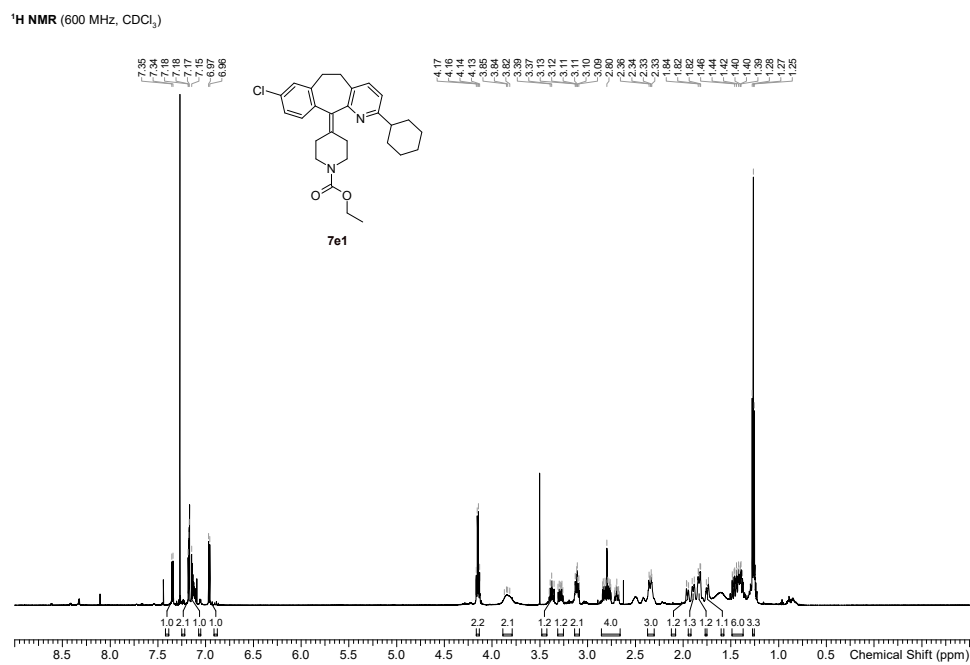


Figure 8: **7e1**, ¹H-NMR spectrum.

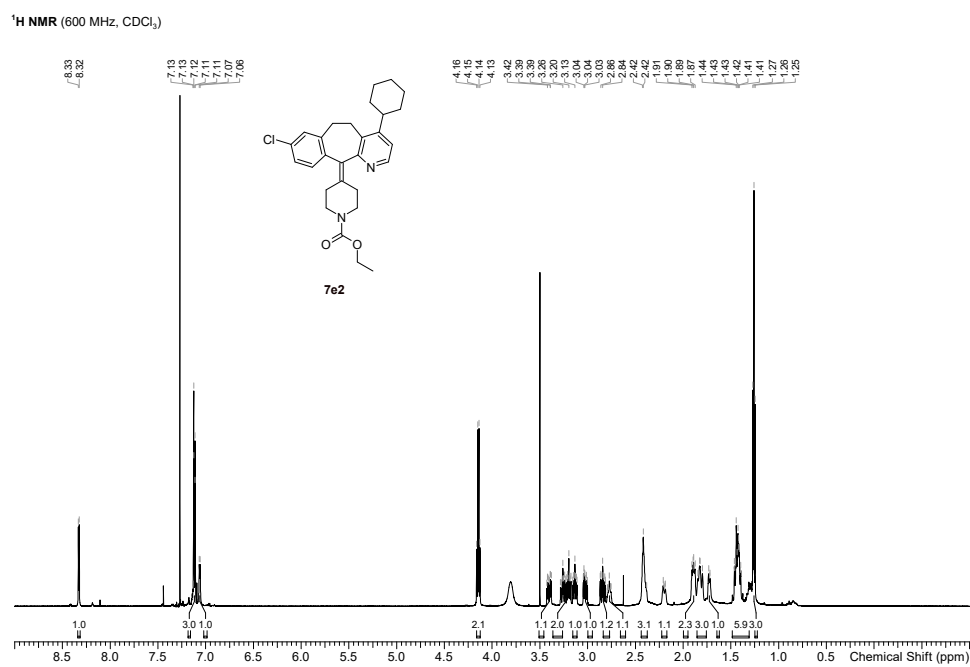


Figure 9: **7e2**, ¹H-NMR spectrum.

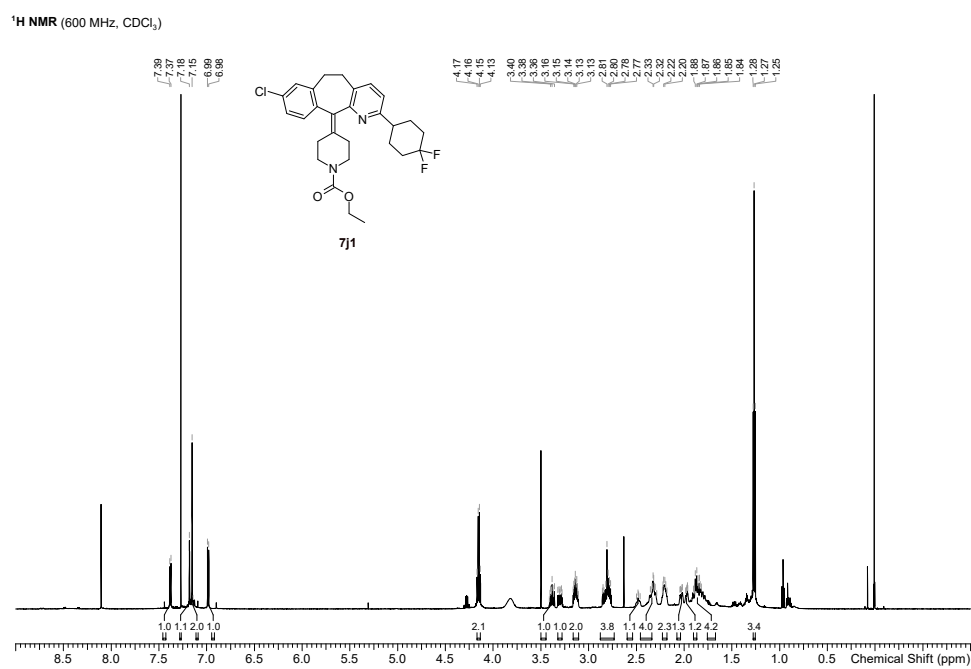
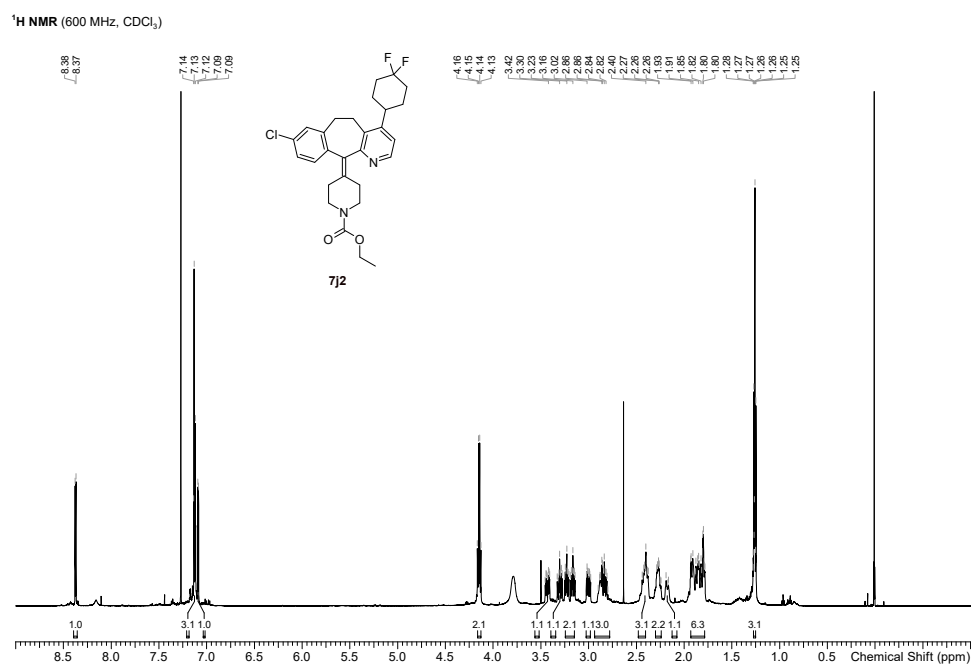


Figure 10: **7j1**, ¹H-NMR spectrum.

Figure 11: **7j2**, ¹H-NMR spectrum.

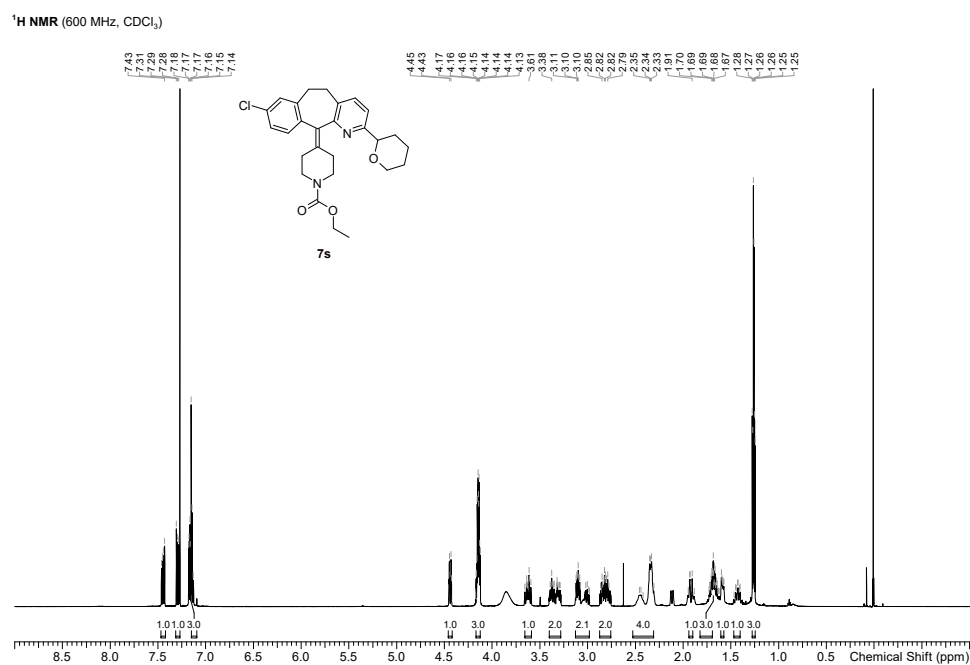
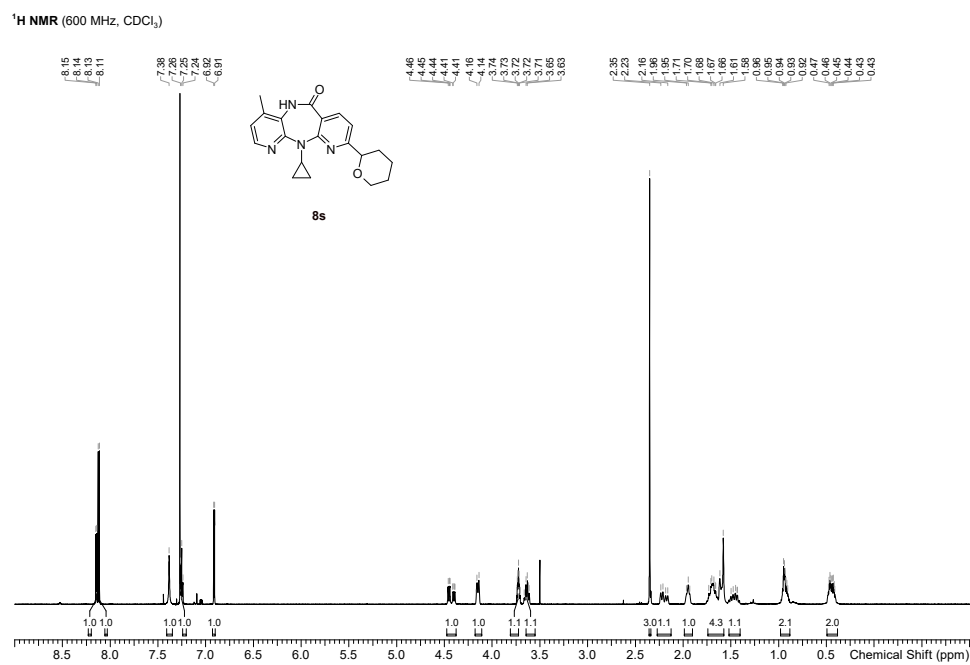


Figure 12: **7s**, ¹H-NMR spectrum.

Figure 13: **8s**, ¹H-NMR spectrum.

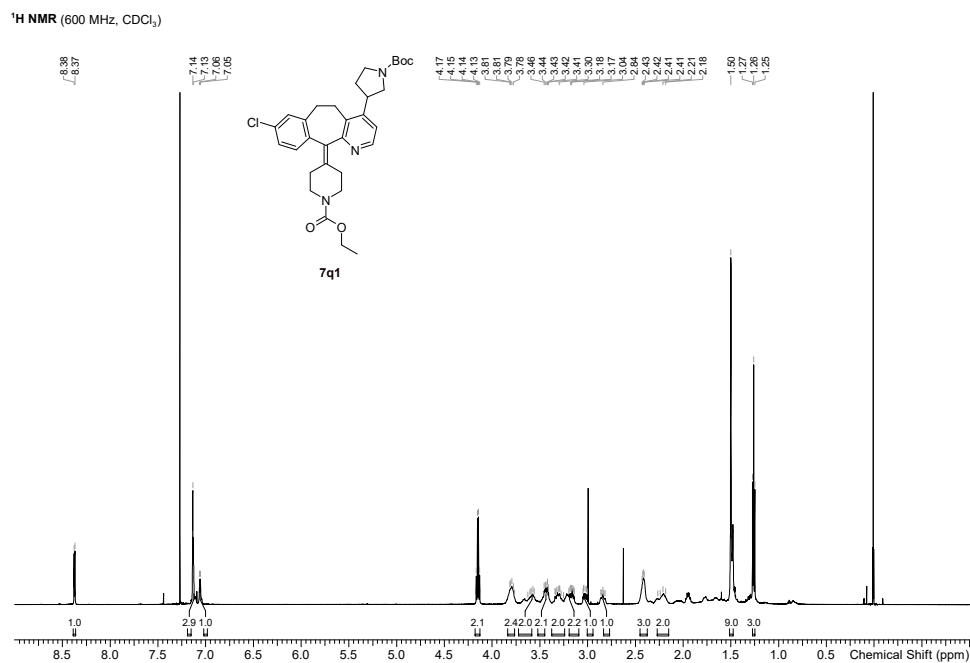
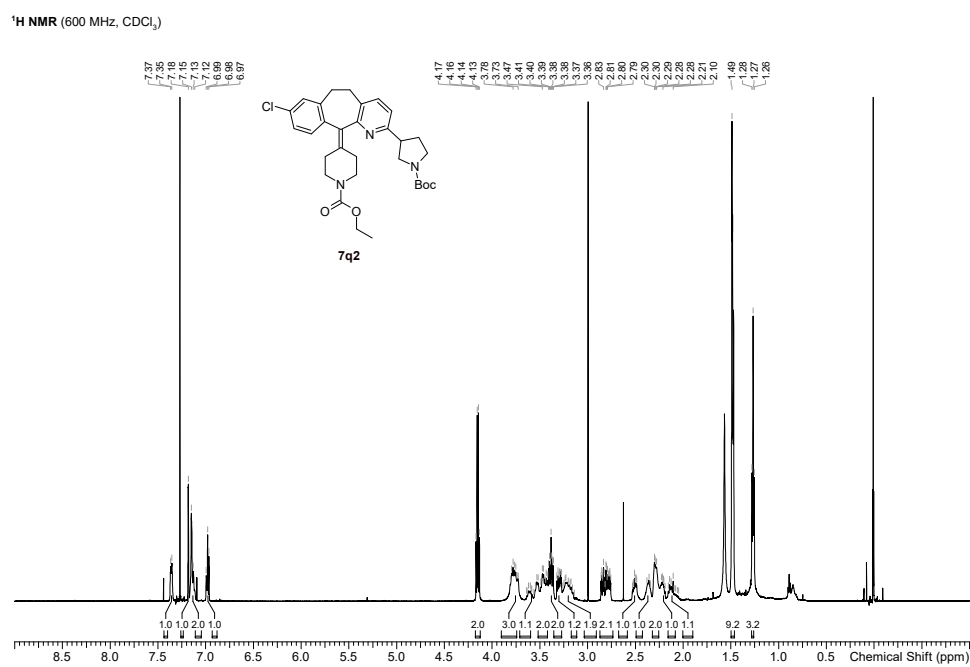


Figure 14: **7q1**, ¹H-NMR spectrum.

Figure 15: **7q2**, ¹H-NMR spectrum.

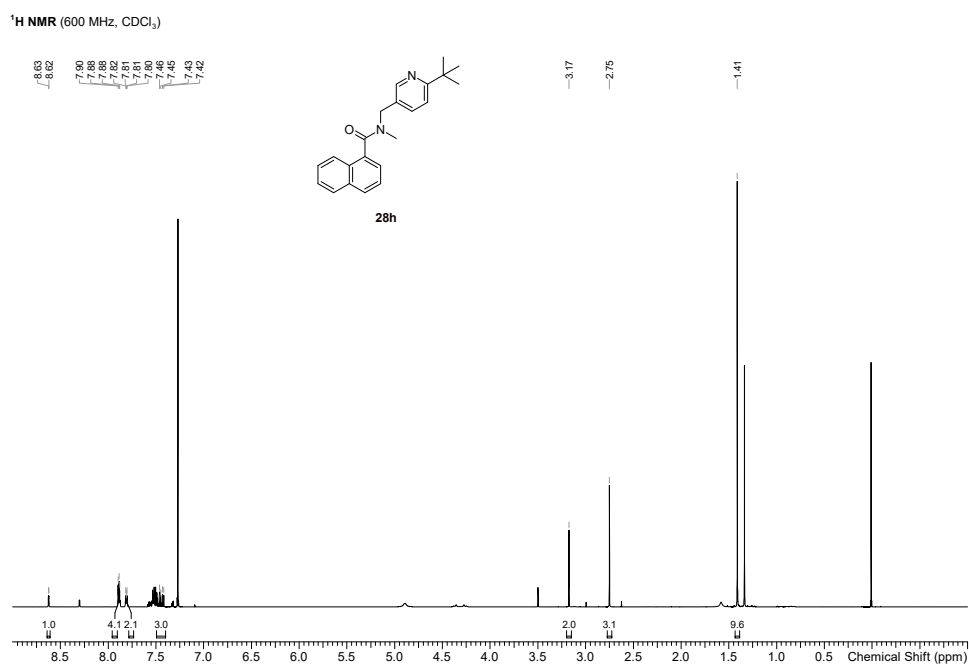


Figure 16: **28h**, ¹H-NMR spectrum.

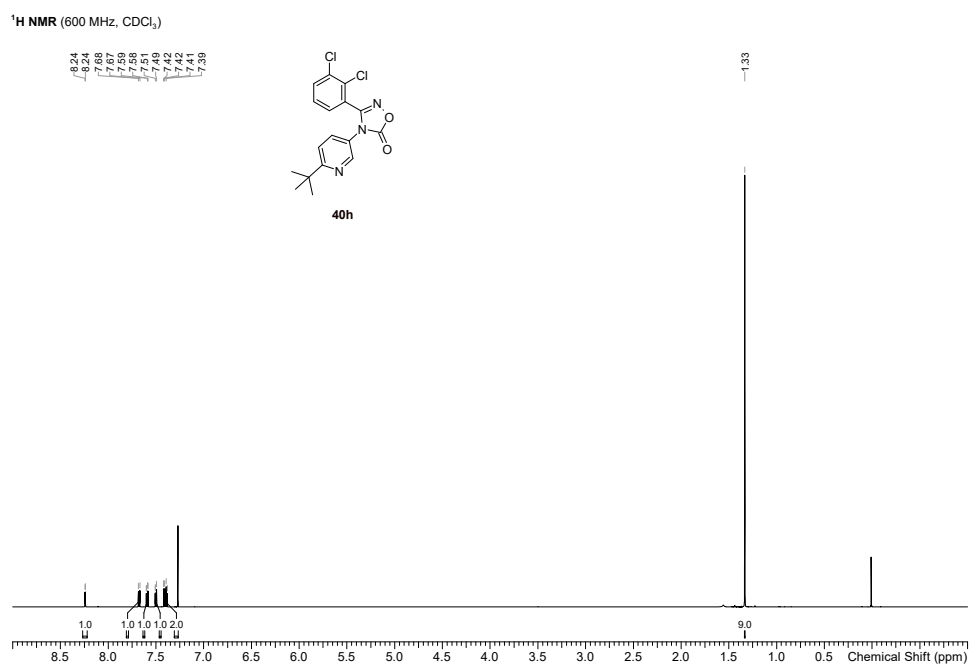
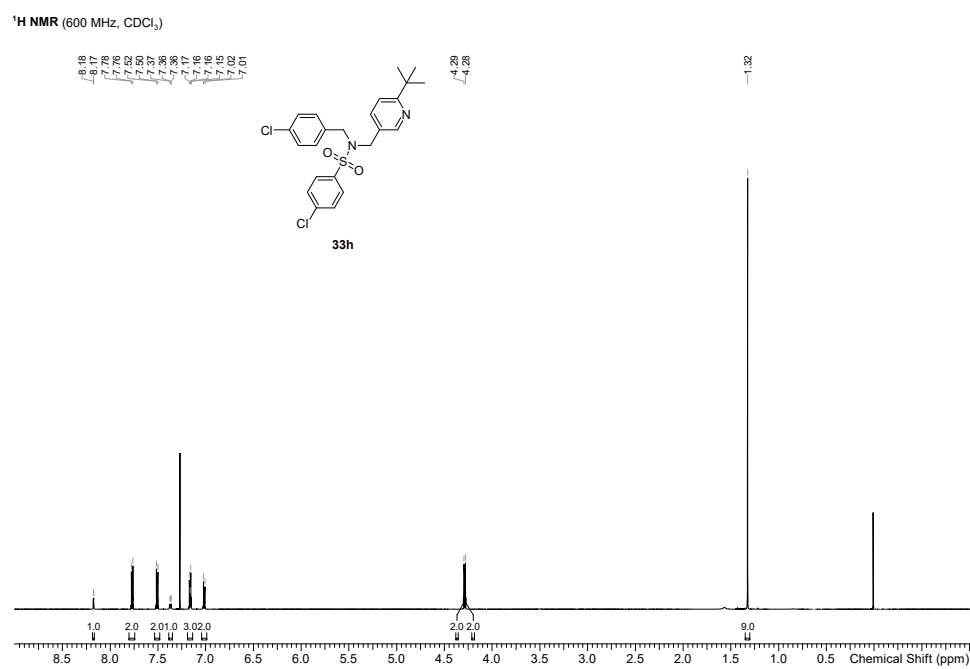


Figure 18: **40h**, ¹H-NMR spectrum.

Figure 19: **33h**, ¹H-NMR spectrum.

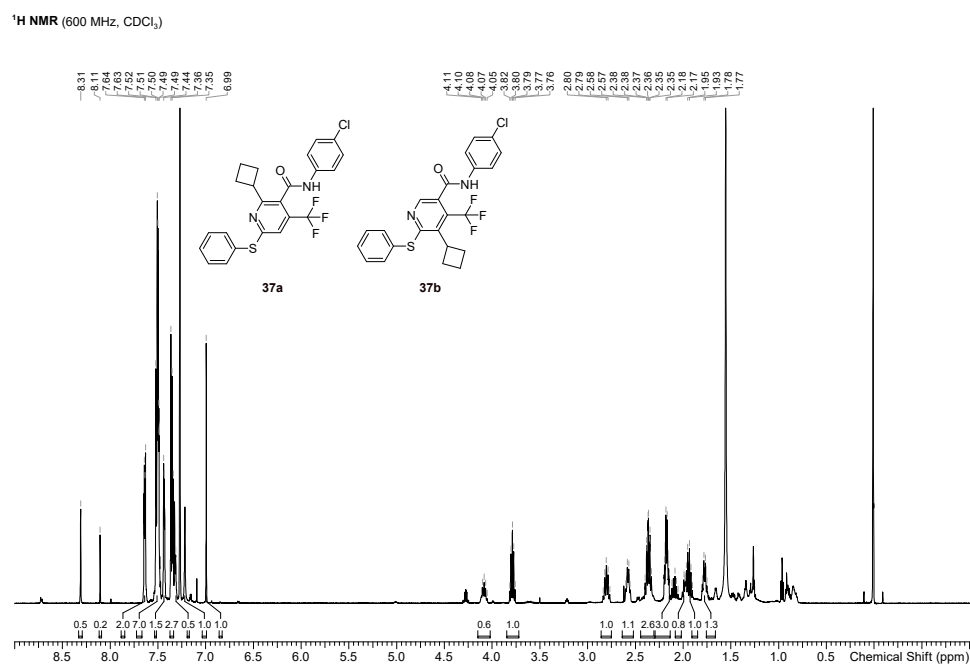
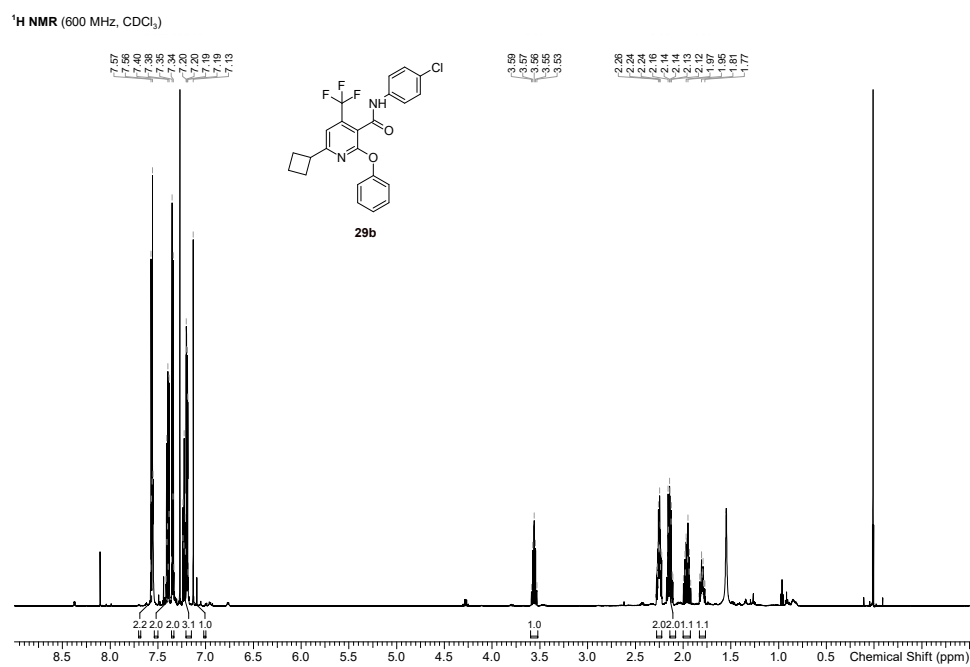
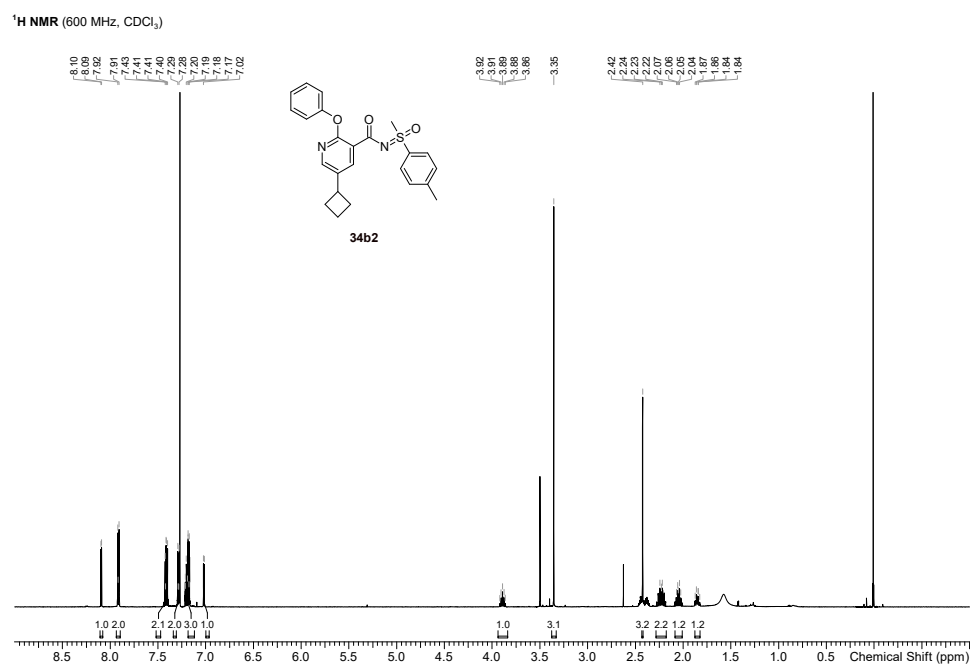


Figure 20: **37b1** & **37b2**, ¹H-NMR spectrum.

Figure 21: **29b**, ¹H-NMR spectrum.

Figure 23: **34b2**, ¹H-NMR spectrum.

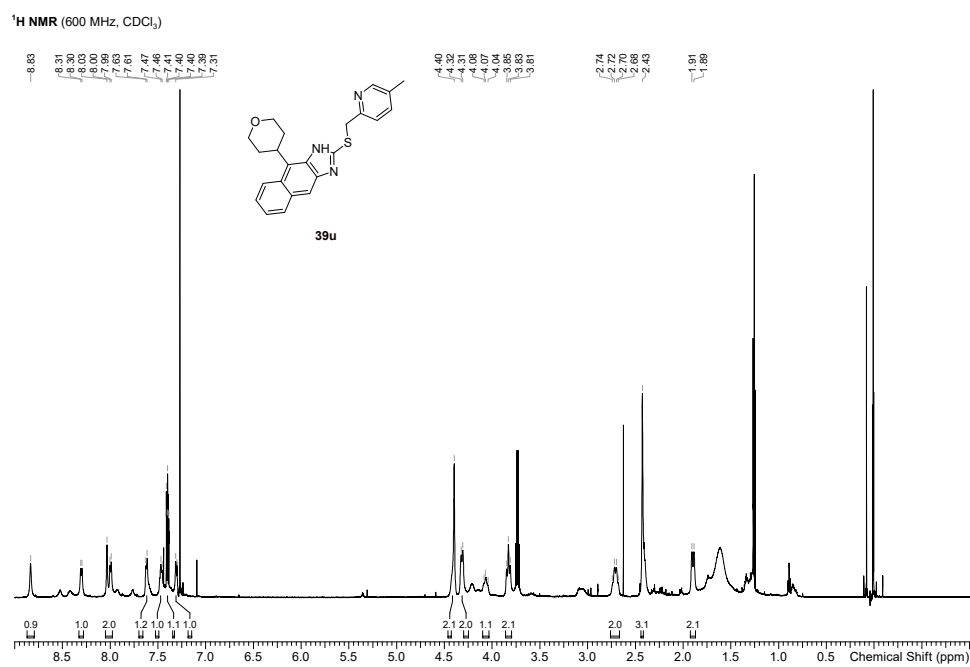
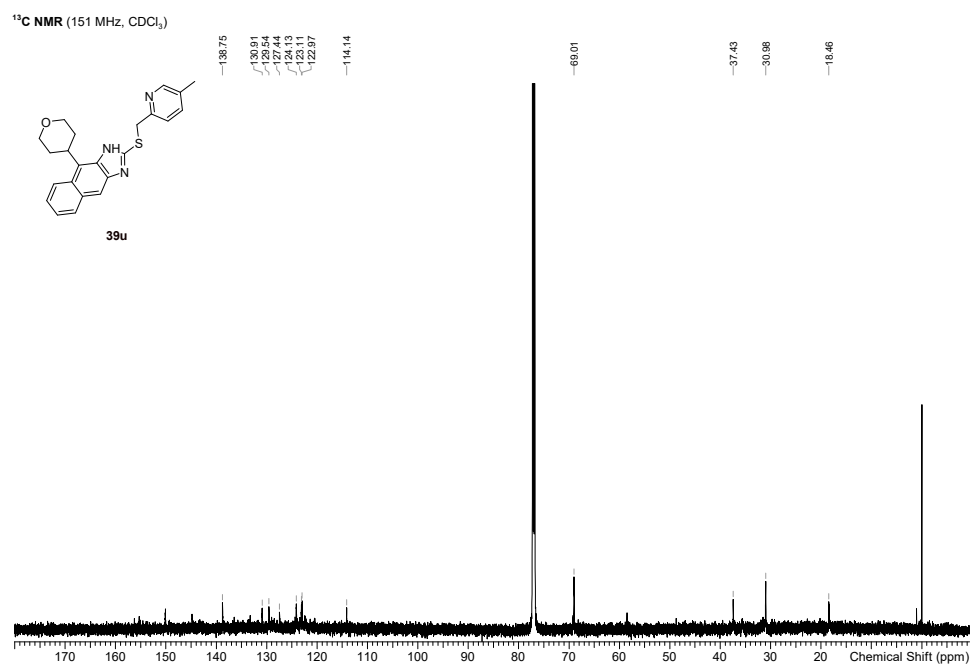


Figure 24: **39u**, ¹H-NMR spectrum.

Figure 25: **39u**, ¹³C-NMR spectrum.

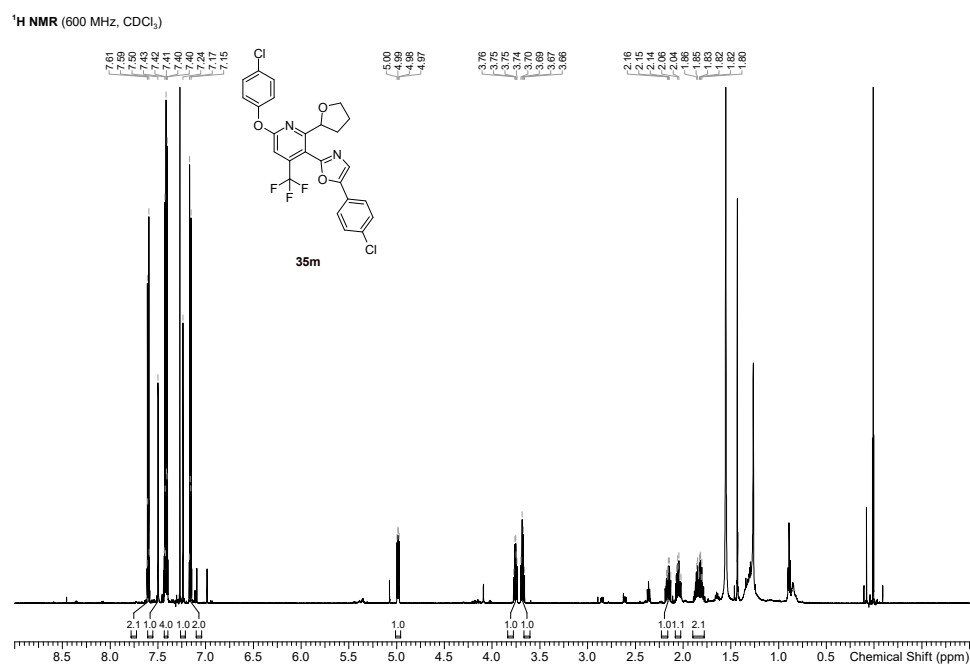
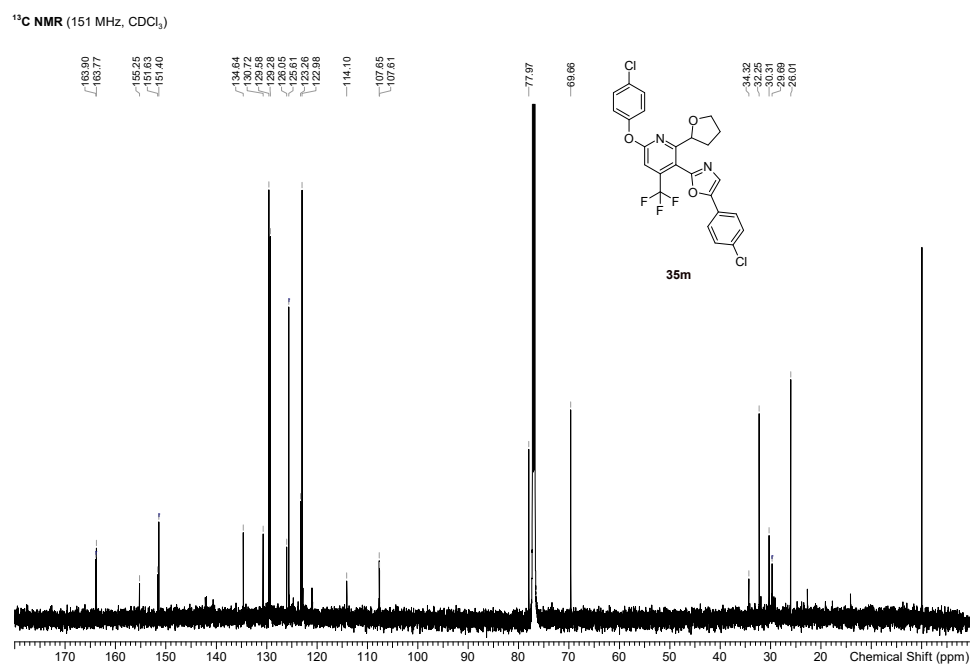


Figure 26: **35m**, ¹H-NMR spectrum.

Figure 27: **35m**, ¹³C-NMR spectrum.

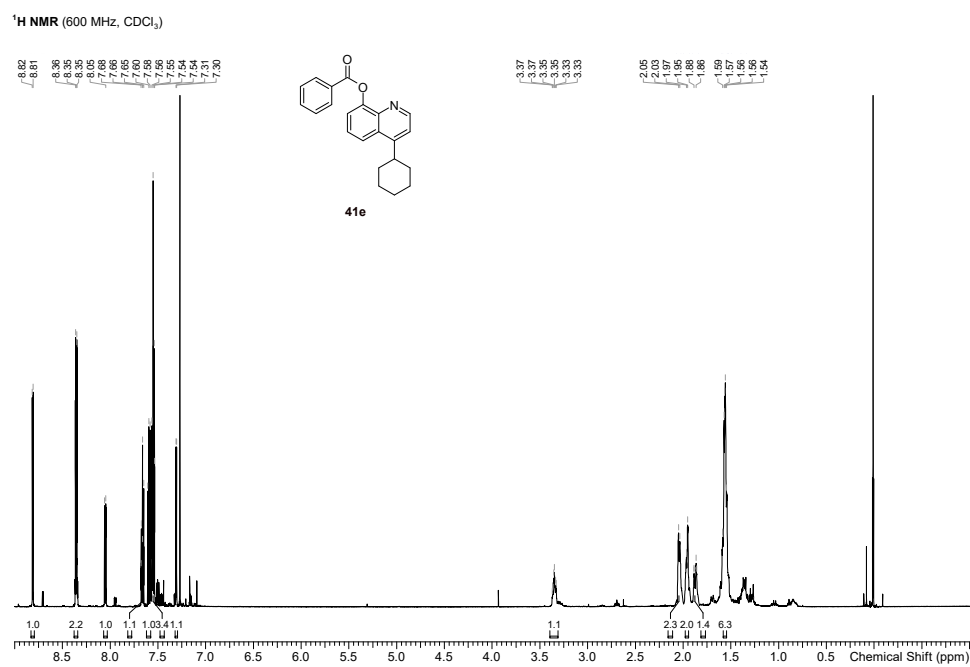
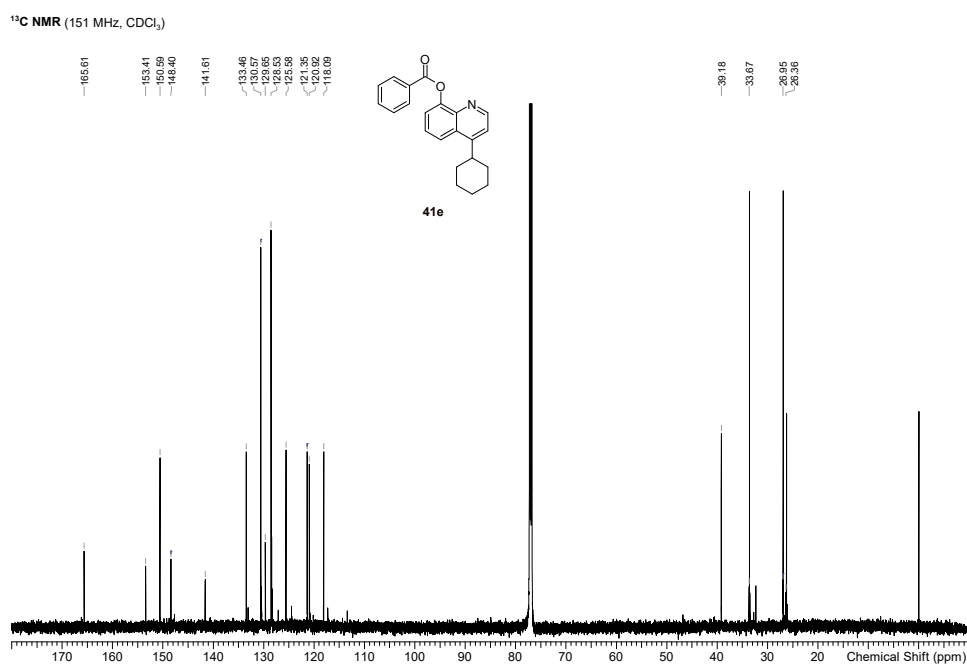


Figure 28: **41e**, ¹H-NMR spectrum.

Figure 29: **41e**, ¹³C-NMR spectrum.

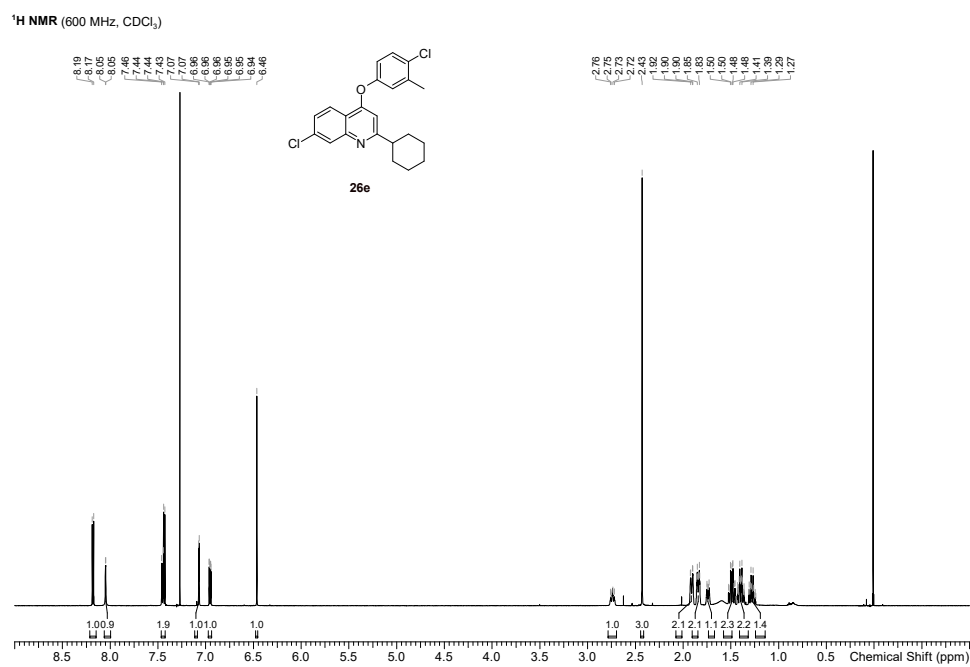


Figure 30: **26e**, ¹H-NMR spectrum.

Discipline is the bridge between dreams and accomplishments.

- Jan Frodeno

7

CONCLUSION AND OUTLOOK

The efficient synthesis of innovative drug molecules to establish structure-activity-relationships (SARs) often remains a challenge in the design-make-test-analyze (DMTA) cycle, a key component in early drug discovery. Modern synthetic methodologies, such as late-stage functionalization (LSF) offer an attractive approach to generating new IP space and modulating pharmacological properties. However, the presence of numerous functional groups within complex drug-like molecules renders straightforward LSF application challenging, often leading to laborious experimentation and failed reactions. To tackle this issue, a semi-automated LSF platform, named DOLPHIN, and a new, simple reaction format termed SURF were designed and implemented. Both tools were applied in case studies to successfully develop machine learning (ML) tools capable of accurate *in silico* reactivity assessment. Thereby, this work contributed to enhancing the compound synthesis efficiency in drug discovery through the strategic application of laboratory automation and artificial intelligence (AI).

Using data-driven and semi-automated workflows, DOLPHIN has proven to efficiently assess and execute LSF transformations on complex drug-like molecules, yet some areas for future improvement remain. Automating literature reaction data extraction through ML and large language models (LLMs) will accelerate the assembly of comprehensive datasets, despite the

current challenges of those models capturing all important parameters. The adoption of the ChemBeads technology for reaction miniaturization could permit the dosing of chemicals in sub-milligram quantities, thereby enabling a broader range of chemistries, including photochemical reactions, which would require the integration of light-emitting modules into the platform. Alternative analytical methods such as MISER, MALDI, DESI, AE-MS, and NMR spectroscopy are set to revolutionize HTE sample processing, necessitating the refinement of the current reaction analysis workflow and visualization tools. The integration of the electronic laboratory journal (ELN), the Google sheet and cloud databases, the robotic systems, and the visualization tool into a unified digital interface would provide a seamless user experience, supporting technology transfer, educational opportunities, and adherence to FAIR data principles.

Future research around SURF should prioritize the refinement of the SURF architecture to enhance its interoperability with emerging data analysis tools and laboratory information management systems, ELNs. Moreover, there is a pressing need to establish robust protocols for the curation and validation of data within SURF to ensure its integrity and reproducibility. As the chemical research community moves towards a more open and collaborative paradigm, the establishment of global standards for data sharing, similar to those in genomics and proteomics, becomes paramount. This will require not only the development of technical solutions but also a cultural shift, supported by policy frameworks and educational initiatives that underscore the value of data stewardship and the ethical implications of data sharing. By fostering an environment that values transparency and collaboration, SURF has the potential to catalyze a new wave of discovery and innovation in chemical synthesis and ML, ultimately propelling the field towards a more efficient and reproducible scientific practice.

To advance the capabilities of ML models developed for predicting reaction outcomes, yields and regioselectivity of C–H borylations towards being more robust and generalizable, future research should address three key areas. Firstly, the continuous generation of data through HTE systems to encompass a broader chemical space is of high importance. This includes extending the scope of iridium-catalyzed borylation reaction conditions, expanding the LSF informer library to cover a more extensive chemical space pertinent to drug molecules, and exploring less conventional transition-metal-catalyzed or metal-free synthetic methodologies to increase reaction condition diversity. Secondly, the impact of integrating advanced featurization techniques that capture complex chemical phenomena, such as transition state energies and 3D molecular interactions, should be investigated. This incorporation of QM descriptors

may offer a more nuanced understanding of reactivity trends and improve prediction accuracy. Thirdly, the development of interpretable ML models is necessary to provide insights into the underlying reaction mechanisms and to identify novel catalysts and reaction conditions. This step will also foster trust and acceptance among chemists, facilitating the practical application of ML predictions in experimental settings.

Similar to the borylation case study, the further enhancement of the Minisci-type C–H alkylation reactivity prediction will require the experimental generation of larger training data sets with a more diverse array of reaction parameters, including alternative oxidants, solvents, and radical precursors, compared to the current single condition screening. Further, systematic exploration of metal salts and the incorporation of photoredox catalysis and electrochemistry could unveil novel, optimized reaction conditions, thereby extending the reaction scope. This could also lead to a broader substrate scope, encompassing a wider range of heterocyclic systems, particularly five-membered rings, which would be instrumental in enabling the model to accurately predict other re-occurring structural motifs prevalent in drug-like molecules. Enhancing featurization with mechanistic insights and regioselectivity prediction combined with improved interpretability of the models, as discussed above, will further support the accuracy and, consequently, the acceptance of these models.

This work at the interface of chemistry, data science, automation, digitalization and ML has highlighted the importance of connecting disciplines to solve challenges in drug discovery. To increase the efficiency of compound synthesis, the highly complex, multi-dimensional problems of chemical synthesis need to be approached with innovative methods from other research fields as well. Even though this thesis provided a first step into a more digitalized and automated approach through DOLPHIN and SURF, many additional studies and research will be needed to herald the digital chemistry age. To keep the digital momentum alive, the next generation of chemists needs to be acquainted with computer science knowledge, made aware of required interdisciplinary collaborations, receive encouraging mentoring and get the freedom to develop ideas outside the already known.

Bibliography

1. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
2. Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
3. Pillai, N., Dasgupta, A., Sudsakorn, S., Fretland, J. & Mavroudis, P. D. Machine learning guided early drug discovery of small molecules. *Drug Discov. Today* **27**, 2209–2215 (2022).
4. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–716 (2004).
5. Ruffolo, R. R. Why has R&D productivity declined in the pharmaceutical industry? *Expert Opin. Drug Discov.* **1**, 99–102 (2006).
6. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).
7. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200 (2012).
8. Schuhmacher, A., Gassmann, O. & Hinder, M. Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.* **14**, 1–11 (2016).
9. Gascón, F., Lozano, J., Ponte, B. & de la Fuente, D. Measuring the efficiency of large pharmaceutical companies: an industry analysis. *Eur. J. Health Econ.* **18**, 587–608 (2017).
10. Terry, C. & Lesser, N. Measuring the return from pharmaceutical innovation 2022. <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/life-sciences-health-care/deloitte-uk-seize-digital-momentum-rd-roi-2022.pdf> (2023).
11. Dörwald, F. Z. *Lead optimization for medicinal chemists: pharmacokinetic properties of functional groups and organic compounds* (John Wiley & Sons, 2012).
12. Mossinghoff, G. J. & Bombelles, T. The importance of intellectual property protection to the American research-intensive pharmaceutical industry. *J. World Bus.* **31**, 38–48 (1996).
13. Saha, C. N. & Bhattacharya, S. Intellectual property rights: An overview and implications in pharmaceutical industry. *J. Adv. Pharm. Technol. Res.* **2**, 88 (2011).
14. Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **42**, 724–733 (2009).
15. Young, R. J. & Leeson, P. D. Mapping the efficiency and physicochemical trajectories of successful optimizations. *J. Med. Chem.* **61**, 6421–6467 (2018).
16. Hoffer, L. *et al.* Integrated strategy for lead optimization based on fragment growing: the diversity-oriented-target-focused-synthesis approach. *J. Med. Chem.* **61**, 5719–5732 (2018).
17. Rossi, T. & Braggio, S. Quality by Design in lead optimization: a new strategy to address productivity in drug discovery. *Curr. Pharmacol.* **11**, 515–520 (2011).

18. Anderson, A. C. Structure-based functional design of drugs: from target to lead compound. *Methods Mol. Biol.*, 359–366 (2012).
19. Zhu, T. *et al.* Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis: Miniperspective. *J. Med. Chem.* **56**, 6560–6572 (2013).
20. Mak, K.-K. & Pichika, M. R. Artificial intelligence in drug development: present status and future prospects. *Drug Discov. Today* **24**, 773–780 (2019).
21. Davis, A. M., Plowright, A. T. & Valeur, E. Directing evolution: the next revolution in drug discovery? *Nat. Rev. Drug Discov.* **16**, 681–698 (2017).
22. Andersson, S. *et al.* Making medicinal chemistry more effective—application of Lean Sigma to improve processes, speed and quality. *Drug Discov. Today* **14**, 598–604 (2009).
23. Plowright, A. T. *et al.* Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discov. Today* **17**, 56–62 (2012).
24. Struble, T. J. *et al.* Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J. Med. Chem.* **63**, 8667–8682 (2020).
25. Cumming, J. G., Davis, A. M., Muresan, S., Haerberlein, M. & Chen, H. Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discov.* **12**, 948–962 (2013).
26. Cherkasov, A. *et al.* QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
27. Avdeef, A. *et al.* PAMPA—critical factors for better predictions of absorption. *J. Pharm. Sci.* **96**, 2893–2909 (2007).
28. Morgenthaler, M. *et al.* Predicting and tuning physicochemical properties in lead optimization: amine basicities. *ChemMedChem* **2**, 1100–1115 (2007).
29. Andrews-Morger, A., Reutlinger, M., Parrott, N. & Olivares-Morales, A. A Machine Learning Framework to Improve Rat Clearance Predictions and Inform Physiologically Based Pharmacokinetic Modeling. *Mol. Pharmaceutics.* **20**, 5052–5065 (2023).
30. Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **9**, 273–276 (2010).
31. Lionta, E., Spyrou, G., K Vassilatis, D. & Cournia, Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* **14**, 1923–1938 (2014).
32. Tosstorff, A. *et al.* A high quality, industrial data set for binding affinity prediction: performance comparison in different early drug discovery scenarios. *J. Comput. Aided Mol. Des.* **36**, 753–765 (2022).
33. Powers, A. S. *et al.* Geometric deep learning for structure-based ligand design. *ACS Cent. Sci.* (2023).
34. Pérez-Benito, L., Casajuana-Martin, N., Jiménez-Rosés, M., Van Vlijmen, H. & Tresadern, G. Predicting activity cliffs with free-energy perturbation. *J. Chem. Theory Comput.* **15**, 1884–1895 (2019).
35. Liu, R. *et al.* Accelerating and Automating the Free Energy Perturbation Absolute Binding Free Energy Calculation with the RED-E Function. *J. Chem. Inf. Model.* (2023).
36. Macdonald, S. J. & Smith, P. W. Lead optimization in 12 months? True confessions of a chemistry team. *Drug Discov. Today* **6**, 947–953 (2001).
37. Santiago, B. G. *et al.* Perspective on high-throughput bioanalysis to support in vitro assays in early drug discovery. *Bioanalysis* **15**, 177–191 (2023).

38. Hillisch, A., Heinrich, N. & Wild, H. Computational chemistry in the pharmaceutical industry: from childhood to adolescence. *ChemMedChem* **10**, 1958–1962 (2015).
39. Zinner, M. *et al.* Toward the institutionalization of quantum computing in pharmaceutical research. *Drug Discov. Today* **27**, 378–383 (2022).
40. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
41. Agrafiotis, D. K., Bandyopadhyay, D., Wegner, J. K. & van Vlijmen, H. Recent advances in chemoinformatics. *J. Chem. Inf. Model.* **47**, 1279–1293 (2007).
42. Lima, A. N. *et al.* Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.* **11**, 225–239 (2016).
43. Lo, Y.-C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).
44. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
45. Ekins, S. *et al.* Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **18**, 435–441 (2019).
46. Elbadawi, M., Gaisford, S. & Basit, A. W. Advanced machine-learning techniques in drug discovery. *Drug Discov. Today* **26**, 769–777 (2021).
47. Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic synthesis: March of the machines. *Angew. Chem. Int. Ed.* **54**, 3449–3464 (2015).
48. Selekman, J. A. *et al.* High-throughput automation in chemical process development. *Annu. Rev. Chem. Biomol. Eng.* **8**, 525–547 (2017).
49. Dimitrov, T., Kreisbeck, C., Becker, J. S., Aspuru-Guzik, A. & Saikin, S. K. Autonomous molecular design: Then and now. *ACS Appl. Mater. Interfaces* **11**, 24825–24836 (2019).
50. Wang, Z., Zhao, W., Hao, G.-F. & Song, B.-A. Automated synthesis: current platforms and further needs. *Drug Discov. Today* **25**, 2006–2011 (2020).
51. Davies, I. W. The digitization of organic synthesis. *Nature* **570**, 175–181 (2019).
52. Angelone, D. *et al.* Convergence of multiple synthetic paradigms in a universally programmable chemical synthesis machine. *Nat. Chem.* **13**, 63–69 (2021).
53. Jana, R., Begam, H. M. & Dinda, E. The emergence of the C-H functionalization strategy in medicinal chemistry and drug discovery. *Chem. Comm.* (2021).
54. Wencel-Delord, J. & Glorius, F. C–H bond activation enables the rapid construction and late-stage diversification of functional molecules. *Nat. Chem.* **5**, 369–375 (2013).
55. Nippa, D. F. *et al.* Late-stage functionalization and its impact on modern drug discovery: Medicinal chemistry and chemical biology highlights. *Chimia* **76**, 258–258 (2022).
56. Bergman, R. G. C–H activation. *Nature* **446**, 391–393 (2007).
57. McMurray, L., O’Hara, F. & Gaunt, M. J. Recent developments in natural product synthesis using metal-catalysed C–H bond functionalisation. *Chem. Soc. Rev.* **40**, 1885–1898 (2011).
58. Gutekunst, W. R. & Baran, P. S. C–H functionalization logic in total synthesis. *Chem. Soc. Rev.* **40**, 1976–1991 (2011).
59. Yamaguchi, J., Yamaguchi, A. D. & Itami, K. C–H bond functionalization: Emerging synthetic tools for natural products and pharmaceuticals. *Angew. Chem. Int. Ed.* **51**, 8960–9009 (2012).

60. Kim, K. E., Kim, A. N., McCormick, C. J. & Stoltz, B. M. Late-stage diversification: A motivating force in organic synthesis. *J. Am. Chem. Soc.* **143**, 16890–16901 (2021).
61. Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016).
62. Blakemore, D. C. *et al.* Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10**, 383–394 (2018).
63. Abrams, D. J., Provencher, P. A. & Sorensen, E. J. Recent applications of C–H functionalization in complex natural product synthesis. *Chem. Soc. Rev.* **47**, 8925–8967 (2018).
64. Moir, M., Danon, J. J., Reekie, T. A. & Kassiou, M. An overview of late-stage functionalization in today's drug discovery. *Expert Opin. Drug Discov.* **14**, 1137–1149 (2019).
65. Hong, B., Luo, T. & Lei, X. Late-stage diversification of natural products. *ACS Cent. Sci.* **6**, 622–635 (2020).
66. Guillemard, L., Kaplaneris, N., Ackermann, L. & Johansson, M. J. Late-stage C–H functionalization offers new opportunities in drug discovery. *Nat. Rev. Chem.* **5**, 522–545 (2021).
67. Castellino, N. J., Montgomery, A. P., Danon, J. J. & Kassiou, M. Late-stage Functionalization for Improving Drug-like Molecular Properties. *Chem. Rev.* (2023).
68. Brückl, T., Baxter, R. D., Ishihara, Y. & Baran, P. S. Innate and guided C–H functionalization logic. *Acc. Chem. Res.* **45**, 826–839 (2012).
69. Börgel, J. & Ritter, T. Late-stage functionalization. *Chem* **6**, 1877–1887 (2020).
70. Stepan, A. F. *et al.* Late-stage microsomal oxidation reduces drug–drug interaction and identifies phosphodiesterase 2A inhibitor PF-06815189. *ACS Med. Chem. Lett.* **9**, 68–72 (2018).
71. Boström, J., Brown, D. G., Young, R. J. & Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discov.* **17**, 709–727 (2018).
72. Meanwell, M., Nodwell, M., Martin, R. & Britton, R. A convenient late-stage fluorination of pyridylic C–H bonds with N-Fluorobenzenesulfonimide. *Angew. Chem. Int. Ed.* **55**, 13244–13248 (2016).
73. Weis, E., Johansson, M. J. & Martín-Matute, B. Late-stage amination of drug-like benzoic acids: Access to anilines and drug conjugates through directed iridium-catalyzed C–H activation. *Chem. Eur. J.* **27**, 18188–18200 (2021).
74. Friis, S. D., Johansson, M. J. & Ackermann, L. Cobalt-catalysed C–H methylation for late-stage drug diversification. *Nat. Chem.* **12**, 511–519 (2020).
75. Yuan, Z. *et al.* Site-selective, late-stage C–H ¹⁸F-fluorination on unprotected peptides for positron emission tomography imaging. *Angew. Chem. Int. Ed.* **57**, 12733–12736 (2018).
76. Xu, P. *et al.* Site-selective late-stage aromatic [¹⁸F] fluorination via aryl sulfonium salts. *Angew. Chem. Int. Ed.* **132**, 1972–1976 (2020).
77. Huang, X. *et al.* Late stage benzylic C–H fluorination with [¹⁸F] fluoride for PET imaging. *J. Am. Chem. Soc.* **136**, 6842–6845 (2014).
78. McCammant, M. S. *et al.* Cu-mediated C–H ¹⁸F-fluorination of electron-rich (hetero) arenes. *Org. Lett.* **19**, 3939–3942 (2017).
79. Roque, J. B., Kuroda, Y., Göttemann, L. T. & Sarpong, R. Deconstructive fluorination of cyclic amines by carbon-carbon cleavage. *Science* **361**, 171–174 (2018).

80. Wan, T. *et al.* Accelerated and scalable C(sp³)-H amination via decatungstate photocatalysis using a flow photoreactor equipped with high-intensity LEDs. *ACS Cent. Sci.* **8**, 51–56 (2021).
81. Lee, C., Seo, H., Jeon, J. & Hong, S. γ -Selective C(sp³)-H amination via controlled migratory hydroamination. *Nat. Comm.* **12**, 5657 (2021).
82. Du, Y.-D., Zhou, C.-Y., To, W.-P., Wang, H.-X. & Che, C.-M. Iron porphyrin catalysed light driven C-H bond amination and alkene aziridination with organic azides. *Chem. Sci.* **11**, 4680–4686 (2020).
83. Simonetti, M., Cannas, D., Just-Baringo, X., Vitorica-Yrezabal, I. & Larrosa, I. Cyclometallated ruthenium catalyst enables late-stage directed arylation of pharmaceuticals. *Nat. Chem.* **10**, 724–731 (2018).
84. Weng, Y. *et al.* Peptide late-stage C(sp³)-H arylation by native asparagine assistance without exogenous directing groups. *Chem. Sci.* **11**, 9290–9295 (2020).
85. Bai, Z. *et al.* Late-stage functionalization and diversification of peptides by internal thiazole-enabled palladium-catalyzed C(sp³)-H Arylation. *ACS Cat.* **11**, 15125–15134 (2021).
86. Schönherr, H. & Cernak, T. Profound methyl effects in drug discovery and a call for new C-H methylation reactions. *Angew. Chem. Int. Ed.* **52**, 12256–12267 (2013).
87. Feng, K. *et al.* Late-stage oxidative C(sp³)-H methylation. *Nature* **580**, 621–627 (2020).
88. Serpier, F. *et al.* Selective methylation of arenes: A radical C-H functionalization/cross-coupling sequence. *Angew. Chem. Int. Ed.* **130**, 10857–10861 (2018).
89. Le-Huu, P., Heidt, T., Claasen, B., Laschat, S. & Urlacher, V. B. Chemo-, regio-, and stereoselective oxidation of the monocyclic diterpenoid β -cembrenediol by P450 BM3. *ACS Cat.* **5**, 1772–1780 (2015).
90. Cheng, L. *et al.* Iron-catalyzed arene C-H hydroxylation. *Science* **374**, 77–81 (2021).
91. Yun, L. *et al.* Selective Oxidation of Benzylic sp³ C-H Bonds using Molecular Oxygen in a Continuous-Flow Microreactor. *Org. Process Res. Dev.* **25**, 1612–1618 (2021).
92. Sarver, P. J. *et al.* The merger of decatungstate and copper catalysis to enable aliphatic C(sp³)-H trifluoromethylation. *Nat. Chem.* **12**, 459–467 (2020).
93. Choi, G., Lee, G. S., Park, B., Kim, D. & Hong, S. H. Direct C(sp³)-H trifluoromethylation of unactivated alkanes enabled by multifunctional trifluoromethyl copper complexes. *Angew. Chem. Int. Ed.* **60**, 5467–5474 (2021).
94. Rey-Rodriguez, R., Retailleau, P., Bonnet, P. & Gillaizeau, I. Iron-Catalyzed trifluoromethylation of enamide. *Chem. Eur. J.* **21**, 3572–3575 (2015).
95. Larsen, M. A. & Hartwig, J. F. Iridium-catalyzed C-H borylation of heteroarenes: Scope, regioselectivity, application to late-stage functionalization, and mechanism. *J. Am. Chem. Soc.* **136**, 4287–4299 (2014).
96. Antropow, A. H., Garcia, N. R., White, K. L. & Movassaghi, M. Enantioselective synthesis of (-)-Vallesine: Late-stage C17-oxidation via complex indole boronation. *Org. Lett.* **20**, 3647–3650 (2018).
97. Gayler, K. M., Kong, K., Reisenauer, K., Taube, J. H. & Wood, J. L. Staurosporine analogs via C-H borylation. *ACS Med. Chem. Lett.* **11**, 2441–2445 (2020).
98. Kim, J. H. *et al.* A radical approach for the selective C-H borylation of azines. *Nature* **595**, 677–683 (2021).

99. Huan, L., Shu, X., Zu, W., Zhong, D. & Huo, H. Asymmetric benzylic C(sp³)-H acylation via dual nickel and photoredox catalysis. *Nat. Comm.* **12**, 3536 (2021).
100. San Segundo, M. & Correa, A. Site-selective aqueous C-H acylation of tyrosine-containing oligopeptides with aldehydes. *Chem. Sci.* **11**, 11531–11538 (2020).
101. Shu, X., Huan, L., Huang, Q. & Huo, H. Direct enantioselective C(sp³)-H acylation for the synthesis of α -amino ketones. *J. Am. Chem. Soc.* **142**, 19058–19064 (2020).
102. Romero, E. *et al.* Enzymatic late-stage modifications: better late than never. *Angew. Chem. Int. Ed.* **60**, 16824–16855 (2021).
103. Lasso, J. D., Castillo-Pazos, D. J. & Li, C.-J. Green chemistry meets medicinal chemistry: A perspective on modern metal-free late-stage functionalization reactions. *Chem. Soc. Rev.* **50**, 10955–10982 (2021).
104. Wang, W., Lorion, M. M., Shah, J., Kapdi, A. R. & Ackermann, L. Late-stage peptide diversification by position-selective C-H activation. *Angew. Chem. Int. Ed.* **57**, 14700–14717 (2018).
105. Liu, R., Li, X. & Lam, K. S. Combinatorial chemistry in drug discovery. *Curr. Opin. Chem. Biol.* **38**, 117–126 (2017).
106. Bunin, B. A. & Ellman, J. A. A general and expedient method for the solid-phase synthesis of 1,4-benzodiazepine derivatives. *J. Am. Chem. Soc.* **114**, 10997–10998 (1992).
107. Kennedy, J. P. *et al.* Application of combinatorial chemistry science on modern drug discovery. *J. Comb. Chem.* **10**, 345–354 (2008).
108. Schmink, J. R., Bellomo, A. & Berritt, S. Scientist-led high-throughput experimentation (HTE) and its utility in academia and industry. *Aldrichimica Acta* **46**, 71–80 (2013).
109. Buitrago Santanilla, A. *et al.* Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).
110. Shevlin, M. Practical high-throughput experimentation for chemists. *ACS Med. Chem. Lett.* **8**, 601–607 (2017).
111. Krska, S. W., DiRocco, D. A., Dreher, S. D. & Shevlin, M. The evolution of chemical high-throughput experimentation to address challenging problems in pharmaceutical synthesis. *Acc. Chem. Res.* **50**, 2976–2985 (2017).
112. Mennen, S. M. *et al.* The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Org. Process. Res. Dev.* **23**, 1213–1242 (2019).
113. Balkenhohl, F., von dem Bussche-Hünnefeld, C., Lansky, A. & Zechel, C. Combinatorial synthesis of small organic molecules. *Angew. Chem. Int. Ed.* **35**, 2288–2337 (1996).
114. Burgess, K., Lim, H.-J., Porte, A. M. & Sulikowski, G. A. New catalysts and conditions for a C-H insertion reaction identified by high throughput catalyst screening. *Angew. Chem. Int. Ed.* **35**, 220–222 (1996).
115. Xiang, X.-D. *et al.* A combinatorial approach to materials discovery. *Science* **268**, 1738–1740 (1995).
116. Szewczyk, J. W., Zuckerman, R. L., Bergman, R. G. & Ellman, J. A. A mass spectrometric labeling strategy for high-throughput reaction evaluation and optimization: Exploring C-H activation. *Angew. Chem. Int. Ed.* **40**, 216–219 (2001).
117. Francis, M. B. & Jacobsen, E. N. Discovery of novel catalysts for alkene epoxidation from metal-binding combinatorial libraries. *Angew. Chem. Int. Ed.* **38**, 937–941 (1999).

118. Taylor, S. J. & Morken, J. P. Thermographic selection of effective catalysts from an encoded polymer-bound library. *Science* **280**, 267–270 (1998).
119. Hinderling, C. & Chen, P. Rapid screening of olefin polymerization catalyst libraries by electrospray ionization tandem mass spectrometry. *Angew. Chem. Int. Ed.* **38**, 2253–2256 (1999).
120. Shultz, C. S. & Krska, S. W. Unlocking the potential of asymmetric hydrogenation at Merck. *Acc. Chem. Res.* **40**, 1320–1326 (2007).
121. Zacuto, M. J., Shultz, S. C. & Journet, M. Preparation of 4-allylisoindoline via a Kumada coupling with allylmagnesium chloride. *Org. Process Res. Dev.* **15**, 158–161 (2011).
122. Chung, C. K. *et al.* Process development of C–N cross-coupling and enantioselective biocatalytic reactions for the asymmetric synthesis of niraparib. *Org. Process Res. Dev.* **18**, 215–227 (2014).
123. McNally, A., Prier, C. K. & MacMillan, D. W. Discovery of an α -amino C–H arylation reaction using the strategy of accelerated serendipity. *Science* **334**, 1114–1117 (2011).
124. DiRocco, D. A. *et al.* Late-stage functionalization of biologically active heterocycles through photoredox catalysis. *Angew. Chem. Int. Ed.* **53**, 4802–4806 (2014).
125. Huff, C. A. *et al.* Photoredox-catalyzed hydroxymethylation of heteroaromatic bases. *J. Org. Chem.* **81**, 6980–6987 (2016).
126. Halperin, S. D. *et al.* Development of a direct photocatalytic C–H fluorination for the preparative synthesis of odanacatib. *Org. Lett.* **17**, 5200–5203 (2015).
127. Yayla, H. G. *et al.* Discovery and mechanistic study of a photocatalytic indoline dehydrogenation for the synthesis of elbasvir. *Chem. Sci.* **7**, 2066–2073 (2016).
128. Belyk, K. M. *et al.* Enantioselective synthesis of (1 R, 2 S)-1-amino-2-vinylcyclopropanecarboxylic acid ethyl ester (Vinyl-ACCA-OEt) by asymmetric phase-transfer catalyzed cyclopropanation of (E)-N-phenylmethyleneglycine ethyl ester. *Org. Process Res. Dev.* **14**, 692–700 (2010).
129. Xiang, B., Belyk, K. M., Reamer, R. A. & Yasuda, N. Discovery and Application of Doubly Quaternized Cinchona-Alkaloid-Based Phase-Transfer Catalysts. *Angew. Chem. Int. Ed.* **126**, 8515–8518 (2014).
130. Humphrey, G. R. *et al.* Asymmetric synthesis of letermovir using a novel phase-transfer-catalyzed aza-Michael reaction. *Org. Process Res. Dev.* **20**, 1097–1103 (2016).
131. Boga, S. B. *et al.* Selective functionalization of complex heterocycles via an automated strong base screening platform. *React. Chem. Eng.* **2**, 446–450 (2017).
132. Knochel, P. & Molander, G. A. *Comprehensive organic synthesis* (Newnes, 2014).
133. Welch, C. J. *et al.* Solving multicomponent chiral separation challenges using a new SFC tandem column screening tool. *Chirality* **19**, 184–189 (2007).
134. Welch, C. J. High throughput analysis enables high throughput experimentation in pharmaceutical process research. *React. Chem. Eng.* **4**, 1895–1911 (2019).
135. Grainger, R. & Whibley, S. A perspective on the analytical challenges encountered in high-throughput experimentation. *Org. Process Res. Dev.* **25**, 354–364 (2021).
136. Boström, J. & Brown, D. G. Stuck in a rut with old chemistry. *Drug Discov. Today* **21**, 701–703 (2016).
137. Nadin, A., Hattotuagama, C. & Churcher, I. Lead-oriented synthesis: a new opportunity for synthetic chemistry. *Angew. Chem. Int. Ed.* **51**, 1114–1122 (2012).

138. Brown, D. G. & Bostrom, J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? Miniperspective. *J. Med. Chem.* **59**, 4443–4458 (2016).
139. Hartwig, J. F. Evolution of C–H bond functionalization from methane to methodology. *J. Am. Chem. Soc.* **138**, 2–24 (2016).
140. Rogge, T. *et al.* C–H activation. *Nat. Rev. Methods Primers* **1**, 43 (2021).
141. Cernak, T. *et al.* Microscale high-throughput experimentation as an enabling technology in drug discovery: Application in the discovery of (Piperidinyl) pyridinyl-1 H-benzimidazole diacylglycerol acyltransferase 1 inhibitors. *J. Med. Chem.* **60**, 3594–3605 (2017).
142. Roughley, S. D. & Jordan, A. M. The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **54**, 3451–3479 (2011).
143. Allen, C. L., Leitch, D. C., Anson, M. S. & Zajac, M. A. The power and accessibility of high-throughput methods for catalysis research. *Nat. Cat.* **2**, 2–4 (2019).
144. Caron, S. & Thomson, N. M. Pharmaceutical process chemistry: Evolution of a contemporary data-rich laboratory environment. *J. Org. Chem.* **80**, 2943–2958 (2015).
145. Bellomo, A. *et al.* Rapid catalyst identification for the synthesis of the pyrimidinone core of HIV integrase inhibitors. *Angew. Chem. Int. Ed.* **51**, 6912–6915 (2012).
146. Zawatzky, K. *et al.* Overcoming “speed limits” in high throughput chromatographic analysis. *J. Chromatogr. A* **1499**, 211–216 (2017).
147. Barhate, C. L. *et al.* Ultrafast chiral separations for high throughput enantiopurity analysis. *ChemComm* **53**, 509–512 (2017).
148. Welch, C. J. *et al.* MISER chromatography (multiple injections in a single experimental run): the chromatogram is the graph. *Tetrahedron: Asymmetry* **21**, 1674–1681 (2010).
149. Farrant, E. Automation of synthesis in medicinal chemistry: Progress and challenges. *ACS Med. Chem. Lett.* **11**, 1506–1513 (2020).
150. Wong, H. & Cernak, T. Reaction miniaturization in eco-friendly solvents. *Curr. Opin. Green Sustain. Chem.* **11**, 91–98 (2018).
151. Christensen, M. *et al.* Automation isn't automatic. *Chem. Sci.* **12**, 15473–15490 (2021).
152. Bahr, M. N., Morris, M. A., Tu, N. P. & Nandkeolyar, A. Recent advances in high-throughput automated powder dispensing platforms for pharmaceutical applications. *Org. Process. Res. Dev.* **24**, 2752–2761 (2020).
153. Lall, M. S. *et al.* Late-stage lead diversification coupled with quantitative nuclear magnetic resonance spectroscopy to identify new structure–activity relationship vectors at nanomole-scale synthesis: application to loratadine, a human histamine H1 receptor inverse agonist. *J. Med. Chem.* **63**, 7268–7292 (2020).
154. Johansson, S. *et al.* AI-assisted synthesis prediction. *Drug Discov. Today* **32**, 65–72 (2019).
155. Eyke, N. S., Koscher, B. A. & Jensen, K. F. Toward machine learning-enhanced high-throughput experimentation. *Trends Chem.* **3**, 120–132 (2021).
156. Farrusseng, D. High-throughput heterogeneous catalysis. *Surf. Sci. Rep.* **63**, 487–513 (2008).
157. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**, 434–443 (2017).
158. Maloney, M. P. *et al.* *Negative Data in Data Sets for Machine Learning Training* 2023.

159. Götz, J. *et al.* High-throughput synthesis provides data for predicting molecular properties and reaction success. *Sci. Adv.* **9**, eadj2314 (2023).
160. Mahjour, B., Shen, Y. & Cernak, T. Ultrahigh-throughput experimentation for information-rich chemical synthesis. *Acc. Chem. Res.* **54**, 2337–2346 (2021).
161. Schmink, J. R. & Krska, S. W. Reversed-polarity synthesis of diaryl ketones via palladium-catalyzed cross-coupling of acylsilanes. *J. Am. Chem. Soc.* **133**, 19574–19577 (2011).
162. Garcia-Losada, P. *et al.* Practical asymmetric fluorination approach to the scalable synthesis of new fluoroaminothiazine BACE inhibitors. *Org. Process Res. Dev.* **22**, 650–654 (2018).
163. Weis, E., Johansson, M., Korsgren, P., Martín-Matute, B. & Johansson, M. J. Merging directed C–H activations with high-throughput experimentation: Development of iridium-catalyzed C–H aminations applicable to late-stage functionalization. *JACS Au* **2**, 906–916 (2022).
164. Walvoord, R. R., Berritt, S. & Kozlowski, M. C. Palladium-catalyzed nitromethylation of aryl halides: an orthogonal formylation equivalent. *Org. Lett.* **14**, 4086–4089 (2012).
165. Knappke, C. E. *et al.* Reductive cross-coupling reactions between two electrophiles. *Chem. Eur. J.* **20**, 6828–6842 (2014).
166. Hansen, E. C. *et al.* New ligands for nickel catalysis from diverse pharmaceutical heterocycle libraries. *Nat. Chem.* **8**, 1126–1130 (2016).
167. Sather, A. C. & Martinot, T. A. Data-rich experimentation enables palladium-catalyzed couplings of piperidines and five-membered (hetero) aromatic electrophiles. *Org. Process Res. Dev.* **23**, 1725–1739 (2019).
168. Lu, X., Roberts, S. E., Franklin, G. J. & Davie, C. P. On-DNA Pd and Cu promoted C–N cross-coupling reactions. *MedChemComm* **8**, 1614–1617 (2017).
169. Schmink, J. R. & Tudge, M. T. Facile preparation of highly-functionalized, nitrogen-bearing diarylmethanes. *Tetrahedron Lett.* **54**, 15–20 (2013).
170. Martinelli, J. R., Watson, D. A., Freckmann, D. M., Barder, T. E. & Buchwald, S. L. Palladium-catalyzed carbonylation reactions of aryl bromides at atmospheric pressure: a general system based on xantphos. *J. Org. Chem.* **73**, 7102–7107 (2008).
171. Lee, J., Schmink, J. R. & Berritt, S. Introduction of Low-Barrier High-Throughput Experimentation in the Undergraduate Laboratory: Suzuki–Miyaura Reaction. *J. Chem. Educ.* **97**, 538–542 (2020).
172. Roberts, R. M. *Serendipity: Accidental discoveries in science* (1989).
173. Hoffmann, R. W. Wittig and his accomplishments: still relevant beyond his 100th birthday. *Angew. Chem. Int. Ed.* **40**, 1411–1416 (2001).
174. Calloway, N. The Friedel–Crafts Syntheses. *Chem. Rev.* **17**, 327–392 (1935).
175. Brown, H. C. & Zweifel, G. Hydroboration. VII. Directive effects in the hydroboration of olefins. *J. Am. Chem. Soc.* **82**, 4708–4712 (1960).
176. Robbins, D. W. & Hartwig, J. F. A simple, multidimensional approach to high-throughput discovery of catalytic reactions. *Science* **333**, 1423–1427 (2011).
177. Logsdon, D. L. *et al.* High-throughput screening of reductive amination reactions using desorption electrospray ionization mass spectrometry. *Org. Process Res. Dev.* **24**, 1647–1657 (2020).
178. Shaabani, S. *et al.* Automated and accelerated synthesis of indole derivatives on a nano-scale. *Green Chem.* **21**, 225–232 (2019).

179. Caldentey, X. & Romero, E. High-throughput experimentation as an accessible technology for academic organic chemists in Europe and beyond. *Chem. Meth.*, e202200059 (2023).
180. Montavon, T. J., Li, J., Cabrera-Pardo, J. R., Mrksich, M. & Kozmin, S. A. Three-component reaction discovery enabled by mass spectrometry of self-assembled monolayers. *Nat. Chem.* **4**, 45–51 (2012).
181. Cabrera-Pardo, J. R., Chai, D. I., Liu, S., Mrksich, M. & Kozmin, S. A. Label-assisted mass spectrometry for the acceleration of reaction discovery and optimization. *Nat. Chem.* **5**, 423–427 (2013).
182. Twilton, J. *et al.* Selective hydrogen atom abstraction through induced bond polarization: Direct α -arylation of alcohols through photoredox, HAT, and Nickel catalysis. *Angew. Chem. Int. Ed.* **57**, 5369–5373 (2018).
183. Badir, S. O., Dumoulin, A., Matsui, J. K. & Molander, G. A. Synthesis of reversed C-acyl glycosides through Ni/photoredox dual catalysis. *Angew. Chem. Int. Ed.* **57**, 6610–6613 (2018).
184. Li, M. *et al.* Transition-metal-free chemo- and regioselective vinylation of azaallyls. *Nat. Chem.* **9**, 997–1004 (2017).
185. Cernak, T. *The Cernak Lab* <https://cernaklab.com/> (2024).
186. Cramer, N. & Coperet, C. *SwissCAT+* <https://swisscatplus.ch/> (2024).
187. Lai, E. Y., Yuan, B., Ackermann, L. & Johansson, M. J. Ruthenium-Catalyzed Aminocarbonylation with Isocyanates Through Weak Coordinating Groups. *Chem. Eur. J.* **29**, e202302023 (2023).
188. Antermite, D. *et al.* Late-stage synthesis of heterobifunctional molecules for PROTAC applications via ruthenium-catalysed C–H amidation. *Nat. Comm.* **14**, 8222 (2023).
189. Uehling, M. R., King, R. P., Krska, S. W., Cernak, T. & Buchwald, S. L. Pharmaceutical diversification via palladium oxidative addition complexes. *Science* **363**, 405–408 (2019).
190. Preshlock, S. M. *et al.* High-throughput optimization of Ir-catalyzed C–H borylation: a tutorial for practical applications. *J. Am. Chem. Soc.* **135**, 7572–7582 (2013).
191. Wills, A. G. *et al.* High-throughput electrochemistry: State of the art, challenges, and perspective. *Org. Process Res. Dev.* **25**, 2587–2600 (2021).
192. Bissonnette, N. B. *et al.* Design of a multiuse photoreactor to enable visible-light photocatalytic chemical transformations and labeling in live cells. *ChemBioChem* **21**, 3555–3562 (2020).
193. Bonfield, H. E. *et al.* The right light: De novo design of a robust modular photochemical reactor for optimum batch and flow chemistry. *ChemPhotoChem* **4**, 45–51 (2020).
194. Sun, A. C. *et al.* A droplet microfluidic platform for high-throughput photochemical reaction discovery. *Nat. Comm.* **11**, 6202 (2020).
195. Pimparkar, K. *et al.* Development of a photochemical microfluidics platform. *J. Flow Chem.* **1**, 53–55 (2011).
196. Lin, S. *et al.* Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, eaar6236 (2018).
197. Nicastrì, M. C., Lehnherr, D., Lam, Y.-h., DiRocco, D. A. & Rovis, T. Synthesis of sterically hindered primary amines by concurrent tandem photoredox catalysis. *J. Am. Chem. Soc.* **142**, 987–998 (2020).

198. Lee, H. *et al.* Photoredox Ni-catalyzed peptide C (sp²)-O cross-coupling: from intermolecular reactions to side chain-to-tail macrocyclization. *Chem. Sci.* **10**, 5073–5078 (2019).
199. Sherwood, T. C. *et al.* Decarboxylative intramolecular arene alkylation using N-(acyloxy) phthalimides, an organic photocatalyst, and visible light. *J. Org. Chem.* **84**, 8360–8379 (2019).
200. Betori, R. C. & Scheidt, K. A. Reductive arylation of arylidene malonates using photoredox catalysis. *ACS Cat.* **9**, 10350–10357 (2019).
201. Rein, J. *et al.* Unlocking the potential of high-throughput experimentation for electrochemistry with a standardized microscale reactor. *ACS Cent. Sci.* **7**, 1347–1355 (2021).
202. Mo, Y., Rughoobur, G., Nambiar, A. M., Zhang, K. & Jensen, K. F. A multifunctional microfluidic platform for high-throughput experimentation of electroorganic chemistry. *Angew. Chem. Int. Ed.* **132**, 21076–21080 (2020).
203. Lehmann, M., Scarborough, C. C., Godineau, E. & Battilocchio, C. An electrochemical flow-through cell for rapid reactions. *Ind. Eng. Chem.* **59**, 7321–7326 (2020).
204. Becker, R., Weber, K., Pfeiffer, T. V., Kranendonk, J. v. & Schouten, K. J. A scalable high-throughput deposition and screening setup relevant to industrial electrocatalysis. *Catalysts* **10**, 1165 (2020).
205. Tu, N. P. *et al.* High-throughput reaction screening with nanomoles of solid reagents coated on glass beads. *Angew. Chem. Int. Ed.* **131**, 8071–8075 (2019).
206. Bahr, M. N. *et al.* Collaborative evaluation of commercially available automated powder dispensing platforms for high-throughput experimentation in pharmaceutical applications. *Org. Process. Res. Dev.* **22**, 1500–1508 (2018).
207. Shen, Y. *et al.* Automation and computer-assisted planning for chemical synthesis. *Nat. Rev. Methods Primers* **1**, 23 (2021).
208. Fermier, A. M. *et al.* Powder dispensing robot for sample preparation. *Analyst* **128**, 790–795 (2003).
209. Gesmundo, N. J., Tu, N. P., Sarris, K. A. & Wang, Y. ChemBeads-enabled photoredox high-throughput experimentation platform to improve C(sp²)-C(sp³) decarboxylative couplings. *ACS Med. Chem. Lett.* **14**, 521–529 (2023).
210. Martin, M. C. *et al.* Versatile methods to dispense submilligram quantities of solids using chemical-coated beads for high-throughput experimentation. *Org. Process Res. Dev.* **23**, 1900–1907 (2019).
211. Impastato, A. C., Brown, J. T., Wang, Y. & Tu, N. P. Readily Accessible High-Throughput Experimentation: A General Protocol for the Preparation of ChemBeads and EnzyBeads. *ACS Medicinal Chemistry Letters* **14**, 514–520 (2023).
212. Brown, J. T., Tu, N. P. & Phelan, R. M. Solid, Noncovalent Formulation of Biocatalysts for Rapid and Accurate Submilligram Dosing to Microtiter Plates. *Org. Process Res. Dev.* **25**, 337–341 (2021).
213. Aguirre, A. L., Loud, N. L., Johnson, K. A., Weix, D. J. & Wang, Y. ChemBead enabled high-throughput cross-electrophile coupling reveals a new complementary ligand. *Chem. Eur. J.* **27**, 12981–12986 (2021).
214. Rago, A. J., Vasilopoulos, A., Dombrowski, A. W. & Wang, Y. Di (2-picolyl) amines as modular and robust ligands for nickel-catalyzed C (sp²)-C(sp³) cross-electrophile coupling. *Org. Lett.* **24**, 8487–8492 (2022).

215. Ward, R. M., Hu, Y., Tu, N. P. & Schomaker, J. M. Solvent effects on the chemo- and site-selectivity of transition metal-catalyzed nitrene transfer reactions: Alternatives to chlorinated solvents. *ChemSusChem*, e202300964 (2023).
216. Piacentini, P., Fordham, J. M., Serrano, E., Hepp, L. & Santagostino, M. Temperature-controlled photoreactors and ChemBeads as key technologies for robust and practical photochemical HTE. *Org. Process Res. Dev.* **27**, 798–810 (2023).
217. Fordham, J. M., Kollmus, P., Cavegn, M., Schneider, R. & Santagostino, M. A “pool and split” approach to the optimization of challenging Pd-catalyzed C–N cross-coupling reactions. *J. Org. Chem.* **87**, 4400–4414 (2022).
218. Kang, K., Loud, N. L., DiBenedetto, T. A. & Weix, D. J. A general, multimetallic cross-Ullmann biheteroaryl synthesis from heteroaryl halides and heteroaryl triflates. *J. Am. Chem. Soc.* **143**, 21484–21491 (2021).
219. Buitrago Santanilla, A., Christensen, M., Campeau, L.-C., Davies, I. W. & Dreher, S. D. P2Et Phosphazene: A mild, functional group tolerant base for soluble, room temperature Pd-catalyzed C–N, C–O, and C–C cross-coupling reactions. *Org. Lett.* **17**, 3370–3373 (2015).
220. Christensen, M. *et al.* Enantioselective Synthesis of α -Methyl- β -cyclopropyldihydrocinnamates. *J. Org. Chem.* **81**, 824–830 (2016).
221. Truppo, M. D., Rozzell, J. D., Moore, J. C. & Turner, N. J. Rapid screening and scale-up of transaminase catalysed reactions. *Org. Process Res. Dev.* **7**, 395–398 (2009).
222. Kempson, J. *et al.* Synthesis Optimization, Scale-Up, and Catalyst Screening Efforts toward the MGAT2 Clinical Candidate, BMS-963272. *Org. Process Res. Dev.* **26**, 1327–1335 (2022).
223. Cook, A., Clément, R. & Newman, S. G. Reaction screening in multiwell plates: high-throughput optimization of a Buchwald–Hartwig amination. *Nat. Protoc.* **16**, 1152–1169 (2021).
224. Rossetti, I. & Compagnoni, M. Chemical reaction engineering, process design and scale-up issues at the frontier of synthesis: Flow chemistry. *J. Chem. Eng.* **296**, 56–70 (2016).
225. Dudukovic, M. P. Reaction engineering: Status and future challenges. *Chem. Eng. Sci.* **65**, 3–11 (2010).
226. Krenkel, H. *et al.* Advancing liquid atmospheric pressure matrix-assisted laser desorption/ionization mass spectrometry toward ultrahigh-throughput analysis. *Anal. Chem.* **92**, 2931–2936 (2020).
227. Blincoe, W. D., Lin, S., Dreher, S. D. & Sheng, H. Practical guide on MALDI-TOF MS method development for high throughput profiling of pharmaceutically relevant, small molecule chemical reactions. *Tetrahedron* **76**, 131434 (2020).
228. Sawicki, J. W., Bogdan, A. R., Searle, P. A., Talaty, N. & Djuric, S. W. Rapid analytical characterization of high-throughput chemistry screens utilizing desorption electrospray ionization mass spectrometry. *React. Chem. Eng.* **4**, 1589–1594 (2019).
229. Wang, Y. *et al.* Acoustic droplet ejection enabled automated reaction scouting. *ACS Cent. Sci.* **5**, 451–457 (2019).
230. Sinclair, I. *et al.* Acoustic mist ionization platform for direct and contactless ultrahigh-throughput mass spectrometry analysis of liquid samples. *Anal. Chem.* **91**, 3790–3794 (2019).
231. DiRico, K. J. *et al.* Ultra-high-throughput acoustic droplet ejection-open port interface-mass spectrometry for parallel medicinal chemistry. *ACS Med. Chem. Lett.* **11**, 1101–1110 (2020).

232. Olson, D. L. *et al.* Microflow NMR: concepts and capabilities. *Anal. Chem.* **76**, 2966–2974 (2004).
233. Silva Elipe, M. V. *et al.* Application of the New 400 MHz High-Temperature Superconducting (HTS) Power-Driven Magnet NMR Technology for Online Reaction Monitoring: Proof of Concept with a Ring-Closing Metathesis (RCM) Reaction. *Org. Process Res. Dev.* **24**, 1428–1434 (2020).
234. Howarth, A., Ermanis, K. & Goodman, J. M. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chem. Sci.* **11**, 4351–4359 (2020).
235. Hardin, J. H. & Smietana, F. R. Automating combinatorial chemistry: A primer on bench-top robotic systems. *Mol. Divers.* **1**, 270–274 (1996).
236. Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
237. Mahjour, B. *et al.* Rapid planning and analysis of high-throughput experiment arrays for reaction discovery. *Nat. Comm.* **14**, 3924 (2023).
238. Mahjour, B., Hoffstadt, J. & Cernak, T. Designing Chemical Reaction Arrays using phac-tor and ChatGPT. *Org. Process Res. Dev.* **27**, 1510–1516 (2023).
239. Roch, L. M. *et al.* ChemOS: orchestrating autonomous experimentation. *Sci. Robot.* **3**, eaat5559 (2018).
240. Roch, L. M. *et al.* ChemOS: An orchestration software to democratize autonomous discovery. *PLoS One* **15**, e0229862 (2020).
241. Mason, J. *et al.* Automated LC-MS analysis and data extraction for high-throughput chemistry. *Digital Discovery* **2**, 1894–1899 (2023).
242. King-Smith, E. *et al.* Probing the chemical ‘reactome’ with high-throughput experimentation data. *Nat. Chem.*, 1–11 (2024).
243. Strieth-Kalthoff, F. *et al.* Machine learning for chemical reactivity: The importance of failed experiments. *Angew. Chem. Int. Ed.* **61**, e202204647 (2022).
244. Mercado, R., Kearnes, S. M. & Coley, C. W. Data sharing in chemistry: lessons learned and a case for mandating structured reaction data. *J. Chem. Inf. Model.* **63**, 4253–4265 (2023).
245. Kearnes, S. M. *et al.* The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
246. Tomczak, J. *et al.* UDM (Unified Data Model) for chemical reactions—past, present and future. *Pure Appl. Chem.* **94**, 687–704 (2022).
247. Wilson, W., Jeffrey, W., *et al.* A vending machine for drug-like molecules—automated synthesis of virtual screening hits. *Chem. Sci.* **13**, 14292–14299 (2022).
248. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
249. Wuest, T., Weimer, D., Irgens, C. & Thoben, K.-D. Machine learning in manufacturing: Advantages, challenges, and applications. *Prod. Manuf. Res.* **4**, 23–45 (2016).
250. Dogan, A. & Birant, D. Machine learning and data mining in manufacturing. *Expert Syst. Appl.* **166**, 114060 (2021).
251. Liakos, K. G., Busato, P., Moshou, D., Pearson, S. & Bochtis, D. Machine learning in agriculture: A review. *J. Sens.* **18**, 2674 (2018).
252. Giri, C., Jain, S., Zeng, X. & Bruniaux, P. A detailed review of artificial intelligence applied in the fashion and apparel industry. *IEEE Access* **7**, 95376–95396 (2019).

253. Woschank, M., Rauch, E. & Zsifkovits, H. A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics. *Sustainability* **12**, 3760 (2020).
254. Zantalis, F., Koulouras, G., Karabetsos, S. & Kandris, D. A review of machine learning and IoT in smart transportation. *Future Internet* **11**, 94 (2019).
255. Stocker, S., Csányi, G., Reuter, K. & Margraf, J. T. Machine learning in chemical reaction space. *Nat. Comm.* **11**, 5505 (2020).
256. Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-assisted analysis in organic synthesis. *Science* **228**, 408–418 (1985).
257. Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **34**, 247–266 (2005).
258. Cook, A. *et al.* Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 79–107 (2012).
259. Corey, E., Wipke, W. T., Cramer III, R. D. & Howe, W. J. Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. *J. Am. Chem. Soc.* **94**, 421–430 (1972).
260. Wipke, W. T. & Dyott, T. M. Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry. *J. Am. Chem. Soc.* **96**, 4825–4834 (1974).
261. Gelernter, H. *et al.* Empirical Explorations of SYNCHEM: The methods of artificial intelligence are applied to the problem of organic synthesis route discovery. *Science* **197**, 1041–1049 (1977).
262. Gelernter, H., Rose, J. R. & Chen, C. Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. *J. Chem. Inf. Comput.* **30**, 492–504 (1990).
263. Hendrickson, J. B. Systematic characterization of structures and reactions for use in organic synthesis. *J. Am. Chem. Soc.* **93**, 6847–6854 (1971).
264. Dugundji, J. & Ugi, I. Computers in Chemistry. *Top. Curr. Chem.* **19**–64 (1973).
265. Gasteiger, J. *et al.* A new treatment of chemical reactivity: Development of EROS, an expert system for reaction prediction and synthesis design in *Organic Synthesis, Reactions and Mechanisms* (1987), 19–73.
266. Jorgensen, W. L. *et al.* CAMEO: a program for the logical prediction of the products of organic reactions. *Pure Appl. Chem.* **62**, 1921–1932 (1990).
267. Satoh, H. & Funatsu, K. SOPHIA, a knowledge base-guided reaction prediction system—utilization of a knowledge base derived from a reaction database. *J. Chem. Inf. Comput.* **35**, 34–44 (1995).
268. Szymkuć, S. *et al.* Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
269. Molga, K., Szymkuc, S. & Grzybowski, B. A. Chemist ex machina: Advanced synthesis planning by computers. *Accounts of chemical research* **54**, 1094–1106 (2021).
270. Lowe, D. M. *Extraction of chemical structures and reactions from the literature* PhD thesis (University of Cambridge, 2012).
271. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
272. Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. & Laino, T. “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).

273. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
274. Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
275. Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
276. Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O. & Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **11**, 154–168 (2020).
277. Jorner, K., Tomberg, A., Bauer, C., Sköld, C. & Norrby, P.-O. Organic reactivity from mechanism to machine learning. *Nat. Rev. Chem.* **5**, 240–255 (2021).
278. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
279. Ertl, P. *et al.* Chemical reactivity prediction: Current methods and different application areas. *Mol. Inform.* **41**, 2100277 (2022).
280. Todeschini, R. & Consonni, V. *Handbook of molecular descriptors* (John Wiley & Sons, 2008).
281. Engel, T. Basic overview of chemoinformatics. *J. Chem. Inf. Model.* **46**, 2267–2277 (2006).
282. Corey, E. J. General methods for the construction of complex molecules. *Pure Appl. Chem.* **14**, 19–38 (1967).
283. Schwaller, P. *et al.* Machine intelligence for chemical reaction space. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **12**, e1604 (2022).
284. Vollmer, J. J. Wiswesser line notation: an introduction. *J. Chem. Educ.* **60**, 192 (1983).
285. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput.* **28**, 31–36 (1988).
286. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput.* **29**, 97–101 (1989).
287. Ash, S., Cline, M. A., Homer, R. W., Hurst, T. & Smith, G. B. SYBYL line notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput.* **37**, 71–79 (1997).
288. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput.* **32**, 244–255 (1992).
289. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics: Volume I: Alphabetical listing / volume II: Appendices, references* (John Wiley & Sons, 2009).
290. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
291. Awale, M. & Reymond, J.-L. Atom pair 2D-fingerprints perceive 3D-molecular shape and pharmacophores for very fast virtual screening of ZINC and GDB-17. *J. Chem. Inf. Model.* **54**, 1892–1907 (2014).
292. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminformatics* **12**, 1–15 (2020).
293. Reutlinger, M. *et al.* Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for ‘orphan’ molecules. *Mol. Inform.* **32**, 133 (2013).
294. Schneider, G., Neidhart, W., Giller, T. & Schmid, G. “Scaffold-hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.* **38**, 2894–2896 (1999).

295. Böhm, H.-J., Flohr, A. & Stahl, M. Scaffold hopping. *Drug Discov. Today* **1**, 217–224 (2004).
296. Renner, S. & Schneider, G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **1**, 181–185 (2006).
297. Schneider, G., Schneider, P. & Renner, S. Scaffold-hopping: how far can you jump? *QSAR Comb. Sci.* **25**, 1162–1171 (2006).
298. Schneider, P. & Schneider, G. Privileged structures revisited. *Angew. Chem. Int. Ed.* **56**, 7971–7974 (2017).
299. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959–5967 (1988).
300. Klebe, G. in *3D QSAR in Drug Design: Recent Advances* 87–104 (Springer, 1998).
301. Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
302. Huang, B. & von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **12**, 945–951 (2020).
303. Christensen, A. S., Bratholm, L. A., Faber, F. A. & Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **152** (2020).
304. Ballester, P. J. & Richards, W. G. Ultrafast shape recognition for similarity search in molecular databases. *Proc. R. Inst. G. B.* **463**, 1307–1321 (2007).
305. Brown, R. D. & Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput.* **37**, 1–9 (1997).
306. Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**, eabe4166 (2021).
307. Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
308. Fujita, S. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *J. Chem. Inf. Comput.* **26**, 205–212 (1986).
309. Hoonakker, F., Lachiche, N., Varnek, A. & Wagner, A. Condensed graph of reaction: considering a chemical reaction as one single pseudo molecule. *Int. J. Artif. Intell. Tools* **20**, 253–270 (2011).
310. Judson, P. N. *et al.* Adapting CHMTRN (chemistry translator) for a new use. *J. Chem. Inf. Model.* **60**, 3336–3341 (2020).
311. Delannée, V. & Nicklaus, M. C. ReactionCode: Format for reaction searching, analysis, classification, transform, and encoding/decoding. *J. Cheminformatics* **12**, 1–13 (2020).
312. Wigh, D. S., Goodman, J. M. & Lapkin, A. A. A review of molecular representation in the age of machine learning. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **12**, e1603 (2022).
313. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **1**, 045024 (2020).
314. Krenn, M. *et al.* SELFIES and the future of molecular string representations. *Patterns* **3** (2022).
315. Moret, M. *et al.* Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Comm.* **14**, 114 (2023).

316. Yuan, W. *et al.* Chemical space mimicry for drug discovery. *J. Chem. Inf. Model.* **57**, 875–882 (2017).
317. Anderson, J. A. *An introduction to neural networks* (MIT press, 1995).
318. Yegnanarayana, B. *Artificial neural networks* (PHI Learning Pvt. Ltd., 2009).
319. Rumelhart, D. E., Widrow, B. & Lehr, M. A. The basic ideas in neural networks. *Commun. ACM* **37**, 87–93 (1994).
320. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
321. Schmidhuber, J. Deep learning in neural networks: An overview. *J. Neural Netw.* **61**, 85–117 (2015).
322. Rumelhart, D. E., Hinton, G. E., Williams, R. J., *et al.* *Learning internal representations by error propagation* 1985.
323. Yoshimori, A., Chen, H. & Bajorath, J. Chemical language models for applications in medicinal chemistry. *Future Med. Chem.* **15**, 119–121 (2023).
324. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process.* **30** (2017).
325. Kamath, U., Graham, K. & Emará, W. *Transformers for Machine Learning: A Deep Dive* (Chapman and Hall/CRC, 2022).
326. Wolf, T. *et al.* *Transformers: State-of-the-art natural language processing in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (2020), 38–45.
327. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
328. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
329. Gupta, A. *et al.* Generative recurrent networks for de novo drug design. *Mol. Inform.* **37**, 1700111 (2018).
330. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inform.* **37**, 1700153 (2018).
331. Merk, D., Grisoni, F., Friedrich, L. & Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Comm. Chem.* **1**, 68 (2018).
332. Arús-Pous, J. *et al.* Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminformatics* **11**, 1–13 (2019).
333. Grisoni, F., Moret, M., Lingwood, R. & Schneider, G. Bidirectional molecule generation with recurrent neural networks. *J. Chem. Inf. Model.* **60**, 1175–1183 (2020).
334. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
335. Griffiths, R.-R. & Hernández-Lobato, J. M. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.* **11**, 577–586 (2020).
336. Kreutter, D., Schwaller, P. & Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **12**, 8648–8659 (2021).
337. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. learn.: Sci. Technol.* **2**, 015016 (2021).
338. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).

339. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
340. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
341. Shilpa, S., Kashyap, G. & Sunoj, R. B. Recent Applications of Machine Learning in Molecular Property and Chemical Reaction Outcome Predictions. *J. Phys. Chem. A* **127**, 8253–8271 (2023).
342. Segler, M. H. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**, 5966–5971 (2017).
343. Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H. & Hernández-Lobato, J. M. A generative model for electron paths. *arXiv preprint arXiv:1805.10970* (2018).
344. Kayala, M. A., Azencott, C.-A., Chen, J. H. & Baldi, P. Learning to predict chemical reactions. *J. Chem. Inf. Model.* **51**, 2209–2222 (2011).
345. Kayala, M. A. & Baldi, P. ReactionPredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. *J. Chem. Inf. Model.* **52**, 2526–2540 (2012).
346. Fooshee, D. *et al.* Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.* **3**, 442–452 (2018).
347. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2**, 725–732 (2016).
348. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. *Adv. Neural Inf. Process.* **30** (2017).
349. Do, K., Tran, T. & Venkatesh, S. *Graph transformation policy network for chemical reaction prediction in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (2019), 750–760.
350. Nikitin, F., Isayev, O. & Strijov, V. DRACON: Disconnected graph neural network for atom mapping in chemical reactions. *PCCP* **22**, 26478–26486 (2020).
351. Nam, J. & Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. *arXiv preprint arXiv:1612.09529* (2016).
352. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Comm.* **11**, 4874 (2020).
353. Wang, L., Zhang, C., Bai, R., Li, J. & Duan, H. Heck reaction prediction using a transformer model based on a transfer learning strategy. *ChemComm* **56**, 9368–9371 (2020).
354. Wu, Y., Zhang, C., Wang, L. & Duan, H. A graph-convolutional neural network for addressing small-scale reaction prediction. *ChemComm* **57**, 4114–4117 (2021).
355. Saebi, M. *et al.* On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **14**, 4997–5005 (2023).
356. Voinarovska, V., Kabeshov, M., Dudenko, D., Genheden, S. & Tetko, I. V. When yield prediction does not yield prediction: an overview of the current challenges. *J. Chem. Inf. Model.* (2023).
357. Horbaczewskyj, C. S. & Fairlamb, I. J. Pd-catalyzed cross-couplings: On the importance of the catalyst quantity descriptors, mol% and ppm. *Org. Process Res. Dev.* **26**, 2240–2269 (2022).
358. Fitzner, M., Wuitschik, G., Koller, R., Adam, J.-M. & Schindler, T. Machine learning C–N couplings: Obstacles for a general-purpose reaction yield prediction. *ACS Omega* **8**, 3017–3025 (2023).

359. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
360. Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **6**, 1379–1390 (2020).
361. Haywood, A. L. *et al.* Kernel methods for predicting yields of chemical reactions. *J. Chem. Inf. Model.* **62**, 2077–2092 (2021).
362. Chuang, K. V. & Keiser, M. J. Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science* **362**, eaat8603 (2018).
363. Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).
364. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
365. Eyke, N. S., Green, W. H. & Jensen, K. F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *React. Chem. Eng.* **5**, 1963–1972 (2020).
366. Rinehart, N. I. *et al.* A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C–N couplings. *Science* **381**, 965–972 (2023).
367. Fitzner, M. *et al.* What can reaction databases teach us about Buchwald–Hartwig cross-couplings? *Chem. Sci.* **11**, 13085–13093 (2020).
368. Ree, N., Göller, A. H. & Jensen, J. H. RegioSQM20: Improved prediction of the regioselectivity of electrophilic aromatic substitutions. *J. Cheminformatics* **13**, 1–9 (2021).
369. Struble, T. J., Coley, C. W. & Jensen, K. F. Multitask prediction of site selectivity in aromatic C–H functionalization reactions. *React. Chem. Eng.* **5**, 896–902 (2020).
370. Guan, Y. *et al.* Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **12**, 2198–2208 (2021).
371. Li, X., Zhang, S.-Q., Xu, L.-C. & Hong, X. Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning. *Angew. Chem. Int. Ed.* **59**, 13253–13259 (2020).
372. Banerjee, S., Sreenithya, A. & Sunoj, R. B. Machine learning for predicting product distributions in catalytic regioselective reactions. *PCCP* **20**, 18311–18318 (2018).
373. Beker, W., Gajewska, E. P., Badowski, T. & Grzybowski, B. A. Prediction of major regio-, site-, and diastereoisomers in diels–alder reactions by using machine-learning: the importance of physically meaningful descriptors. *Angew. Chem. Int. Ed.* **58**, 4515–4519 (2019).
374. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
375. Kireev, D. B. Chemnet: A novel neural network based method for graph/property mapping. *J. Chem. Inf. Comput.* **35**, 175–180 (1995).
376. Baskin, I. I., Palyulin, V. A. & Zefirov, N. S. A neural device for searching direct correlations between structures and properties of chemical compounds. *J. Chem. Inf. Comput.* **37**, 715–721 (1997).
377. Merkwirth, C. & Lengauer, T. Automatic generation of complementary descriptors with molecular graph networks. *J. Chem. Inf. Model.* **45**, 1159–1168 (2005).

378. Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process.* **28** (2015).
379. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148** (2018).
380. Unke, O. T. & Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
381. Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J. & Schneider, G. Δ -Quantum machine-learning for medicinal chemistry. *PCCP* **24**, 10775–10783 (2022).
382. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273 (2022).
383. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
384. Axelrod, S. & Gomez-Bombarelli, R. Molecular machine learning with conformer ensembles. *Mach. Learn.: Sci. Technol.* **4**, 035025 (2023).
385. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3**, 1023–1032 (2021).
386. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
387. Jiménez-Luna, J., Skalic, M., Weskamp, N. & Schneider, G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *J. Chem. Inf. Model.* **61**, 1083–1094 (2021).
388. Li, Y., Vinyals, O., Dyer, C., Pascanu, R. & Battaglia, P. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324* (2018).
389. De Cao, N. & Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973* (2018).
390. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9**, 10752 (2019).
391. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. *Adv. Neural Inf. Process.* **34**, 9405–9415 (2021).
392. Heinen, S., von Rudorff, G. F. & von Lilienfeld, O. A. Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space. *J. Chem. Phys.* **155** (2021).
393. Bragato, M., von Rudorff, G. F. & von Lilienfeld, O. A. Data enhanced Hammett-equation: reaction barriers in chemical space. *Chem. Sci.* **11**, 11859–11868 (2020).
394. Von Rudorff, G. F., Heinen, S. N., Bragato, M. & von Lilienfeld, O. A. Thousands of reactants and transition states for competing E2 and S2 reactions. *Mach. Learn.: Sci. Technol.* **1**, 045026 (2020).
395. Stuyver, T. & Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: performance, generalizability, and explainability. *J. Chem. Phys.* **156** (2022).
396. Qiu, J. *et al.* Selective functionalization of hindered meta-C–H bond of o-alkylaryl ketones promoted by automation and deep learning. *Chem* (2022).
397. Zaheer, M. *et al.* Deep sets. *Adv. Neural Inf. Process.* **30** (2017).
398. Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network. *Adv. Neural Inf. Process.* **32** (2019).

399. Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* **55**, 39–53 (2015).
400. Schneider, N., Stiefl, N. & Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *J. Chem. Inf. Model.* **56**, 2336–2346 (2016).
401. Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
402. Sayle, R. A., Mayfield, J. W., Lagerstedt, I. & Pirie, R. *NextMove Software Pistachio* <https://www.nextmovesoftware.com/index.html> (2024).
403. Elsevier. *Reaxys* <https://www.reaxys.com> (2024).
404. Services, C. A. *SciFinder* <https://scifinder.cas.org> (2024).
405. Chemistry, T. *Science of Synthesis* <https://www.thieme.de/en/thieme-chemistry/about-science-of-synthesis-54781.htm> (2024).
406. Jiang, S. *et al.* When SMILES smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access* **9**, 85071–85083 (2021).
407. Toniato, A., Schwaller, P., Cardinale, A., Geluykens, J. & Laino, T. Unassisted noise reduction of chemical reaction datasets. *Nat. Mach. Intell.* **3**, 485–494 (2021).
408. Polanczyk, G., De Lima, M. S., Horta, B. L., Biederman, J. & Rohde, L. A. The worldwide prevalence of ADHD: A systematic review and metaregression analysis. *Am. J. Psychiatry* **164**, 942–948 (2007).
409. Buchwald, H. *et al.* Bariatric surgery: A systematic review and meta-analysis. *JAMA* **292**, 1724–1737 (2004).
410. Fisch, C. & Block, J. Six tips for your (systematic) literature review in business and management research. *Manag. Rev. Q.* **68**, 103–106 (2018).
411. Thomé, A. M. T., Scavarda, L. F. & Scavarda, A. J. Conducting systematic literature review in operations management. *Prod. Plan. Control.* **27**, 408–420 (2016).
412. Huynh, M.-T., Nippa, M. & Aichner, T. Big data analytics capabilities: Patchwork or progress? A systematic review of the status quo and implications for future research. *Technol. Forecast. Soc. Change* **197**, 122884 (2023).
413. Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P. & Kyriakidou, O. Diffusion of innovations in service organizations: Systematic review and recommendations. *Milbank Q.* **82**, 581–629 (2004).
414. Elsevier. *Scopus* [www.https://scopus.com/search/](https://scopus.com/search/) (2024).
415. Clarivate. *Web of Science* <https://www.webofscience.com/wos/woscc/basic-search> (2024).
416. Li, J., Burnham, J. F., Lemley, T. & Britton, R. M. Citation analysis: Comparison of Web of Science®, scopus™, SciFinder®, and Google Scholar. *JERML* **7**, 196–217 (2010).
417. Rajan, K., Zielesny, A. & Steinbeck, C. DECIMER 1.0: Deep learning for chemical image recognition using transformers. *J. Cheminformatics* **13**, 1–16 (2021).
418. McDaniel, J. R. & Balmuth, J. R. Kekule: OCR-optical chemical (structure) recognition. *J. Chem. Inf. Comput.* **32**, 373–378 (1992).
419. Valko, A. T. & Johnson, A. P. CLiDE Pro: The latest generation of CLiDE, a tool for optical chemical structure recognition. *J. Chem. Inf. Model.* **49**, 780–787 (2009).
420. Park, J. *et al.* Automated extraction of chemical structure information from digital raster images. *Chem. Cent. J.* **3**, 1–16 (2009).

421. Ibison, P. *et al.* Chemical literature data extraction: The CLiDE Project. *J. Chem. Inf. Comput.* **33**, 338–344 (1993).
422. Qian, Y., Guo, J., Tu, Z., Coley, C. W. & Barzilay, R. RxnScribe: A sequence generation model for reaction diagram parsing. *arXiv preprint arXiv:2305.11845* (2023).
423. Wilary, D. M. & Cole, J. M. ReactionDataExtractor 2.0: A deep learning approach for data extraction from chemical reaction schemes. *J. Chem. Inf. Model.* **63**, 6053–6067 (2023).
424. Chen, T., Saxena, S., Li, L., Fleet, D. J. & Hinton, G. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852* (2021).
425. Guo, J. *et al.* Automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model* **62**, 2035–2045 (2021).
426. Nippa, D. F. *et al.* Simple User-Friendly Reaction Format. *ChemRxiv* (2023).
427. Nippa, D. F. *et al.* Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *Nat. Chem.* **16**, 239–248 (2024).
428. Hartwig, J. F. Borylation and silylation of C–H bonds: a platform for diverse C–H bond functionalizations. *Acc. Chem. Res.* **45**, 864–873 (2012).
429. Wang, M. & Shi, Z. Methodologies and strategies for selective borylation of C–Het and C–C bonds. *Chem. Rev.* **120**, 7348–7398 (2020).
430. Li, Y. & Wu, X.-F. Direct C–H bond borylation of (hetero) arenes: evolution from noble metal to metal free. *Angew. Chem. Int. Ed.* **59**, 1770–1774 (2020).
431. Tian, Y.-M., Guo, X.-N., Braunschweig, H., Radius, U. & Marder, T. B. Photoinduced borylation for the synthesis of organoboron compounds: Focus review. *Chem. Rev.* **121**, 3561–3597 (2021).
432. Neeve, E. C., Geier, S. J., Mkhaliid, I. A., Westcott, S. A. & Marder, T. B. Diboron (4) compounds: from structural curiosity to synthetic workhorse. *Chem. Rev.* **116**, 9091–9161 (2016).
433. Fyfe, J. W. & Watson, A. J. Recent developments in organoboron chemistry: old dogs, new tricks. *Chem* **3**, 31–55 (2017).
434. Suzuki, A. Recent advances in the cross-coupling reactions of organoboron derivatives with organic electrophiles, 1995–1998. *J. Org. Chem.* **576**, 147–168 (1999).
435. Blakey, S. B. & MacMillan, D. W. The first Suzuki cross-couplings of aryltrimethylammonium salts. *J. Am. Chem. Soc.* **125**, 6046–6047 (2003).
436. Campeau, L.-C. & Hazari, N. Cross-coupling and related reactions: Connecting past success to the development of new reactions for the future. *J. Organomet. Chem.* **38**, 3–35 (2018).
437. Saito, Y., Segawa, Y. & Itami, K. Para-C–H borylation of benzene derivatives by a bulky iridium catalyst. *J. Am. Chem. Soc.* **137**, 5193–5198 (2015).
438. Oeschger, R. *et al.* Diverse functionalization of strong alkyl C–H bonds by undirected borylation. *Science* **368**, 736–741 (2020).
439. Fier, P. S., Luo, J. & Hartwig, J. F. Copper-mediated fluorination of arylboronate esters. Identification of a Copper(III)fluoride complex. *J. Am. Chem. Soc.* **135**, 2552–2559 (2013).
440. Furuya, T. & Ritter, T. Fluorination of boronic acids mediated by Silver(I)triflate. *Org. Lett.* **11**, 2860–2863 (2009).
441. Furuya, T., Kaiser, H. M. & Ritter, T. Palladium-mediated fluorination of arylboronic acids. *Angew. Chem. Int. Ed.* **47**, 5993–5996 (2008).

442. Mazzotti, A. R., Campbell, M. G., Tang, P., Murphy, J. M. & Ritter, T. Palladium (III)-catalyzed fluorination of arylboronic acid derivatives. *J. Am. Chem. Soc.* **135**, 14012–14015 (2013).
443. Liskey, C. W., Liao, X. & Hartwig, J. F. Cyanation of arenes via iridium-catalyzed borylation. *J. Am. Chem. Soc.* **132**, 11389–11391 (2010).
444. Malapit, C. A., Reeves, J. T., Busacca, C. A., Howell, A. R. & Senanayake, C. H. Rhodium-catalyzed transnitration of aryl boronic acids with dimethylmalononitrile. *Angew. Chem. Int. Ed.* **128**, 334–338 (2016).
445. Luo, Y. *et al.* Copper-mediated cyanation of aryl boronic acids using benzyl cyanide. *Tetrahedron* **69**, 8400–8404 (2013).
446. Litvinas, N. D., Fier, P. S. & Hartwig, J. F. A general strategy for the perfluoroalkylation of arenes and arylbromides by using arylboronate esters and [(phen)CuRF]. *Angew. Chem. Int. Ed.* **51**, 536–539 (2012).
447. Ye, Y., Künzi, S. A. & Sanford, M. S. Practical method for the Cu-mediated trifluoromethylation of arylboronic acids with CF₃ radicals derived from NaSO₂CF₃ and tert-butyl hydroperoxide (TBHP). *Org. Lett.* **14**, 4979–4981 (2012).
448. Zhang, S.-L. & Bie, W.-F. Isolation and characterization of copper(III)trifluoromethyl complexes and reactivity studies of aerobic trifluoromethylation of arylboronic acids. *RSC Adv.* **6**, 70902–70906 (2016).
449. Murphy, J. M., Liao, X. & Hartwig, J. F. Meta halogenation of 1,3-disubstituted arenes via iridium-catalyzed arene borylation. *J. Am. Chem. Soc.* **129**, 15434–15435 (2007).
450. Szumigala, R. H., Devine, P. N., Gauthier, D. R. & Volante, R. Facile synthesis of 2-bromo-3-fluorobenzonitrile: An application and study of the halodeboration of aryl boronic acids. *J. Org. Chem.* **69**, 566–569 (2004).
451. Hume, P., Furkert, D. P. & Brimble, M. A. Regioselective Iridium(I)-catalysed remote borylation of oxygenated naphthalenes. *Tetrahedron Lett.* **53**, 3771–3773 (2012).
452. Tang, Y.-L., Xia, X.-S., Gao, J.-C., Li, M.-X. & Mao, Z.-W. Direct bromodeboration of arylboronic acids with CuBr₂ in water. *Tetrahedron Lett.* **64**, 152738 (2021).
453. Zhang, G., Lv, G., Li, L., Chen, F. & Cheng, J. Copper-catalyzed halogenation of arylboronic acids. *Tetrahedron Lett.* **52**, 1993–1995 (2011).
454. Woźniak, Ł. *et al.* Catalytic enantioselective functionalizations of C-H bonds by chiral iridium complexes. *Chem. Rev.* **120**, 10516–10543 (2020).
455. Oro, L. A. & Claver, C. *Iridium catalysts for organic reactions* (Springer Nature, 2021).
456. Parry, D. M. Closing the loop: Developing an integrated design, make, and test platform for discovery. *ACS Med. Chem. Lett.* **10**, 848–856 (2019).
457. Nippa, D. F. *et al.* Identifying opportunities for late-stage C-H alkylation with high-throughput experimentation and in silico reaction screening. *Commun. Chem.* **6**, 256 (2023).
458. Minisci, F., Bernardi, R., Bertini, F., Galli, R. & Perchinummo, M. Nucleophilic character of alkyl radicals—VI: A new convenient selective alkylation of heteroaromatic bases. *Tetrahedron* **27**, 3575–3579 (1971).
459. Minisci, F., Vismara, E. & Romano, U. Silver-mediated oxidative decarboxylation of carboxylic acids by peroxocompounds new sources of carbon-centered radicals for heteroaromatic substitution. *Tetrahedron Lett.* **26**, 4803–4806 (1985).
460. Duncton, M. A. Minisci reactions: Versatile CH-functionalizations for medicinal chemists. *MedChemComm* **2**, 1135–1161 (2011).

461. Proctor, R. S. & Phipps, R. J. Recent advances in Minisci-type reactions. *Angew. Chem. Int. Ed.* **58**, 13666–13699 (2019).
462. Tang, R.-J., Kang, L. & Yang, L. Metal-free oxidative decarbonylative coupling of aliphatic aldehydes with azaarenes: Successful Minisci-type alkylation of various heterocycles. *ASC* **357**, 2055–2060 (2015).
463. Ren, P., Salihu, I., Scopelliti, R. & Hu, X. Copper-catalyzed alkylation of benzoxazoles with secondary alkyl halides. *Org. Lett.* **14**, 1748–1751 (2012).
464. McCallum, T., Jouanno, L.-A., Cannillo, A. & Barriault, L. Persulfate-enabled direct C-H alkylation of heteroarenes with unactivated ethers. *Synlett* **27**, 1282–1286 (2016).
465. Zhang, L. & Liu, Z.-Q. Molecular oxygen-mediated Minisci-type radical alkylation of heteroarenes with boronic acids. *Org. Lett.* **19**, 6594–6597 (2017).
466. Fujiwara, Y. *et al.* Practical and innate carbon–hydrogen functionalization of heterocycles. *Nature* **492**, 95–99 (2012).
467. Antonchick, A. P. & Burgmann, L. Direct selective oxidative cross-coupling of simple alkanes with heteroarenes. *Angew. Chem. Int. Ed.* **52**, 3267–3271 (2013).
468. Kan, J., Huang, S., Lin, J., Zhang, M. & Su, W. Silver-catalyzed arylation of (hetero)arenes by oxidative decarboxylation of aromatic carboxylic acids. *Angew. Chem. Int. Ed.* **54**, 2199–2203 (2015).
469. Sato, N. & Matsuura, T. Studies on pyrazines. Part 32. Synthesis of trisubstituted and tetrasubstituted pyrazines as ant pheromones. *J. Chem. Soc., Perkin Trans.*, 2345–2350 (1996).
470. A. AmrollahiBiyouki, M., AJ Smith, R., Bedford, J. J. & Leader, J. P. Hydroxymethylation and Carbamoylation of Di-And Tetramethylpyridines Using Radical Substitution (Minisci) Reactions. *Synth. Commun.* **28**, 3817–3825 (1998).
471. Xie, Z.-F., Ootsu, K. & Akimoto, H. Convergent approach to water soluble camptothecin derivatives. *Bioorganic Med. Chem. Lett.* **5**, 2189–2194 (1995).
472. Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: Increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**, 6752–6756 (2009).
473. Lovering, F. Escape from flatland 2: Complexity and promiscuity. *MedChemComm* **4**, 515–519 (2013).
474. Ingold, K. & DiLabio, G. A. Bond strengths: The importance of hyperconjugation. *Org. Lett.* **8**, 5923–5925 (2006).
475. Galloway, J. D., Mai, D. N. & Baxter, R. D. Silver-catalyzed Minisci reactions using selectfluor as a mild oxidant. *Org. Lett.* **19**, 5772–5775 (2017).
476. Sutherland, D. R., Veguillas, M., Oates, C. L. & Lee, A.-L. Metal-, photocatalyst-, and light-free, late-stage C-H alkylation of heteroarenes and 1,4-quinones using carboxylic acids. *Org. Lett.* **20**, 6863–6867 (2018).
477. Bosset, C. *et al.* Minisci-photoredox-mediated α -heteroarylation of *N*-protected secondary amines: Remarkable selectivity of azetidines. *Org. Lett.* **20**, 6003–6006 (2018).
478. Wang, Q.-Q. *et al.* Electrocatalytic Minisci acylation reaction of *N*-heteroarenes mediated by NH_4I . *Org. Lett.* **19**, 5517–5520 (2017).