Application of machine learning for the discovery and optimization of new nanocarrier formulations

Felix Sieber-Schäfer

2025

Dissertation zur Erlangung des Doktorgrades

der Fakultät für Chemie und Pharmazie

der Ludwig-Maximilians-Universität München

Application of machine learning for the discovery and optimization of new nanocarrier formulations

Felix Sieber-Schäfer

aus

Regensburg, Deutschland

2025

Erklärung und Eidesstattliche Versicherung

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Frau Prof. Dr. Olivia M. Merkel betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfsmittel erarbeitet.

München, den 27.11.2025

Felix Sieber-Schäfer

Dissertation eingereicht am: 27.11.2025

1. Gutachterin: Prof. Dr. Olivia M. Merkel

2. Gutachter: Prof. Dr. Peter M. Tessier

Mündliche Prüfung am: 13.01.2026

Für meine Familie

Table of Content

List of Abbreviations

1H NMR	Proton nuclear magnetic resonance spectroscopy
QSTR	Quantitative structure-transfection relationship
5-CV	Five-fold cross-validation
AA	All-atom
AF647	Alexa Fluor 647
AI	Artificial intelligence
AL	Active learning
ALI	Air-liquid interface
ANOVA	Analysis of variance
AP	Polymer derived from 5-aminopentan-1-ol
ASGPR	Asialoglycoprotein receptor
ASO	Antisense oligonucleotide
ATP	Adenosine 5′-triphosphate
AUC	Area under the curve
BALF	Bronchoalveolar lavage fluid
BG	Bisphenol A glycerolate
CCK-8	Cell Counting Kit-8
CCO	Central composite design for maximized orthogonality

cDNA	Complementary DNA
CG-MD	Coarse-grained Molecular Dynamics
CLSM	Confocal laser scanning microscopy
CMV	Cytomegalovirus
COPD	Chronic obstructive pulmonary disease
CPP	Critical process parameter
CQA	Critical quality attribute
CTB	CellTiter Blue
CV	Cross-validation
DA	1,4-butanediol diacrylate
DAPI	4′,6-Diamidino-2-phenylindole
DAR	Diacrylate ratio
DCM	Dichloromethane
DEPC	Diethyl pyrocarbonate
DMEM	Dulbecco’s Modified Eagle’s Medium
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DoE	Design of Experiments
dsRNA	Double-stranded RNA

DSPC	1,2-Distearoyl-sn-glycero-3-phosphocholine
DT	Decision Tree
DTT	Dithiothreitol
EC50	Half-maximal effective concentration
EDTA	Ethylenediaminetetraacetic acid
EE	Encapsulation efficiency
EF	Enrichment factor
eGFP	Enhanced green fluorescent protein
EGFR	Epidermal growth factor receptor
FBS	Fetal bovine serum
FDA	U.S. Food and Drug Administration
FLuc	Firefly luciferase
FoMAML	First-order model-agnostic meta-learning
FRR	Flow-rate ratio
FSL	Few-shot learning
GA	Genetic Algorithm
GalNAc	N-acetylgalactosamine
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GCN	Graph Convolutional Network
GPC	Gel permeation chromatography

H&E	Hematoxylin and eosin
hATTR	Hereditary transthyretin-mediated amyloidosis
HBG	HEPES-buffered glucose
HEK-293T	Human embryonic kidney cells
HeLa	Human cervix carcinoma cell line
HEPES	4-(2-Hydroxyethyl)-1-piperazineethanesulfonic acid
hPCLS	Human precision-cut lung slices
HT	High-throughput
HTS	High-throughput screening
IFN-β	Interferon beta
IL-6	Interleukin-6
IVIS	In vivo imaging system
KD	Knockdown
LAX	Lipo-xenopeptides
LDH	Lactate dehydrogenase
LGBM	Light Gradient Boosting Machine
LNP	Lipid nanoparticle
LOOCV	Leave-one-out cross-validation
MACCS	Molecular ACCess System
MAE	Mean absolute error

MAML	Model-agnostic meta-learning
MCP-1	Monocyte chemoattractant protein-1
MD	Molecular dynamics
MFI	Median fluorescence intensity
ML	Machine learning
MMFF	Merck Molecular Force Field
mRNA	Messenger RNA
MS ESI	Mass spectrometry electrospray ionization
MSD	Mean square displacement
MTT	3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2H-tetrazolium bromide
Mw	Weight-average molar mass
N/P	Molar ratio of amines (N) to phosphate groups (P)
N2a	Murine neuroblastoma cell line
NA	Nucleic acid
NLP	Natural language processing
OA	Oleylamine
OD	Optical density
OFAT	One factor at a time
P(SpOABAE)	Poly-spermine-co-oleylamine beta-aminoesters
P/S	Penicillin/streptomycin

PBAE	Poly(β-amino ester)
PBS	Phosphate-buffered saline
PDI	Polydispersity index
PEI	Polyethylenimine
PFA	Paraformaldehyde
PLL	Poly-L-lysine
PLO	Poly-L-ornithine
PME	Particle Mesh Ewald
POPC	1-Palmitoyl-2-oleoyl-phosphatidylcholine
qPCR	Quantitative polymerase chain reaction
QSAR	Quantitative structure–activity relationship
QSTR	Quantitative structure–transfection relationship
RF	Random forest
RISC	RNA-induced silencing complex
RLU	Relative light units
RNA	Ribonucleic acid
RNAi	RNA interference
RNase	Ribonuclease
RSM	Response surface method
RT	Room temperature

SA	Synthetic Accessibility
SD	Standard deviation
SHAP	SHapley Additive exPlanations
siGAPDH	siRNA targeting GAPDH
siGFP	siRNA targeting EGFP
siNC	Scrambled/negative-control siRNA
siRNA	Small interfering RNA
SMD	Steered Molecular Dynamics
SP	Spermine
SVR	Support Vector Regressor
TBS	Tri-Boc-spermine (tris(tert-butoxycarbonyl)spermine)
TDA	Tetradecylamine
TEER	Transepithelial electrical resistance
TFA	Trifluoroacetic acid
TFE-IDA	N-(trifluoroethyl)iminodiacetyl
TFR	Total flow rate
TL	Transfer learning
TNF-α	Tumor necrosis factor alpha
UMAP	Uniform Manifold Approximation and Projection
UTR	Untreated controls

vHTS	Virtual high-throughput screening
WO	Water-octanol

Contribution to Publications

Chapter II – Design of Experiments Grants Mechanistic Insights into the Synthesis of Spermine-Containing PBAE Copolymers

Adrian P. E. Kromer, Felix Sieber-Schäfer, Johan Farfan, Olivia M. Merkel

This chapter was published as a research article in *ACS Applied Materials & Interfaces*, 2024 (doi: 10.1021/acsami.4c06079). Adrian P. E. Kromer and I contributed as shared first authors. We jointly performed the synthesis and experiments, with additional support from Johan Farfan. The Design of Experiments analysis and stability assays were carried out by Adrian P. E. Kromer. I developed and evaluated the PeakFinder script. The manuscript was written collaboratively by Adrian P. E. Kromer and myself. Olivia M. Merkel reviewed the manuscript and provided scientific guidance throughout the project.

Chapter III – Machine Learning on a Small Orthogonal Polymer Library Reveals Functional Insights and Optimizes PBAE Copolymer Synthesis and Performance

Felix Sieber-Schäfer, Adrian P. E. Kromer, Müge Molbay, Simone Carneiro, Min Jiang, Anny Nguyen, Joschka Müller, Johan Farfan Benito, and Olivia M. Merkel

This chapter is currently under review as a research article in *Biomaterials*. Adrian P. E. Kromer and I contributed as shared first authors. The in vivo experiments were performed by Adrian P. E. Kromer, Müge Molbay, Simone Carneiro, Min Jiang, Anny Nguyen, and Joschka Müller. Most of the in vitro experiments were conducted by Adrian P. E. Kromer and Johan Farfan Benito. I designed and implemented all code, and executed and evaluated the machine learning analyses. The manuscript was written collaboratively by Adrian P. E. Kromer and myself. Müge Molbay reviewed the manuscript. Olivia M. Merkel reviewed the manuscript and provided scientific guidance throughout the project.

Chapter IV – Machine Learning-Enabled Polymer Discovery for Enhanced Pulmonary siRNA Delivery

Felix Sieber-Schäfer, Min Jiang, Adrian Kromer, Anny Nguyen, Müge Molbay, Simone P. Carneiro, David Jürgens, Gerald Burgstaller, Bastian Popper, Benjamin Winkeljann, Olivia M. Merkel

This chapter was published as a research article in *Advanced Functional Materials*, 2025 (doi: 10.1002/adfm.202502805). Min Jiang and I contributed as shared first authors. Min

Jiang conducted the in vitro assays, nanoparticle formulation, and ex vivo experiments. I generated the dataset, performed the polymer synthesis as well as the GPC and NMR measurements, and carried out all machine learning-related tasks. Adrian P. E. Kromer, Anny Nguyen, Müge Molbay, and Simone P. Carneiro performed the in vivo experiments. Gerald Burgstaller provided the hPCLS samples, and Bastian Popper acquired the H&E-stained images of animal lung tissue. The manuscript was written collaboratively by Min Jiang and myself. Benjamin Winkeljann and Olivia M. Merkel reviewed the manuscript and provided scientific guidance throughout the project.

Chapter V – From Bits to Bonds - High throughput virtual screening of RNA nanocarriers using a combinatorial approach of Machine Learning and Molecular Dynamics

Felix Sieber-Schäfer, Jonas Binder, Tim Münchrath, Katharina M. Steinegger, Min Jiang, Benjamin Winkeljann, Wolfgang Friess, Olivia M. Merkel

This chapter was published as a research article in *Journal of the American Chemical Society*, 2025 (doi: 10.1021/jacs.5c12694). Jonas Binder and I contributed as shared first authors. Jonas Binder performed all molecular dynamics simulations and implemented several modules within the main script. I performed conceptualisation, carried out all machine learning-related tasks and implemented several modules in the main script. Tim Münchrath and I conducted all validation experiments and synthesized the polymers. Katharina M. Steinegger performed the parametrization of the bead matrix. Min Jiang conducted the in vitro cell culture experiments. The manuscript was written collaboratively by Jonas Binder and myself. Benjamin Winkeljann, Wolfgang Friess, and Olivia M. Merkel reviewed the manuscript and provided scientific guidance throughout the project.

Chapter VI – Capturing Molecular Motion by Integrating MD-Derived Descriptors into Predictive Machine Learning Models for RNA delivery

Felix Sieber-Schäfer, Nora Martini, Sophie Thalmayer, Tobias Burghardt, Melina Grau, Ernst Wagner, Benjamin Winkeljann, Olivia M. Merkel

This chapter is about to be submitted to *ACS Nano*. Nora Martini and I contributed as shared first authors. Nora Martini performed all molecular dynamics-related work. I was responsible for the conceptualization of the study and for all machine learning–related tasks. Sophie Thalmayer, Tobias Burghardt, and Melina Grau carried out the synthesis and experimental data curation. The manuscript was written collaboratively by Nora Martini and myself. Benjamin Winkeljann and Olivia M. Merkel reviewed the manuscript. Ernst Wagner, Benjamin Winkeljann and Olivia M. Merkel, provided scientific guidance throughout the project.

Chapter VII – Meta-Learning as a Promising Strategy for Lipid Nanoparticle Optimization and Ionizable Lipid Discovery

Felix Sieber-Schäfer, Lasse Hagedorn, Leon Reger, Katharina Möbius, Benjamin Winkeljann, Olivia M. Merkel

This chapter is about to be submitted to *ACS Nano Letters*. Lasse Hagedorn and I contributed as shared first authors. Lasse Hagedorn and Katharina Möbius performed the synthesis of all tested ionizable lipids. I was responsible for the conceptualization of the study and for all machine learning-related tasks. Lasse Hagedorn, Leon Reger, and Katharina Möbius carried out the nanoparticle formulations. Leon Reger conducted all in vitro cell culture assays. The manuscript was written collaboratively by Lasse Hagedorn and myself. Benjamin Winkeljann and Olivia M. Merkel reviewed the manuscript and provided scientific guidance throughout the project.

Chapter I - Introduction

1 XNA in Therapy

The emergence of nucleic acids as therapeutics has fundamentally reshaped how we understand and treat disease. Targets long considered undruggable are becoming tractable, opening mechanism-based options for patients like recent successful treatments of β-thalassemia[1] and heterozygous familial hypercholesterolemia[2] could show. While the first promising trials were conducted deoxyribonucleic acid (DNA)-based[3,4], the most rapid advances in recent years have come from ribonucleic acid (RNA), whose diverse roles in gene regulation make it a particularly versatile therapeutic substrate. RNA modalities, including messenger RNA (mRNA), small interfering RNA (siRNA), and antisense oligonucleotides (ASOs), enable direct, programmable modulation of gene expression with tunable duration, driving a broad wave of innovation across indications (Figure I.1).

During the COVID-19 pandemic, mRNA platforms provided a definitive proof-of-concept for the rapid development and deployment of novel vaccines[5,6]. As a therapeutic modality, mRNA enables transient, in vivo expression of defined proteins and is therefore well suited to protein-replacement strategies. Ongoing clinical programs are evaluating mRNA for monogenic metabolic disorders such as propionic acidemia[7], methylmalonic acidemia[8], and ornithine transcarbamylase deficiency[9].

While mRNA has shown clear success in protein-replacement therapy, gene-silencing approaches are suited to diseases driven by toxic, mutant, or dysregulated gene expression. Two widely used oligonucleotide modalities are antisense oligonucleotides (ASOs) and small interfering RNAs (siRNAs). ASOs are single-stranded, chemically modified oligomers that act either by RNase H1-mediated cleavage of the target RNA or by steric blocking to modulate pre-mRNA splicing or inhibit translation. These mechanisms occur primarily in the nucleus and can also operate in the cytoplasm[10]. In contrast, siRNAs are double stranded RNAs that engage the RNA-induced silencing complex (RISC). After guide-strand loading into Argonaute-2, the complex cleaves complementary cytosolic mRNA, leading to its degradation and durable gene silencing[11].

Although nucleic-acid therapeutics are among the most promising drug classes of this century, delivery remains the principal hurdle. Ubiquitous endo- and exonucleases, rapid renal clearance, and innate immune recognition can degrade or eliminate nucleic acids before cellular uptake. Chemically modified ASOs and N-acetylgalactosamine (GalNAc)–conjugated siRNAs can often be dosed subcutaneously without a vector for hepatocyte targets via asialoglycoprotein receptor (ASGPR)[12], whereas most other modalities (e.g., mRNA, plasmid DNA) and extrahepatic siRNA delivery still require dedicated delivery systems.

Viral vectors leverage evolved entry mechanisms and showed great success by offering the vector for the first gene therapy ever approved[13], but face constraints related to immunogenicity[14], manufacturing complexity[15], and payload limits[16], motivating alternative strategies. Non-viral carriers offer modular design, scalable manufacturing, and opportunities for targeting while mitigating several safety concerns associated with viral delivery.

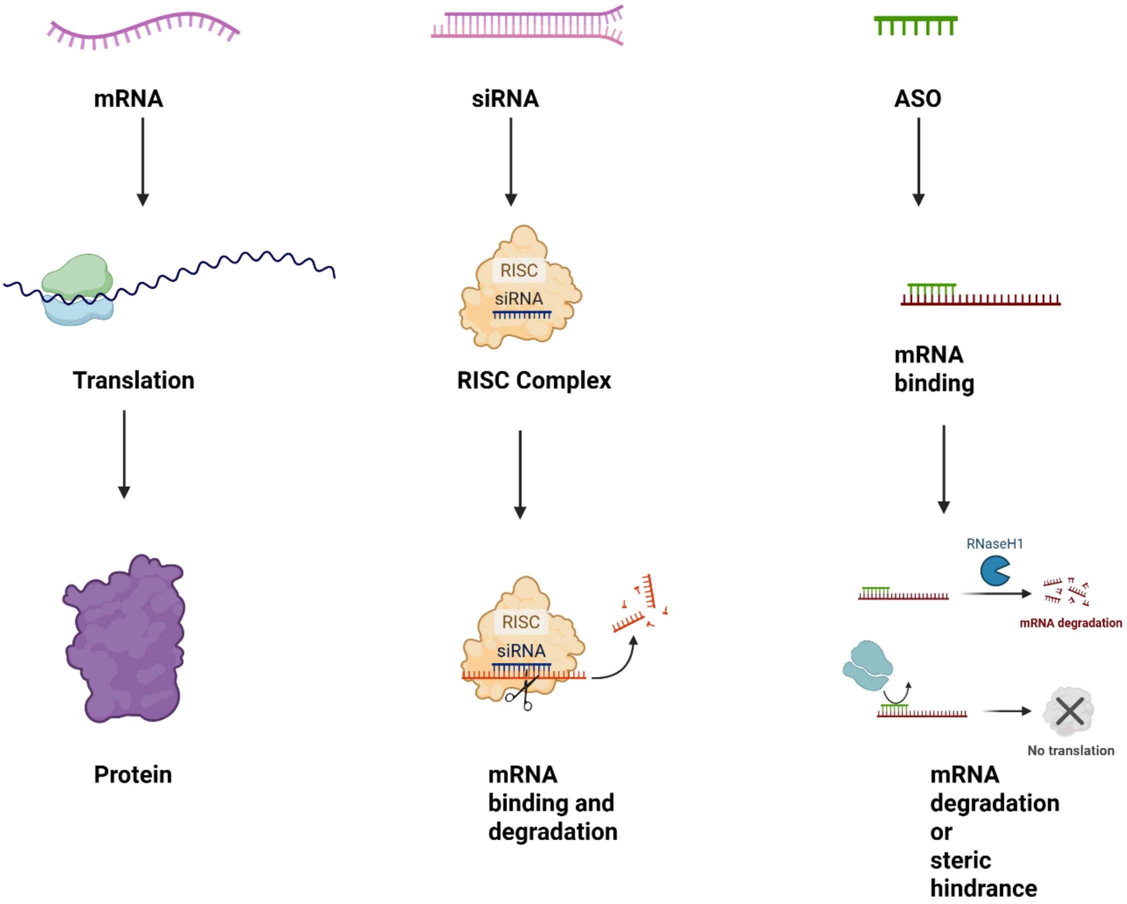


Figure I.1: Therapeutic functionalities for different nucleic acids. mRNA (left) is translated by cytosolic ribosomes to produce protein. siRNA (middle) is loaded into Argonaute to form the RNA-induced silencing complex (RISC), which directs sequence-specific cleavage and degradation of complementary mRNA. Antisense oligonucleotides (ASOs, right) bind target RNA to induce RNase H–mediated degradation or sterically block key processes such as translation or splicing.

2 Nanocarriers

Early non-viral nucleic-acid delivery in the 1960s used liposomes, simple polycations such as poly-L-lysine (PLL) and poly-L-ornithine (PLO)[17], and calcium phosphate precipitation[18]. Today the field is dominated by polycationic polymers and lipid assemblies, especially lipid nanoparticles (LNPs).

Polymers are well-studied vehicles for encapsulating and delivering nucleic acids. Their chemical tunability, architectural control, and colloidal stability are major advantages and have led to a wide variety of polymeric nanocarriers over the years[17,19]. Polyethylenimine (PEI), discovered as carrier 1995, was first used for efficient DNA transfection[20] and later adapted for mRNA[21] and siRNA[22]. Its high density of protonatable amines enables strong condensation and protection and supports endosomal escape via hypothesized proton-sponge effect, but this same feature is linked to cytotoxicity[23,24] and the lack of biodegradability[25] further raises safety concerns. Poly(β-amino esters) (PBAEs), which were introduced as gene carrier by Lynn and Langer in 2000[26], offer a biodegradable alternative. Their ester bonds hydrolyze into small by-products, and the chemistry is highly tunable via side-chain, backbone, and end-group modifications and through formulation choices. In addition, many PBAEs exhibit buffering capacity near endosomal pH, which can aid endosomal escape while maintaining a more favourable safety profile[27]. While promising, cytotoxicity concerns and reproducibility issues still limit the application of polymeric nanocarriers in clinical trials[28]. By contrast, lipid-based systems like liposomes and especially lipid nanoparticles (LNPs) have achieved the fastest clinical progress, exemplified by patisiran (Onpattro)[29], approved in 2018 for polyneuropathy in adults with hereditary transthyretin-mediated amyloidosis (hATTR), and by the LNP-based mRNA COVID-19 vaccines BNT162b2[6] and mRNA-1273[5]. Lipid nanoparticles (LNPs) typically comprise four components: an ionizable lipid that complexes the nucleic acid and promotes endosomal escape via pH-triggered protonation, a helper phospholipid, cholesterol (or a related sterol) to modulate membrane packing, and a PEG-lipid for steric shielding and colloidal stability[30]. While this platform is highly successful, optimizing four interdependent constituents remains non-trivial and remains a challenge especially in optimization. Lipo-xenopeptides offer a compelling alternative: like polymers they enable one-component formulations, yet, thanks to solid-phase synthesis, their molecular weight and composition are precisely defined. Thalmayer et al.[31] demonstrated stable lipopolyplex formation with promising in-vitro and in-vivo performance across multiple cargos. Still, safety metrics have not yet reached clinically relevant thresholds, and scaling production while preserving sequence fidelity may be challenging. Nanocarriers face multiple, interlocking hurdles: maintaining chemical and colloidal stability within a defined window, achieving an appropriate pK_a range for charge switching[32], and overcoming poorly understood mechanisms like endosomal escape and subsequent cytosolic release ultimately govern efficacy[33,34]. Manufacturing, transport, and storage further shape performance through process and logistics variables (e.g., mixing regime, sterile filtration, lyophilization, and cold-chain requirements)[35–37]. Consequently, nucleic-acid formulation with nanocarriers is a multidimensional, multi-objective optimization problem spanning molecular design, process engineering, and use-context parameters (route of administration, dose, repeat dosing, and target tissue), all of which interact to determine required outcomes.

3 Discovery and Optimization of RNA Formulations

As described above formulation optimization is a multi-stage problem coupling chemistry, mixtures into a single, high-dimensional problem. While performance is driven by the properties of the chemical compounds and the cargo, the formulation process itself plays a major role as well. The manufacturing of nanoparticles, especially LNPs, is typically carried out with microfluidic devices[38,39], where chip architecture and process conditions determine particle characteristics such as size, polydispersity index (PDI), encapsulation efficiency and biodistribution[40,41]. Whereas manufacturing conditions like flow-rate ratio (FRR) and total flow rate (TFR) can usually be treated as continuous variables, molecular identities are more complex and clearly multidimensional. Optimization therefore often treats them as discrete choices for simplicity, risking a loss of chemically relevant information. Compounding this, the ratios of components used in the formulation are crucial and add further complexity.

There are three conventional approaches commonly used for formulation optimization. The classical lab-scale route is adjustment of one factor at a time (OFAT). While this enables sequential tuning, from chemistry to formulation and subsequent post-processing such as drying, OFAT ignores factor interactions and almost invariably misses global optima, which is critical in any true optimization. High-throughput screening (HTS) addresses this by sampling many potential carriers and, ideally, varying process conditions in parallel[27,42,43]. However, HTS requires equipment that is not available in every laboratory and can be material-intensive when the experimental grid is narrow. Moreover, selection of the screening grid often relies more on educated guesses than on systematic, quantitative design and is therefore prone to bias. Design of Experiments (DoE) reduces the number of required experiments by using statistical designs that balance information gain against experimental effort while post-analysis then fits response surfaces that provide process insight and help identify sweet spots[44]. Additionally, the opportunity to use different designs like full factorial, latin hypercube and mixture design, to just mention a few, increases the flexibility when solving different types of problems. Although DoE is frequently considered a gold standard in industrial formulation work, it benefits from prior knowledge of the process, and the often complex, high-dimensional response landscape can still demand a large number of experiments. In practice, DoE is therefore often applied to individual sub-tasks only[45,46].

Nevertheless, data-driven decision-making is both relevant and beneficial. Machine learning (ML), as a branch of artificial intelligence (AI), offers ways to uncover patterns in complex processes and is an attractive tool that can be adapted to the needs of formulation science, potentially enabling global optimization as well as explainability (Figure I.2). In the next section, ML is briefly introduced, and Section 6 outlines how ML helps treat molecules as informative data, an aspect typically lacking in the conventional methods described above.

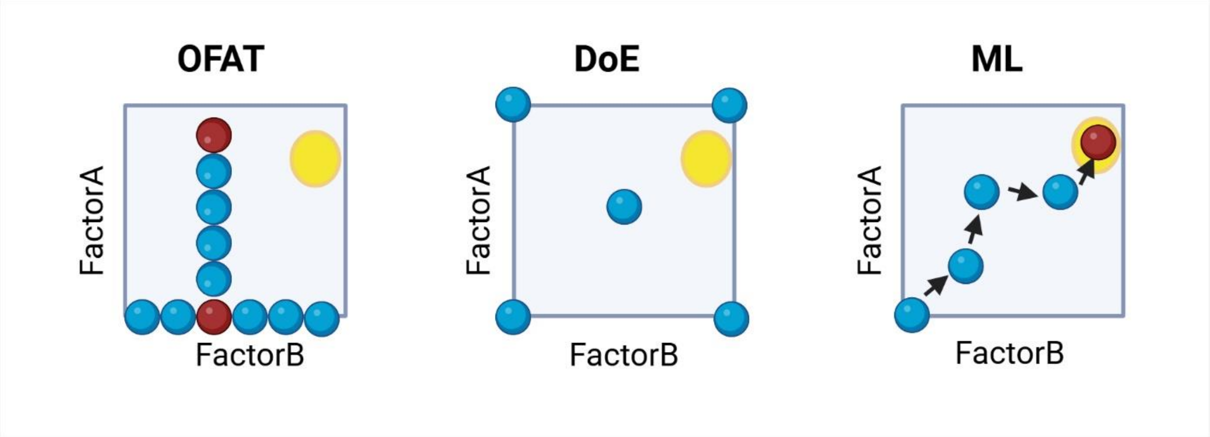


Figure I.2: Comparing optimization strategies. Classical one-factor-at-a-time (OFAT) varies a single variable while holding others constant, often missing interactions and trapping the search near local optima. Design of Experiments (DoE) systematically samples the factor space, enabling interpolation and estimation of interactions via response surfaces. Data-driven and machine-learning workflows build on these data to iteratively propose new experiments, improving efficiency and increasing the likelihood of identifying the global optimum.

4 Machine Learning

Machine learning is, strictly speaking, an intersection of software development and data science. It designs algorithms that learn from data to forecast outcomes for unseen instances. A model learns to predict a target value, often called the label **y**, based on available information represented by known variables or features **X**. During training the model receives a dataset with known y values and attempts to predict them. The error is computed with a loss function, and the model is optimized to reduce this loss. In general, machine learning can be viewed as a process whose goal is to minimize a loss. This description refers to supervised learning, which is one of the largest areas of applied machine learning and the focus of this section. It is also important to distinguish regression, where the target is continuous, from classification, where the target consists of discrete classes.

The usual workflow begins with careful data cleaning to remove errors and duplicates, followed by a split into training and test sets. This split is essential for assessing generalizability beyond the data seen during training and for detecting overfitting. Overfitting occurs when a model learns noise in the training set and then performs poorly on new data. After the split, the model is trained and its hyperparameters are tuned. Hyperparameters are settings that are chosen before training rather than learned during training. A common approach is K fold cross validation, where the training data is divided into K folds. Each fold is used once for evaluation while the remaining folds are used for training, and this procedure is repeated across all candidate hyperparameter settings. The process is illustrated in Figure I.3.

Model choice also matters. The No Free Lunch theorem[47] states that no single model is universally superior and that the best choice depends on the data. Linear regression fits a linear relationship by learning a parameter vector that minimizes a squared loss. For small datasets with moderate dimensionality one can solve directly with the normal equation. The linear model can be extended by mapping features into polynomial bases. This can improve accuracy but also raises the risk of overfitting. For classification, logistic regression applies a sigmoid function to produce probabilities and then assigns classes using a threshold.

Tree based models split data into leaves using a loss such as the Gini impurity for classification or the squared error for regression. Individual decision trees are flexible but can overfit, which motivates regularization. Strong regularization can then underfit. Two ensemble strategies address this tension. Bagging, as in Random Forests[48], trains many trees on resampled data or feature subsets and averages their predictions. Boosting trains trees sequentially, each one focusing on the errors of the previous model. Prominent examples for boosting are XGBoost[49], LightGBM[50] and CatBoost[51].

Kernels provide another elegant route to nonlinearity. A kernel defines a similarity between points that corresponds to an inner product in an implicit feature space. This idea enables algorithms that depend only on inner products to model complex relationships without explicitly constructing high dimensional features. Examples include the support vector machine and kernel ridge regression.

Artificial neural networks, especially deep and transformer-based models, are the most widely used approach for state-of-the-art results in vision, natural language processing (NLP), and speech, and they underpin current generative-AI systems deployed across industry[52]. A neural network consists of layers of units connected by weights. Each unit aggregates inputs, multiplies them by learnable weights, adds a bias, and applies a nonlinear activation function. Training proceeds by computing a loss on a sample or a batch of samples, then updating the weights using gradient descent with backpropagation[53]. Many architectures exist for specific data types, including convolutional neural networks for images[54] and graph neural networks for relational data[55]. Neural networks in general often excel with large datasets[56], though they can also be effective with small datasets when carefully designed and regularized[57–59] .

The final topic in this section is active learning, which is particularly useful for laboratory workflows. Active learning uses model predictions and the estimated uncertainty to select new experiments that are expected to be informative. An acquisition function balances exploration of uncertain regions and exploitation of promising candidates. This strategy can accelerate tasks such as optimizing nanoparticle uptake[60] , guiding molecular design for material[61] discovery or optimizing chemical synthesis reactions[62].

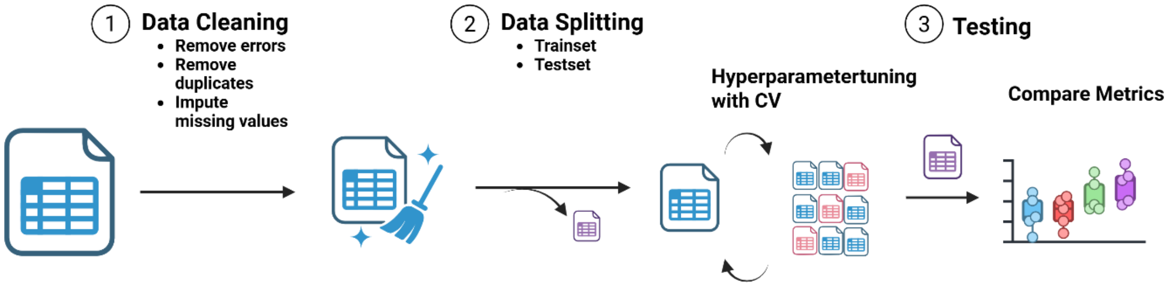


Figure I.3: Simplified data workflow. After assembly, records are cleaned by removing errors and duplicates and imputing missing values. The curated dataset is split into training and test sets. The training set is used for feature engineering, model selection, and hyperparameter tuning via cross-validation, then the final model is fit on the full training data. Performance is evaluated once on the hold-out test set and summarized for comparison.

5 Machine Learning in Molecular Sciences

Machine learning can in principle be applied to almost any task if the data are reliable. Working with molecules is more demanding because the model needs an input representation that exposes chemically meaningful patterns. The act of turning a molecule into a machine readable vector is called featurization. A common strategy is to encode molecules as binary vectors known as fingerprints. Molecular ACCess System (MACCS) keys[63] use a fixed dictionary of structural motifs and set a bit to one when the motif is present and to zero when it is absent. Pharmacophore fingerprints[64] emphasize features that drive receptor interactions such as hydrogen bond donors and acceptors, aromatic systems, positive or negative centers, and their pairwise distances on the molecular graph. Morgan fingerprints[65] capture local neighborhoods by enumerating circular subfragments around each atom up to a chosen radius and mapping them to a bit vector through a deterministic hash. This approach is efficient and expressive, with collisions as the main limitation. Many other fingerprints exist, including ones that incorporate three dimensional information or encode protein ligand interactions, and the examples here are only illustrative. Molecular descriptors provide another route. Instead of presence or absence of patterns they summarize properties as numbers. Simple descriptors include molecular weight or formal charge. Intermediate ones rely on estimated surfaces and volumes, for example topological polar surface area[66]. More complex families arise from matrices built on the molecular graph or on three dimensional coordinates. Examples include descriptors derived from the adjacency matrix, BCUT eigenvalue descriptors from Burden matrices[67], or WHIM descriptors[68] that summarize the covariance of atom coordinates possibly weighted by charges or masses.

The rise of neural networks popularized graph based encodings. A molecule can be viewed as a graph with atoms as nodes and bonds as edges. Graph neural networks learn atom level and bond level representations through message passing[69] or attention[70] and train end to end so that the learned encoding directly supports the prediction task through backpropagation. Modern variants can learn from unlabeled data[71], incorporate three dimensional information in an equivariant way[72], or model higher order interactions[73].

Featurizing whole formulations is even more challenging because one usually works with mixtures at specific ratios. The representation must capture both the identity of each component and its proportion. One can append mixture ratios to the molecular vectors, embed the ratio into each component feature before aggregation, or let the model learn how to combine components through a permutation invariant set encoder with ratio based weights[74].

In drug delivery the manufacturing process often matters as well. Process variables such as flow rate ratio and total flow rate influence particle size, polydispersity, and surface potential, so a practical feature set needs to include both chemistry and process. Designing such joint representations remains a central challenge for machine learning in formulation science.

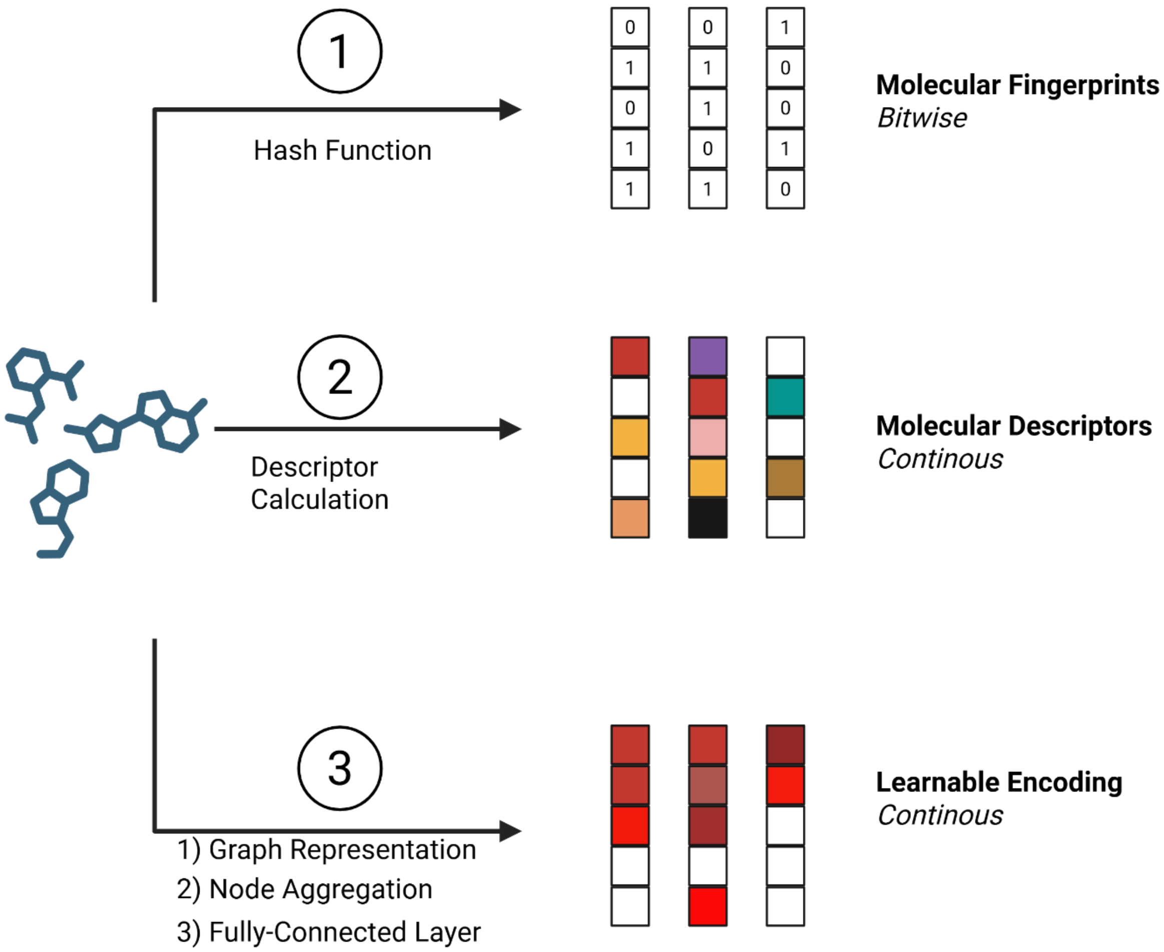


Figure I.4: Molecular featurization. Illustrative encodings of molecules: (1) hashed binary fingerprints capturing substructures and topology, (2) physicochemical and structural descriptors computed from 2D/3D properties, and (3) learned representations from graph neural networks that operate directly on the molecular graph.

6 Aim of the Thesis

The aim of this thesis is to explore how machine-learning (ML) workflows can be systematically integrated into the formulation and discovery of RNA nanocarriers. While ML is highly efficient, it is also data-dependent and requires substantial effort in experimental design, data preparation, and iterative refinement. By investigating and critically assessing different data-driven approaches across multiple case studies, this work aims to identify where and how ML can meaningfully accelerate nanocarrier development.

Chapter II highlights the potential of a classical data-driven approach, Design of Experiments (DoE), to control and explain the synthesis and behaviour of PBAEs as polymeric carriers for siRNA delivery. In addition, a custom data-driven method is established to estimate blend characteristics in step-growth polymerisation.

Chapter III demonstrates the benefits of applying ML pipelines to the same PBAE dataset, leveraging prior data when new labels become available or when the data no longer fit the original DoE. ML-based optimisation of synthesis parameters is investigated for its potential to improve key in vitro and in vivo readouts.

Chapter IV describes the integration of historical data into carrier discovery workflows, followed by the synthesis and in vitro/in vivo validation of the prioritised candidates, illustrating how legacy datasets can guide new formulation efforts.

Chapter V presents the development of a novel software framework that optimises PBAEs in silico by combining delivery-specific molecular dynamics (MD) challenges with experimentally calibrated ML optimisation and validates the resulting predictions experimentally.

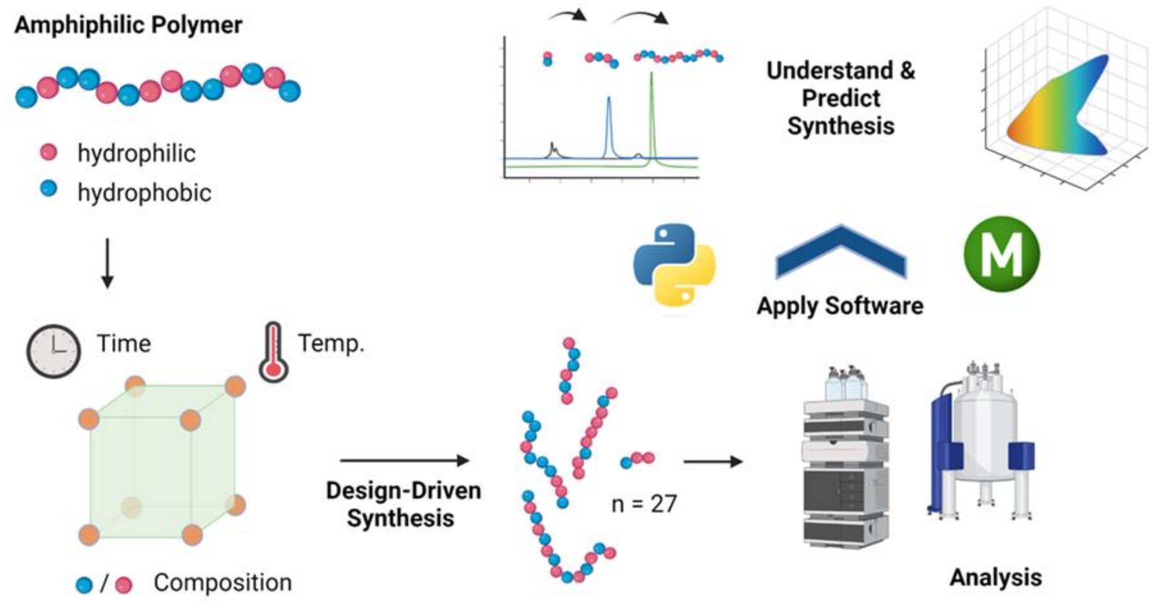
Chapter VI further explores MD/ML integration by introducing 4D QSTR (quantitative structure–transfection relationship), an approach that aggregates dynamic molecular information across MD frames and allows the identification of significant events by comparing different MD challenges, time windows, and data-splitting strategies.

Chapter VII investigates meta-learning as a potential solution to batch effects when merging heterogeneous, noisy datasets and evaluates its use in active-learning workflows for formulation discovery. Furthermore, novel lipids are synthesised and tested to demonstrate the practical relevance of these methods in very low-data regimes.

Chapter VIII summarizes the findings and provides additional remarks, conclusions and a brief outlook on potential future directions.

Chapter II - Design of Experiments Grants Mechanistic Insights into the Synthesis of Spermine-Containing PBAE Copolymers

1 Graphical Abstract



2 Abstract

Successful therapeutic delivery of siRNA with polymeric nanoparticles seems a promising but not vastly understood and complicated goal to achieve. Despite years of research, no polymer-based delivery system has been approved for clinical use. Polymers, as a delivery system, exhibit considerable complexity and variability, making their consistent production a challenging endeavor. However, a better understanding of the polymerization process of polymer excipients may improve reproducibility and material quality for more efficient use in drug products. Here, we present a combination of Design of Experiment and Python-scripted data science to establish a prediction model, from which important parameters can be extracted that influence the synthesis results of poly-beta-amino esters (PBAEs), a common type of polymers used preclinically for nucleic acid delivery. We synthesized a library of 27 polymers, each one at different temperatures, with different reaction times and educt ratios using an orthogonal central composite (CCO-) design. This design allowed a detailed characterization of factor importances and interactions using a very limited amount of experiments. We characterized the polymers by analyzing the resulting composition by 1H-NMR and the size distribution by GPC measurements. To further understand the complex mechanism of block polymerization in a one-pot synthesis, we developed a python script that helps to understand possible step-growth steps. We successfully developed and validated a predictive response surface and gathered a deeper understanding of the synthesis of polyspermine-based amphiphilic PBAEs.

Keywords: DOE, Python, polymer synthesis, polyplexes, siRNA, drug delivery

3 Introduction

Since the SARS-CoV-19 pandemic, the delivery of ribonucleic acid (RNA) by nanoparticles has become an ever more rapidly developing field of research. Up to now, the clinically approved drug delivery systems for RNA drugs are all based on Lipid Nanoparticles (LNP) technology[75,76]. However, LNPs face problems with regard to storage and stability[77] and encapsulate only a very low drug load of approximately 4% w/w[78]. Polymeric delivery systems, such as poly(beta)aminoesters (PBAEs), that were initially designed by the group of Robert Langer in 2000[26] represent a reasonable and well-studied alternative. In general, this type of polymer is easy to synthesize and in the past, end-capped homopolymers[79] and co-polymers[80] showed promising transfection on DNA[81], mRNA[82] and siRNA[83] in *in vitro* and *in vivo* models[80]. However, synthesis of polymers, especially copolymers is hard to control [84] and often leads to a mixture of different molecular weight and composition species[85]. This is undesirable, since these factors decrease reproducibility on the one hand but govern the ability to deliver the cargo to target cells[86] and the level of toxicity[87,88] on the other hand. Furthermore, they complicate a clean correlation between species and activity. Therefore, a strategy is needed that helps control and reveal the underlying mechanisms of step-growth polymerization and help understand the process. To do so, often dozens of experiments are needed to interpret and predict all the possible influencing factors.

For many years the help of Design of Experiment (DoE)[89] has been used to decrease the number of necessary experiments to address a problem and to help analyze important factors as well as define predictive models that can design an accurate response surface that is used to make assumptions about future experiments and helps therefore to reduce the waste of resources and to improve sustainability of chemical synthesis.

In recent years, the combination of data science and high throughput synthesis allowed for a significant knowledge gain in the field of nanomedicine[90–92]. This approach can be extremely useful since it allows for optimized decision in situations, where it is rather complicated to understand the mechanistic insights of how nanocarrier design influences the delivery of cargo[93]. DoE can also be applied here to guide scientists in designing the experiments to achieve optimization and valuable insights into complex processes[94,95]. In our work, we aim to use these tools to face difficult tasks in polymeric delivery such as controlling and understanding the synthesis of amphiphilic co-polymers[96] and their molecular weight distribution[85].

To demonstrate how data science can be used to understand and facilitate complicated scientific questions such as the controlled synthesis of block co-polymers for the encapsulation of RNA, we synthesized spermine- and oleylamine-modified PBAE-based co-polymers using DoE to iterate over a variable space with reasonable ranges for synthesis parameters including temperature, reaction time and the ratio of monomers, that influence the characteristics of the synthesized materials[97] [98]. Spermine was chosen as a body-own polycation to enhance RNA encapsulation efficiency and oleylamine to introduce hydrophobicity into the resulting polyplexes to facilitate the endosomal escape, demonstrated by previous work from our group[99]. As readout, we selected the final composition of blocks in the resulting polymer and different results from the size measurements of the polymer. For analysis we used multiple linear regression to generate a Response Surface Model and made use of different estimators that allow insights into the variables, which were most important for the prediction. To gather more information about possible structures, we designed a Python script that proposes possible polymeric compositions for Gel-Permeation-Chromatography (GPC) peak sequences. This approach was chosen to help interpret the often quite hard to analyze GPC chromatograms of co-polymers. Finally, we developed an assay that is able to mimic intracellular unpackaging of siRNA from polyplexes. This work presents a method to handle limited data effectively by using DoE and open source python libraries to facilitate the understanding and the analysis of complex synthesis mechanisms.

4 Methods and Materials

4.1 Materials

Di-tert-butyl decarbonate, oleylamine, spermine, dimethylformamide (99,5% pure) and SYBR Gold Nucleic Acid Gel Stain were purchased from Fischer Scientific (Hampton, NH, USA). Ethyl trifluoroacetate, sodium chloride, heparin sodium salt 180 USP units/mg and Triton-X 100% solution were bought from Sigma Aldrich (Taufkirchen, Germany) and 1,4-butanendiol diacrylate was obtained from TCI Chemical Industry Co., LTD (Tokio, Japan). Triflouroacetic acid (99,9%, extra pure) was purchased from Acros Organics (Geel, Belgium). Methanol-d6 was obtained from Deutero (Kastellaun, Germany). Dichlormethane, methanol, ammonia, potassium permanganate, magnesium sulfate, acetone, pentane and formic acid (>99% pure) were purchased from VWR Chemicals (Ismaning, Germany).

4.2 Triboc-spermine synthesis

Tris(tert-butoxycarbonyl)spermine, abbreviated as tri-Boc-spermine (TBS) was synthesized as described elsewhere[100]. In brief, spermine (1 eq) was dissolved in methanol and stirred at -78 °C, ethyl trifluoroacetate (1 eq) was added dropwise subsequently and stirred at - 78 °C for 1 h, then 0 °C for 1 h. Without isolation, di-tert-butyl dicarbonate (4 eq) was added dropwise to the solution and stirred at room temperature for 2 days. Finally, the solution was adjusted to a pH above 11 by 25% ammonia and stirred overnight to cleave the trifluoroacetamide protecting group. The mixture was then evaporated under vacuum and the residue was diluted with dichloromethane (DCM) and washed with distilled water and saturated sodium chloride aqueous solution. The DCM phase was finally dried by magnesia sulfate and concentrated to give the crude product. The crude product was purified by column chromatography (CH2Cl2\MeOH\NH3, aq. 7:1:0.1, SiO2, KMnO4; Rf = 0.413). TBS was isolated and characterized by 1H nuclear magnetic resonance spectroscopy (^{1}H-NMR).

4.3 Polymer synthesis and characterization

Poly-spermine-*co*-oleylamine beta-aminoesters (P(SpOABAE)) were synthesized based on a previously described approach[101]. Briefly, TBS as hydrophilic monomer, oleylamine (OA) as hydrophobic monomer and 1,4-butanendiol diacrylate (DA) were mixed in different molar ratios in dimethylformamide (DMF) resulting in total concentrations of 300 mg/mL. Polymers were stirred at different temperatures and for different durations (Compare Table II.1). After the respective reaction time, mixtures were transferred to petri dishes to evaporate the solvent. The subsequent deprotection of the polymer was carried out in a mixture of 20 ml dichloromethane (DCM) and 1 ml trifluoroacetic acid (TFA) for 100 mg polymer, followed by stirring for 2 hours at room temperature. In the following, DCM/TFA was evaporated and the dry deprotected product was precipitated 3 times in pentane using acetone to dissolve the precipitate (Figure II.1a). Supernatants were discarded and the final precipitate was dried for 2 days under vacuum (room temperature, 20 mbar). Final polymers were characterized by ^{1}H-NMR (Figure II.S1) and GPC. Measurements were performed with an Agilent aqueous GPC using a PSS Novema max Lux 100A followed by two PSS Novema max Lux 3000A columns. The chromatographic system and calibration standards were set up according to pre-analysis from Agilent Technologies on P(SpOABAE) polymers. Measurements were performed at 40°C in 0.1 M sodium chloride solution supplemented with 0.3% formic acid. Samples were prepared at 4 g/L and measured at a flow rate of 1 mL/min. Molar mass distributions were obtained through the Agilent WinGPC software against pullulan calibration standards in the range of 180 Da to 1450 kDa. A daisy-chain detector setup of an Agilent 1260 VWD was used followed by an Agilent 1260 GPC/SEC MDS and ending with an Agilent 1260 RID.

4.4 Design of Experiment

A Response Surface Method (RSM)[102] was applied using the MODDE® Pro 13.0.2 (Sartorius Data Analytics, Göttingen, Germany) software. Briefly, four critical process parameters (CPP) at three levels were chosen based on their theoretical impact on the critical quality attributes (CQA) of molecular weight and final subunit ratio. The four CPPs were i) reaction temperature (set to 80°, 100° or 120° Celsius), ii) reaction time (set to 24h, 48h or 72h), iii) initial molar OA ratio, defined as the molar ratio of primary amines from OA to the overall number of primary amines (set to 0.30; 0.55 or 0.80), and iv) the ratio between the diacrylate (DA) and the total theoretical number of primary amines (0.80; 1.00 or 1.20). A Central Composite Design for maximized Orthogonality (CCO) was chosen using a starpoint distance of 1.55[103]. Three center points were added to evaluate the process stability (Figure II.1b+c). Statistical significance was determined by ANOVA and defined by p-values below 0.05. Predictions with 95% confidence intervals were generated based on fitted, significant RSM model terms.

4.5 PeakFinder software

To gather more insights into the polymerization process, a program was written using Python3 programming language (version 3.11.5). Pandas (version 2.0.3) was used for data handling. The molecular weights of the monomer units are used as input data in the code together with information about the single peak maxima (Mp), the associated component ratio (obtained from NMR spectra), an error range, a maximal iteration parameter and a boolean expression parameter if endcapping with diacrylate is possible or not. Based on this information, possible polymer structures are calculated for each peak and the program outputs the sequence of monomer combinations that fits the data best.

4.6 Species isolation via spin columns

To isolate a single polymer species represented by a GPC peak, polymers were dissolved at 4 mg/mL in the mobile phase. 1 mL of solution was transferred to 30 kDa cutoff Vivaspin 6 centrifugal concentrator columns from Sartorius (Göttingen, Germany). Samples were concentrated at 8000 g for 15 min. The concentrated samples were diluted to 1 mL with fresh mobile phase. This procedure was repeated three times. Final samples were measured using the before mentioned GPC method.

4.7 Particle formation with siRNA

Polymers were dissolved in cell culture grade DMSO at a concentration of 25 mg/mL. Nanoparticles were prepared at a ratio of protonated amines in the polymer to negatively charged phosphates in the siRNA backbone (N/P Ratio) of 10. Polymer stocks and siRNA (IDT, Leuven, Belgium) were diluted in 10 mM Hepes Buffer pH 5.4 to equal volumes before mixing. Mixing was done using an Integra Voyager 125 µL pipette (Integra Biosciences, Zizers, Switzerland), resulting in final concentrations of 500 nM siRNA. After mixing, particles were incubated for 90 minutes at room temperature to allow proper particle formation. The hydrodynamic diameter (DH) and polydispersity index (PDI) of the obtained nanoparticles were determined by dynamic light scattering. Therefore, a Zetasizer Ultra series (Malvern Instruments, U.K.) was used running 3 measurements per sample at a backscatter angle of 173°.

4.8 Stability

The stability of the resulting nanoparticles was evaluated by a modified polyanion competition assay[104]. Briefly, differently concentrated mixtures of Triton-X and heparin were applied to release the siRNA from the nanoparticles. In a black 384-well plate, 10 µL nanoparticle suspension was mixed with 20 µL of stress solution with the respective concentration level. Seven different concentrations plus a blank were used per nanoparticle suspension. After adding the stress solutions, plates were sealed to avoid evaporation and incubated at 37°C at 150 rpm for 1h. Afterwards 5 µL of a 4x SYBR Gold dye was added to the mixture and incubated for 5 minutes in the dark. Finally, the fluorescence was measured using a TECAN Spark plate reader (TECAN, Männedorf, Switzerland) plate reader at 492 nm excitation and 537 nm emission wavelength. Using the GraphPad Prism5 2007 Software, a nonlinear fit was performed to calculate the EC50 values of each polymer relative to the maximum released siRNA in each sample.

Table II.1: Experimental setup of the CCO-design (left) with reaction time in hours, temperature in °C, initial molar OA ratio, defined as the molar ratio of primary amines from OA to the overall number of primary amines, and the ratio between the diacrylate and the total theoretical number of primary amines. Results of the CCO-design (right) with Final OA ratio in percent, Mn and Mw in Da, PDI without a unit and >33 kDa and < 2kDa in percent.

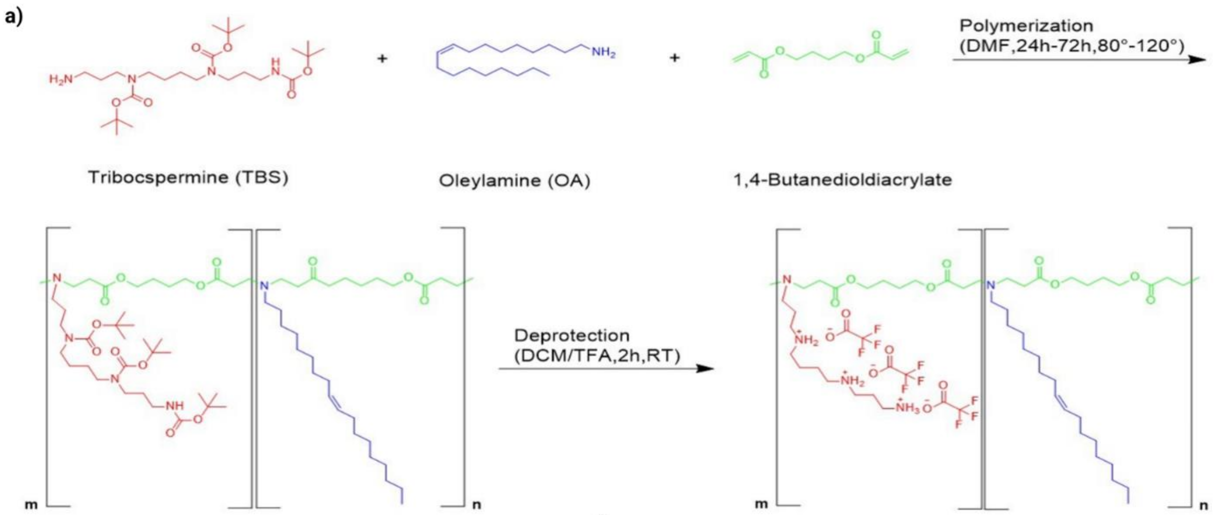
Exp No	Time	*Tmp*	*OA Initial*	*DA*	*Final OA*	*Mw*	*Mn*	*PDI*	*> 33kDa*	*< 2kDa*
1	24	80	30	0.8	0.412591	26878	15106	1.7792	31.3	0
2	72	80	30	0.8	0.438533	30610	16007	1.9123	38.53	0
3	24	120	30	0.8	0.391529	20022	10174	1.9681	22.5	0.5
4	72	120	30	0.8	0.486896	34815	14699	2.3686	50.08	0
5	24	80	80	0.8	0.827116	49961	25683	1.9453	79.39	0
6	72	80	80	0.8	0.78296	46782	22113	2.1156	75.4	0
7	24	120	80	0.8	0.787716	50345	21866	2.3025	80.24	0.26
8	72	120	80	0.8	0.761166	46650	15674	2.9763	72.58	1.35
9	24	80	30	1.2	0.417113	30380	15028	2.0215	42.15	0

10	72	80	30	1.2	0.360606	30299	14755	2.0535	40.72	0
11	24	120	30	1.2	0.323504	24760	11493	2.1544	31.97	0.59
12	72	120	30	1.2	0.274581	23140	10066	2.2987	30.68	1.36
13	24	80	80	1.2	0.703628	72203	38998	1.8515	91.11	0
14	72	80	80	1.2	0.728641	94201	45229	2.0828	92.62	0
15	24	120	80	1.2	0.73199	69166	35215	1.9641	90.32	0.3
16	72	120	80	1.2	0.691716	61153	26180	2.3359	85.15	0.6
17	10.88	100	55	1	0.608254	52849	29004	1.8221	81.8	0
18	85.12	100	55	1	0.631649	48145	22212	2.1675	76.11	0
19	48	69.07	55	1	0.62282	57537	34471	1.6691	86.7	0
20	48	130.9	55	1	0.576394	38442	13251	2.901	60.31	1.06
21	48	100	16.33	1	0.334521	22595	12648	1.7865	21.6	0
22	48	100	93.66	1	0.915431	171040	68271	2.5053	95.25	0.1
23	48	100	55	0.690	0.667118	30209	11942	2.5297	48.08	1.06
24	48	100	55	1.309	0.445172	43643	19900	2.1931	69.81	0
25	48	100	55	1	0.728655	51106	25095	2.0365	80.19	0
26	48	100	55	1	0.601359	50364	25522	1.9734	79.52	0
27	48	100	55	1	0.561223	50238	25043	2.0061	78.6	0

5 Results and Discussion

5.1 Controlling the synthesis via DoE

The two most important CQAs controlling the nucleic acid delivery performance of a polymer are the molecular weight distribution[87,88] and the composition of the polymer itself[105]. In case of amphiphilic spermine-modified PBAEs, previous studies showed that the ratio of hydrophobic side chains[99] plays a major role in the transfection efficiency of PBAE copolymers [101]. Additionally, it was shown for numerous PBAEs that the molecular weight plays vital functions in governing the performance as well as toxicity[106]. Therefore, the main goal of this study was to establish a synthesis route which would allow the precise prediction and control over the final constitution of the P(SpOABAE) polymers. By using the CCO, the design space, which was investigated, was maximized and by investigating 5 levels for each factor (Figure II.1b) the prediction strength was increased (Table II.1).



b)

Factors	Number of levels	Units	Factor levels studied				
Initial OA Ratio	5	%	16,3	30	55	80	93,7
Reaction time	5	h	10,9	24	48	72	85,1
Temperature	5	°C	69	80	100	120	131
Diacrylate ratio	5		0,69	0,8	1	1,2	1,31

c)

CQA	Sum of squares	Degrees of freedom	Mean square (Variance)	F	p-value
Final OA Ratio	0,757067	4	0,189267	160,615	$< 1 \times 10^{-3}$
Mw	0,942607	4	0,235652	31,7042	$< 1 \times 10^{-3}$
Mn	0,948579	4	0,237145	28,0392	$< 1 \times 10^{-3}$
PDI	0,0515111	4	0,0128778	6,32118	0,002
+33kDa	13108,3	4	13108,3	38,9963	$< 1 \times 10^{-3}$

Figure II.1: a) Overview of the applied synthesis for the used poly(beta aminoesters). Polymerization was carried out using different timepoints, temperatures and component ratios. b) Factors used for the CCO design. c) the CQAs selected as readout together with the data from ANOVA.

After performing the synthesis and analysis, the responses (Figure II.1c) were fitted using multiple linear regression. For the CQA final OA ratio, a strong regression of R_2 = 0.968 and a high validity of $Q_2 = 0.948$ were found indicating a strong model (Figures II.2a and II.S2). In the next step, the factors, which had been the most relevant for the model fit were investigated. By choosing a CCO, the factor strengths for linear as well as quadratic model terms, together with interactions between different CPPs was estimated. For the final OA ratio, only three model terms showed a p-value below 0.05 and were deemed significant (Figure II.S7). Unsurprisingly, the most relevant CPP was the initial OA ratio with a scaled and centered coefficient of 18.3%. Also, according to expectations, the temperature and reaction time did not impact the final OA ratio significantly. Surprisingly, the two other significant CPPs were the linear and quadratic diacrylate ratio with coefficients of -4.8% and -2.6% (Figures II.3a and II.S7). Although they were less relevant, it is still unexpected that this CPP can influence the final OA ratio. A potential reason for this observation might be the calculation approach chosen to determine the final OA ratio (eq.II.1). In this approach, the diacrylate backbone is taken into account in the formula and thereby naturally impacts the final results.

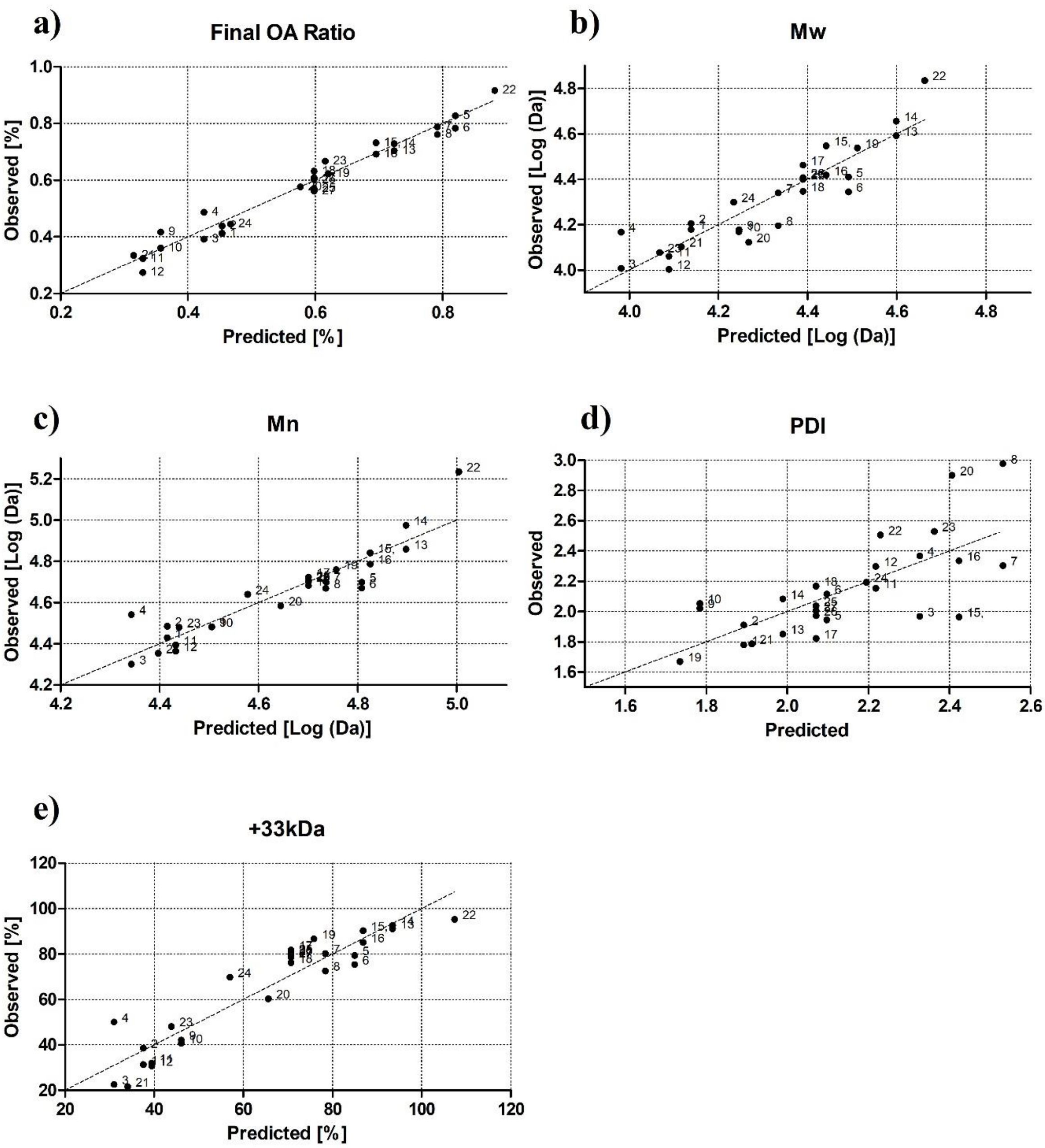


Figure II.2: Observed vs Predicted plot for a) final OA Ratio (R_2=0.97), b) Mw (R_2=0.85), c) Mn (R_2=0.84), d) PDI (R_2=0.53) and e) >33 kDa (R_2=0.88) for the CCO-design generated with 27 polymers.

$$OA\ Ratio = \frac{I_{(0.9ppm)}}{n_{H(terminal\ group,0.9\ ppm)}} x \frac{n_{H(backbone,4.2ppm)}}{I_{(4.2ppm)} - \left(n_{H(terminal\ group,0.9\ ppm)}\right) x\ \frac{I_{(0.9ppm)}}{n_{H(terminal\ group,0.9\ ppm)}}}$$

(eq.II.1)

In contrast to other polymerization mechanisms, the step-growth Michael-addition did not lead to a single polymer species but rather a mixture of several distinctive peaks. This finding will be further discussed below. To evaluate the presence of unreacted monomers the, numerical percentage of species below 2,000 Da (<2 kDa) was determined (Table II.1). Since the DoE can only interpret discrete numerical values, a way to make our library “interpretable” for the DoE algorithms had to be found. Therefore, several specific CQAs rather than a single molar mass distribution were added. To start, the overall Mn, Mw, PDI of the polymer as well as the numerical percentage of the polymer species above 33,000 Da (>33 kDa) were analyzed and introduced. For each CQA except for the PDI, a model with a regression above R_2 = 0.84 and a cross-validation value above Q_2 = 0.75 were found (Figure II.2 b-e, II.3 b-e, II.S8-II.S11). This outcome confirmed that the model was able to understand the synthesis and which CPPs govern the polymerization mechanisms. Surprisingly, the main factor controlling the three responses of Mn, Mw and >33 kDa was the OA ratio. Since the PDI of polymers is calculated by dividing the Mw by the Mn, this CQA is susceptible to error propagation. This problem is reflected in higher scatters in the observed vs predicted plot (Figure II.2 d) and higher standard deviations in the coefficient plot (Figure II.3 d).

Reaction time was not significant for any of the responses and temperature only played a minor role on the Mn.

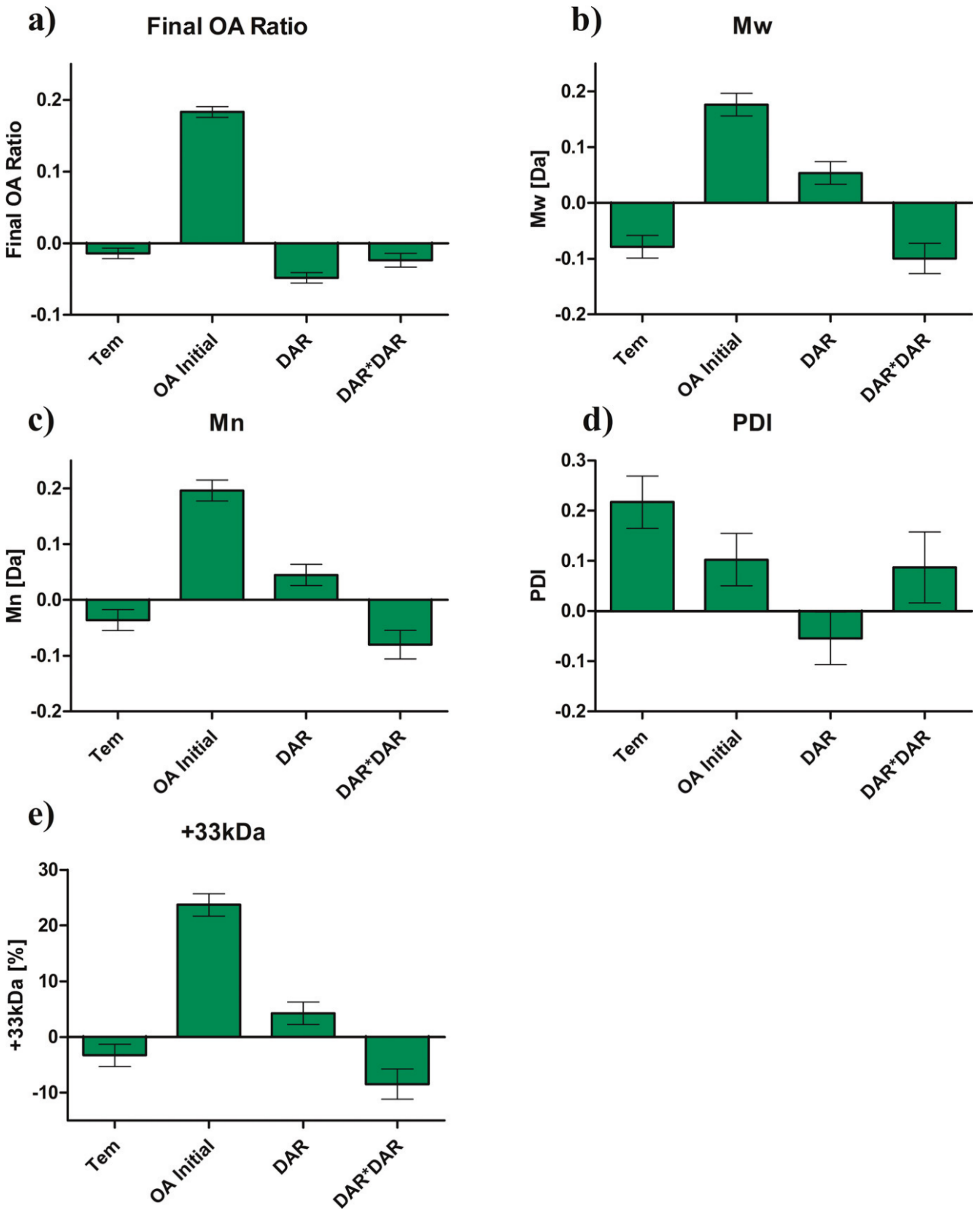
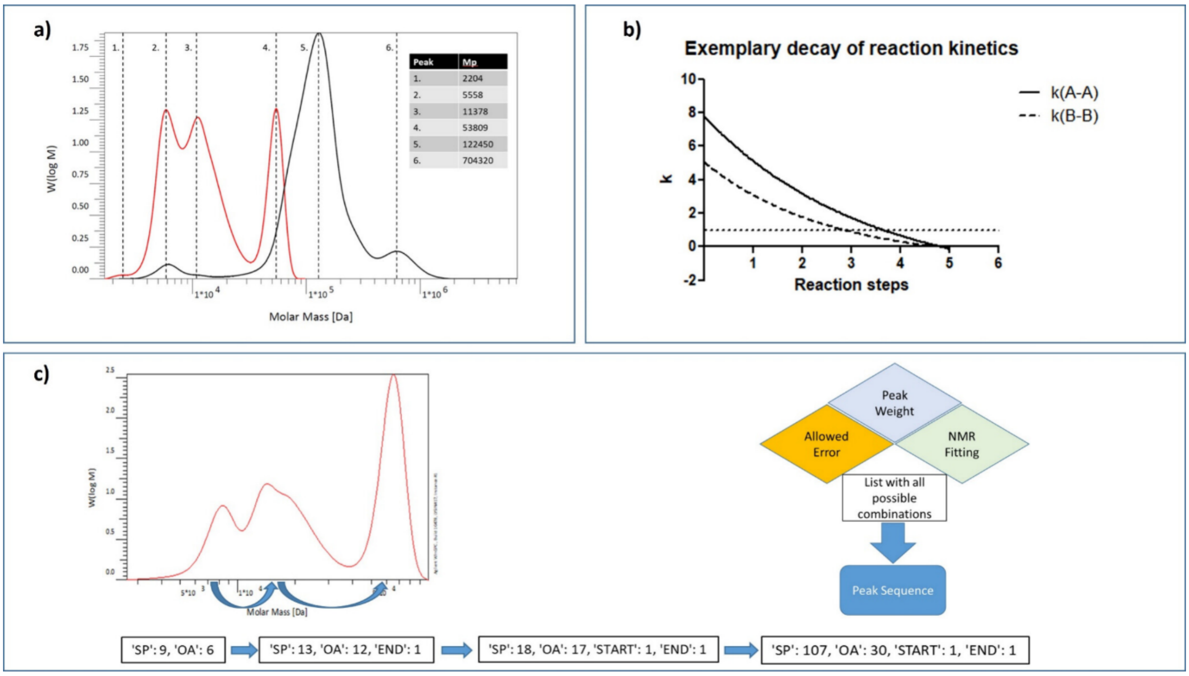


Figure II.3: Model coefficients for a) final OA Ratio (R_2=0.97; Q_2=0.95), b) Mw (R_2=0.85; Q_2=0.77), c) Mb (R_2=0.84; Q_2=0.75), d) PDI (R_2=0.53; Q_2=0.29) and e) >33 kDa (R_2=0.88; Q_2=0.81) for the CCO-design generated with 27 polymers.

5.2 Understanding key mechanisms

The initial hypothesis was that the molecular weight of the polymers would be mainly governed by the reaction time and temperature following common consensus[107]. However, the presented data suggest a more complex mechanism. Since the analyses showed that the main factor governing the large >33 kDa species was the OA ratio, it was concluded that the reaction kinetics of OA was faster than the kinetics of the TBS subunits. A faster reaction of hydrophobic subunits was already reported in literature[84]. However, it was observed that the maximum size of the >33 kDa species correlated with the OA ratio as well (Figure II.3e). This could not be explained with faster kinetics alone. Analyzing all GPC data more extensively showed that all polymers had a characteristic sequence in which the peaks occurred (Figure II.4a). This was explained by the mechanism of step-growth polymerization.



Peak	Mp
1.	2204
2.	5558
3.	11378
4.	53809
5.	122450
6.	704320

'SP': 9, 'OA': 6 → 'SP': 13, 'OA': 12, 'END': 1 → 'SP': 18, 'OA': 17, 'START': 1, 'END': 1 → 'SP': 107, 'OA': 30, 'START': 1, 'END': 1

Table II.4: a) Exemplary GPC peaks and Mp weights of polymer 3 (red) and 22 (black) in an overlay molar mass distribution. b) Exemplary decay of different reaction kinetics as a function of already occurred reaction steps. c) The PeakIdentifier tries to give the researcher an assumption, starting from the molecular weight distribution in GPC data, about peak sequences. On the right a schematic overview illustrates how the PeakIdentifier attempts to match individual peaks and the peak sequence using the available data. At the bottom, an example sequence proposed by the PeakIdentifier for the molecular weight distribution above is shown. The units and the corresponding numbers suggest the peak compositions that matches the data best.

In step-growth polymerization, monomers undergo simultaneous parallel reactions to form dimers, which subsequently engage in further parallel reactions to produce tetramers and subsequent oligomeric species[108]. Interestingly, in co-polymers the same mechanism applies with the difference that three kinetics are occurring in parallel. The kinetics of two building blocks of the same type reacting with each other (kA-A, kB-B) and the kinetics of two different building blocks reacting with each other (kA-B, kB-A). Additionally, each reaction slows down exponentially, with the number of reactions (r) that have already occurred. With this behavior, the following relation could be drawn:

kA-A (r=1) > kA-A (r=2) >…> kA-A (r=n)

(eq.II.2)

kB-B (r=1) > kB-B (r=2) >...> kB-B (r=n)

(eq.II.3)

kA-B (r=1) > kA-B (r=2) >...> kA-B (r=n)

(eq.II.4)

kB-A (r=1) > kB-A (r=2) >...> kB-A (r=n)

(eq.II.5)

Together with the finding that the OA homopolymerization kinetics are faster than TBS homopolymerization kinetics, a new hypothesis was established.

It was proposed that the reaction reaches its thermodynamic equilibrium after a certain amount of steps after which the reaction kinetics decrease to a level where statistically no more reactions occur, for example, where a certain threshold was reached. How many reactions it takes, for example, and how long the polymers become before the threshold is reached is hence governed by the initially faster kinetics (kA-A). In this case the kinetics and initial amount of OA (Figure II.4b).

Although the relationship between the >33 kDa species and the initial OA content may be explained by this hypothesis, one needs to take into account that in theory only one single species of varying size should have arisen from each synthesis. The fact that one can simultaneously observe all different stages of the step-growth polymerization underlined the reversibility of the Michael-addition (Figure II.4a)[109].

$$(A) + (B) \rightleftharpoons (AB) \rightleftharpoons (ABAB) \rightleftharpoons (ABABABAB)$$

(eq.II.6)

The reversibility indicated that all stages of the step-growth synthesis are in equilibrium with each other. The equilibrium that the reactions reaches (eq.II.6) is, according to these findings, governed by the ratio between faster reacting OA and slower reacting TBS (Figure II.4b).

A deeper investigation of the impact of the diacrylate (Figure II.3b+c and II.5b+c) showed that the Carother's equation[110] also held true for these polymers, showcasing that a diacrylate ratio of 1.0 leads to the largest polymers.

To incorporate the new hypothesis into the data set, an in-house software package was written.

The software aimed to mimic the block-copolymer step-growth reaction, which was expected in this system. Therefore, the absolute Mw of single building blocks was combined together with an error term, to allow variance. This step was repeated for every peak in the chromatogram, which led to a list of all possible peak sequences. Finally, peak sequences were matched with the corresponding peak-weight and the polymer block composition data obtained from NMR to match the most suitable peak sequences. The software then outputs the peak sequence with the best match. To increase the likelihood that the sequence matched the data, the program was constrained to select only sequences that assumed a growth in single building blocks. Additionally, end capping with diacrylate was only possible when there was an excess in the amount of diacrylate used for synthesis.

It was important to note that the function did not apply any further physicochemical steps to calculate a matching sequence and the results were calculated from the obtained data. Therefore, high data quality was a major assumption of the program.

Figure II.4c shows an example for the PeakIdentifier from sample number 10. The error range was set to 15 % to allow for the absolute combined monomers to vary with this value from the proposed combination, and the NMR ratio was set to 38.42 [%]. The PeakIdentifier suggested a scenario where Oleylamine (OA) and Triboc-spermine (TBS) react with equal probability. This assumption was based on the understanding that although OA reacts more quickly (due to faster kinetics), TBS is available in greater concentration within the reaction mixture, balancing the reaction likelihood between the two. The last peak observed might be the result of a subsequent synthesis reaction, where the higher concentration of TBS in the sample prompts the oligomers to undergo a reaction. What was shown clearly, is that the PeakIdentifier explained possible step-growth reactions in combination with different kinetics. It has to be mentioned that the PeakIdentifier provided a range of possibilities, but since the program worked with absolute data one had to make sure to precisely select a reasonable error range.

To validate the software (Figure II.S12), two single peak fractions were isolated using spin columns. To verify a successful isolation, GPC was measured again (Figure II.S13). The NMR results from the isolated fractions were compared to the PeakIdentifier results. From the NMR data for polymer 16, an 89.29% OA ratio was observed in the isolated peak at 67,750 Da and for polymer 17, 62.0% OA monomer was found in the isolated peak at 62,877 Da. The PeakIdentifier calculated 124 OA units to 9 Spermine units, which corresponds to a ratio of 93.2% for peak 16 and 75 OA units to 46 Spermine units, which is precisely 62.0% for peak 17. We consider a delta in the estimation and the real ratio of under 5% as successful, which was satisfied for both polymers tested (3.91% for 16 and 0 for 17). Based on this example it was shown that the PeakIdentifier allows for a quite precise estimation of possible polymer fractions within this synthesis.

Another observation that was made was the presence of a side product appearing around 8 ppm in the NMR (Figure II.S14). However, a correlation between the intensity of the NMR peaks of this impurity and the temperature could be shown. Furthermore did the DoE approach allow us to find the optimal setpoints to avoid the generation of these side products in the first place (Figure II.S15). This highlights how DoE did not only improve the understanding of the step-growth synthesis process but also how the most robust setpoints could be identified to achieve the best results.

Interestingly, within the selected range, reaction time did not show any influence on the readout parameters. This result could be caused by the fact that the equilibrium of the polymerization process was already in a stable state after a short period of time and was not further influenced by longer reactions. Despite the fact that high temperature led to the mentioned side products and a possible reversibility in Michael addition reaction, it did surprisingly not show any influence on the polymer size parameters.

5.3 Prediction

After the fitting of the model, a response surface for the entire design space was generated (Figure II.5a-e). To validate the model, three different polymers with varying final OA Ratios of 40%, 50% and 60% (Table II.S1) were predicted. The reasoning behind these setpoints was to spread through the design space as far as possible to validate a wide range. Additionally, the predictions for the molecular weights were validated with the same polymers. Having gained a deeper understanding of the complexity of our polymerization process, it was all the more surprising how well the model did not just fit the already generated data but also predicted the validation data (Figure II.6 and Table II.S1).

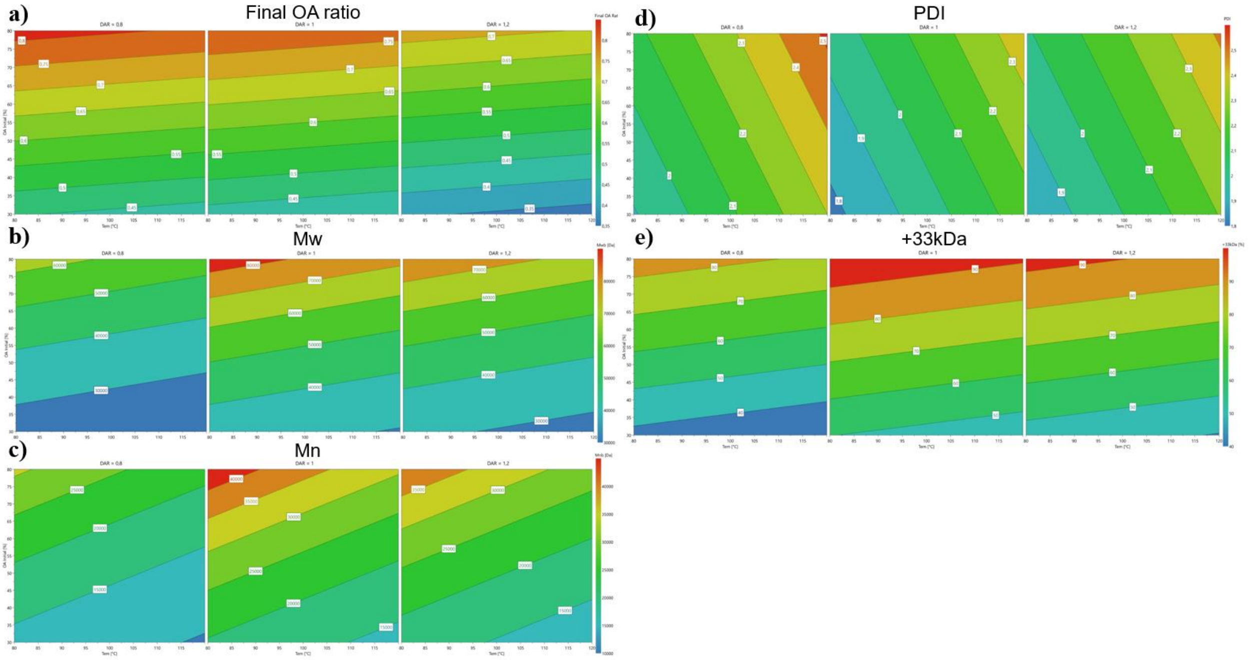


Figure II.5: 3 Dimension plot of the Response Surface of a) the final OA Ratio, b) the Mw, c) the Mn d) the PDI, and e) the >33kDa model fitted from the CCO-design of 27 polymers showing the impact of the diacrylate ratio (left 0.9, center 1.0 and right 1.2), initial molar OA ratio, and temperature.

The model was capable of accurately predicting the final OA ratio as well as the molecular weight of the respective polymers. This dataset confirmed that with DoE even highly complex mechanisms such as the showcased co-polymerization mechanism can be understood and controlled, allowing a precise manufacturing of new desired polymers. With this approach it is possible to synthesize any desired polymer in the design space without any further trial and error studies, as it is the common approach in polymer synthesis[111].

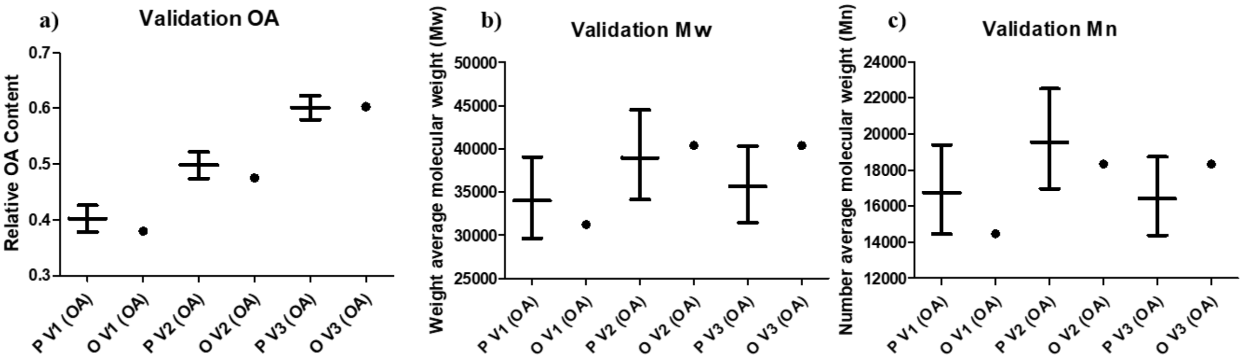


Figure II.6: Prediction (P, Error Bars) and observed values (dots) for the validation of a) the OA ratios, b) the Mw values and c) the Mn values of three validation polymers.

5.4 Stability

As previously shown[99], amphiphilic PBAE-based spermine copolymers can mediate highly effective gene silencing when they are used for siRNA formulation and delivery. To confirm that the entire design space has relevance to subsequent performance tests, it was investigated if all polymers formed nanoparticles, encapsulated and finally released siRNA. As shown in Figure II.S16 and II.S17, all polymers were able to form stable particles, which encapsulated the entire amount of the provided siRNA. Through the new stability assay, assumptions about the strength of the intra-particular forces stabilizing the particles were additionally made. This allowed the investigation of which polymers would form the most and least stable particles. Polymer 5 and 6 formed the most stable particles and polymer 16 formed the least stable particles. The strongest correlations for the stability of the particles were found for the synthesis temperature (Figure II.7b), DA ratio (Figure II.7d), and the PDI of the resulting nanoparticles (Figure II.7f). More precisely did a lower DA ratio and a lower temperature during the synthesis lead to more stable nanoparticles. For the synthesis time (Figure II.7a) and the initial OA ratio (Figure II.7c), no clear trends could be found. Similarly, the hydrodynamic diameter of the nanoparticles did not show a clear trend. Polymer 14 formed much larger particles than all other polymers but showed comparable stability (Figure II.S16+II.S17). Additionally, the difference in deviation of the EC_50 values showed a relation to the synthesis parameters (Figure II.7b+II.7d), indicating controllability by carefully choosing the proper settings. These parameters can become very important for subsequent *in vitro* and *in vivo* studies. Further analysis showed that the stability correlated with the PDI of the nanoparticles, indicating that less homogenous particles are harder to break up (Figure ff).

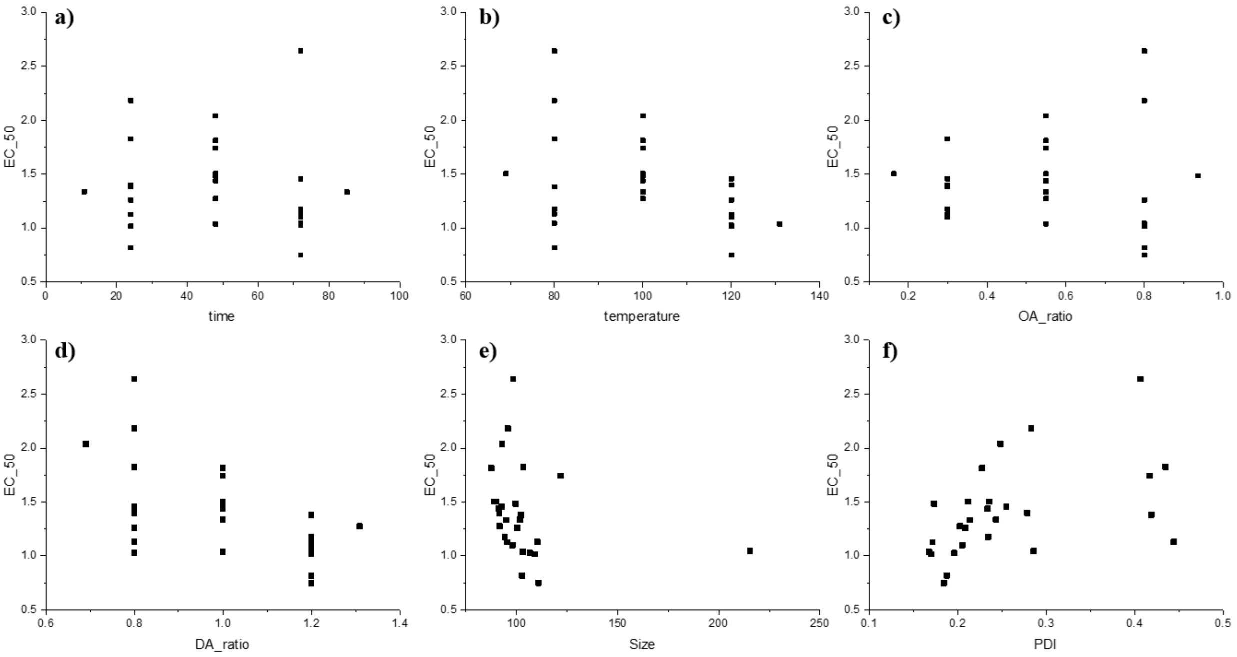


Figure II.7: Stability values (EC_50) derived from the stability assay plotted against the initial CPP from the CCO-design being a) the time of reaction, b) the temperature of the reaction, c) the initial OA ratio and d) the DA ratio as well as the DLS data with e) the hydrodynamic diameter of the tested particles and f) the PDI of the tested particles.

6 Conclusion

This study highlighted the value of DoE as a tool to gain deeper mechanistic understanding of PBAE-based copolymer synthesis. Besides the revelation of key parameters controlling the synthesis of P(SpOABAE), a model that accurately predicts the outcome of a synthesis approach was established. According to our knowledge, this is the first report of a model that is capable of predicting molecular weight as well as building block ratios of copolymers. In combination with the PeakIdentifier Software, a detailed picture of any synthesized copolymer can be generated. As a deep understanding of the used polymers is the first step for any scientific study, we are confident that these findings will prove valuable for other scientists in the search of a more controlled material generation.

7 Acknowlegements

This work was supported by ERC-2022-COG-101088587 to Olivia Merkel and by the Future Cluster CNAT-M, funded by the Federal Ministry of Education and Research. Graphics were partially generated using BioRender

8 Supplementary Information

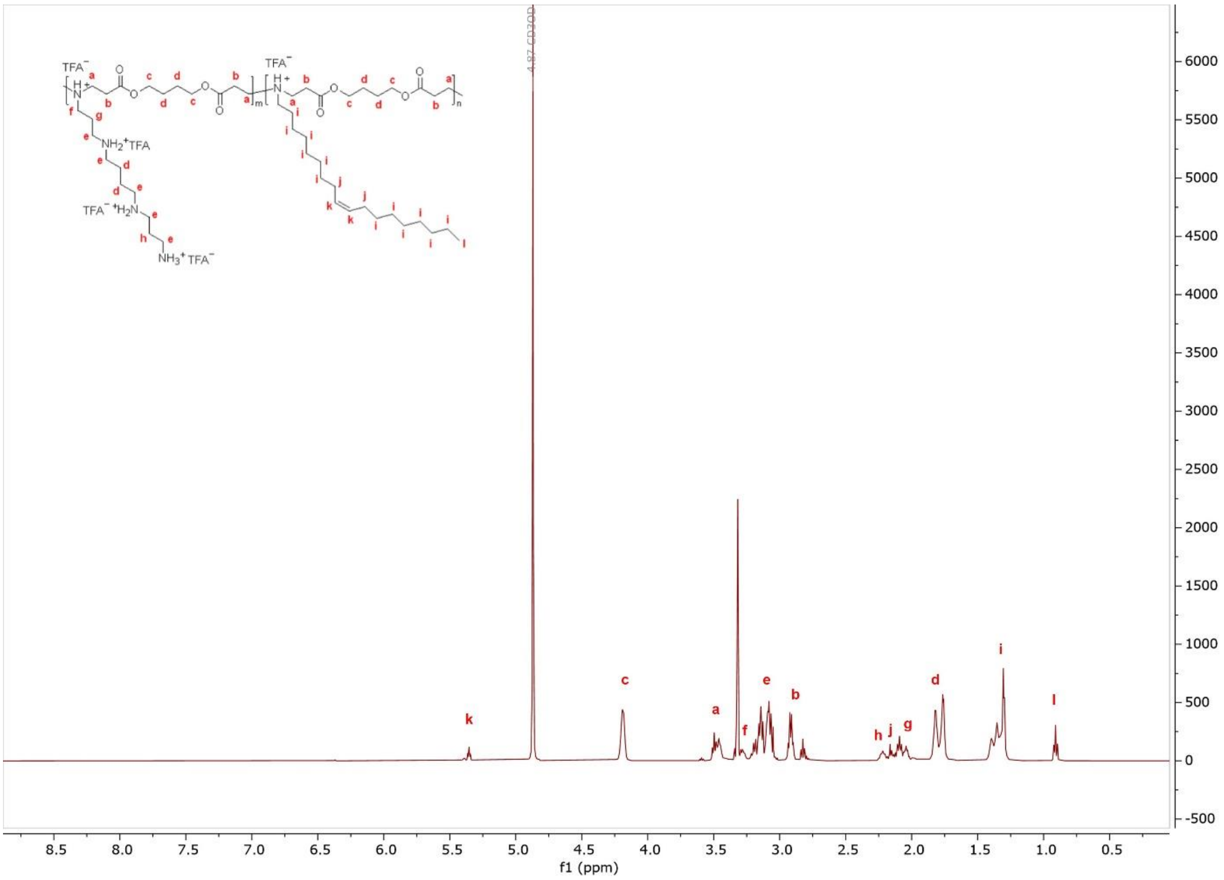


Figure II.S1: Exemplary ^{1}H-NMR of the resulting Poly-spermine-co-oleylamine beta-aminoesters after synthesis and purification

Final OA Ratio	*DF*	*SS*	*MS (variance)*	*F*	*p*	*SD*
Total	*26*	*9.53312*	*0.366658*			
Constant	*1*	*8.75131*	*8.75131*			
Total corrected	*25*	*0.781813*	*0.0312725*			*0.17684*
Regression	*4*	*0.757067*	*0.189267*	*160.615*	***0.000***	*0.435048*
Residual	*21*	*0.0247461*	*0.00117839*			*0.0343276*

Lack of Fit	*20*	*0.0239407*	*0.00119703*	*1.48616*	***0.578***	*0.0345982*
(Model error)						
Pure error	*1*	*0.000805451*	*0.000805451*			*0.0283805*
(Replicate error)						
	N = 26	*Q2 =*	*0.948*	*Cond. no. =*	*2.832*	
	DF = 21	*R2 =*	*0.968*	*RSD =*	*0.03433*	
		R2 adj. =	*0.962*			

Figure II.S2: ANOVA table of the final OA Ratio from the CCO-design of 27 polymers.

II.:

Mwb~	DF	SS	MS (variance)	F	p	SD
Total	27	582.11	21.5596			
Constant	1	581.004	581.004			
Total corrected	26	1.10613	0.0425434			0.206261
Regression	4	0.942607	0.235652	31.7042	**0.000**	0.48544
Residual	22	0.163522	0.00743283			0.0862139
Lack of Fit	20	0.16349	0.0081745	506.302	**0.002**	0.0904129

(Model error)						
Pure error	*2*	*3.2291e-05*	*1.61455e-05*			*0.00401815*
(Replicate error)						
	N = 27	*Q2 =*	*0.768*	*Cond. no. =*	*2.731*	
	DF = 22	*R2 =*	*0.852*	*RSD =*	*0.08621*	
		R2 adj. =	*0.825*			

Figure II.S3: ANOVA table of the Mw from the CCO-design of 27 polymers.

Mnb~	DF	SS	MS (variance)	F	p	SD
Total	27	503.413	18.6449			
Constant	1	502.278	502.278			
Total corrected	26	1.13465	0.0436402			0.208902
Regression	4	0.948579	0.237145	28.0392	**0.000**	0.486975
Residual	22	0.186067	0.0084576			0.0919652
Lack of Fit	20	0.186026	0.00930132	456.658	**0.002**	0.0964434
(Model error)						
Pure error	2	4.07365e-05	2.03682e-05			0.00451312
(Replicate error)						
	N = 27	Q2 =	0.747	Cond. no. =	2.731	
	DF = 22	R2 =	0.836	RSD =	0.09197	
		R2 adj. =	0.806			

Figure II.S4: ANOVA table of the Mn from the CCO-design of 27 polymers.

PDI~	DF	SS	MS (variance)	F	p	SD
Total	27	2.96082	0.10966			
Constant	1	2.86449	2.86449			
Total corrected	26	0.0963304	0.00370502			0.0608688
Regression	4	0.0515111	0.0128778	6.32118	**0.002**	0.11348
Residual	22	0.0448193	0.00203724			0.0451358
Lack of Fit	20	0.0447258	0.00223629	47.8424	**0.021**	0.0472894
(Model error)						
Pure error	2	9.34858e-05	4.67429e-05			0.00683688
(Replicate error)						
	N = 27		Q2 = 0.288	Cond. no. = 2.731		
	DF = 22		R2 = 0.535	RSD = 0.04514		
		R2 adj. = 0.450				

Figure II.S5: ANOVA table of the PDI from the CCO-design of 27 polymers.

>33 kDa	***DF***	***SS***	***MS (variance)***	***F***	***p***	***SD***
Total	*27*	*126153*	*4672.33*			
Constant	*1*	*111196*	*111196*			
Total corrected	*26*	*14957.1*	*575.273*			*23.9848*
Regression	*4*	*13108.3*	*3277.08*	*38.9963*	***0.000***	*57.2458*
Residual	*22*	*1848.78*	*84.0356*			*9.16709*
Lack of Fit	*20*	*1847.51*	*92.3754*	*144.963*	***0.007***	*9.61121*
(Model error)						
Pure error	*2*	*1.27447*	*0.637236*			*0.798271*
(Replicate error)						
	N = 27	*Q2 =*	*0.806*	*Cond. no. =*	*2.731*	
	DF = 22	*R2 =*	*0.876*	*RSD =*	*9.167*	
		R2 adj. =	*0.854*			

Figure II.S6: ANOVA table of the >33 kDa fraction from the CCO-design of 27 polymers.

Final OA Ratio	***Coeff. SC***	***Std. Err.***	***P***	***Conf. int(±)***
Constant	*0.600624*	*0.0107169*	*2.34777e-24*	*0.022287*
Tmp	*-0.0141401*	*0.00752962*	*0.074348*	*0.0156587*
OA Initial	*0.183216*	*0.00752961*	*7.19313e-17*	*0.0156587*
DAR	*-0.0481132*	*0.00752962*	*2.46714e-06*	*0.0156587*
DAR*DAR	*-0.0255963*	*0.0104307*	*0.0229448*	*0.0216919*

N = 26 *Q2 = 0.948* *Cond. no. = 2.832*

DF = 21 *R2 = 0.968* *RSD = 0.03433*

R2 adj. = 0.962

Confidence = 0.95

Figure S7. Coefficient table (Scaled and Centered) for final OA-Ratio model from the fitted CCO-design.

Mwb~	***Coeff. SC***	***Std. Err.***	***P***	***Conf. int(±)***
Constant	*4.70064*	*0.025692*	*1.65088e-36*	*0.0532832*

		5		
Tem	*-0.0362279*	*0.0189106*	*0.0684845*	*0.0392185*
OA Initial	*0.196177*	*0.0189106*	*6.16452e-10*	*0.0392185*
DAR	*0.0447507*	*0.0189106*	*0.0271844*	*0.0392185*
DAR*DAR	*-0.0802959*	*0.0254828*	*0.00463685*	*0.0528483*

N = 27 *Q2 = 0.768* *Cond. no. = 2.731*

DF = 22 *R2 = 0.852* *RSD = 0.08621*

R2 adj. = 0.825

Confidence = 0.95

Figure II.S8: Coefficient table (Scaled and Centered) for Mw model from the fitted CCO-design.

Mnb~	Coeff. SC	Std. Err.	P	Conf. int(±)
Constant	4.38982	0.0274064	3.07204e-35	0.0568377
Tem	-0.0786417	0.0201722	0.000772216	0.0418347
OA Initial	0.176359	0.0201722	1.31122e-08	0.0418347
DAR	0.0536849	0.0201722	0.0142611	0.0418347
DAR*DAR	-0.0996573	0.0271827	0.00135652	0.0563738

N = 27	Q2 = 0.747	Cond. no. = 2.731
DF = 22	R2 = 0.836	RSD = 0.09197
	R2 adj. = 0.806	
		Confidence = 0.95

Figure II.S9: Coefficient table (Scaled and Centered) for Mn model from the fitted CCO-design.

PDI~	Coeff. SC	Std. Err.	P	Conf. int(±)
Constant	0.310812	0.0134508	6.38448e-17	0.0278955

Tem	*0.0424165*	*0.00990035*	*0.000301253*	*0.0205322*
OA Initial	*0.0198192*	*0.00990035*	*0.0577805*	*0.0205322*
DAR	*-0.00893745*	*0.00990035*	*0.376442*	*0.0205322*
DAR*DAR	*0.0193637*	*0.0133411*	*0.160772*	*0.0276678*

N = 27	*Q2 = 0.288*	*Cond. no. = 2.731*
DF = 22	*R2 = 0.535*	*RSD = 0.04514*
	R2 adj. = 0.450	
		Confidence = 0.95

Figure II.S10: Coefficient table (Scaled and Centered) for PDI model from the fitted CCO-design.

>33 kDa	Coeff. SC	Std. Err.	P	Conf. int(±)
Constant	70.7029	2.73187	5.74094e-18	5.66558
Tem	-3.29656	2.01076	0.115342	4.17009
OA Initial	23.7096	2.01076	5.56369e-11	4.17009
DAR	4.24879	2.01076	0.0461756	4.17009
DAR*DAR	-8.48071	2.70957	0.00487188	5.61934

N = 27 Q2 = 0.806 Cond. no. = 2.731

DF = 22 R2 = 0.876 RSD = 9.167

R2 adj. = 0.854

Confidence 0.95 =

Figure II.S11: Coefficient table (Scaled and Centered) for >33 kDa model from the fitted CCO-design.

```
Pseudocode of the function:
Algorithm PeakIdentifier
Input: chromatogram_peaks, mw_building_blocks, error_term, peak_weights, NMR_data,end-cap bool
Output: best_matching_sequence

1. Initialize all_sequences as an empty list
2. For each peak in chromatogram_peaks do:
    2.1 Calculate adjusted_mw = mw_building_blocks + error_term + end-cap bool
    2.2 Generate all possible sequences for the peak using adjusted_mw
    2.3 Add generated sequences to all_sequences
3. Initialize best_match_score as negative infinity
4. Initialize best_matching_sequence as None
5. For each sequence in all_sequences do:
    5.1 Calculate match_score for sequence based on peak_weights and NMR_data
    5.2 If match_score > best_match_score then:
        5.2.1 Update best_match_score to match_score
        5.2.2 Update best_matching_sequence to sequence
6. Return best_matching_sequence
```

Figure II.S12: PeakIdentifier Pseudo code explaining the function of the PeakIdentifier. The code is used to match GPC and NMR data to the chromatogram and is expected to help identifying peaks and peak sequences of step-growth polymerization products.

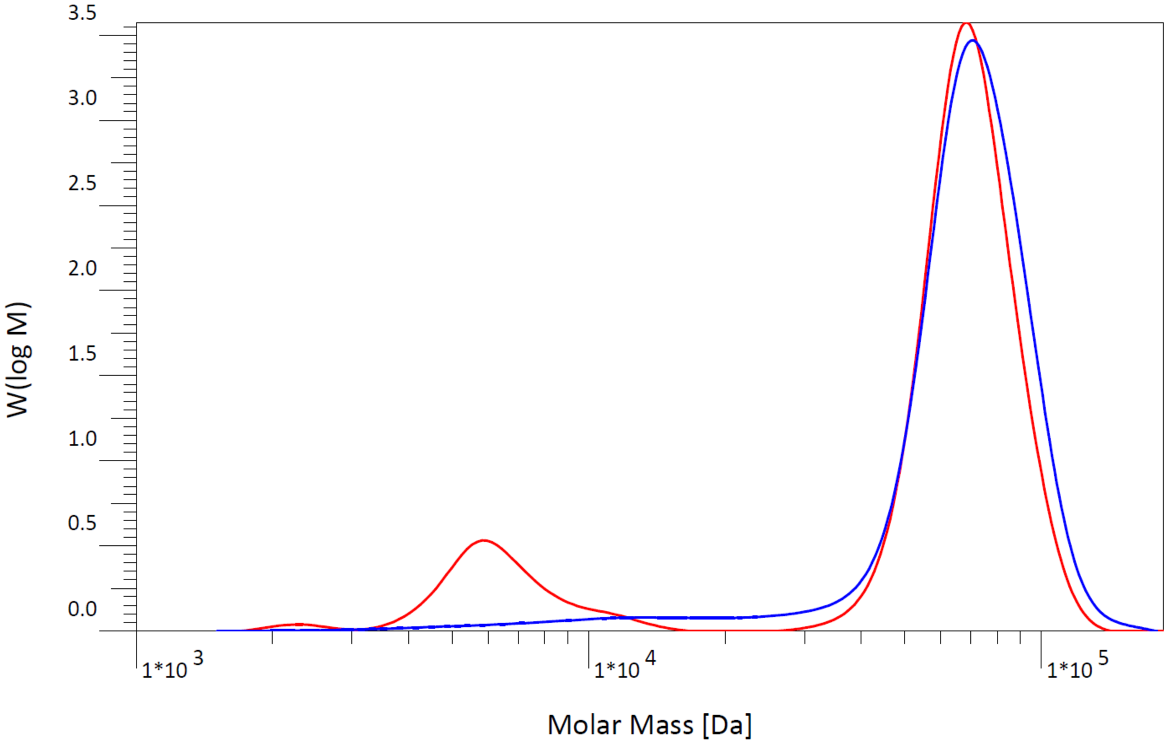


Figure II.S13: Molar mass distribution of Polymer 16 before (red) and after (blue) 3 purification steps in a 30.000 Da MWCO spin column.

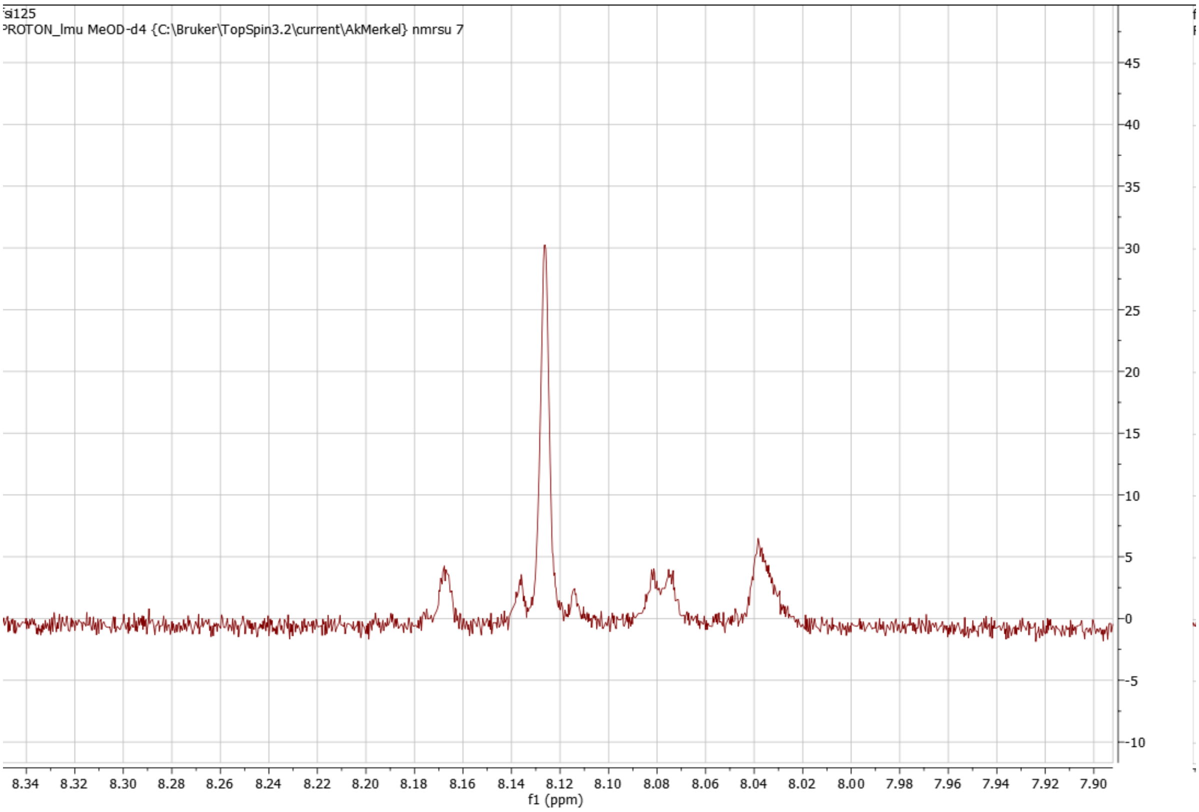


Figure II.S14: ^{1}H-NMR spectrum of temperature dependent side products after 8 ppm.

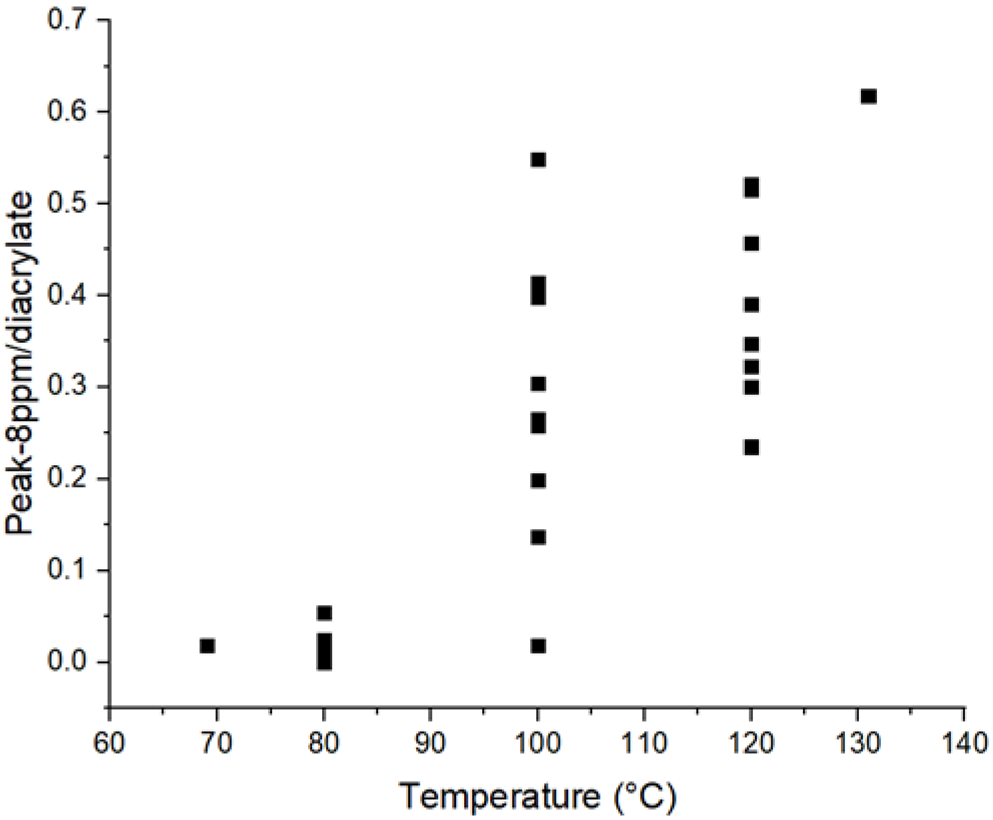


Figure II.S15: Correlation between side products (NMR species at 8 ppm) and reaction temperature.

Table S1: Validation settings and results for three validation polymers. CQA predictions are shown with 95% confidence intervals from lower (L) to upper (U) limit and results are shown in observed (O) columns.

Polymer	Time	Tem	OA	DAR	OA (L)	OA (U)	OA (O)	Mw (L)	Mw (U)	Mw (O)	Mn (L)	Mn (U)	Mn (O)	PDI (L)	PDI (U)	PDI (O)	+33k Da (L)	+33k Da (U)	+33k Da (O)
V1	48	100	38	1.2	0.378	0.426	**0.380**	29638	39042	**31235**	14457	19398	**14472**	1.89	2.18	**2.16**	43.99	56.71	**46.28**
V2	48	100	41	1	0.474	0.522	**0.475**	34131	44505	**40404**	16966	22519	**18342**	1.86	2.14	**2.20**	51.30	63.55	**61.80**
V3	48	100	52	0.8	0.580	0.623	**0.603**	31497	40349	**40295**	14386	18736	**18338**	2.04	2.32	**2.20**	49.41	60.85	**61.71**

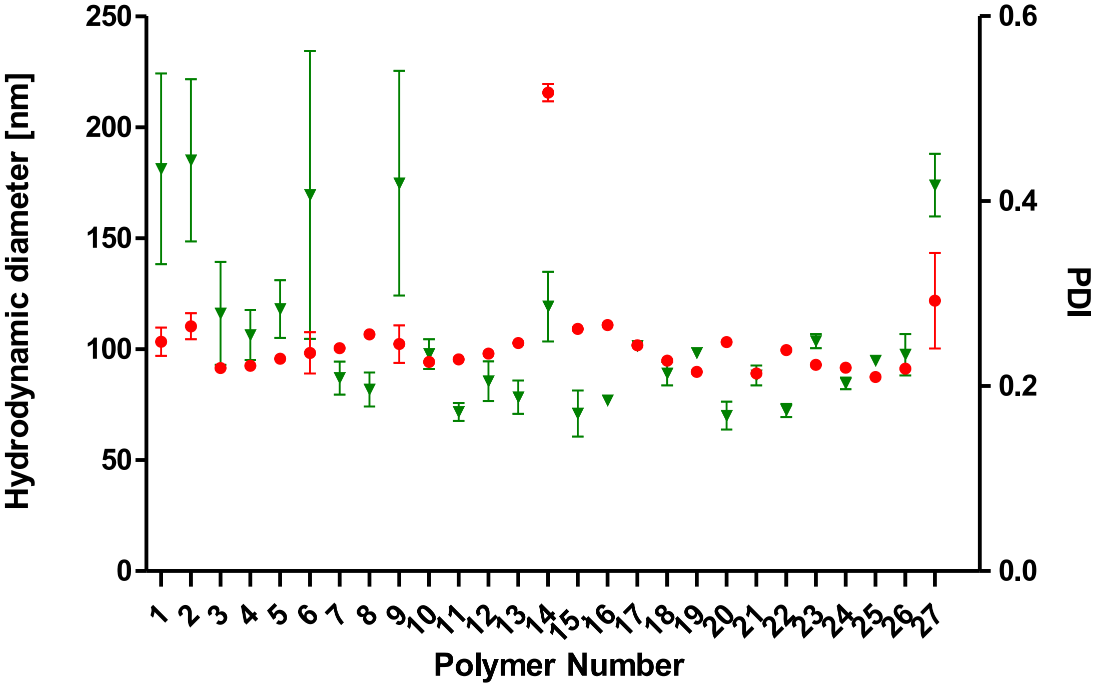


Figure II.S16: Dynamic light scattering data of hydrodynamic diameter (red circles) and polydispersity index (green triangle) of siRNA containing particles used for the stability assay.

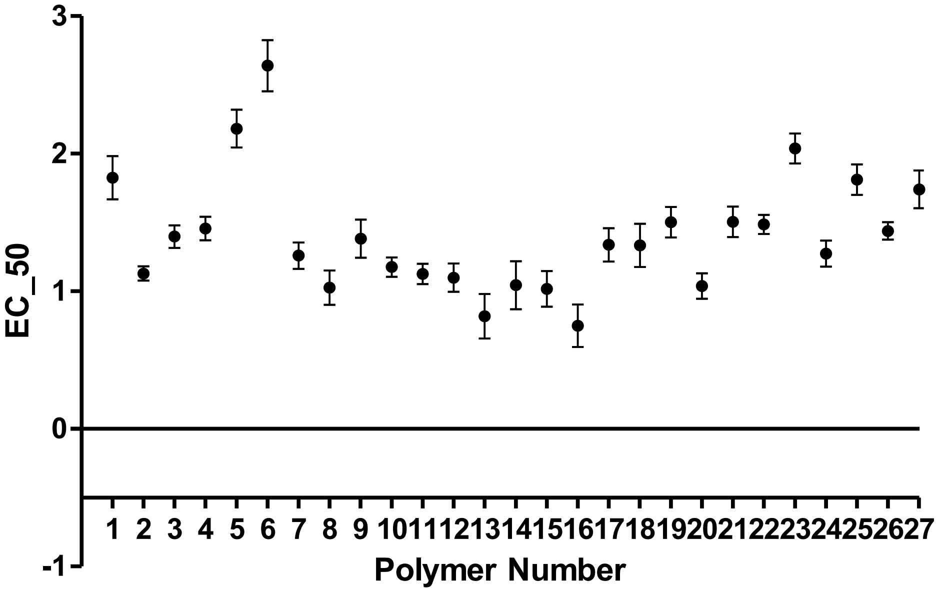
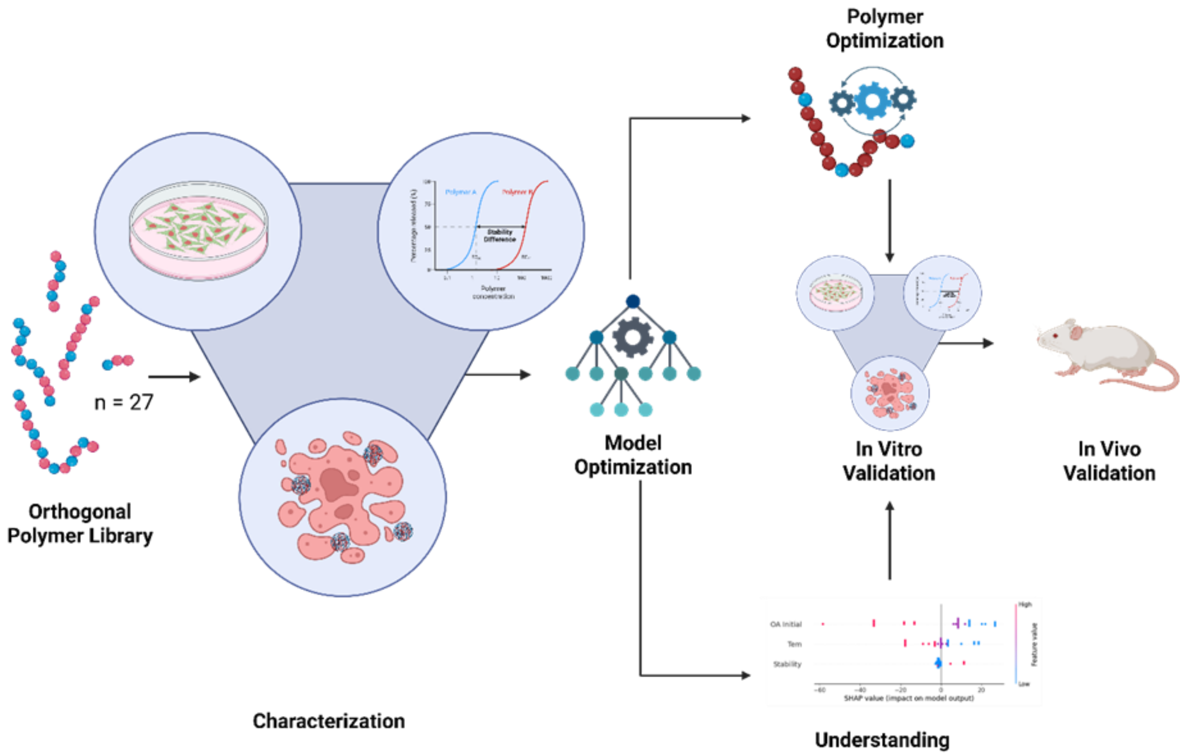


Figure II.S17: EC_50 values for siRNA containing nanoparticles generated with different polymers and determined by Heparin and Triton-X competition assay (n=3)

Chapter III - **Machine Learning on an Orthogonal Polymer Library Reveals Governing Factors and Optimizes PBAE Copolymers' Synthesis and Performance**

9 Graphical Abstract



10 Abstract

Pulmonary siRNA delivery is a promising therapeutic approach for future pandemics and many non-infectious lung diseases. Polymeric nanocarriers, especially poly-beta aminoesters are an easily tunable and versatile delivery system to protect RNA from degradation. To maneuver the vast chemical space and generate control and understanding, an orthogonal polymer library of amphiphilic-spermine-based poly-beta-aminoesters was investigated for gene knockdown, toxicity and particle stability.

Subsequently, a Nested-Leave-One-Out Cross Validation approach was chosen to screen different machine learning models allowing to capture useful information within the limited dataset. Analyzing key factors governing the particle performance identified too high intra-particle stability as a disadvantage for successful gene knockdown. This finding facilitated improved model performance through a few-shot learning approach. Leveraging these combined and optimized models, a novel polymer candidate was predicted and subsequently validated in vitro. A superior knockdown and toxicity profile as well as stability trends were confirmed. In vivo experiments, however, highlighted the lack of in-vitro-in-vivo correlation after model optimization for in vitro performance. Nonetheless, reduced in vivo immunogenicity was achieved through the chosen approach.

Keywords: PBAE polymers, siRNA Delivery, Machine Learning, Orthogonal Library, in vivo – in vitro correlation

11 Introduction

RNA-based therapeutics are rapidly transforming modern medicine, demonstrating profound impact across diverse therapeutic areas. The global pandemic highlighted the critical role of mRNA vaccines as a leading-edge biotechnological solution[6,112] for proactive disease prevention. While the success of mRNA vaccines is undeniable, the therapeutic potential of RNA extends considerably beyond prophylactic applications. Harnessing the inherent versatility of RNA's biological functions opens up a wide spectrum of therapeutic possibilities, reflecting their fundamental role in cellular processes. One potential therapeutic approach is the use of short interfering RNA (siRNA) for target gene silencing. This regulatory RNA is built intracellularly by slicing double stranded RNA (dsRNA) molecules into 20-25 nucleotide long sections and leading to mRNA degradation via an enzyme complex called "RNA induced silencing complex" (RISC). This mechanism could unlock a promising pulmonary antiviral therapeutic strategy for future pandemics[113]. Since RNAs are prone to degradation after injection into a patient due to ubiquitously expressed RNase enzymes, they need to be protected. For this purpose, various nanocarriers, generated from different materials and compositions, are used. Intensively investigated carriers for performing successful delivery are polymeric delivery systems such as PEI[114], PLGA[115,116] or PBAEs[80,117]. Although all are established materials, only the latter provides high cargo condensation while being biodegradable at the same time[86], making PBAEs well-suited for RNA delivery.

As the tremendous amount of potential chemical structures enables infinitely many possibilities of tailoring polymers for each individual use case[118], a strategy is needed, for researchers to design a carrier system that suits their needs faster than with a classical trial-and-error approach. One potential way to do so is rational design using human knowledge [119–121]. While promising, this requires a large amount of expertise and may lead to human errors due to biases and limited capability of extrapolating beyond experience. Another strategy used, is the screening of big libraries[81,122]. This allows for the discovery of a broad chemical space and has already led to the discovery of high-performing carrier systems. However, while being promising on the one hand, this method can only be applied if abundant resources, time and workforce are available which is not applicable for many labs. For this purpose, drug delivery research has started to implement more systematic attempts such as design of experiments (DoE), a method where an a-priori design space is set up, helping in systematically discovering a huge space without performing unnecessary experiments. Even though this method established itself as the gold standard in industry for most optimization tasks[44], it provides a rigid scaffold limited by the pre-selected design region and data points.

Machine learning (ML) is a powerful method that can overcome this limitation by allowing for a nearly infinite flexibility in data analysis, optimization and prediction, which makes it an increasingly integral component of modern drug discovery pipelines[123,124]. In recent years, several groups have contributed towards potential applications of ML in designing drug delivery systems[125,126]. However, ML is known to be heavily dependent on both data quantity and quality, which is a problem in the field of polymeric drug delivery, where data is often sparse or too heterogenous to use. Current contributions in the field predominantly focus on either machine learning (ML)-assisted high-throughput screening[127] or the utilization of existing datasets[128]. However, these approaches present inherent limitations, particularly within academic research settings. High-throughput screening infrastructure is often unavailable or impractical for many research questions, while sufficiently large and diverse datasets, capable of enabling robust predictive modeling, remain scarce, especially in comparison to the data abundance available for small molecules.

Here a new method is introduced, where ML is used within a previously synthesized small dataset of spermine-based amphiphilic poly-beta aminoesters (PBAEs)[129]. The data obtained by using a DoE design allowed for precise synthesis and a deeper understanding of the process itself. Subsequently, it is used to optimize PBAE capability for successful gene knockdown while maintaining low cytotoxicity. Additionally, an approach is presented to tackle the low-data problem using a nested-leave-one-out cross validation loop to design a robust algorithm to predict synthesis conditions that enable the polymerization of a new lead candidate that outperformed the current benchmark. Furthermore, it was shown that machine learning is the method of choice when incorporating additional information about data due to the flexibility in designing few-shot modelFinally, a deeper understanding of feature-relations was generated, by performing feature ablation studies and investigating SHAPley[130] values for the models. To translate the theoretical work into a practical set-up and to show the strengths but also the limitations of machine learning in this context, subsequently the optimized nanocarrier was initially tested in vitro. Here, the performance of the algorithm was validated and key findings about particle stability were confirmed. Testing the in-vitro-in-vivo-correlation, gene knockdown and toxicity as well as immunogenicity were investigated in mice.

12 This work lays the ground for researchers to make optimal use of limited data and helps in predicting and understanding new delivery systems without extensive and ineffective screening.

13 Materials and Methods

13.1 Materials

Dicer substrate double-stranded siRNA targeting enhanced green fluorescent protein (eGFP) (siGFP, 25/27mer), and scrambled, negative control siRNA (siNC, 25/27mer) were purchased from IDT (Integrated Technologies, Inc., Leuven, Belgium). Sequences and additional information are provided in the Supporting Information, Table S1. HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), ethyl trifluoroacetate, sodium chloride, Tris-EDTA buffer solution 100×, RPMI 1640 medium, Triton X-100, heparin sodium salt from porcine intestinal mucosa, heat-inactivated fetal bovine serum (FBS), penicillin/streptomycin solution (P/S), geneticin (G418), Dulbecco’s phosphate-buffered saline (PBS), cOmplete™ Mini EDTA-free protease-inhibitor-cocktail were obtained from Sigma-Aldrich (Darmstadt, Germany). Branched polyethyleneimine (PEI) (5 kDa, Lupasol G100) was a kind gift from BASF (Ludwigshafen, Germany). Di-*tert*-butyl decarbonate, oleylamine, spermine, dimethylformamide (99,5% pure), Lipofectamine 2000, OPTI-MEM serum reduced medium, 0.05% trypsin-EDTA, Alexa Fluor 647 NHS ester, and a SYBR Gold Nucleic Acid Gel Stain 10,000X concentrate in DMSO and siMMP7 were purchased from Thermo Fisher Scientific (Schwerte, Germany). 1,4-Butanendiol diacrylate was obtained from TCI Chemical Industry Co., Ltd. (Tokyo, Japan). Trifluoroacetic acid (99,9%, extra pure) was purchased from Acros Organics (Geel, Belgium). Methanol-d6 was obtained from Deutero (Kastellaun, Germany). Dichloromethane, methanol, ammonia, potassium permanganate, magnesium sulfate, acetone, pentane, and formic acid (>99% pure) were purchased from VWR Chemicals (Ismaning, Germany).

13.2 Data Preprocessing

Experimental data was saved in Excel format and was transformed in a pandas dataframe. The features were defined as Time ("Time"), Temperature("Tem"), initial Oleylamin content ("OA"), Diacrylate ratio ("DAR"). As target values we defined Gene Expression, Toxicity and Stability. Note that Stability was used as additional feature in a few-shot approach when predicting Gene Expression and Stability. Subsequently data was scaled using a MinMaxScaler. In this complete dataset, no values were missing.

13.3 Nested-CV-Loop

The selection of an appropriate model is a critical step in running a predictive machine-learning pipeline. Because we are dealing with data scarcity, we used only algorithms that are known to perform well with limited data. Each model was placed in a single scikit-learn pipeline together with a Min–Max scaler to avoid information leakage. We employed a nested cross-validation scheme: first, 15 % of the data was split off as a hold-out set, which was evaluated only after hyper-parameter optimization. To ensure that the hold-out set represented the distribution of the training data, we discretized the continuous target into five equal-frequency (quantile) bins and stratified the train–test split on those bins. In the inner loop, 100 randomly chosen hyper-parameter configurations were assessed for each model using leave-one-out cross-validation (LOOCV). After ten outer-loop repetitions, the model with the lowest mean absolute error (MAE) and its associated optimal hyper-parameters were selected for subsequent optimization.

Zero Shot vs Few-Shot Model

To compare whether certain additional experimental data can help in predicting others, we investigated the influence of the experimentally determined colloidal stability of the nanoparticle suspension. To do so, we included experimental stability values as additional features into the gene expression and toxicity models. Since we experienced a threshold-like behavior of Gene Expression and stability, the stability data was binarized after normalization.

13.4 Feature Ablation

To investigate the influence of the single features and whether they influence the predictive power of the model, feature ablation experiments were executed. For this purpose, we iteratively removed features and compared the performance across all LOOCV splits as absolute mean error with a base model containing all features. When exceeding the error threshold, the feature was assumed to just add noise to the model and was rated irrelevant.

13.5 Optimized Model Comparison

13.6 Model evaluation included a comparison of the optimized models against a simple mean predictor baseline, providing a straightforward benchmark. This dummy model always predicts the average value of the training set's target variable. The MAE achieved by the baseline model was contrasted with that of our few-shot and zero-shot models.

13.7 Model Interpretation

Model interpretation was performed using SHAP (SHapley Additive exPlanations) values to quantify each feature's contribution to the difference between the model's prediction and the expected value, providing insights into model behavior and enabling identification of critical features. To visualize feature importance for the zero-shot and few-shot models, we employed beeswarm plots. Furthermore, waterfall plots were used to illustrate the decision-making process of the models. Finally, feature relationships were investigated using scatter plots of SHAP values against their corresponding feature values.

13.8 Prediction Pipeline

Parameter prediction was performed using a combinatorial approach. Specifically, we generated discrete parameter ranges and combined these ranges to create an exhaustive list of possible parameter settings. These settings were then evaluated using the zero-shot models. The resulting performance metrics were stored in a data frame and subsequently sorted using a hierarchical sorting strategy. This allowed us to identify parameter configurations that maximize gene knockdown while minimizing toxicity.

13.9 Triboc-Spermine Synthesis

Tri*tert*-butyl carbonyl spermine, abbreviated as tri-Boc-spermine (TBS) was synthesized as described elsewhere[100]. Briefly, spermine (1 equiv) was dissolved in methanol and stirred at −78 °C before ethyl trifluoroacetate (1 equiv) was added dropwise. Subsequently, the mixture was stirred at −78 °C for 1 h and then at 0 °C for 1 h. Without isolation, di*tert*-butyl dicarbonate (4 equiv) was added dropwise to the solution and stirred at room temperature for 2 days. Finally, the solution was adjusted to a pH above 11 by 25% ammonia and stirred overnight to cleave the trifluoroacetamide protecting group. The solvent in the mixture was then evaporated under vacuum, and the residue was diluted with dichloromethane (DCM) and washed with distilled water and saturated sodium chloride aqueous solution. The DCM phase was finally dried by magnesia sulfate and concentrated to give the crude product. The crude product was purified by column chromatography (CH_2Cl_2\MeOH\NH_3, aq 7:1:0.1, SiO_2, $KMnO_4$; R_f = 0.413). TBS was isolated and characterized by ^{1}H nuclear magnetic resonance spectroscopy (^{1}H NMR).

13.10 Polymer Synthesis and Characterization

Poly spermine-*co*-oleylamine beta-aminoesters (P(SpOABAE)) were synthesized based on a previously described approach[114]. Briefly, TBS as a hydrophilic monomer, oleylamine (OA) as a hydrophobic monomer, and 1,4-butanendiol diacrylate (DA) as backbone were mixed in different molar ratios in dimethylformamide (DMF), resulting in total concentrations of 300 mg/mL. After the respective reaction time, mixtures were transferred to Petri dishes to evaporate the solvent. The subsequent deprotection of the polymer was carried out in a mixture of 20 mL of dichloromethane (DCM) and 1 mL of trifluoroacetic acid (TFA) for 100 mg of polymer, followed by stirring for 2 h at room temperature. In the following, DCM/TFA was evaporated and the dry deprotected product was precipitated 3 times in pentane using acetone to dissolve the precipitate. Supernatants were discarded, and the final precipitate was dried for 2 days under vacuum (room temperature, 20 mbar). The synthesis process is depicted in Figure III.1A. Final polymers were characterized by ^{1}H NMR and GPC. Measurements were performed with an Agilent aqueous GPC using a PSS Novema Max Lux 100A followed by two PSS Novema Max Lux 3000A columns. The chromatographic system and calibration standards were set up according to preanalysis from Agilent Technologies on P(SpOABAE) polymers. Measurements were performed at 40 °C in a 0.1 M sodium chloride solution supplemented with 0.3% formic acid. Samples were prepared at 4 g/L and measured at a flow rate of 1 mL/min. Molar mass distributions were obtained through the Agilent WinGPC software against pullulan calibration standards in the range of 180 Da to 1450 kDa. A daisy-chain detector setup of an Agilent 1260 VWD was used, followed by an Agilent 1260 GPC/SEC MDS and ending with an Agilent 1260 RID.

13.11 Gene Knockdown

H1299 stably expressing eGFP were seeded on 48-well or 24-well plates at a density of 5,000 or 10,000 cells per well in 1640 RPMI supplemented with 10% FCS and 1% Penicilin/Streptomycin, respectively. Nanoparticles were prepared at N/P ratio 10 encapsulating either siGFP or siNC RNA, and cells were transfected 24h after seeding in triplicates with 10 or 20 pmol siRNA per well. After 48 hours, median fluorescence intensity (MFI) was recorded using a BD LSR Fortessa using the BD FACSDivaTM Software and counting 10,000 events. Gene knockdown was calculated as the ratio between MFI of cells treated with siGFP NPs and siNC NPs.

13.12 Cell Viability

Cell viability and toxicity were tested simultaneously using a CellTiter Blue (CTB) and Lactate dehydrogenase (LDH) assay. In 96-well plates, 5,000 16HBE14o- cells were seeded. After 24 hours, the polymer library was tested in triplicates. Each polymer was tested at 8 different concentrations between 1 and 500 µg/mL. After 48 hours of incubation, 50 µL supernatant of each well was transferred to a fresh plate and LDH was quantified following the manufacturers protocol. Briefly, to each well 50 µL of freshly resuspended reagent mix was added, and the plates were incubated in the dark for 30 min. Afterwards, 50 µL stop solution was added into each well and absorbance was measured.

For the CTB assays, the cell containing wells were filled up with 30 µL of fresh media and 20 µL CTB and incubated for 4h. Afterwards, absorbance was measured at 570 and 600 nm.

Using JMP 17 pro, sigmoidal curve fits were generated through all concentrations and repetitions of the CTB and LDH assays, and turning points were calculated and defined as IC50 values.

13.13 Determination of attractive forces between siRNA and polymers

A previously reported stability assay was used to determine the attractive forces between siRNA and polymers. The stability values for the input library were reported in the same publication[129]. Following this protocol, nanoparticle stability was investigated using heparin and triton-X. Briefly, 10 µL nanoparticle suspension was treated with 20 µL of 8 different concentrations of a mixture of heparin and triton-X in a black 384-well plate (Greiner Bio-One, Frickenhausen, Germany). As reference, siRNA solutions resembling the concentrations of NPs were treated with the same concentrations of heparin and triton-X.

Plates were sealed and incubated for 1h at 37°C at 250 rpm. Afterwards 5 µL of a 4x SYBR Gold solution were added to each well and mixed by pipetting. After 5 minutes of incubation fluorescence was measured at 492/20 nm excitation wavelength and 537/20 nm emission wavelength. Comparing the fluorescence intensity of the treated nanoparticle solution to the respective siRNA solutions' intensity, a release percentage was calculated. Fitting the released percentage against the used concentration of heparin and triton-X, using Prism5 software, an EC50 value was calculated. This value was defined as the concentration at which half of all siRNA is released from the nanoparticle suspension.

13.14 Animal Treatment Protocol

Female BALB/c mice, aged 6-8 weeks, were purchased from Charles River Laboratories. The mice were housed in a controlled facility for 14 days to acclimatize, with a 12-hour light/dark cycle. All animal procedures were approved by the Government of Upper Bavaria and conducted in accordance with approved protocols.

Mice were intratracheally instilled with 1 nmol of siRNA encapsulated at N/P 10 with either the previous lead candidate or the new ML-2 polymer, administered through intratracheal instillation under ketamine/xylazine anesthesia. As control, equivalent volume of 25kDa hyperbranched PEI polyplexes encapsulating the same amount of siRNA was applied as well as unencapsulated siRNA or pure formulation buffer. All formulations were tested with either siRNA targeted against murine Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) or negative control (NC). Mice were euthanized 24 hours after application mice through cardiac blood collection.

Lungs were flushed twice with 500 µL of PBS buffer containing 2 mM EDTA and one cOmplete™, Mini, EDTA-free protease-inhibitor-cocktail tablet per 10 mL to collect the bronchoalveolar lavage fluid (BALF). Briefly, solutions were injected into the trachea and subsequently recollected. A second 500 µL of the same PBS solution was instilled and recollected. The collected BALF was centrifuged for 5 minutes at 500 g. The supernatant was frozen at -20°C and stored at -80°C until further analysis.

Lungs were subsequently perfused with 20 mL of 0.9% sterile sodium chloride. To do so, the vena cava inferior was cut and the solution injected into the left ventricle. After sufficient perfusion, one lung lobe from each treatment group was dissected, fixed in 4% paraformaldehyde (PFA) for at least 24 hours, and then embedded in paraffin for histological analysis via H&E staining.

The remaining lung lobes and undissected lungs were stored at 1 mL RNAlater™ Stabilization Solution, frozen and stored at -20°C until further analysis.

13.15 In Vivo Gene Knockdown

GAPDH gene knockdown in mouse lungs was determined through qPCR. RNA was isolated from mouse lungs using Lysing Matrix D tubes containing 1.4 mm Zirconium-Silicate spheres from MP Biomedicals and a TRIzol/chloroform isolation protocol. Briefly, mouse lungs were thawed on ice and transferred to the lysing tubes. After the transfer, 1 mL of TRIzol was added to each tube. Using a Tissue Lyzer the samples were homogenized. The RNA was isolated through chloroform precipitation. After centrifugation, the aqueous phase was washed with molecular grade isopropanol followed by ethanol. The final RNA pellets were dissolved in RNase free water and concentrations were determined. Using a high-capacity cDNA reverse transcription kit (Thermo Fisher Scientific), complementary DNA (cDNA) was prepared. Finally, quantitative real-time PCR (qRT-PCR) was performed applying an iTaq Universal SYBR Green Supermix (Bio-Rad, Feldkirchen, Germany) on a StepOnePlus system (Thermo Fisher Scientific). Beta-Actin was used as the reference gene with Mm_GAPDH_3_SG primers (Qiagen) for GAPDH and Mm_ACTB_2_SG (Qiagen) primers specific for mouse β-actin. For normalization of GAPDH levels, the ΔΔCt method was applied.

13.16 In Vivo Biodistribution and Cell Uptake

To investigate the biodistribution and cellular uptake 6–8-week-old BALB/c mice were treated with 1 nmol of siRNA fluorescently labeled with a AF647 label as described previously. siRNA was either applied unformulated or encapsulated into the previous lead candidate or ML-2 polymer. After 24 hours, mice were sacrificed, and bladders, lungs, livers, kidneys, spleens, and the hearts were collected. Using an IVIS Lumina III (PerkinElmer, Shelton, CT, USA) fluorescence intensity in these organs was measured.

For further analysis, lungs were dissociated using a gentleMACS tissue Dissociator (Miltenyi Biotec, Bergisch Gladbach, Germany) together with gentleMACS C (Miltenyi Biotec, Bergisch Gladbach, Germany) tubes following the manufacturers protocol. Cell suspensions were incubated with PBS solution containing Zombie UV™ and afterwards stained with FITC anti-mouse CD45, BUV395 anti-mouse CD3, Vioblue anti-mouse CD4, APC-Cyanine7 anti-mouse CD8, PE-Cyanine7 anti-mouse F4/80, BUV605 anti-mouse CD11c, BV785 anti-mouse CD326, PE/Dazzle™594 anti-mouse CD170 and

PerCP/Cyanine5.5 anti-mouse CD19 for 30 min at 4°C. The stained cells were measured using a Cytek® Aurora (San Diego, California, USA) implemented with autofluorescence extraction for the detection of cellular uptake (Figure III.S1).

13.17 BALF Cytokine Measurements

Cytokines from collected BALF solutions were quantified using a LEGNEDplex^{TM} Mouse Inflammation Panel (Biolegend, San Diego, California, USA) following the manufacturers protocol and an Attune NxT flow cytometer (ThermoFisher Scientific, Waltham, MA USA). Results are reported as total detected concentration and as relative induction compared to the highest induction for each individual cytokine.

14 Results and Discussion

14.1 Library Performance Evaluation

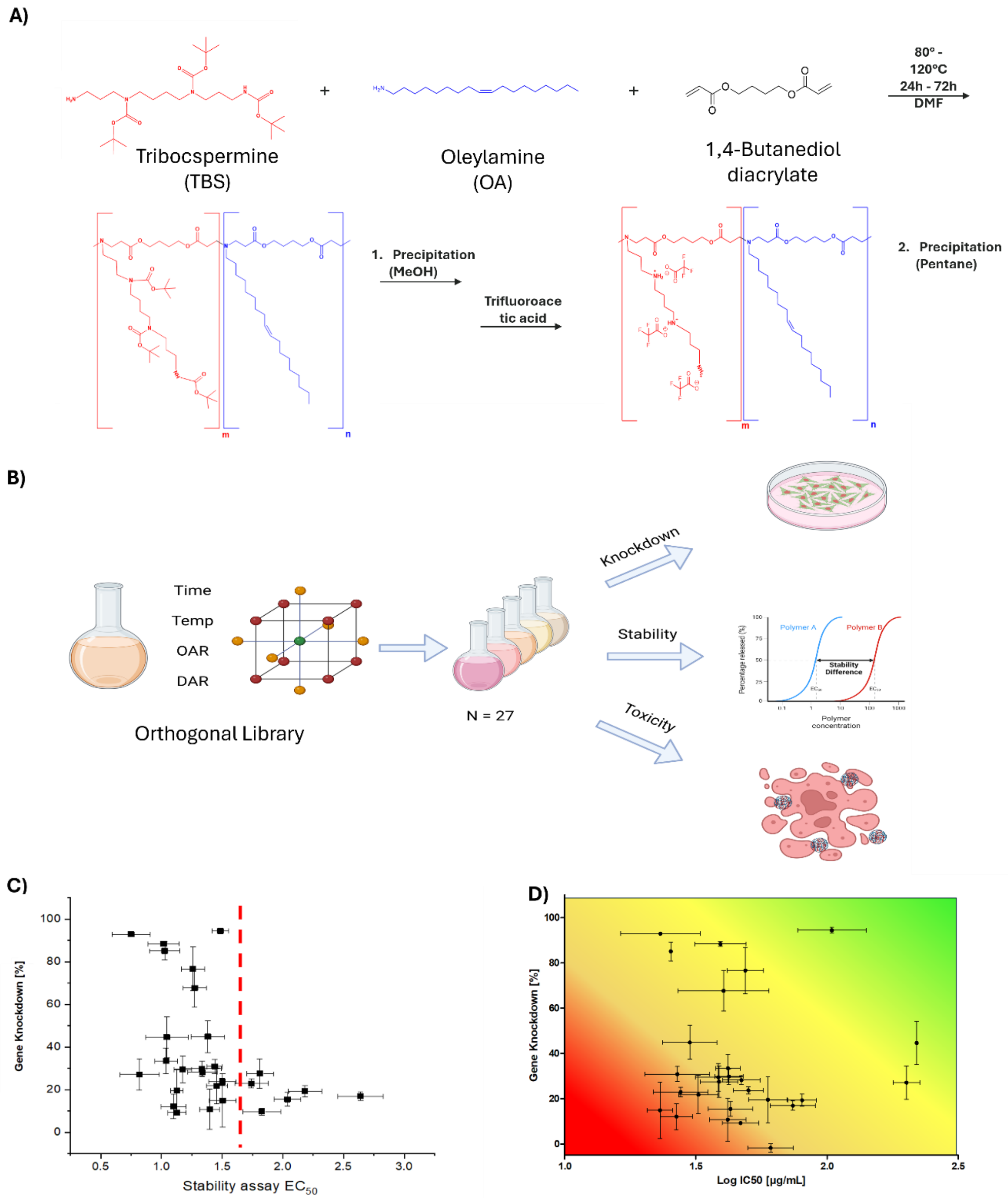


Figure III.1: Workflow of the screening process applied in this study. A) Synthesis approach of the applied PBAE polymers B) A previously reported library generated through DoE and varying key synthesis parameters was tested for knockdown efficiency, stability and toxicity. C) Gene Knockdown correlated against previously reported stability of particles and D) against cell viability determined via CTB. Error bars depict SD for gene knockdown and SD of the fit for EC_{50} and IC_{50} with n=3.

The aim of this work was the investigation and optimization of key parameters governing the performance of PBAE polymers as siRNA delivery vehicles *in vitro* and *in vivo* for pulmonary therapy. We therefore utilized a previously reported library of 27 differently synthesized PBAE polymers (Figure III.1A)[129]. The library was generated through a Central Composite Orthogonal design optimizing the synthesis parameters of total synthesis time, synthesis temperature, oleyl amine ratio, being the ratio of the two sidechains, and diacrylate ratio, being the ratio of the sidechains to the backbone (Figure III.1B). All factors were investigated over 5 levels and with all resulting polymers, nanoparticles were successfully formulated. Nanoparticle stability was already reported[23].

To complement the previously reported data set, nanoparticles were tested for gene knockdown in an H1299 eGFP-expressing lung cell line by encapsulating and delivering siRNA against eGFP. The results were plotted against the previously reported stability values (Figure III.1C). Interestingly, an apparent division threshold was found within the data set. Above this threshold, the particles appeared to lose their functionality *in vitro*. This was unexpected since the common consensus suggests that particles need a certain stability to not lose their integrity before reaching the endosome. In contrast, the data presented here suggest that the major bottleneck for the investigated PBAE nanoparticles was not premature particle disintegration but rather excessively strong intraparticular stabilizing forces. Since only below the found threshold a successful gene knockdown above 90% was observed, it was hypothesized that at too high EC50 values, particles did not disintegrate within the endolysosomal pathway to release their siRNA cargo and mitigate gene knockdown. This hypothesis was underscored by the observation that above the identified threshold, the highest achieved gene knockdown effects were below 30%. Previous studys reported similar observations, implying that polyplexes lose potency if the intraparticular stabilizing forces become too strong to release the cargo[131]. On the other hand, weakening the intraparticular forces can increase the nanoparticles performance[132]. Therefore, a clear design criterion for next generation polymers was stated. The criterion was that nanoparticle stability needed to be lower than an EC_{50} value of 1.6, in order to successfully release the siRNA within the endosome.

In the next step, cytotoxicity and cell viability of the polymers from the library were investigated in pulmonary epithelial cells by the means of CTB and LDH assays (Figure III.S2). A correlation comparison between both IC50 results showed that the tested polymers were well tolerated in a range from 25 to 175 mg/mL and the results from CTB and LDH correlated strongly with each other (Figure III.S3+S4). As expected, polymers exhibiting higher toxicity also showed a greater negative impact on cell viability, and vice versa. Furthermore, this finding enabled a reduction in experimental workload and cost since a single assay was sufficient to reliably assess polymer safety. CTB assays resulted in a slightly lower IC50 value than LDH assays (Figure III.S3). Moving forward, for these reasons CTB was chosen as main readout.

To finally evaluate the performance of the polymer library, gene knockdown was plotted against the IC50 values determined via CTB (Figure III.1D). This showed another surprising finding, which was the successful decoupling of toxicity from efficiency of the nanoparticle system. One of the biggest challenges for RNA delivery is the "*efficiency/safety dilemma*", where higher transfection efficiency is often associated with increased cytotoxicity. The root cause is most likely associated to the membrane disruptive potential of the carrier system. A certain membrane fusogenicity is necessary for endosomal escape, while excessive disruption of endolysosomal compartments or cellular membranes can trigger immunogenicity, apoptosis and toxicity[133–135]. It was therefore a remarkable finding that the investigated library contained a polymer with exceptional gene knockdown as well as superior safety profiles (Figure III.1D, green area).

14.2 Nested CV Approach

Building upon the nested cross-validation framework described before[136], we implemented a similar approach with specific modifications tailored to our low-data context (Figure III.2A). First, recognizing the limitations of complex models in data-scarce settings, we opted to exclude the neuronal network component present in the referenced methodology. Second, to ensure the hold-out set was representative of the training data distribution, we stratified the dataset based on the target variable, dividing the data into five bins prior to splitting. This stratification ensured that each fold maintained a similar target distribution to the overall dataset. Furthermore, within the inner cross-validation loop, we employed LOOCV. LOOCV was chosen to maximize the training data available for each inner fold, which is particularly advantageous when working with limited datasets. In our experiments, we trained models to predict two distinct target variables: Gene Expression post-treatment and Toxicity, quantified as IC50 (see Methods section for details). We also investigated the potential benefit of incorporating additional nanoparticle characteristics, specifically stability, as input features. Consistent with the nomenclature used in[31], we refer to the variant that includes the additional stability descriptor as the "few-shot" model, even though only one extra feature is added; this usage follows the prior work’s feature-augmentation context and should not be confused with the standard few-shot/one-shot paradigms that describe limited numbers of training examples. While we observed improved results for the few-shot approach for all Gene Expression models (Figure III.2B), addition of stability did not seem to have a big impact on the IC50 value (Figure III.2C). The only model that slightly improved was the DecisionTree (DT). However, its performance was still poorer than that of the best zero shot-model, which was the RandomForest (RF) with an MAE of 0.3673. For the Gene Expression model, XGBoost outperformed other models (MAE of 14.18). However, for the few-shot model, the Support Vector Regressor (SVR) was slightly better. Good performance of an SVR with low data and non-linear interactions was already seen previously[137]. Among the best performing model class, we picked the best hyperparameter-setting for the most robust models (Figure III.S5), which were further optimized in the next steps.

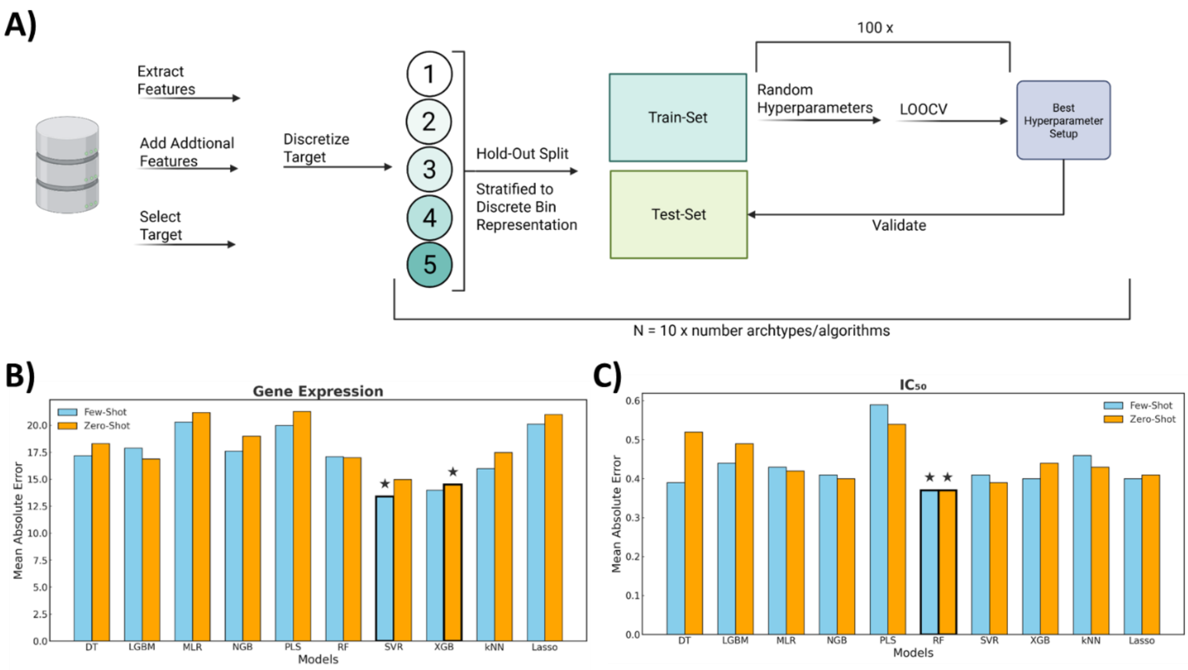


Figure III.2: Nested-Leave-One-Out Cross-Validation Approach A) Machine learning pipeline where data is preprocessed and subsequently categorized to allow for stratified splitting of holdout data. The train set is used to tune each algorithm with a random hyperparameter search and leave-one-out validation. The process is repeated ten times and the mean absolute error is calculated to obtain the most robust model. B) Mean Absolute Error of multiple models tested for Gene Expression with the ML pipeline. Few-Shot models (blue) with stability measurements of nanoparticles included. The models marked with an asterisk and a bold frame are the most robust models selected for optimization. C) Mean Absolute Error of multiple models tested for IC50 with the ML pipeline. Few-Shot models (blue) with stability measurements of nanoparticles included. The models marked with an asterisk and a bold frame are the most robust models selected for optimization.

14.3 Feature Ablation Experiment

To further optimize model performance and enhance process understanding, we conducted a feature ablation experiment (Figure III.S6). In this experiment, we evaluated the performance of each model, assessed via Leave-One-Out Cross-Validation (LOOCV), by iteratively removing individual features. For the Toxicity model, feature ablation revealed no statistically significant performance differences; only a marginal increase in MAE was observed when removing Temperature for the zero-shot model and Time for the few-shot model. Conversely, for the Gene Expression model, we observed that ablating Time and Diacrylate-Ratio (DAR) improved zero-shot model performance. In contrast, DAR remained important for the few-shot model. These findings align with our prior work, which indicated a limited impact of reaction time on polymer characteristics.

14.4 SHAP Analysis

To gain deeper insights into model decision-making, we calculated SHAP (SHapley Additive exPlanations) values for all models (see Figure III.3A and III.3B). The SHAP analysis generally corroborated the findings from the feature ablation experiment. Furthermore, it elucidated feature importance for predicting high knockdown/low gene expression, suggesting a requirement for high oleylamine content (OA Initial) and elevated Temperature (Tem) in the zero-shot model. In contrast, the few-shot model's SHAP values reflected the stability threshold identified previously. For the IC50 prediction, Temperature emerged as a significant parameter, with lower temperatures associated with reduced toxicity, while higher OA Initial concentrations appeared favorable. This observation may be attributed to the potential formation of a side-product at elevated temperatures, as documented in our earlier publication[129]. Stability, however, exhibited no influence on predicted toxicity (Figure III.3B). It is important to note that SHAP values represent model interpretations rather than ground truth and, given the weaker predictive performance of the IC50 model, these results require cautious interpretation. Detailed SHAP plots for all models and features and correlation plots between SHAP values and features are provided in the Supplementary Information (Figure III.S7 and III.S8).

14.5 Final Model Performance and Baseline Comparison

To demonstrate the final model performance, we benchmarked both the zero-shot and few-shot models against a dummy baseline model (see Methods section). Additionally, we visualized the results in predicted-versus-real plots (Figure III.S9). The Gene Expression zero-shot model exhibited promising performance, achieving a MAE of 10.59 and a Pearson correlation coefficient (r) of 0.8494 in the predicted-versus-real plot (Figure III.S3A and Figure III.9A). The incorporation of stability as a feature further enhanced predictive performance (MAE= 7.605, r=0.9078), underscoring the existence of a stability threshold above which particle stability is too high to release the cargo into the cytosol, what was already observed in earlier work (Figure III.3A and Figure III.S9B). For the Toxicity model, performance improvements over the baseline (MAE of 0.2816 versus MAE of 0.3476) were observed, and a correlation between predicted and experimental values was evident for the zero-shot model (r= 0.3605, Figure III.3B and Figure III.S9C). However, no significant difference was found between the zero-shot and few-shot models (Figure III.3B and Figure III.S9D), further supporting the conclusion that stability does not substantially influence the toxicity of the nanocarrier system.

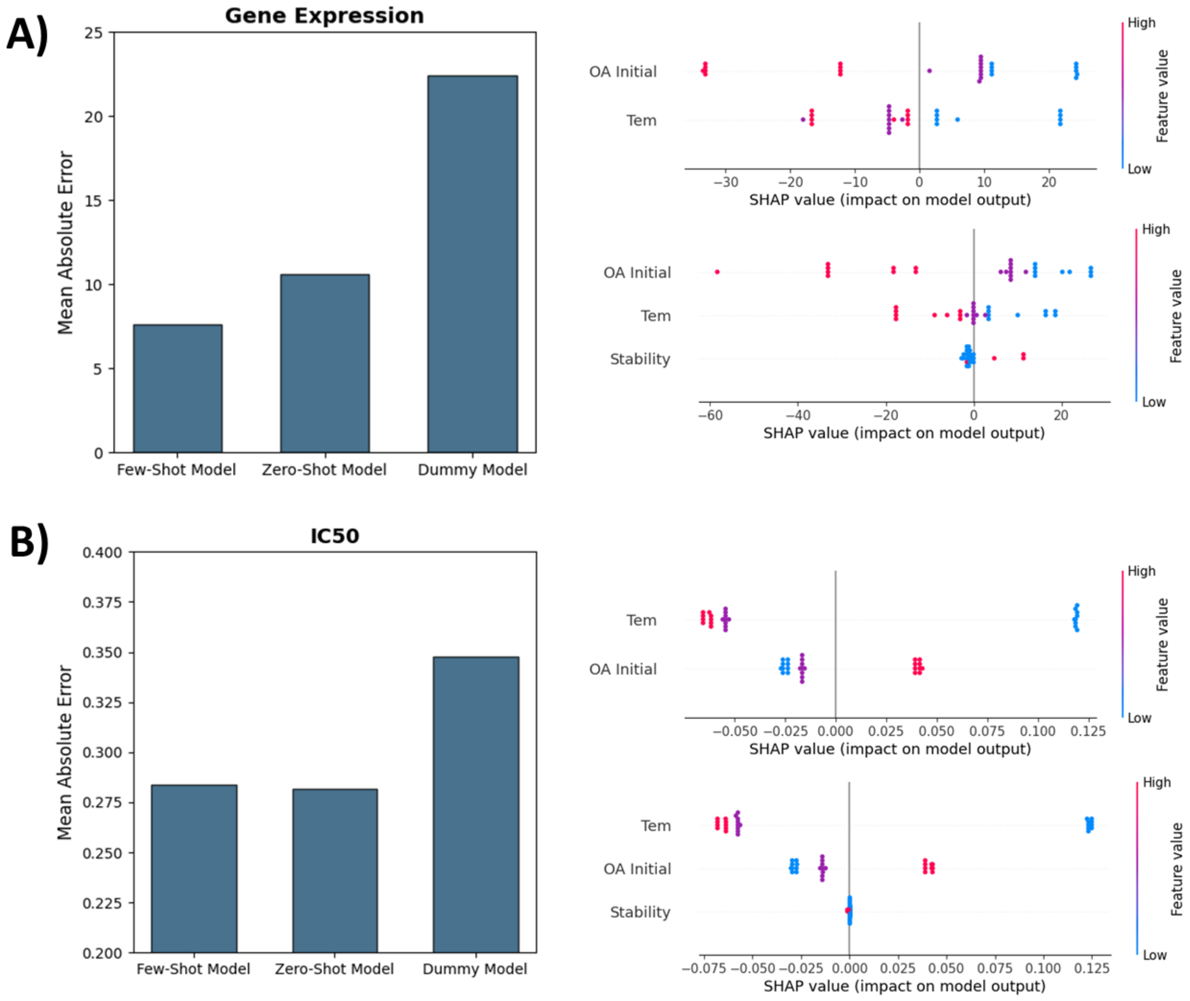


Figure III.3: Optimized Model Characteristics A) Gene Expression MAE Comparison of optimized Few-Shot and Zero-Shot Models with a Dummy-Baseline Model evaluated with LOOCV above: SHAP values of Zero-Shot

model and below: few-shot model B) IC50 MAE Comparison of optimized Few-Shot and Zero-Shot Models with a Dummy-Baseline Model evaluated with LOOCV upper panel: SHAP values of Zero-Shot model and lower panel: few-shot model.

14.6 End-to-End Prediction Pipeline and Validation

To ultimately validate the utility of machine learning with limited data, for predicting novel formulations, we constructed an end-to-end prediction pipeline (Figure III.4A). This pipeline involved generating all feasible combinations within physically plausible feature ranges and employing our zero-shot models as an independent multi-output model to predict Gene Expression/Knockdown and Toxicity. Given the superior predictive power of the Gene Expression model, we implemented a hierarchical sorting strategy, prioritizing high knockdown followed by low toxicity. The model-predicted optimal polymer, termed ML-2 and characterized by 95% OA Initial and synthesis at a Temperature of 130°C, was subsequently synthesized (see Methods section), analyzed (see Figure III.S10 and Figure III.S11), and experimentally validated. To further highlight the model's decision path, we added additional SHAP waterfall plots (see Figure III.S12 and Figure III.S13), confirming the results from the full model's beeswarm plot.

14.7 Machine Learning-Derived Polymer Evaluation *in vitro*

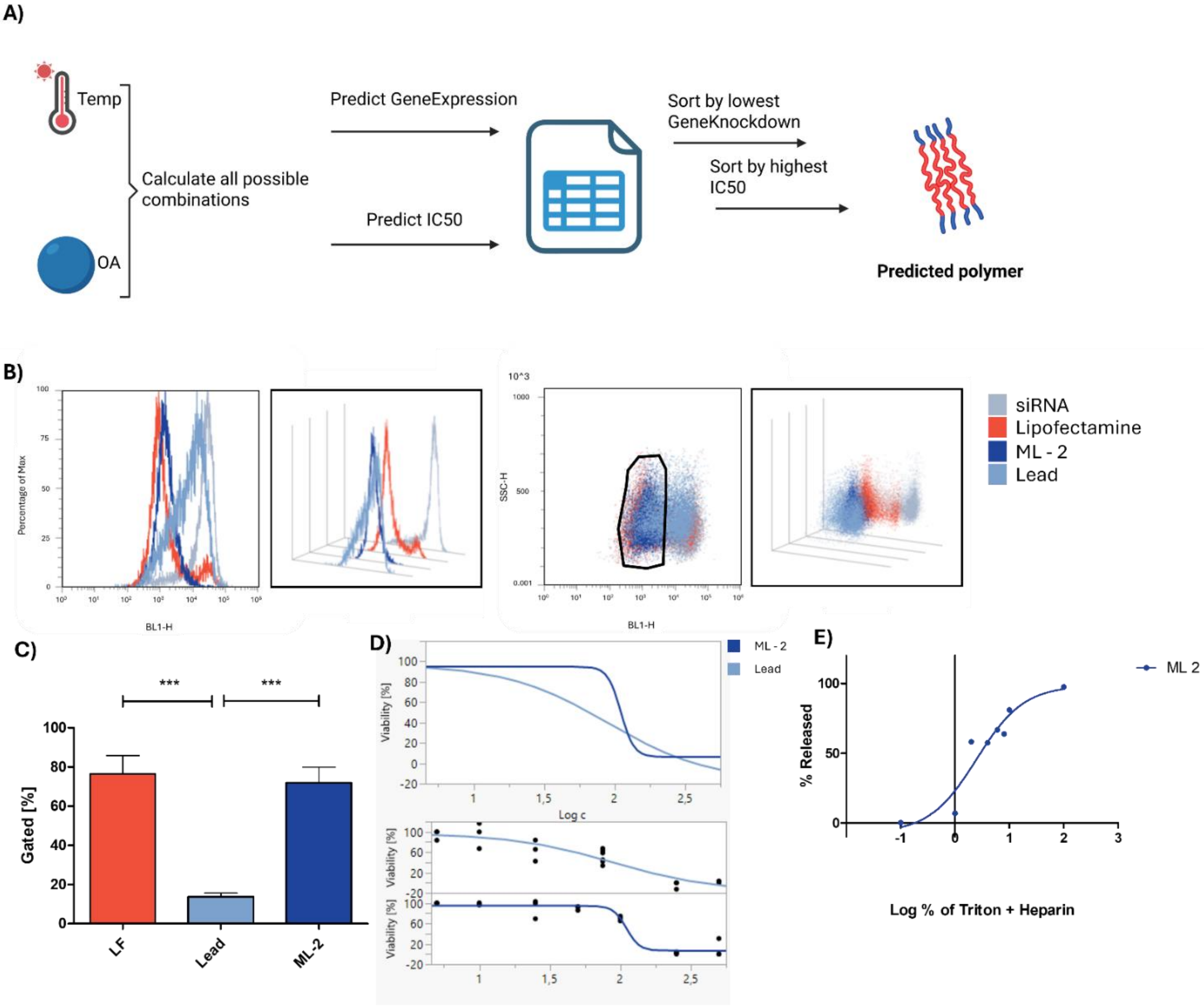


Figure III.4: In vitro performance evaluation and comparison of optimized PBAEs. A) Overview of the prediction pipeline for the optimized polymer, B) Histogram and Dot plot of H1299 eGFP cells treated with Lipofectamine 2000, ML-2 or the previous lead candidate encapsulating siGFP siRNA, and C) percentage of gated cells with nearly complete knockdown of eGFP with N=3 (*** depicting a p ≤ 0.001). D) Toxicity of ML-2 and lead candidate determined via CTB assay with n=3, and E) stability of ML-2 determined through Heparin and Triton-x competition. Dots depict mean of n=3.

To validate the performance of the new ML-2 polymer as pulmonary delivery agent, it was compared against a previously reported lead candidate[132] derived from classical trial and error synthesis optimization. In the following this polymer will be referred to as “Lead” candidate. Besides different synthesis settings, these two polymers mainly differ in their OA ratio, with the predicted ML-2 having a higher ratio at 93% and the previous Lead polymer a lower at 75%. To investigate if the new ML-2 polymer was indeed superior in performance, a gene knockdown experiment in H1299 eGFP cells was conducted. As shown in Figure

III.4 B) ML-2 did indeed mediate a more potent gene knockdown than the Lead polymer and seemingly a more complete downregulation than Lipofectamine 2000 (Figure III.4 B). The median fluorescence intensity did not differ significantly between Lipofectamine 2000 and ML-2 (Figure III.S14). To get a more detailed view on the differences on the polymers' performances, the dot plots of the cell populations were compared via the gated percentage (Figure III.4 B +C). ML-2 was clearly superior to the Lead polymer but showed again no statistical difference compared to Lipofectamine 2000. The Lead polymer on the other hand showed a large cell population with a non-complete gene knockdown. This indicates that the lead polymer does not reach saturation of cytosolic siRNA delivery unlike ML-2. This difference of saturation is also depicted in the gated percentage (Figure III.4 C) and clearly shows the superior efficiency of ML-2 compared to the Lead polymer.

A major downside of the previous Lead candidate is the early onset of toxicity as can be seen from the CTB curve (Figure III.4 D). Even though the IC_{50} value of the Lead polymer is in an excellent range with 89 µg/mL, the early onset of the curve decline indicates that toxicity can already occur at much lower concentrations. ML-2 showed a superior IC_{50} value, although in a comparable range with an IC_{50} value of 109 µg/mL. However, additionally to a higher IC_{50} value, the curve decline was also much steeper indicating a much later "onset of toxicity" at higher concentrations. This finding confirmed the potential of the machine learning approach since ML-2 showed to have better efficiency and safety profiles than the previous lead candidate.

Finally, to prove our previous findings, we determined the stability of the ML-2 nanoparticles (Figure III.4 E), which was in the expected range, below the above-described threshold necessary for successful gene delivery.

14.8 Machine Learning-Derived Polymer Evaluation *in vivo*

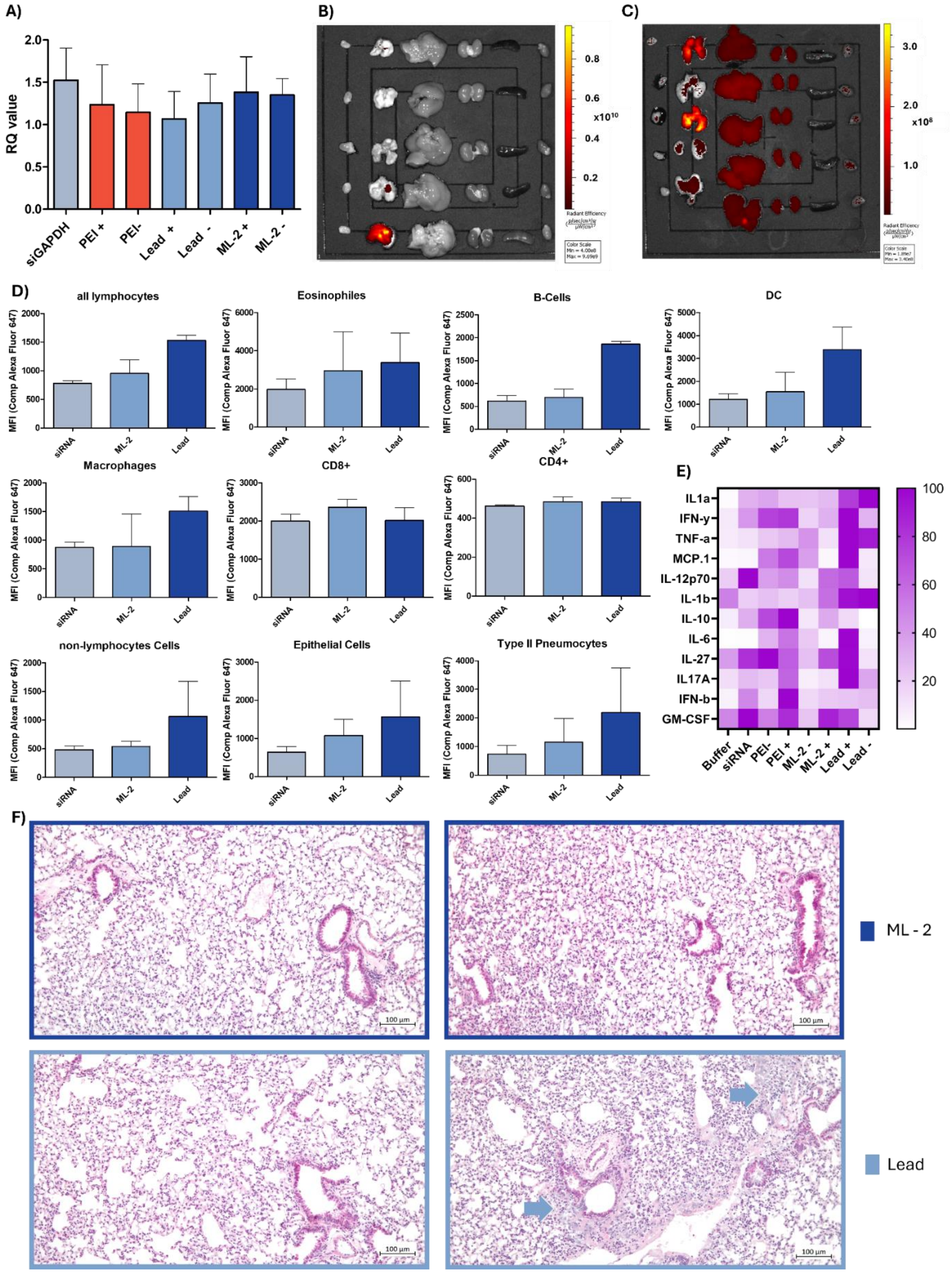


Figure III.5: *In vivo* results of the lead and ML-2 comparison. A) Fold-change of GAPDH against β-actin determined by ΔΔCt method with buffer only as reference standard. B) Fluorescence intensity measurements of bladder, lungs, liver, kidneys, spleen, and heart (from left to right) 24 hours after intratracheal instillation of 1 nmol siRNA encapsulated into lead (top three) and ML-2 (bottom two) polymer, or C) 1 nmol of pure siRNA. D) Flow cytometric analysis of cell suspension generated from mouse lungs through tissue grinders. E) Cytokine expression measured in BALF samples, normalized to the respective maximum value. F) Tissue slices from mouse lungs treated with ML-2 (top) encapsulating siGAPDH (left) and siNC (right) and lead polymer (bottom) encapsulating siGAPDH (left) and siNC (right).

In order to investigate if the superior properties of ML-2 would translate into an *in vivo* model both polymers were applied to female BALB/c mice intratracheally. Unfortunately, no clear gene knockdown for ML-2 was observed as well as just a slight reduction in gene expression for the Lead polymer (Figure III.5. A). This could be associated with the GAPDH housekeeping gene, which plays a crucial role in cell metabolism. A forced downregulation via e.g. siRNA can lead to upregulation of the gene translation as compensation, which is reflected by the observation, that PEI did not mediate a gene downregulation either. Additionally, the loss of efficacy moving from *in vitro* to *in vivo* models is not unprecedented. Another reason for this poor in-vitro-in-vivo correlation could be the challenging barriers in intratracheal applications such as the presence of respiratory mucus and the bronchoalveolar architecture. To investigate this hypothesis, we tested the Lead polymer in an air-liquid- interface (ALI) cell culture model of mucus producing CALU-3 cells where a similar loss in efficacy was observed (Figure III.S15.). This shows that the bronchial mucus forms a major barrier neglected by the machine learning algorithm utilized here. Although the mucus hampers the delivery of the nanoparticles to the lung cells, a considerable retention within the lungs (Figure III.4 B) was still observed compared to blank siRNA (Figure III.4 C), which was rapidly distributed throughout the entire body. A deeper investigation of the uptake into lung cells through flow cytometry showed that especially the Lead polymer mediates a considerable uptake in most cell types (Figure III.4D and Figure III.S16). For a therapeutic effect, uptake into epithelial and type II pneumocytes, the most relevant and most prevalent cell types, is commonly aimed for. In both cell types, the Lead polymer enabled a superior uptake compared to the ML-2 polymer, but both were increased compared to pure siRNA. A negative correlation between polymer hydrophobicity and mucus penetration might be the reason for the superior uptake for the Lead compared to ML-2 polymer. Since the second optimization task of the algorithm was toxicity, the *in vivo* compatibility was investigated next. To exclude false positive results, polymers were tested for endotoxins and confirmed to be endotoxin free (Figure III.S17). BALF Cytokines showed partially higher levels after treatment with the Lead polymer than after administration of PEI polyplexes (Figure III.4 E and Figure III.S18). Treatment with the ML-2 polyplexes, on the other side, resulted in comparable cytokine levels as measured after administration of free siRNA or Buffer alone, indicating high biocompatibility. These findings were complemented by the tissue slices prepared from treated lungs, where only for the Lead polymer immune cell invasions were observed, whilst ML-2 was comparable to pure siRNA application (Figure III.4 F and Figure III.S19). These results show the successful improvement of safety and tolerability of the predicted PBAE. One reason could be the more stealth-like properties mediated through the higher hydrophobicity. Especially in macrophages and DCs, the uptake of ML2 was comparable to pure siRNA indicating an evasion of immune recognition, which can also be seen in the low levels of TNF-α, IL-6 and IL-27 (Figure III.4 E and F).

15 Conclusion

This study successfully demonstrated the efficiency of machine learning for extracting valuable insights from well-structured data, even with limited datasets. Furthermore, the successful synthesis of an optimized nanocarrier using predicted conditions validates the Nested-Leave-One-Out Cross Validation approach as a valuable tool for developing generalizable prediction for the selected feature space. Feature analysis also proved crucial for gaining deeper mechanisti However, the model's exclusive reliance on *in vitro* data resulted in predictions that did not fully translate to the complexities of *in vivo* environments. Therefore, future research incorporating *in vivo* data from the early stages of optimization is essential to develop more robust and clinically translatable predictive models, ultimately leading to improved therapeutic outcomes.

16 Data Availability

All experimental data and the Python code used are available upon request. The data used to fit and validate the Machine Learning models are shown in Figure III.S20.

17 Competing Interests

Olivia Merkel is a Co-founder of RNhale GmbH, Scientific Advisory Board Member of Coriolis Pharma, AMW, and Corden Pharma as well as a consultant for PARI Pharma, AbbVie Deutschland, and Boehringer-Ingelheim International. Adrian Kromer and Felix Sieber-Schäfer are consultants for AMW.

18 Acknowledgments

This work was supported by ERC-2022-COG-101088587 to O.M.M. and by the German Federal Ministry of Education and Research (BMBF) in the framework of the Cluster4Future program (Cluster for Nucleic Acid Therapeutics Munich, CNATM, Project ID: 03ZU1201AA). Graphics were partially generated using BioRender.

19 Supplementary Information

Table S1: adjusted from Zimmermann et al, doi: 10.1016/j.jconrel.2022.09.021. Sequences of siRNAs used in the study. Nt = nucleotides; GFP = green fluorescence protein; NC = negative control; GAPDH = housekeeping gene GAPDH; A = Adenine; C = Cytosine; G = Guanine; U = Uracil; T = Thymine; p = phosphate residue; lower case bold letters = 2´-deoxyribonucleotides; capital letters = ribonucleotides; underlined capital letters = 2´-O-methylribonucleotides.

Name	**Sense strand (5’-3’)**	**Antisense strand** (3’-5’)	**Length (nt)**	
			Sense	**Antisense**
siGFP	pACCCUGAAGUUCAUCUGCACCAC**cg**	ACUGGGACUUCAAGUAGACGUGGUGGC	**25**	**27**
siNC	pCGUUAAUCGCGUAUAAUACGCGU**at**	CAGCAAUUAGCGCAUAUUAUGCGCAUAp	**25**	**27**
siGAPDH	pGGUCGGAGUCAACGGAUUUGGUC**gt**	UUCCAGCCUCAGUUGCCUAAACCAGCA	**25**	**27**
siGAPDH (MM)	pAGCAUCUCCCUCACAAUUUCCAU**cc**]	ACUCGUAGAGGGAGUGUUAAAGGUAGG	**25**	**27**

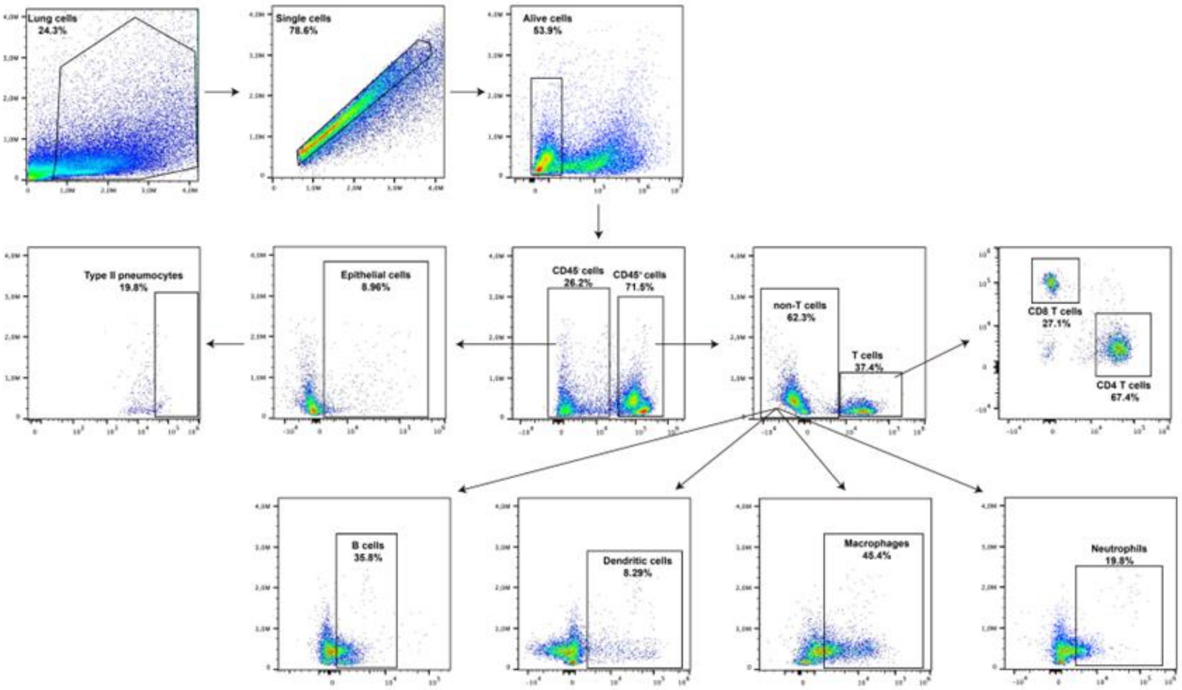
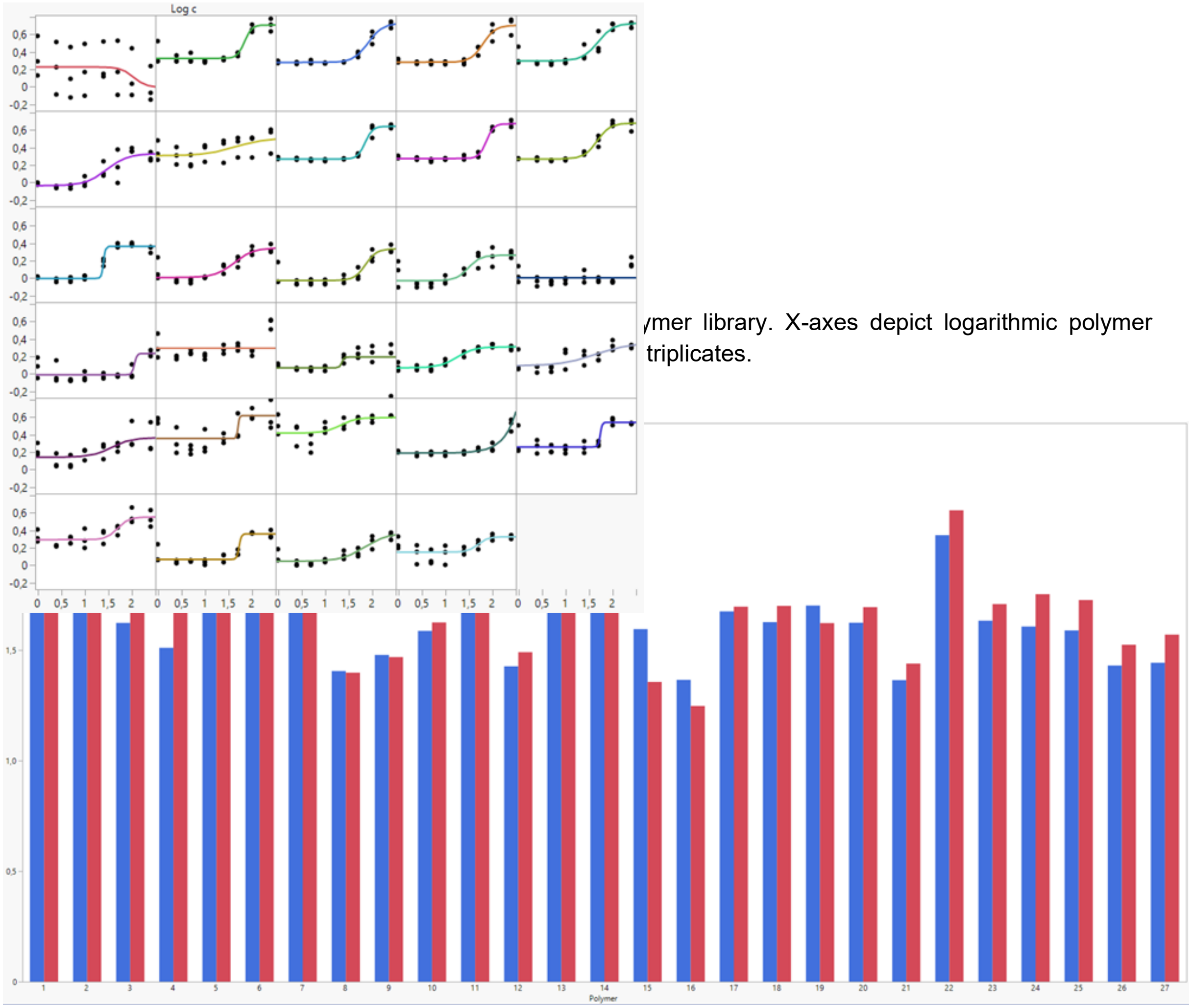


Figure III.S1: Gating strategy for flow cytometric analysis of single cell suspensions obtained from mouse lungs.



ymer library. X-axes depict logarithmic polymer
triplicates.

Figure III.S3: Comparison of CTB (blue) and LDH (red) IC50 values for polymer library.

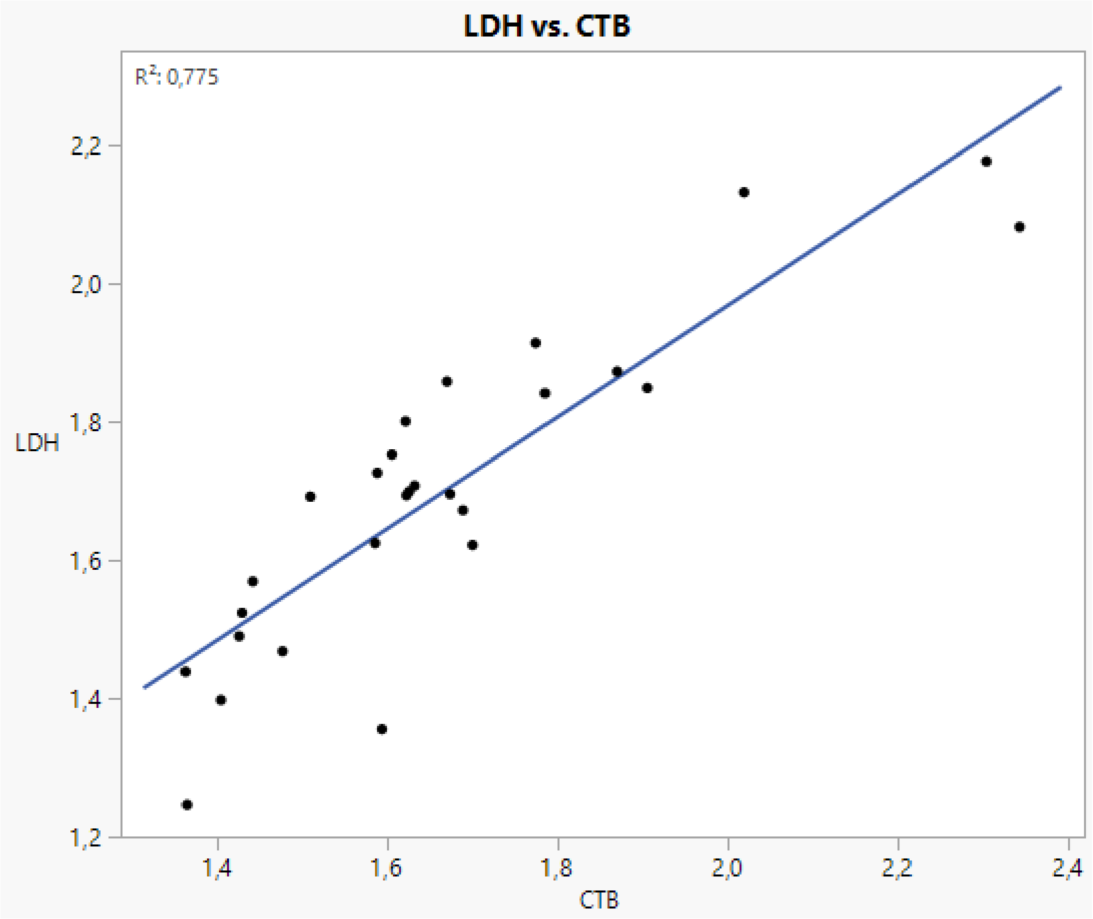


Figure III.S4: Correlation between IC50 values determined via LDH and CTB.

GeneExpression ZeroShot:

XGBRegressor(learning_rate=0.2, max_bin=None,
max_cat_threshold=None, max_cat_to_onehot=None, max_delta_step=4,
max_depth=4, max_leaves=None, min_child_weight=2.0, missing=nan,
monotone_constraints=None, n_estimators=100)

GeneExpression FewShot:

SVR(C=1, degree=4, epsilon=0.2, kernel='poly', shrinking=False)

IC50 ZeroShot:

RandomForestRegressor(ccp_alpha=0.005, criterion='absolute_error',
min_samples_leaf=4, oob_score=True)

IC50 FewShot:

RandomForestRegressor(ccp_alpha=0, criterion='absolute_error',
min_samples_leaf=4, min_samples_split=8, oob_score=True)

Figure III.S5: Model and Hyperparameter Settings after evaluation

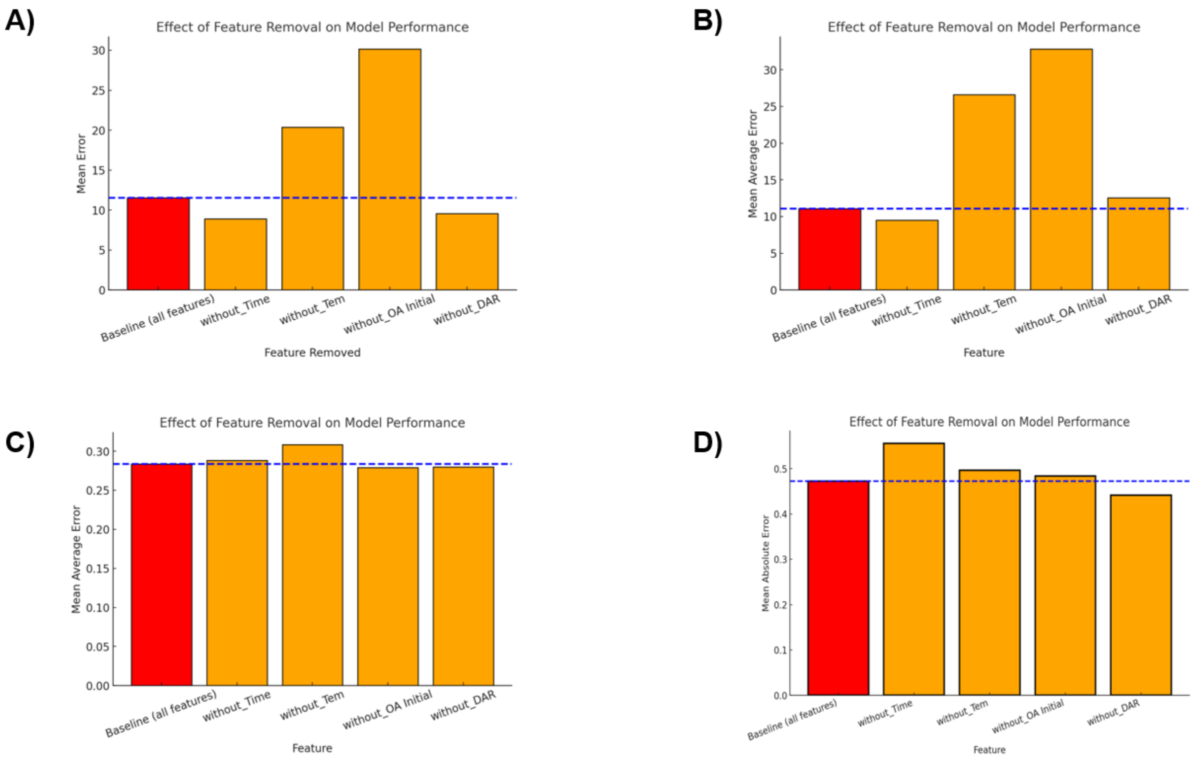


Figure III.S6: Feature ablation study for A) Zero-Shot Gene Expression B) Few-Shot Gene Expression C) Zero-Shot IC50 D) Few-Shot IC50.

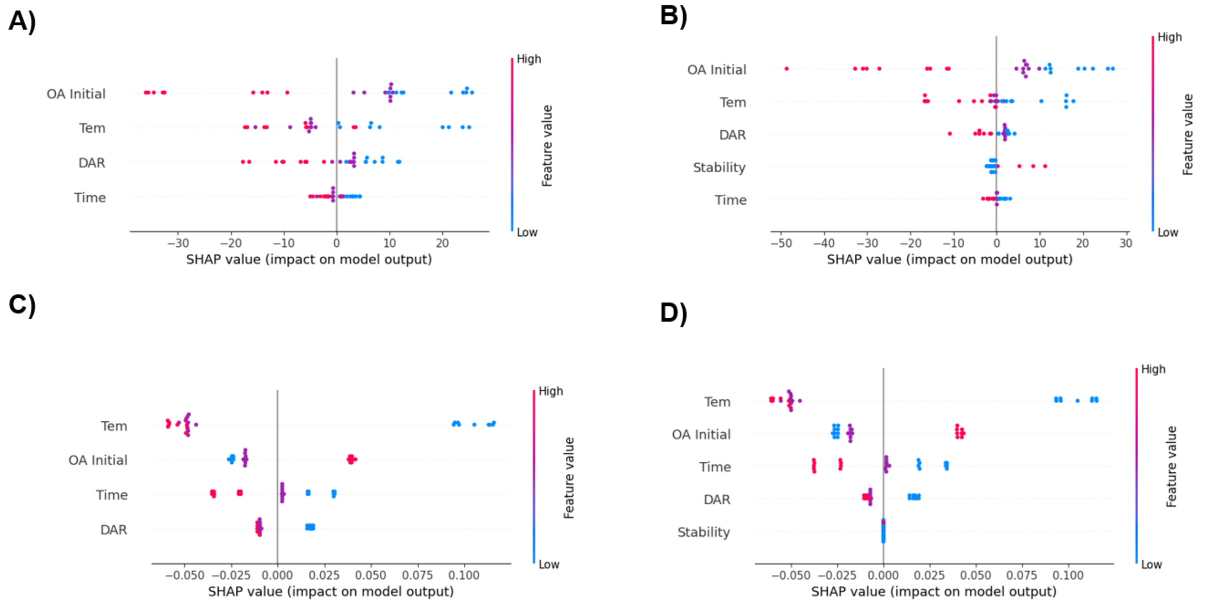


Figure III.S7: SHAP results with all features A) Zero-Shot Gene Expression B) Few-Shot Gene Expression C) Zero-Shot IC50 D) Few-Shot IC50.

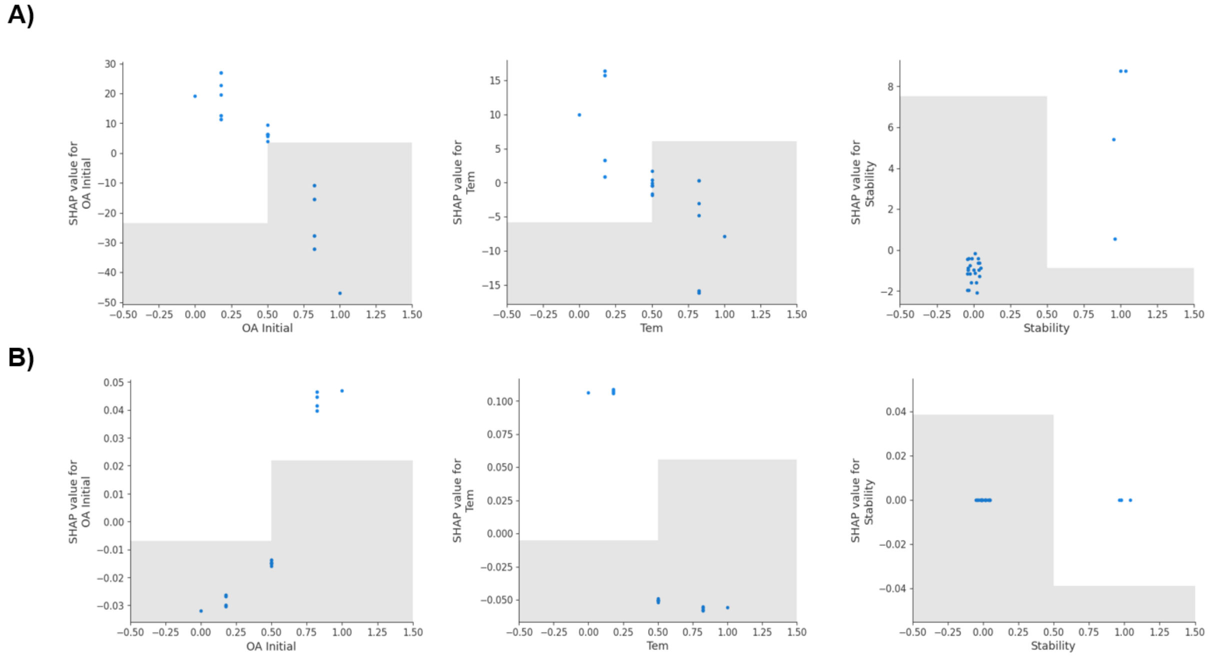


Figure III.S8: Scatter plots of SHAP values and used features after the feature ablation study for the Few-shot model for A) the Gene Expression Model and B) the IC50 Model.

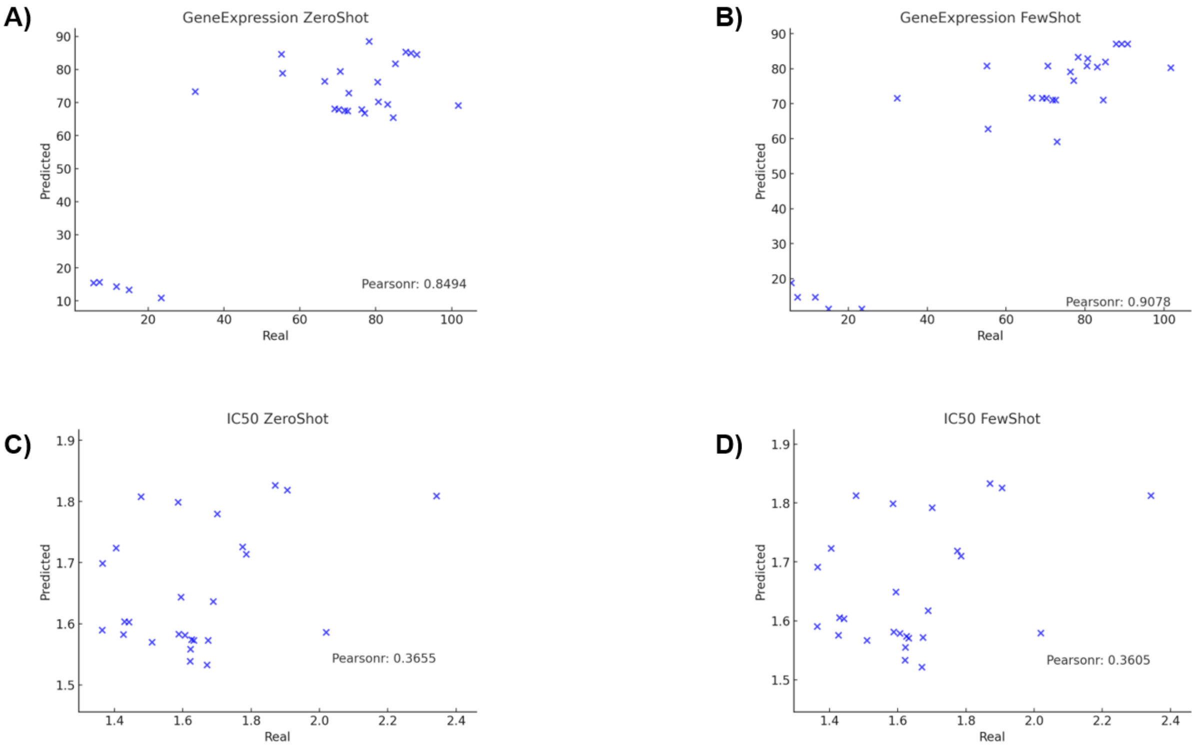


Figure III.S9: Predicted vs Real Scatter Plots for A) Zero-Shot Gene Expression B) Few-Shot Gene Expression C) Zero-Shot IC50 D) Few-Shot IC50.

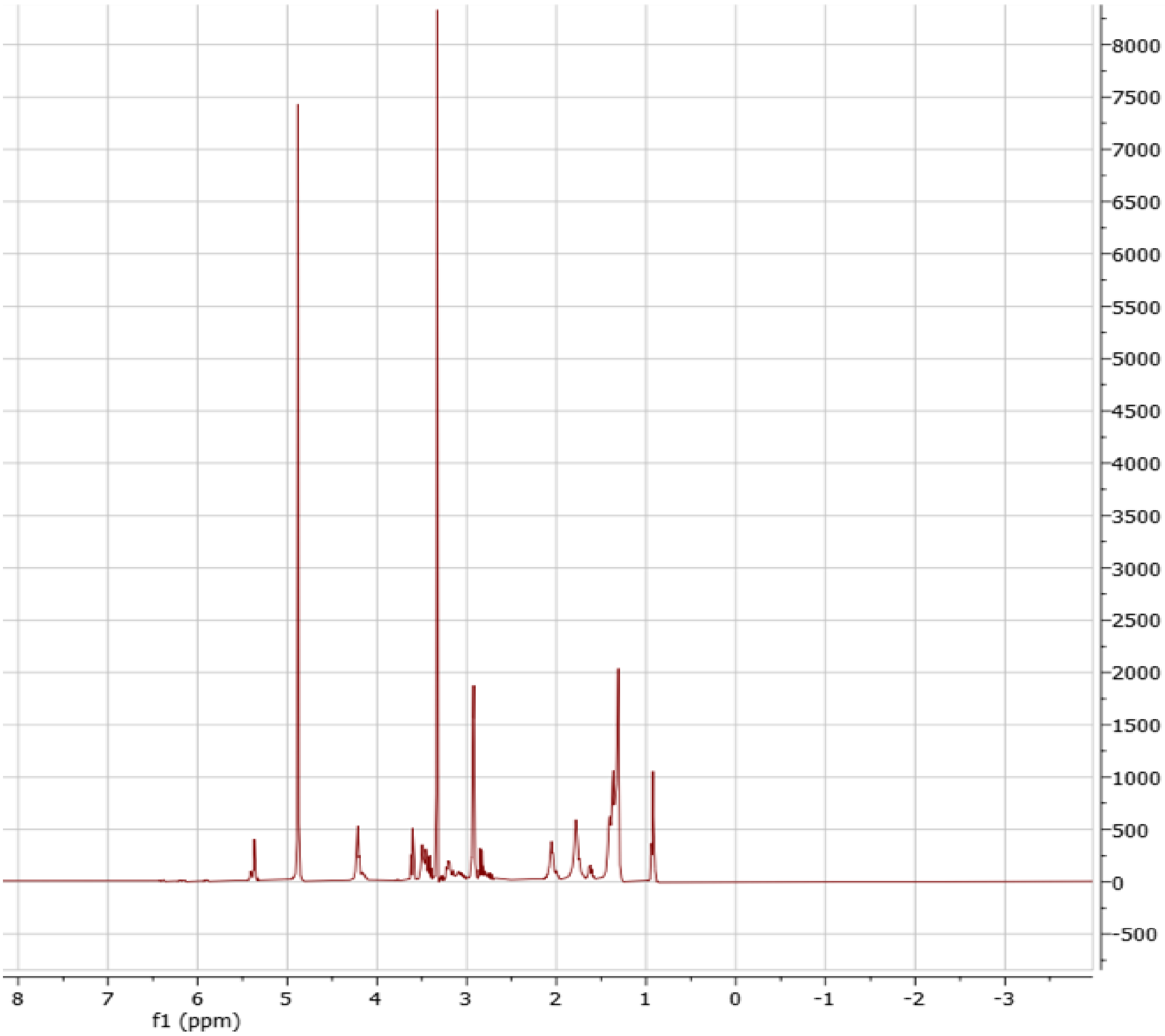


Figure III.S10: 1H-NMR of the ML-optimized polymer ML-2.

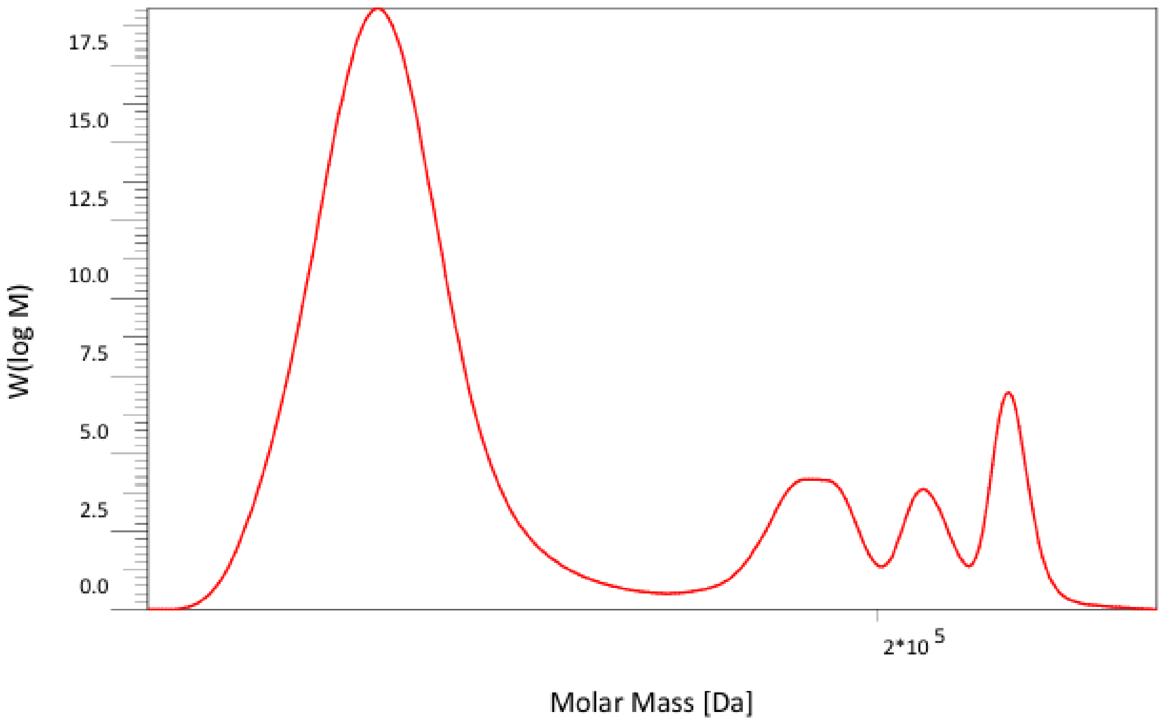


Figure III.S11: GPC measurement of the ML-optimized polymer ML-2.

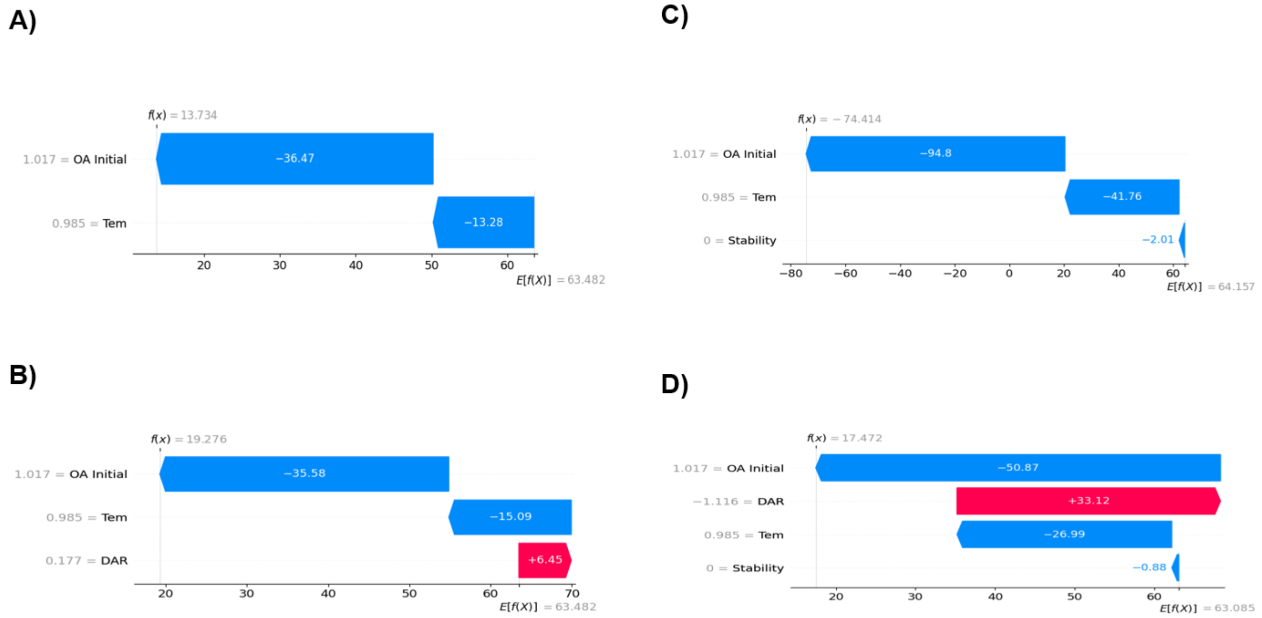


Figure III.S12: Single Point Prediction of optimized polymer for the Gene Expression Models with A) Zero-Shot after feature ablation B) Zero-Shot before feature ablation C) Few-Shot after feature-ablation D) Few-shot before feature ablation.

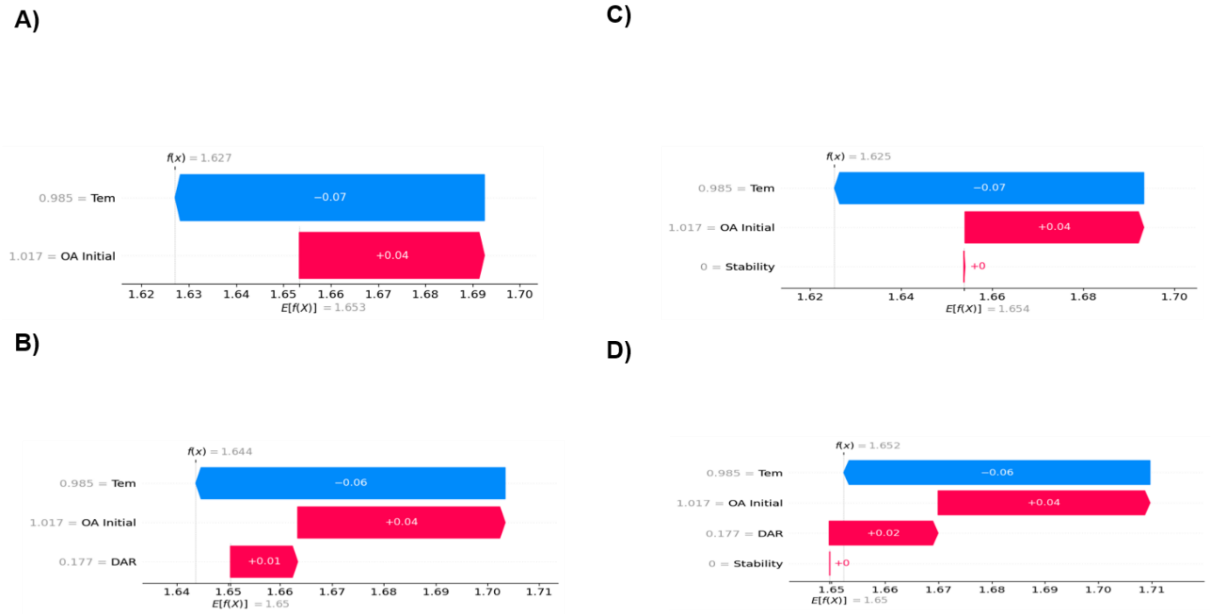


Figure III.S13: Single Point Prediction of optimized polymer for the IC50 Models with A) Zero-Shot after feature ablation B) Zero-Shot before feature ablation C) Few-Shot after feature-ablation D) Few-shot before feature ablation.

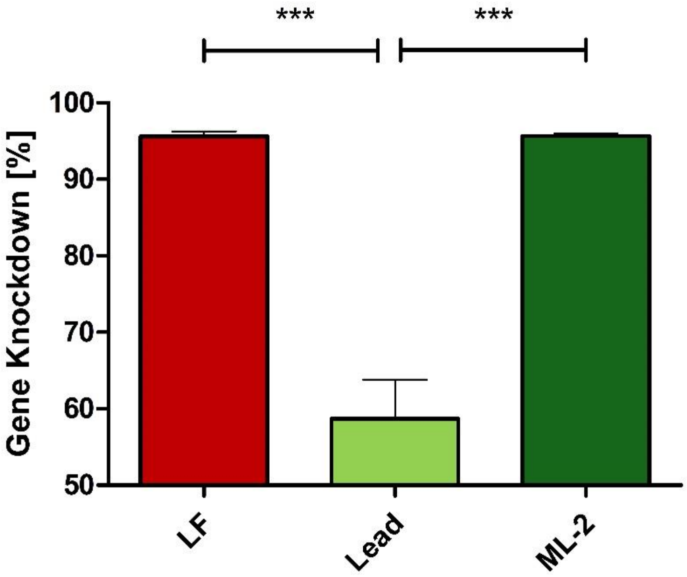


Figure III.S14: Gene Knockdown calculated from the median fluorescence intensity comparing H1200 eGFP cells treated with pure siGFP (for LF) or nanoparticles encapsulating siNC against siGFP with N=3 (*** depicting a p ≤ 0.001).

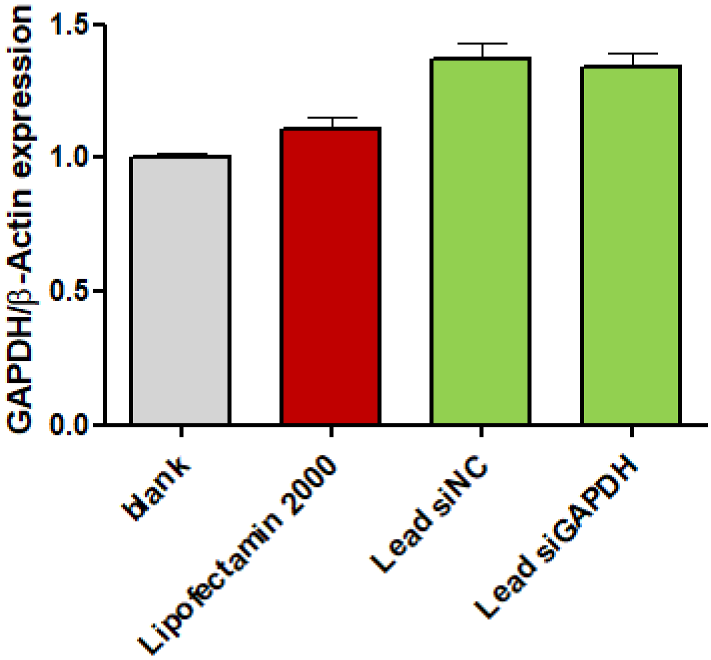


Figure III.S15: GAPDH gene expression determined via qPCR in air-liquid-interface-cultured CALU-3 cells[27] after treatment with Lipofectamine or Lead polymer encapsulating siNC or siGAPDH. No sequence-dependent significant difference was found (n=3).

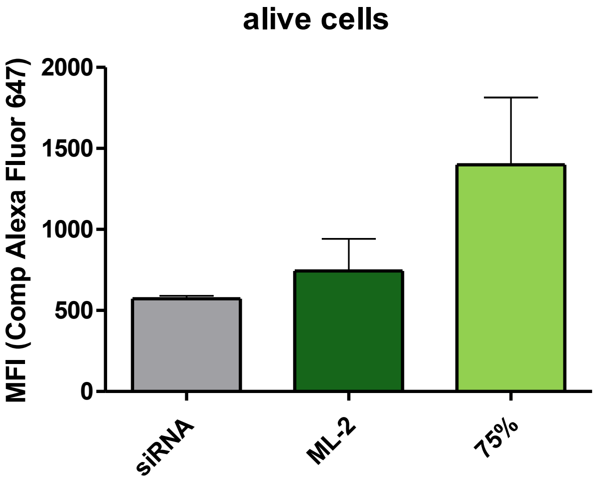


Figure III.S16: MFI of all alive cells measured from mouse lung single cell suspensions.

Endotoxin test using a LAL-reaction (S)

To ensure an endotoxin free synthesis product polymers were investigated using the Endosafe® Endochrome-K™ Kinetic Chromogenic (KCA) LAL Endotoxin Detection Reagent (Charles River, Sulzfeld, Germany). Briefly, A calibration curve was prepared from the kits reference sample in duplicates in a range from 0.05 to 5 IU/mL. Polymer samples of the lead candidate and ML-2 were prepared in two concentrations of 0.1 and 0.01 mg/mL in duplicates. One sample of each polymer concentration was spiked with endotoxin references to a final concentration of 0.5 I.U./mL, while the other sample was used without any further modification. To 100 mL of the respective samples, 100 µL of freshly resuspended LAL-reagent was added. After 5 minutes of incubation at 37°C, sample absorbance was measured with a plate reader at 374 nm (TECAN Spark, TECAN, Männedorf, Switzerland. At 37°C all samples were measured every 15 seconds at the same seconds for 30 minutes. No increase above an absorbance value of 1 after 30 minutes was interpretated as an Endotoxin Concentration below the LoD for the kit and stated as “Endotoxin-free”.

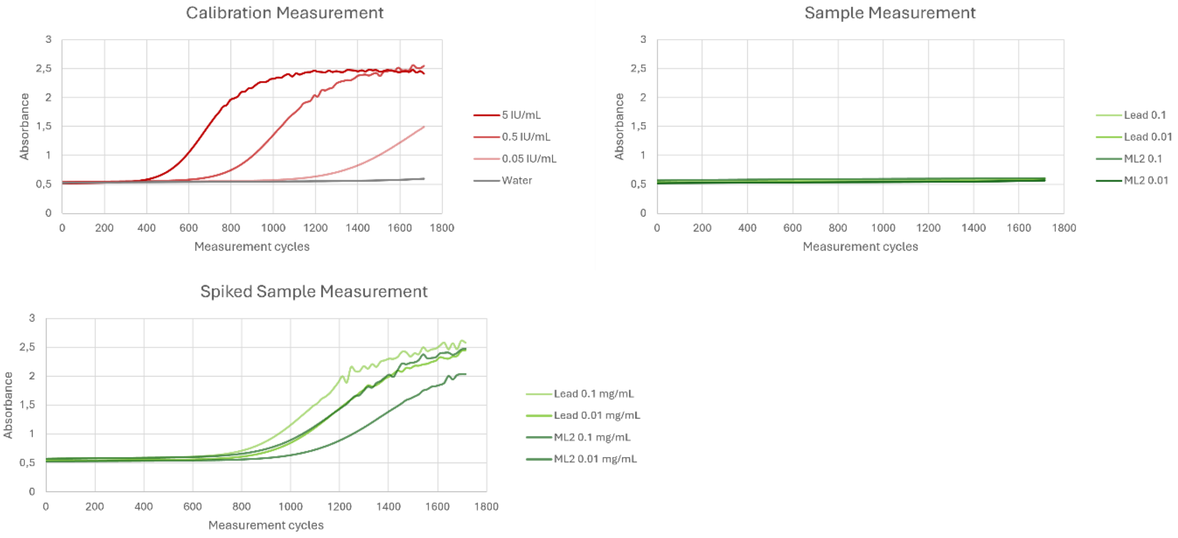


Figure III.S17: LAL Endotoxin Detection results showing the calibration measurement of pure endotoxin standards (left, top), samples spiked with 0.5 IU/mL endotoxin standard (left, bottom) and samples without any modification (right).

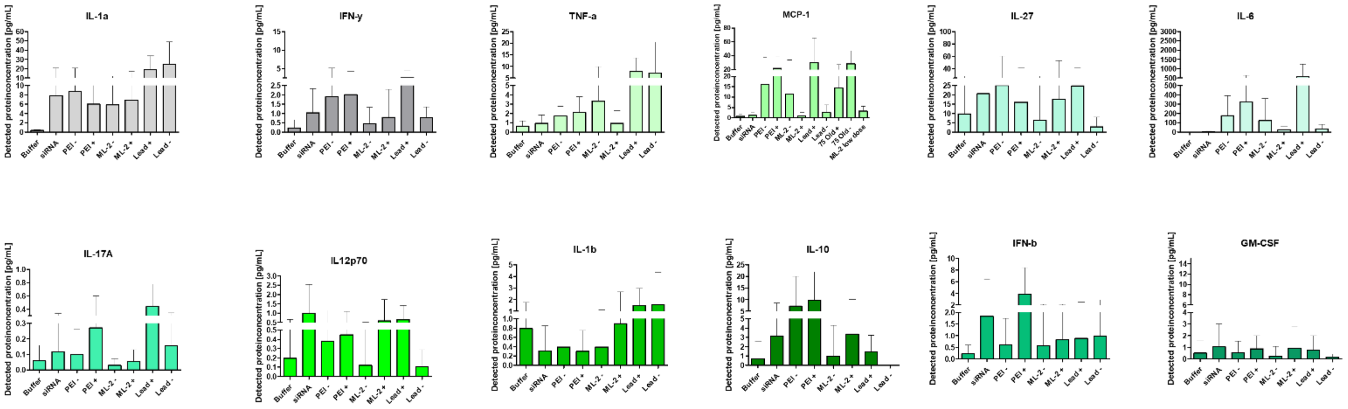


Figure III.S18: Cytokine quantification from BALF samples using the ELISA Inflammation Panel, reported as pg/mL.

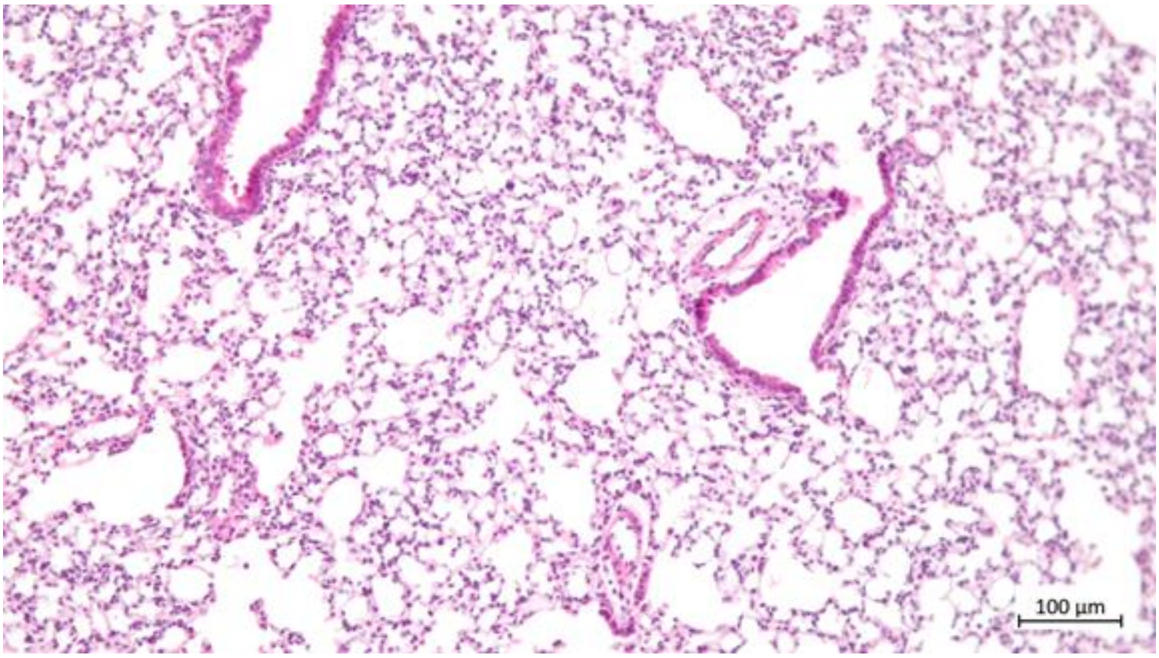
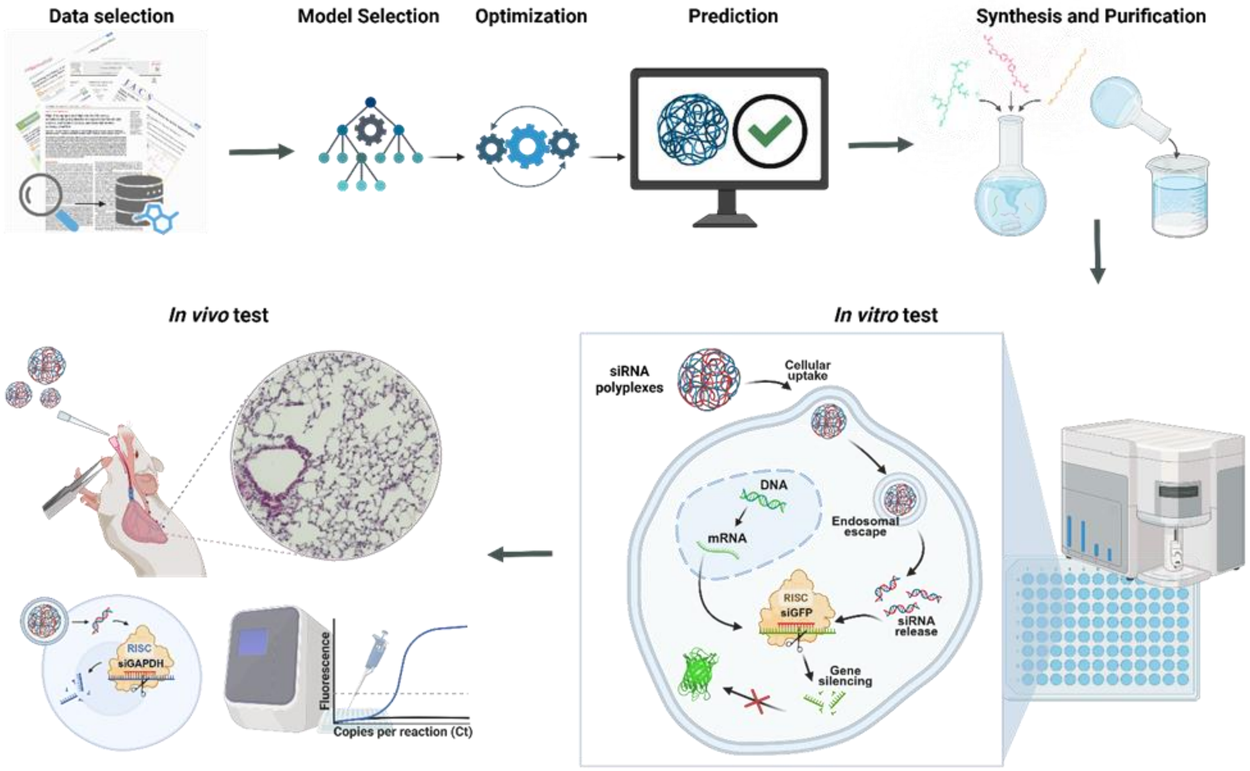


Figure III.S19: H&E staining of tissue slice obtained from a mouse lung treated with buffer containing only siRNA.**III.S**

Chapter IV - Machine Learning-Enabled Polymer Discovery for Enhanced Pulmonary siRNA Delivery

1 Graphical Abstract



2 Abstract

Nucleic acid therapeutics are poised to revolutionize the clinical treatment of diseases once considered undruggable. Although these therapeutic approaches hold significant promise, delivering the nucleic acid cargo remains challenging due to susceptibility to nuclease degradation. Among all carrier systems, polymers stand out for their high tunability and cost-effectiveness. However, their flexible structure greatly expands the chemical space, making experimental exploration both costly and time-consuming. Leveraging published data and machine learning methods provides a valuable strategy to address these issues. The present study demonstrates a way to merge data from multiple sources and use this information to identify a new polyester that effectively delivers siRNA into lung cells. The newly discovered polymer was further examined in ex vivo experiments and tested in a mouse model. The results indicate that a polymer capable of silencing specific genes in vivo can be discovered through machine learning, circumventing an extensive trial-and-error process in the search for novel materials.

Keywords: Machine learning; polymer discovery; siRNA delivery; pulmonary delivery; polymeric nanoparticle; polyester; Poly(beta)aminoester

3 Introduction

Therapeutic nucleic acids (NAs) are one of the most promising innovations in clinical research. A huge number of diseases that were previously considered undruggable, such as hypercholesterolemia[138] or Huntington's disease[139] can now be treated effectively through this groundbreaking approach to therapy. Since the discovery of NAs by Friedrich Miescher[140] in 1868, extensive research has been conducted aiming to translate this technology into actual medicines. It was in 1998, when the first NA-based drug, vitravene®, received approval by the FDA for the treatment of cytomegalovirus (CMV) retinitis. However, as of 2024, only 20 further applications have been approved[141]. One reason for the slow progress may be that NAs, and particularly ribonucleic acid (RNA) is unstable in the bloodstream and rapidly degraded by ubiquitous RNases. To circumvent this limitation, it became common practice to encapsulate RNA into carrier systems that protect the cargo from enzymatic degradation and help to guide the NAs to the desired tissue. In this context, lipid nanoparticles (LNPs) have become increasingly popular. As of today, three LNP-based RNA therapeutics have received FDA market approval, namely the SARS-CoV-2 vaccines Comirnaty, and Spikevax as well as Onpattro, a therapy for hereditary transthyretin-mediated (hATTR) amyloidosis. However, LNPs have been associated with certain concerns, including their potential to trigger inflammation[142], immunogenicity[143] and challenges with long-term storage[77]. Therefore, polymeric carrier systems have been proposed as an alternative to circumvent the limitations of lipid-based carriers[118]. Polymer materials are generally very flexible and can be modified with optimized chemical structures. This enables simplified adaptation as, in contrast to LNPs, only a single component needs to be adapted. To condense polyanionic RNA via electrostatic interactions, polymers need to contain protonable groups. In many cases, amines are introduced in the polymer structure to fulfil this role[144]. However, nanoparticles formulated with polycationic materials may cause safety issues if they are unable to be properly excreted, resulting in their accumulation within the body. Polycations remaining in the body can interact non-specifically with intracellular proteins and peptides, which may affect their functionality, and lead to cytotoxicity. This limitation can be addressed by introducing biodegradability into the polymer, which is very common in polyester structures. In addition to biodegradability, the inclusion of specific structural motifs, such as hydrophobic segments, is essential in enhancing the functional properties of the polymer. These hydrophobic motifs[145] are usually included in forms of amphiphilic block copolymers[80,96] to shield excessive electrostatic interactions. Additionally, these additions supplement the base polymer with hydrophobic properties for interactions with RNA and biological membranes, which support better performance in cellular uptake, endosomal escape and many more.

Unfortunately, understanding the exact structure activity relationship between block copolymers and successful delivery of cargo is highly complex and far from trivial. This complexity is amplified by the thousands of potential variations in polymer architecture, composition, and environmental interactions, as well as the fact that synthesizing these polymers is both time-intensive and requires significant material resources, adding to the challenge of systematic exploration. Yet, it is exactly this understanding that is necessary to design new high performing and safe carrier systems. In recent years, the development and application of artificial intelligence algorithms have significantly increased. These algorithms might help to uncover the underlying patterns differentiating successful from unsuccessful block copolymers and facilitate the virtual screening of potential candidates before synthesis. Machine learning (ML) models that could be used to make this possible, are highly data driven and therefore dependent on available experimental data. While ML is already broadly used for polymeric property predictions such as Tg[146] or dielectric constant[147], not much work is published on using ML models for the design of new amphiphilic polymeric nanocarriers. Pioneering work in this field was conducted by the groups of Green[128] and Reineke[127]. Both used high-throughput synthesis and screening methods to collect data and make predictions for unseen combinations. The need for the availability of high throughput screening opportunities is however limiting the wider use of these approaches. Furthermore, the authors relied solely on machine learning applied to a single type of polymer, which inherently limits the exploration of the broader chemical space and restricts the potential to uncover diverse structure-activity relationships.

Here, we show how the discovery of new polymeric nanocarriers can be guided with a prediction model trained on literature data for different kinds of polyesters. In this work, we emphasize pulmonary siRNA delivery to the lungs as a demonstration of our approach, while noting that it could equally be applied to other therapeutic cargo and targets, following a similar strategy. We collected >600 different polyester structures used for siRNA delivery from previous publications and trained multiple ML models with the corresponding gene silencing data. To obtain insights into polymeric siRNA delivery, we investigated key factors that drive successful delivery of cargo. Our lead model was then used to synthesize a novel amphiphilic polymer, which was subsequently tested for its performance of delivering siRNA. Starting with *in vitro* testing we progressively increased biological complexity by evaluating the polymer in an air-liquid-interface model followed by *ex vivo* human Precision-Cut-Lung-Slices (hPCLS). These models reflect critical challenges in pulmonary RNA delivery, including RNase activity, the mucus barrier and tight junctions in respiratory epithelium. Finally, we evaluated the polymer’s safety for pulmonary administration and its ability to facilitate gene knockdown in an *in vivo* murine model.

Our approach offers an easy-to-use method for designing new nanocarriers by utilizing historical data. Additionally, we demonstrate how data from a broader chemical space can be used to identify polymeric properties essential for successful delivery. To the best of our knowledge, we are the first to synthesize an amphiphilic polymer for siRNA delivery using ML, thereby contributing to a deeper understanding of RNA delivery via polymeric nanocarriers.

4 Results and Discussion

4.1 Generalizable Machine Learning Framework

A primary goal of this study is to empower researchers lacking HTS capabilities to employ ML on existing literature data. Our methodology achieves this by systematically integrating information from diverse sources into a unified dataset. However, compiling data from literature presents an inherent challenge: balancing the scope of chemical diversity. On the one hand, sufficient diversity is desirable for training models that yield generalizable insights into structure-property relationships. On the other hand, literature datasets are often sparse compared to HTS data. Including systems with widely divergent chemical structures or fundamentally different delivery mechanisms introduces significant noise. With limited data points, this can easily overwhelm the underlying patterns related to a specific delivery strategy, preventing the ML model from effectively learning the relevant mechanisms. Therefore, our approach necessitates carefully constraining the literature search to a 'comparable chemical space'—focusing on systems sharing core structural similarities and presumed mechanisms. This focused scope enhances the signal-to-noise ratio, enabling the model to identify meaningful correlations even from limited data. We illustrate this methodology using a curated dataset of amphiphilic polyester structures, representing a class with comparable underlying chemistry.

Converting molecular structures into a format readable for a ML algorithm is a prerequisite for ML applications in the chemical space, and several methodologies have been proposed.[65,148,149]. Commonly employed fingerprints or SMILES rely on purely structural information, limiting their use for a generalization as required here. This limitation can be overcome using representation as molecular graphs or molecular descriptors[150]. Unfortunately however, using descriptors alone also does not necessarily lead to a good generalization since high dimensional representations are prone to overfitting[151]. Thus, we used a Tree-Based feature reduction to eliminate descriptors that did not contribute to the overall prediction of the model. To ensure valid representations of polymeric data, each of the polymer building blocks (hydrophilic, hydrophobic, endcapping), was separately encoded, and the ratio information was embedded by multiplying each descriptor with this ratio factor. The molecular weight and the cell type used in the original dataset were added to the sample. The latter was achieved using one-hot encoding, a method that converts categorical features into binary vectors, enabling their representation in machine learning models. To minimize the noise that is introduced by the experimental data and especially by merging datasets of different origin, we decided to use a binary binning approach to turn the regression problem, using the reported gene silencing percentages, into a classification problem. We selected a gene knockdown efficiency of 50% as threshold to separate the formulations into two different classes, reflecting our primary goal of assessing whether synthesizing a polymer is worthwhile rather than focusing on exact gene silencing values.

Utilizing binary classification generally enhances interpretability, simplifies the analysis and effectively addresses data imbalance.

Using the prepared dataset, we first compared different ML algorithms (Figure IV.S1A). To address the imbalance in the dataset, balanced accuracy/mean recall was used to handle potential model biases towards the major class and a RandomOverSampler was used to guarantee balanced training. The data was split into a train/test set at a ratio of 80:20 and 100 models were trained using each algorithm. The LGBMClassifier[50] showed the best performance (0.8217 balanced accuracy) and was therefore selected for further optimization. We then compared different resampling strategies (Figure IV.S2), with SMOTEEN[152] showing the best balanced accuracy (0.8309). After tuning using hyperopt (Figure IV.S3), additional feature reduction was performed, where eleven features lead to the best model performance (Figure IV.1A). This process was visualized using UMAP, revealing how feature reduction minimized gaps in the chemical properties space (Figure IV.1B). This approach was aimed to reduce the risk of overfitting while limiting the physicochemical information required to encode molecular structures. This ultimately facilitated the integration of different datasets and the generalization of unseen structures. The eleven most important features, using SHAP are shown in Figure IV.S4. The tuned LGBMClassifier was finally evaluated using 100 stratified train-test splits of 80/20 and showed a mean balanced accuracy of 0.8462 on the validation sets (Figures IV.1C and IV.1D). Afterwards, the model was trained on the entire dataset and used for the prediction task. The full workflow is also visualized in Figure IV.S5.

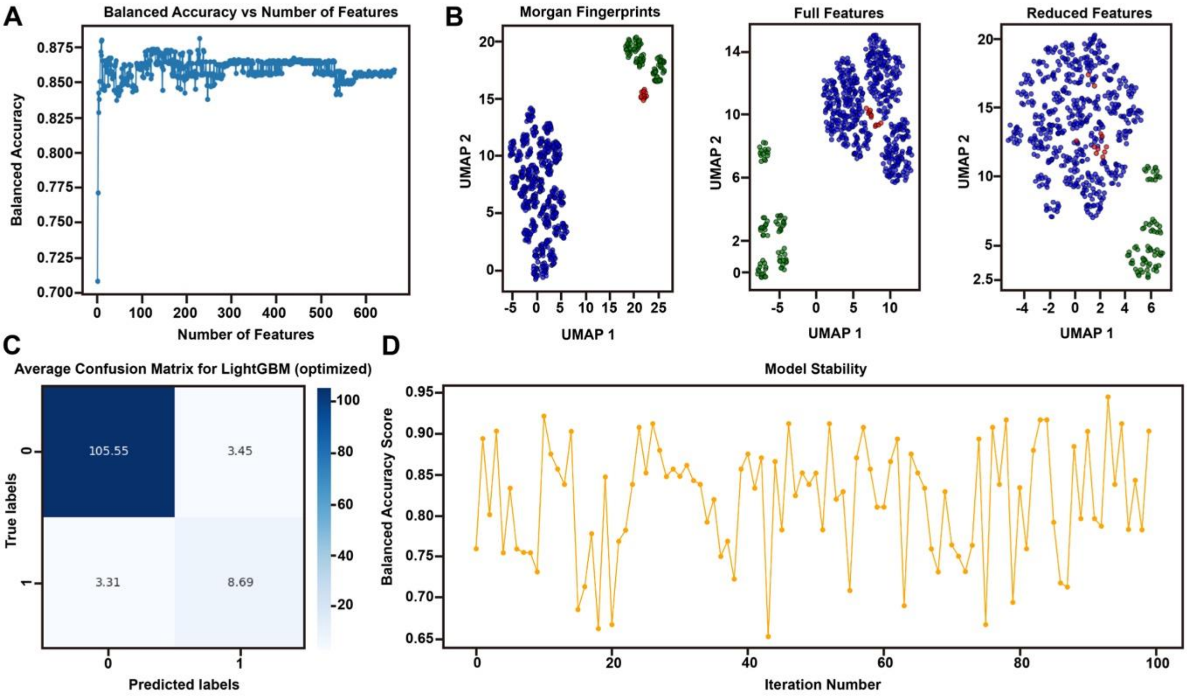


Figure IV.1: Insights into the ML Process A) Overview of the iterative feature reduction using LightGMBC feature reduction. B) Comparison of feature space visualization with UMAP using featurization with MorganFingerprints (left), all RDkitDescriptors and self-defined features (middle), reduced feature set of the eleven most important features (right). C) Confusion matrix of the fully optimized LightGBMC. Evaluated on the training set of 100 train-test splits and averaged accordingly. D) Evaluation of Model stability over 100 train-test splits ranging from 0.65 to 0.95 balanced accuracy.

4.2 Experimental Validation

To experimentally validate the trained classifier, novel polymers were rationally designed from available precursors via established synthetic routes. Given our group's significant expertise in synthesizing and characterizing poly(beta-amino ester)s (PBAEs), this class of polymers was selected as the focus for the validation set. To the best of our knowledge, all selected polymers are unpublished structures. The classifier predicted their potential knockdown efficiency. Based on these predictions, three polymers expected to exhibit low efficiency and three expected to exhibit high efficiency were selected for chemical synthesis and subsequent *in vitro* evaluation. Their schematic structures were shown in Figure IV.S6, with specific chemical structures provided in Figures IV.S7-S12. siRNA was formulated with these polymers at an N/P ratio of 10, and polyplexes were characterized regarding size, polydispersity and zeta potential, as presented in Figure IV.S13. Gene silencing efficiency was assessed in both enhanced green fluorescent protein (eGFP)-stably expressing H1299 cells (using siRNA targeting eGFP) and A549 cells (using siRNA targeting epidermal growth factor receptor (EGFR)). Consistent with the predictions, all three polymers anticipated to have low efficiency demonstrated negligible knockdown efficiencies (Figure IV.S14). However, the polymer OA-BG, comprising full oleylamine (OA) modification with bisphenol A glycerolate (BG) as its backbone, predicted as a high-efficiency candidate, failed to achieve the 50% knockdown threshold, reaching only 31.88% eGFP knockdown in H1299-eGFP cells and 24.62% eGFR knockdown in A549 cells. These results represented approximately 30% of the knockdown efficiency achieved by Lipofectamine 2000 and thus OA-BG was considered a false positive. In contrast, the other two polymers predicted to be high-performing, SP/TDA-BG (spermine/tetradecylamine with the BG backbone) and SP/OA-BG (spermine/oleylamine with the BG backbone), successfully demonstrated the predicted high knockdown efficiencies (91.82% and 96.17% eGFP knockdown, respectively). Overall, five out of six polymers were correctly classified, resulting in an experimental validation accuracy of 0.8333, which closely aligns with the classifier's estimated performance metric of 0.8462 (Section 2.1).

4.3 Characterization of Polymer and siRNA-loaded Polyplexes

Following the experimental validation in Section 2.2, among the polymers tested, SP/TDA-BG demonstrated high transfection efficiency, in agreement with the classifier's prediction. Given its promising performance, we selected SP/TDA-BG as a model polymer for further systematic investigation into the relationship between its structural characteristics and biological activity. Although the machine learning model specifically suggested a 50:50 SP:TDA ratio as optimal, inspired by the transfection cliffs theory[153], we sought to investigate how minor deviations from this composition might impact transfection performance, as such effects are not necessarily captured by the machine learning model[154]. Hence, we synthesized the corresponding PBAE polymers following the synthetic procedure shown in Figure IV.2A, adjusting the molar ratios of cationic monomer spermine and lipophilic monomer tetradecylamine from 40% to 60%, which were further confirmed by ^{1}H NMR analysis (Figure IV.S15). In addition, referring to our previous work on efficient siRNA delivery via amphiphilic PBAEs incorporating SP and OA with 1,4-butanediol diacrylate as the backbone[99,155], we also selected PBAE SP0.3/OA0.7 as a benchmark for comparative evaluation in our study.

The polymers were then complexed with siRNA at different N/P ratios. It is worth noting that, in the used dataset, N/P ratios were always set to at least 15 to ensure complete siRNA encapsulation and corresponding effectiveness. However, in our experimental work, we aimed to minimize polymer use, to particularly improve *in vivo* tolerability, based on our previous studies confirming efficient gene silencing and encapsulation at lower N/P ratios[100,155]. Therefore, we initiated screening from an N/P ratio of 3, increasing up to 10. Specifically, we assessed the physicochemical properties of the formed polyplexes, including size, size distribution and zeta potential. Most polyplexes formed with diameters ranging from 50 to 300 nm and acceptable PDI values around 0.2 (Figure IV.2B). Examining the zeta potential, a significant change was observed between N/P ratios of 3 and 5, particularly in case of polyplexes prepared with PBAEs SP0.5/TDA0.5 and SP0.4/TDA0.6, which displayed noticeable charge reversal (Figure IV.2C). Incomplete or unstable encapsulation of siRNA at N/P 3 (Figure IV.S17) could explain this observation. This near-neutral flipping zeta potential also revealed colloidal instability as evidenced by the extremely large size exceeding 2000 nm in case of polyplexes prepared with PBAE SP0.3/OA0.7 at N/P 5. When the ratio was increased to N/P 7 and N/P 10, the siRNA was completely encapsulated and the polyplexes were more stable in size.

Although stable formation of polyplexes is important for siRNA delivery, appropriate siRNA release is equally critical for successful gene silencing as the final action site will be in the cytoplasm, where the released siRNA cargo from polyplexes should bind with the RNA-induced silencing complex (RISC) to fulfill its function. Therefore, we investigated siRNA release from polyplexes in the presence of Triton X and heparin, which will competitively interfere hydrophobic and electrostatic interactions, respectively. After a non-linear fitting of released siRNA to the added interferents, EC50 values revealed that the release of equal amounts of siRNA from the polyplexes required higher concentrations of Triton X and heparin (6.2% vs. 5.2%) when the spermine ratio in the polymers increased from 40% to 60% (Figure IV.2D). The EC50 value for SP0.3/OA0.7 polyplexes was even higher (12.1%), demonstrating the tightest binding between siRNA and the polymer in our study. The binding strength effectively protected siRNA from degradation by RNase, as all formulations retained more than 90% siRNA content after incubation with the enzyme. In contrast, free siRNA lost 99% of its integrity when treated with the same amount of RNase (Figure IV.2E).

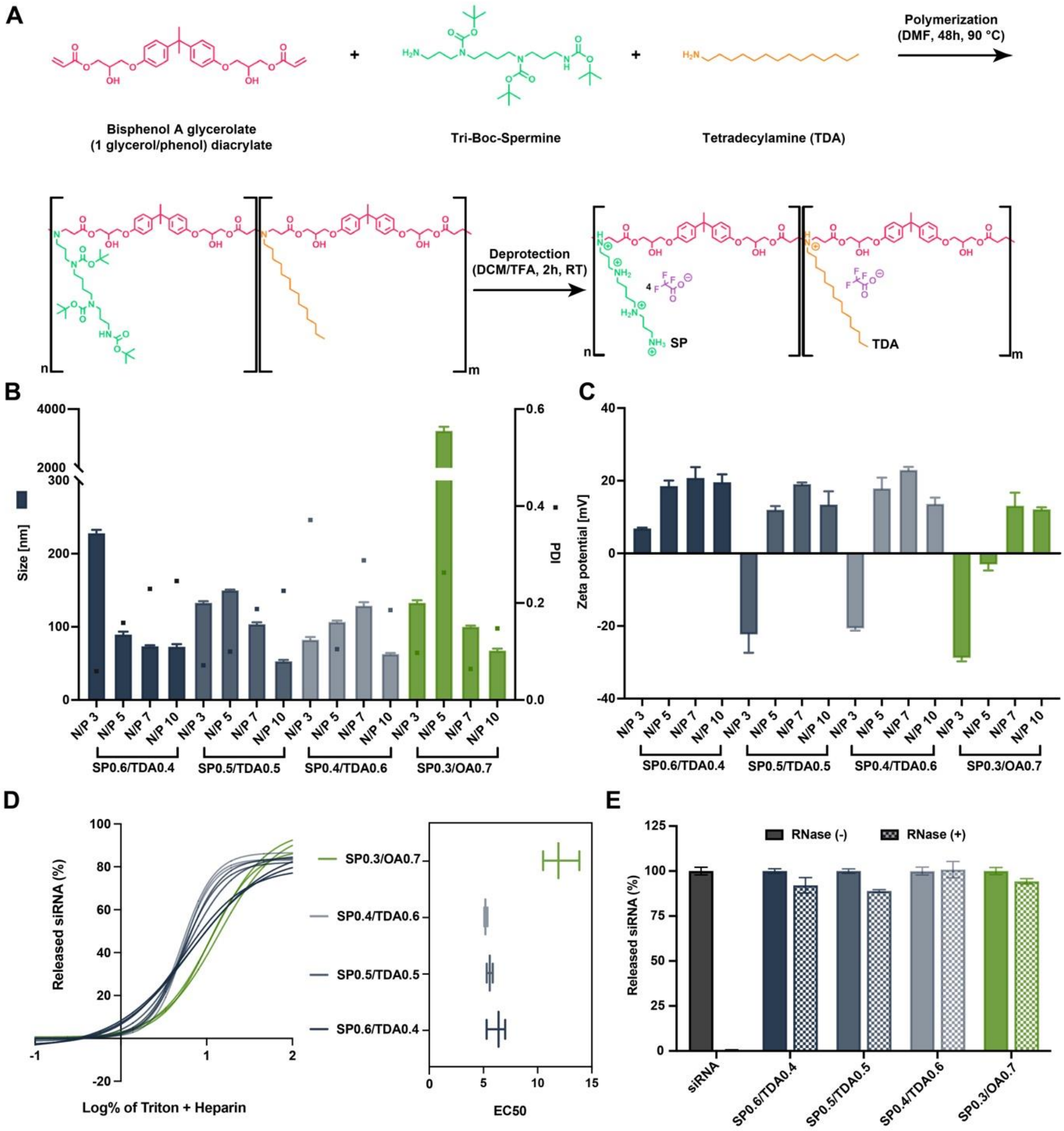


Figure IV.2: Characterization of synthetic polymers and siRNA-loaded polyplexes. (A) Synthesis procedure and the structure of SP/TDA-BG PBAE polymers. (B) Hydrodynamic diameter (represented by bar graph), polydispersity (represented by symbol), and (C) zeta potential of siRNA-loaded polyplexes prepared at different N/P ratios. (D) siRNA release from polyplexes at N/P ratio of 10 in the presence of Triton X and heparin using SYBR Gold assay, and EC50 values obtained by non-linear fitting analysis of released siRNA to added interfering substances. (E) RNase protection assay of polyplexes prepared at an N/P ratio of 10. Polyplexes were firstly treated with RNase at 37°C for 30 min, followed by RNase deactivation by heating to 70°C for 30 min. After incubation with Triton X and heparin, released siRNA was quantified using SYBR Gold assay. Results are presented as mean ± SD, n=3.

4.4 *In Vitro* Performances: Cytotoxicity, Cell uptake and Knockdown Effects

We initially evaluated the safety profile of our polyplexes by assessing the viability of H1299 cells exposed to increasing polymer concentrations. The cell counting kit (CCK-8) assay showed a dose-dependent trend in cell viability. Notably, even at the highest N/P ratio of 20, cell viability remained above 80%. When the N/P ratio was reduced to 10, the viability of H1299 cells consistently reached 90-95% in all groups (Figure IV.3A). Therefore, all following experiments were conducted at an N/P ratio of 10 or lower. Next, we performed a wider uptake screening of polyplexes formulated from N/P 3 to N/P 10 in H1299 cells. With increasing N/P ratio, the uptake of all polyplex formulations was improved (Figure IV.3B). Quenching the fluorescent signal on the cell surface with trypan blue, only resulted in a negligible decrease in the detected mean fluorescence intensity (MFI), indicating internalization of the polyplexes rather than non-specific adsorption on the surface. Furthermore, the knockdown effects of enhanced green fluorescent protein (eGFP) in H1299 cells stably expressing eGFP were consistently exceeding 94% in all polyplexes formulated at N/P ratios > 3 (Figure IV.3C).

The uptake of polyplexes at N/P 10 in A549 cells mirrored the trends observed in H1299 cells, with reduced uptake observed when either SP or TDA proportions exceeded 60% (Figure IV.3D). This aligns with the mechanism of adsorptive endocytosis which is generally associated with polyplex uptake[156]. For highly hydrophilic cationic polymers such as poly(ethyleneimine) (PEI) and poly(L-lysine) (PLL), uptake primarily relies on electrostatic interaction with cell membrane[157,158]. Hydrophobic modifications, however, have been shown to enhance uptake through interactions with lipids and membrane proteins[159,160]. Similarly, Rui et al. reported that increasing PBAE hydrophobicity initially boosted uptake before declining, regardless of whether delivering siRNA, mRNA or DNA[80]. In our study, PBAE SP0.5/TDA0.5 polyplexes achieved the highest uptake, with an MFI > 80,000. This indicates that a balance of electrostatic and hydrophobic interactions is crucial for optimal delivery.

Importantly, improved cellular uptake does not always correlate with stronger transfection. Although siRNA-loaded PBAE SP0.3/OA0.7 polyplexes showed superior internalization in A549 cells, transfection efficiency was lower than expected and inferior to the performance observed in H1299/eGFP cells (Figure IV.3E). This discrepancy may be attributed to the differences in siRNA lengths used for targeting eGFP (52 nucleotides) and EGFR (42 nucleotides) or differences in cell-type specific intracellular processing. Meanwhile, the slower release of siRNA that we observed in SP0.3/OA0.7 polyplexes may be another reason (Figure IV.2D). Notably, despite lower uptake of PBAE SP0.4/TDA0.6 polyplexes, their knockdown efficiency (53.4%), was comparable to SP0.5/TDA0.5 (51.2%). This observation might be explained by the efficient endosomal escape, which we investigated utilizing the Galectin-8 (Gal8) assay[161]. In brief, Gal8 binds glycans exposed upon endosomal membrane disruption, enabling quantification of endosomal escape using Gal8-mRuby-expressing cells[31,80]. The average number of Gal8-mRuby3 punctate fluorescent spots increased from 5.43 to 16.25 per cell as the lipophilic TDA content was increased from 40% to 60% (Figure IV.3F). This finding underscored that lipophilic components enhanced hydrophobic interactions with membranes, leading to structural instability of the membrane and disruption[162]. As a result of this disruption, polyplexes were able to escape the endosome, releasing siRNA into the cytoplasm to bind RISC, cleave target mRNA, and achieve effective knockdown of protein translation.

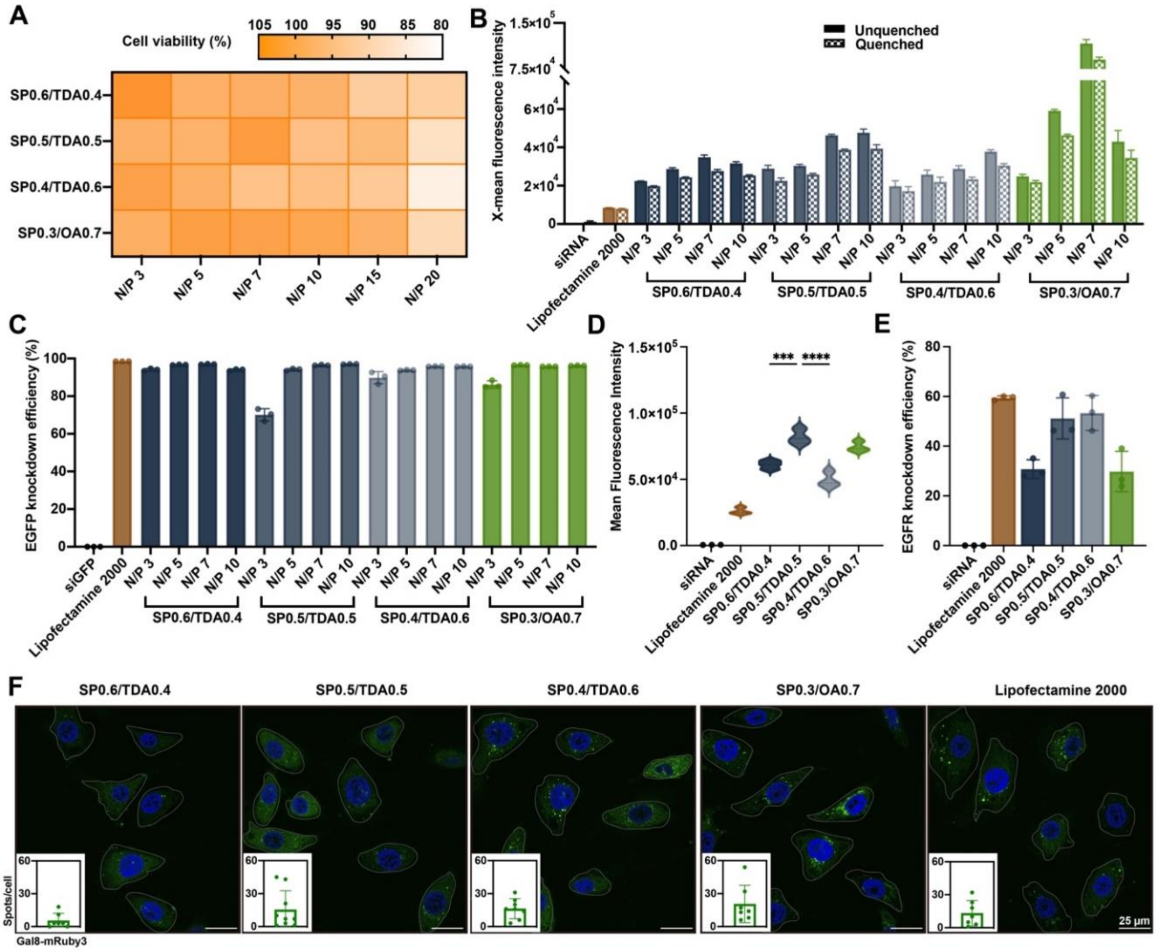


Figure IV.3: *In Vitro* performances of siRNA-loaded polyplexes. (A) Viability of H1299 cells after treatment with polyplexes formulated at different N/P ratios. (B) Cellular uptake of polyplexes containing Alexa Fluor 647-labeled siRNA in H1299 cells, presented as mean fluorescence intensity. (C) EGFP knockdown efficiency of polyplexes in H1299/eGFP cells. (D) Cellular uptake of polyplexes containing Alexa Fluor 647-labeled siRNA in A549 cells. (Data are presented as mean ± SD, *n* = 3; ***$p < 0.001$, ****$p < 0.0001$, one-way ANOVA) (E) EGFR knockdown efficiency of polyplexes in A549 cells. (F) Fluorescent spots of Gal8-mRuby3 in genetically modified Hela cells after 4 h of treatment with different polyplexes. Green color represents Gal8-mRuby3, nuclei are shown in blue. Scale bar, 25 µm. Quantification of Gal8-mRuby3 dots was performed by the Fuji plug-in of Image J, and data are presented as mean ± SD.

4.5 Mucus Penetration on ALI model and Gene Silencing in hPCLS

For pulmonary delivery, the mucus layer on the surface of the respiratory tract poses a significant barrier to effective siRNA delivery[163,164]. If an RNA-carrier interacts excessively with mucus, it will not be able to penetrate this barrier during the time of mucus turnover, leading to its clearance from the lung before cellular internalization. Additionally, tight junctions between respiratory epithelial cells further act as another barrier to paracellular transport of siRNA[165,166]. To evaluate the ability of our PBAEs-siRNA formulations to overcome these lung-specific barriers, we used an air-liquid interface (ALI) culture of Calu-3 cells. Under ALI conditions, Calu-3 cells differentiate into a pseudostratified epithelium, produce mucus and cilia-like microvilli, and thus closely mimic the *in vivo* respiratory tract environment[167,168]. As shown in Figure IV.4A, we obtained images by laser confocal laser scanning microscopy (CLSM), labelling mucus (green), cell nuclei (blue) and siRNA (red). Importantly, the mucus was largely distributed above the nuclei in all samples, confirming the successful establishment of a cell monolayer with mucus on the air-exposed side. When treated with free siRNA, signals from the siRNA were barely detectable. In the Lipofectamine 2000 control group, a very weak red signal was observed across the mucus layer toward the cell layer. For the ALI cells treated with PBAE SP0.3/OA0.7 polyplexes, the red signal was significantly increased but mainly distributed within the mucus layer. In contrast, strong red signals were observed in the samples treated with PBAE SP/TDA polyplexes, with a wide distribution extending from the mucus layer to the cellular nuclei. However, a slight decrease in the red signal was observed across the cell monolayer as the lipophilic TDA ratio increased in the PBAE polymers. As previously reported, the long mucin proteoglycans chains present in mucus entangle, usually forming hydrophobic domains and hydrophilic channels in the network. This periodic hydrophobic domains have been shown to interact with hydrophobic particles or particles exhibiting hydrophobic moieties[163,169]. For polyplexes with comparable electrical properties, this hydrophobic affinity may cause polyplexes with higher ratio of lipophilic monomers, either OA or TDA, to be restricted in diffusion. Overall, CLSM images showed that amphiphilic PBAEs consisting of SP/TDA were able to penetrate mucus and mediate sufficient uptake in epithelial cells.

Further increasing the biological complexity, we evaluated the gene silencing effects of our polyplexes in human Precision-Cut-Lung-Slices (hPCLS) (Figure IV.4B). hPCLS are widely recognized as a powerful tool for investigating drug responses in an environment that accurately reflects the complexity of the human lower respiratory tract. hPCLS maintain the native lung architecture, which includes the respiratory parenchyma and small airways, as well as a variety of lung-resident cells, including type I and II alveolar cells, bronchial epithelial cells, endothelial cells, and immune cells[170]. After 48 h of siGAPDH transfection in hPCLS, the gene silencing effects were evaluated by measuring the downregulation of the housekeeping gene GAPDH as previously described[171]. In this proof-of-concept study, GAPDH was chosen as a target gene only to evaluate the delivery efficiency, and it will be replaced with an aberrant gene for treating specific diseases in future applications. In addition, the hPCLS used in our study were derived from non-lesional regions and were in principle free of abnormal genes. As a result, qPCR analysis of the extracted RNA from the slices showed that the average GAPDH/β-Actin ratio was approximately 1.0 in the free siGAPDH-treated group, while in the Lipofectamine 2000-treated group, this ratio dropped significantly to 0.71 (Figure IV.4C). SP0.6/TDA0.4 and SP0.4/TDA0.6 polyplexes enabled a slight decrease of GAPDH gene expression in the hPCLS, with reductions of 16.3% and 20.1%, respectively. Overall, SP0.5/TDA0.5 polyplexes demonstrated the highest gene silencing efficiency, achieving a 43.7% reduction of the GAPDH level, confirming the need for balancing cationic and hydrophobic content in the PBAE nanocarriers for efficient pulmonary delivery.

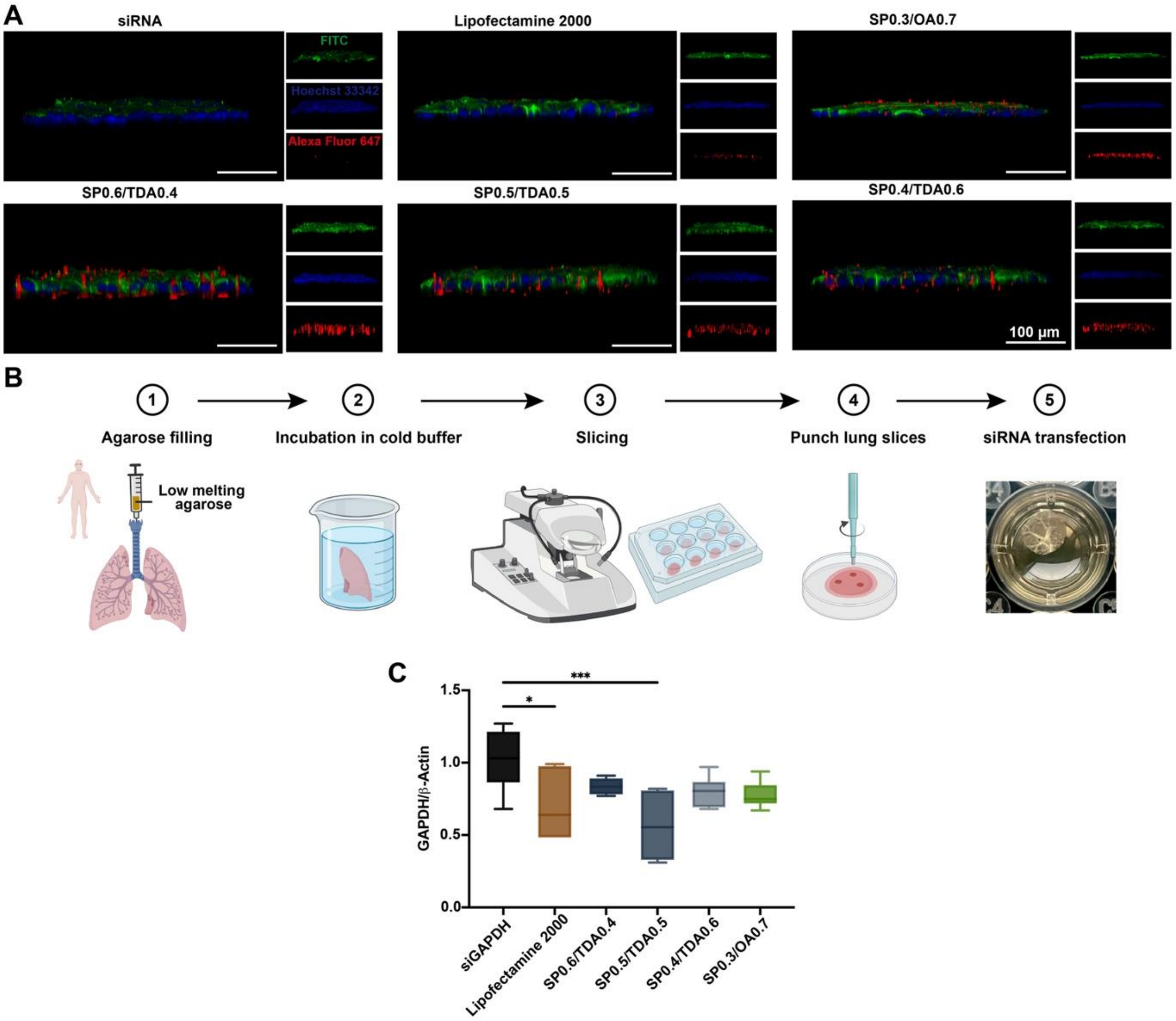


Figure IV.4: Mucus penetration assay and *ex vivo* knockdown of house-keeping gene GAPDH. (A) Mucus penetration of polyplexes in air-liquid interface (ALI) culture of Calu-3 cells 24 h after transfection. Red color represents Alexa Fluor 647 labeled siRNA, nuclei are shown in blue and mucus layer in green. Scale bar, 100 µm. (B) Schematic diagram of preparation of human precision cut lung slices (hPCLS). (C) GAPDH gene knockdown efficiency in hPCLS transfected with different formulations. The experiments were performed in technical triplicates and data are presented as mean ± SD, n = 2; **p* < 0.05, ****p* < 0.001, one-way ANOVA.

4.6 *In Vivo* performance: Biodistribution, Biocompatibility and Knockdown Effects after Pulmonary Delivery

Based on *in vitro* and *ex vivo* results, we selected PBAE SP0.5/TDA0.5 to move further to *in vivo* studies. Alexa Fluor 647-labeled siRNA was loaded into polyplexes and delivered via intratracheal instillation (Figure IV.5A). Compared to free siRNA, polyplexes demonstrated significantly higher retention and internalization in the lung (Figures IV.5B and IV.S19), with an 82.3-fold increase in fluorescence intensity (Figure IV.5C). As observed previously[172], when administered as polyplexes, some siRNA entered systemic circulation via pulmonary capillaries, accumulating in the liver before metabolism as evidenced by the signal detected in the liver and kidneys, respectively. Due to the complex architectural structure in the lung, polyplexes may face challenges in reaching the respiratory zone, which cannot be accurately evaluated by *IVIS* imaging. CLSM images revealed that polyplexes containing pHrodo red-labeled siRNA, represented in red color, have successfully reached not only the lower respiratory tract but also the respiratory zone (Figure IV.5D). Furthermore, polyplex uptake was observed in various cell types within the lung, and the corresponding flow cytometric gating strategy is shown in Figure IV.*S*20. The higher MFI in polyplex-treated alive lung cells (average 2136) was consistent with *IVIS* imaging results (Figure IV.5E). High MFI detected in dendritic cells (average 4941), macrophages (average 7773), and eosinophils (average 1348) highlighted strong phagocytosis potential in the lung, which generally poses a challenge for pulmonary siRNA delivery. Interestingly, the uptake of polyplexes in both CD4$^+$ and CD8$^+$ T cells remained low, being beneficial for avoiding adverse immune activation and in line with the need for targeting ligands for efficient T cell transfection[173]. Importantly, lung epithelial cells, particularly type II pneumocytes, are often related to the progression of respiratory diseases, such as chronic obstructive pulmonary disease (COPD), lung cancer, lung fibrosis, and pneumonia[174–176]. The uptake of polyplexes in epithelial cells, particularly in type II pneumocytes was 9.46-fold and 7.61-fold higher, respectively, when compared to free siRNA. These results suggest the potential of siRNA therapy based on our carrier system for treating respiratory diseases in the future and underline the need for nanocarriers in pulmonary delivery.

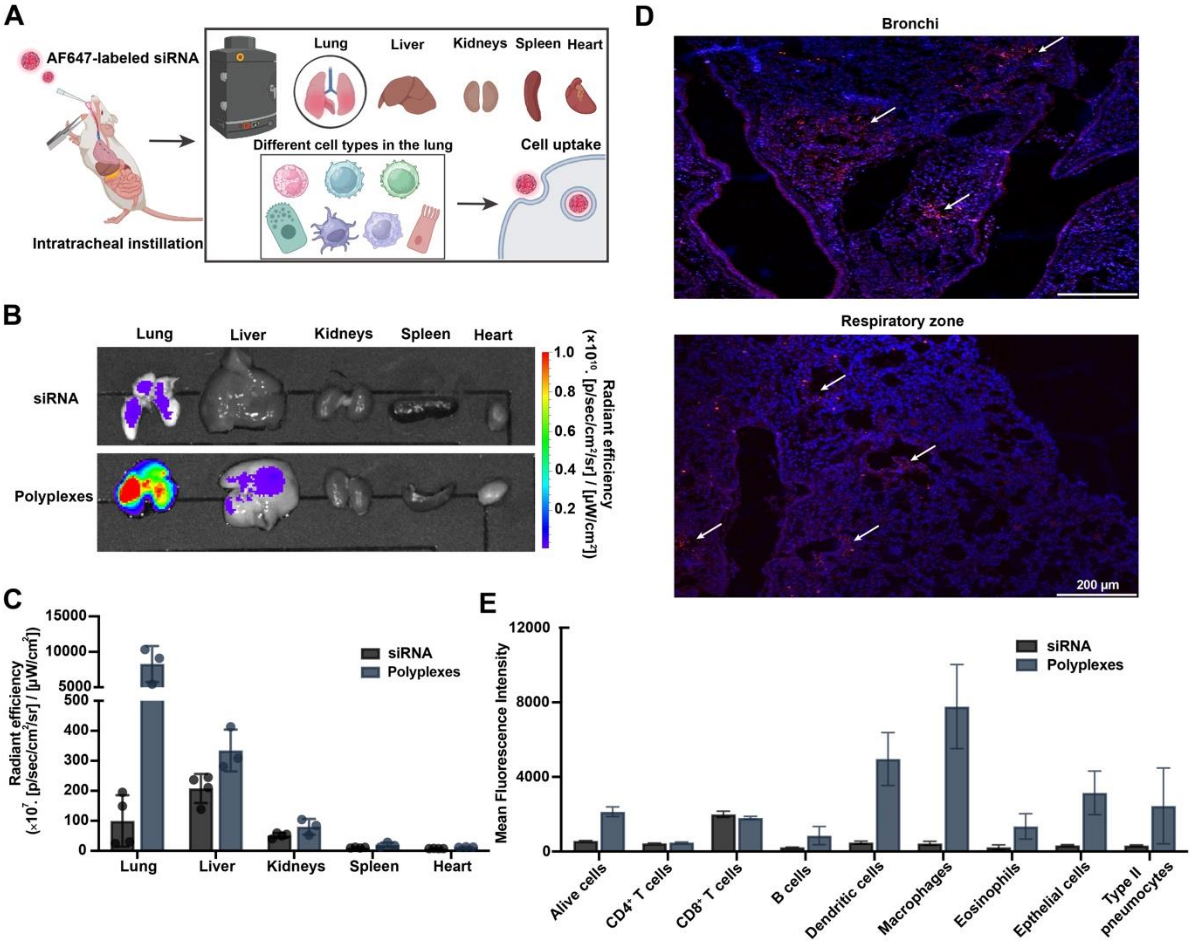


Figure IV.5: *In vivo* biodistribution in the organs and cellular uptake in the lung. (A) Schematic diagram of in vivo distribution investigation after intratracheal instillation of polyplexes containing Alexa Fluor 647-labeled siRNA. (B) Representative organ distributions in mice that received free siRNA or siRNA-loaded polyplexes, respectively. (C) Quantification of fluorescence intensity of Alexa Fluor 647 labeled siRNA distributed in the organs. (D) Distribution of polyplexes containing pHrodo red-labeled siRNA in different lung regions. White arrows indicate polyplexes. Scale bar, 200 μm. (E) Mean fluorescence intensity of AF647-labeled siRNA in different cell types in the lung. Data are presented as mean ± SD, n=3.

Next, we evaluated the siRNA knockdown efficiency in the lung and performed safety assessment. PEI25k, as well-established control, presents reliable transfection efficiency in gene delivery and has been widely used in previous studies focusing on polymer-based carriers[177,178]. Due to its known cytotoxicity, PEI25k is also used as a positive control in safety evaluations and was therefore included in our *in vivo* test. RNA extracted from the lungs treated with different formulations was analyzed via qPCR. In the control group that received buffer only, the average GAPDH/β-Actin ratio was 1.03 (Figure. IV.6A). In mice treated with free siGAPDH, the ratio increased to 1.44, with a broader standard deviation of 0.39, demonstrating that free siRNA did not achieve GAPDH gene silencing. In contrast, siGAPDH-loaded PBAE polyplexes showed a significant 30.4% reduction of the GAPDH/β-Actin ratio when compared to negative control siRNA-loaded PBAE polyplexes. In the mice treated with PEI-siGAPDH, the GAPDH/β-Actin ratio oppositely increased to a broad range from 1.21 to 2.69, likely due to severe lung inflammation as hematoxylin and eosin (H&E) staining revealed noticeable inflammatory cell infiltration, alveolar wall thickening, and disruption of the alveolar architecture in these mice (Figure. IV.6B). Conversely, lung tissue structures in PBAE polyplex-treated mice were well-preserved, with clear alveolar spaces and negligible alveolar wall thickening as observed in buffer- and free siRNA-treated groups, which suggested minimal lung tissue damage or inflammation in these mice. Consistent with the H&E staining results, levels of inflammatory cytokines, i.e., IL-6, MCP-1, IFN-β, TNF-α in BALF were significantly higher in PEI-siGAPDH and PEI-siNC treated mice when compared to other groups (Figure. IV.6C). In particular, IL-6 was detected at the highest concentration among all cytokine types, in the PEI-siGAPDH group, with an average value of 322 pg/mL, which was 9.4-fold and 64.4-fold higher than PBAE-siGAPDH and free siGAPDH groups, respectively. However, PBAE polyplexes treatment did not abnormally elevate cytokine levels, which remained comparable to blank and free siRNA-treated mice on most indicators.

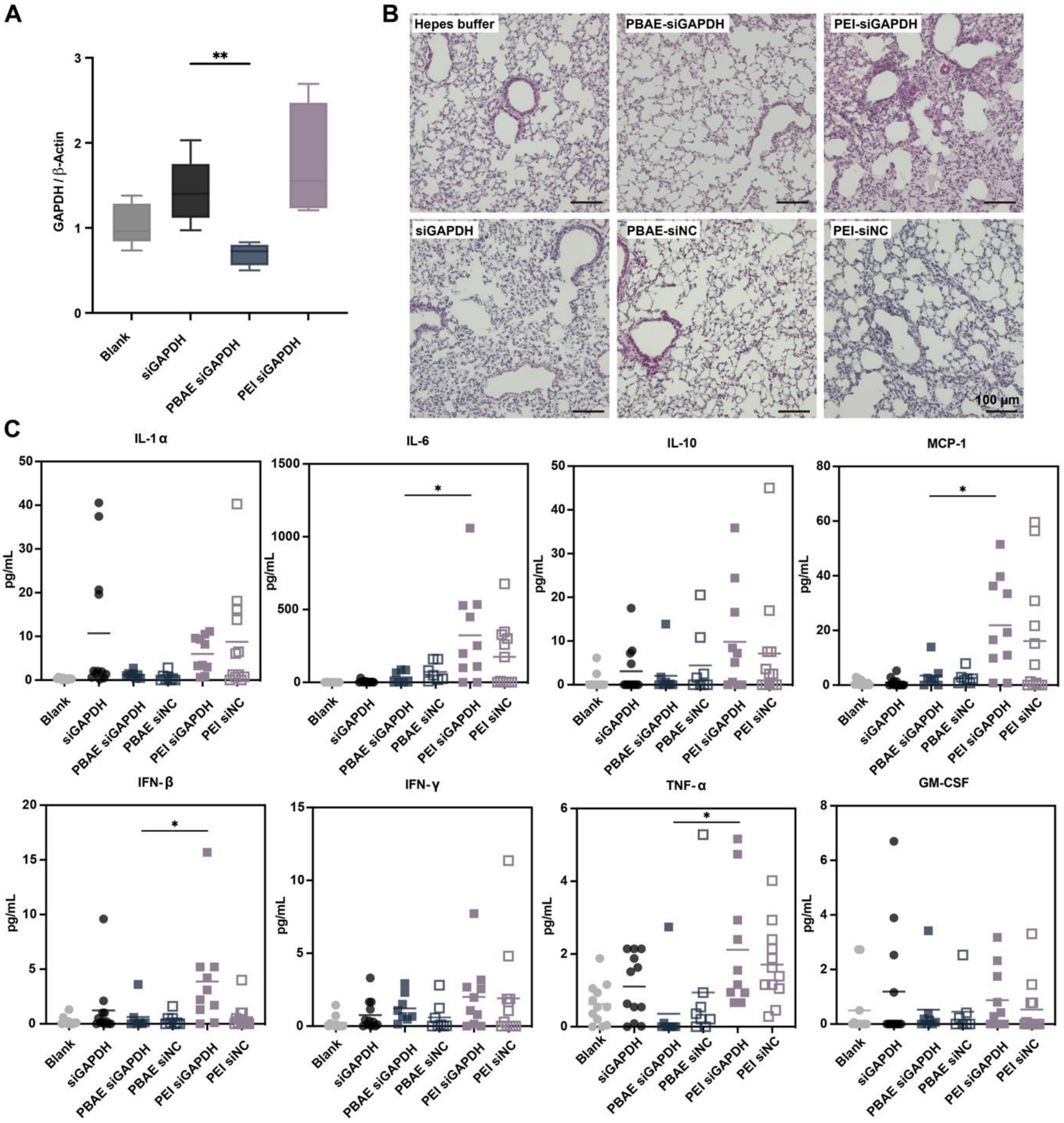


Figure IV.6: *In vivo* GAPDH gene silencing and safety evaluations**.** (A) GAPDH gene silencing effects of polyplexes in the lung, data are presented as mean ± SD, n=6 for Blank and siGAPDH groups, n=4 for PBAE siGAPDH and PEI siGAPDH groups, ***p* < 0.01, Student's t-test. (B) H&E-stained lung sections collected from mice treated with different formulations. Scale bar, 100 μm. (C) Inflammatory cytokine levels in the bronchoalveolar lavage fluid (BALF) collected from mice treated with different formulations. The experiments were performed in technical duplicates and data are presented as mean ± SD, n=6 for Blank and siGAPDH groups, n=4 for PBAE siGAPDH and PEI siGAPDH groups, **p* < 0.05, one-way ANOVA.

5 Study Limitations and Data Scarcity

The study presented here demonstrates an elegant, literature-driven strategy for screening polymeric gene-delivery candidates and yields promising results on a newly synthesised validation set. Nevertheless, several limitations must be acknowledged so that readers can appreciate the scope of our conclusions.

First, although focusing on polyesters is a sensible starting point, essential details like copolymerisation patterns, block lengths, architecture, dispersity, and molecular weight variation are rarely reported, and even when they are, they are seldom provided in a standardized and machine-readable format. As a result, descriptors based on idealised repeat units capture only a fraction of the true physicochemical diversity. Ongoing standardisation efforts that mandate sharing raw chromatograms and NMR spectra may eventually allow direct ingestion of this information into machine-learning pipelines, but such data are not yet widely available.

Second, data sparsity is a major hurdle. Whereas proteins and small molecules benefit from extensive databases, experimentally characterised polymeric gene-delivery systems are scarce. We therefore limited the chemical space to structurally similar polyesters and used a UMAP projection solely as a qualitative coverage check. Predictions outside this region must be treated with caution, because extrapolating far from the training manifold typically yields unreliable results. A rigorous, quantitative safeguard was not implemented here for three practical reasons: (1) no curated set of truly out-of-domain polymers yet exists for calibration; (2) distance estimates are highly sensitive to the chosen descriptor space; and (3) alternative distance metrics and thresholding schemes can give conflicting signals when data are sparse. As larger, standardised data sets emerge, these challenges should become tractable, enabling formal applicability-domain filters to accompany future models.

Third, biological context also matters. One-hot encoding of cell lines allows within-set predictions but offers no mechanistic insight and cannot guarantee accuracy for cell types absent from the training data. Future work could explore lineage- or transcriptome-derived embeddings to improve transferability.

Finally, although RDKit descriptors efficiently encode molecular structure, they are not optimised for human-interpretable structure-function insight. Graph-based neural network representations may provide traceable, learnable features and can be backmapped to their structure[123,179] once larger, standardised data sets become available.

6 Conclusions

This study provides an efficient approach for utilizing literature data to train a ML model for predicting suitable polymeric delivery systems. By employing straightforward strategies, we successfully merged multiple different datasets with different carrier systems. The trained in silico model was validated to be accurate by assessing *in vitro* gene silencing outcomes when siRNA was delivered using polymers that the ML model predicted to be effective or ineffective. Among the tested polymers, one candidate PBAE SP/TDA-BG was selected for detailed investigations of its structural characteristics and biological performance. This polymer, with its balanced hydrophilic and hydrophobic moieties combined with a biodegradable backbone, overcame key biological barriers in pulmonary siRNA delivery. Remarkably, it achieved efficient *in vivo* gene silencing without detectable adverse effects. These findings highlight the capability of the ML model to significantly reduce the need for extensive experimental screening efforts and associated resource costs and ethical considerations. Our study also provides conceptual insights into the complex processes of polymeric siRNA delivery, which emphasizes the transformative role of ML in optimizing delivery systems. While current limitations include a constrained dataset, which makes it difficult to extrapolate to novel polymer types, this challenge could be mitigated as more data becomes available. With expanded datasets, data-intensive methods, such as Deep Generative Models, could aid the design of entirely new materials for future nanomedicine applications.

7 Experimental Section

7.1 Data processing and Machine Learning

Structural data was collected from literature references[80,180,181] on 605 polymers that had been employed for siRNA delivery before, reflecting a range of different polyester types. Chemical structures were created using ChemDraw (version 22.2.0). All data related tasks were performed using Python (version 3.11.5). Molecule sanitizing, embedding and MMFF force field optimization as well as Molecular Descriptor and Morgan Fingerprint calculation were performed using the widely adopted cheminformatics library RDKit (version 2024.09.1). Each block monomer was encoded separately and the respective component ratio was incorporated in the descriptors by multiplying them with the weighted ratio of copolymer blocks following Kim et al.[182]. Gene knockdown (KD) performance was categorized into two groups: KD < 50%, and > 50%. Additional data, including monomer ratios, cell types, and molecular weight (Mw), were incorporated. Data was cleaned removing multiple entries and columns that contain NaN (Not a number), followed by normalization of features using StandardScaler class from sklearn (version 1.6.0). Various models (SVM, KNN, RF, XGB, LGBM, NaiveBayes) with weighted sampling due to dataset imbalance, were evaluated. The lead model (LGBM) was fine-tuned with hyperopt (version 0.2.7). Important features were calculated using SHAP values and a TreeExplainer class. Irrelevant features were excluded from the dataset, using the integrated feature_importance method in LGBM. Data was split into training and test sets, stratified by KD classes (20% test set ratio). The trained model was applied to assess new, unpublished polymer formulas, identifying one high-performing polymer selected for synthesis. Additionally, waterfall plots were calculated for the predicted polymer using SHAP library version (version 0.46.0). The following Python libraries were used for data handling and plotting: Sklearn, Imblearn (0.13.0), Pandas (2.1.4), Numpy (1.26.4), Seaborn (0.13.2), Matplotlib (3.9.0).

7.2 Chemicals

Ethyl trifluoroacetate, tetradecylamine, oleylamine, 4-Amino-1-butanol, 1,4-butanediol diacrylate and bisphenol A diglycidyl ether diacrylate were purchased from Sigma Aldrich (Taufkirchen, Germany). Di-tert-butyl dicarbonate, spermine and SYBR Gold Nucleic Acid Gel Stain were bought from Fisher Scientific (Hampton, NH, USA).

7.3 Synthesis of Tri-boc-spermine

Tris(tert-butoxycarbonyl)spermine, abbreviated as tri-Boc-spermine, was synthesized as described elsewhere[100]. In brief, spermine (1 eq) was dissolved in methanol and stirred at - 78 °C, ethyl trifluoroacetate (1 eq) was subsequently added dropwise and stirred at - 78 °C for 1 h, then at 0 °C for 1 h. Without isolation, di-tert-butyl dicarbonate (4 eq) was added dropwise to the solution and stirred at room temperature (RT) for 2 days. Finally, the solution was adjusted to a pH > 11 by 25% ammonia and stirred overnight to cleave the trifluoroacetamide protecting group. The mixture was then evaporated under vacuum and the residue was diluted with dichloromethane (DCM) and washed with distilled water and saturated sodium chloride aqueous solution. The DCM phase was finally dried by magnesia sulfate and concentrated to give the crude product. The crude product was purified by column chromatography (CH_2Cl_2\MeOH\NH_3, aq. 7:1:0.1, SiO_2, $KMnO_4$; Rf = 0.413). Tri-Boc-spermine was isolated and characterized by 1H nuclear magnetic resonance spectroscopy (1H-NMR).

7.4 Synthesis of PBAE

The synthesis involved dissolving hydrophilic amine in dimethylformamide (DMF) and adding lipophilic amine and the diacrylate backbone (1.2 eq). The reaction was sealed, heated and kept at 90°C for 48 h, then cooled to RT. DMF was evaporated, and the solid polymer was solubilized in DCM. Deprotection of the triboc-spermine containing polymers was achieved by the dropwise addition of Trifluoroacetic acid (TFA) to a final concentration of 5% v/v, cleaving the Boc groups. The reaction was stirred at RT for two h. To obtain the deprotected polymer, the solvent was evaporated. For all polymers the solid was purified by precipitating it in diethyl ether followed by a centrifugation step (1250 rpm for 2 min). The procedure was repeated three times. The final product was dried under vacuum and characterized using 1H-NMR.

7.5 Gel Permeation Chromatography (GPC)

GPC was performed with an Agilent aqueous GPC using a PSS Novema max Lux 100A followed by two PSS Novema max Lux 3000A columns. The chromatographic system and calibration standards were set up according to pre-analysis by Agilent Technologies. Measurements were performed at 40°C in 0.1 M sodium chloride solution supplemented with 0.3% formic acid. Samples were prepared at 4 g/L and measured at a flow of 1 mL/min. Molar Mass distributions were obtained through the Agilent WinGPC Software against pullulan calibration standards in the range of 180 Da to 1450 kDa. A daisy-chain detector setup of an Agilent 1260 VWD was followed by an Agilent 1260 GPC/SEC MDS and ended with an Agilent 1260 RID.

7.6 Preparation of Polyplexes

To prepare PBAE-siRNA polyplexes, the polymer stock solution was diluted to various concentrations with diethyl pyrocarbonate (DEPC) treated water. Next, an equal volume of a specific amount of siRNA diluted in 10 mM HEPES buffer (pH 5.4) was added, and the mixture was incubated at RT for 30 min to obtain siRNA-loaded polyplexes at different N/P ratios. The N/P ratio represents the molar ratio between the polymer amine groups (N) and the siRNA phosphate groups (P), and the amount of polymer required for different N/P ratios was calculated using the following formula:

m (polymer in pg) = n siRNA (pmol) x N/P x number of nucleotides siRNA x M protonable unit (g/mol)

The number of nucleotides for asymmetric 25/27mer siRNA was set to 52, while in EGFR siRNA, only 42 nucleotides were present. The protonable units for each polymer were calculated by dividing the molar mass of the repeating unit by the number of protonable amines within each repeating unit.

7.7 Characterization of polyplexes

Particle size, polydispersity index (PDI) and zeta potential of PBAEs-siRNA polyplexes were determined using a Zetasizer Ultra (Malvern Instruments, Malvern, UK). All measurements were conducted using a 10 mM HEPES buffer as dispersant. Results are expressed as mean ± standard deviation (SD) over three measurements.

The encapsulation efficiency of siRNA was determined using SYBR gold assays. In brief, 15 µL of PBAEs-siRNA polyplexes were added into a 384-well plate, then 5 µL of a 4X SYBR Gold solution were added to each well and incubated for 15 min protected from light at RT. Fluorescence intensity was measured using a plate reader (Tecan, Männedorf, Switzerland) with excitation and emission wavelength set at 492 nm and 555 nm, respectively. An equal amount of free siRNA was used as 100% value for calculating the unencapsulated siRNA in different polyplex samples.

7.8 siRNA release assay

SYBR Gold assay was performed to investigate siRNA release from polyplexes under different conditions. First, PBAEs-siRNA polyplexes at an N/P ratio of 10 were prepared as described under 4.6. Polyplexes containing 10 pmol of siRNA were incubated with serial dilutions of Triton X and heparin in a 384-well plate for 30 min at 37°C. Then, 10 µL of a 4X SYBR Gold solution were added to each well and incubated for 15 min. The results were measured as described under 5.7.

7.9 RNase protection assay

PBAE-siRNA polyplexes at an N/P ratio of 10 were prepared as previously described. A total of 50 µL of the respective formulations containing 50 pmol of siRNA, was incubated with 1 µg RNase A (Sigma-Aldrich, Taufkirchen, Germany) for 30 min at 37°C. As a control group, 50 pmol of free siRNA was included, either treated with 1 µg of RNase, or left untreated as a 100% reference value for calculating the degraded siRNA. Subsequently, the RNase was deactivated by heating to 70 °C for 30 min. To release the RNA, 1% Triton

X and 4 IU heparin was added and incubated for 30 min at 37°C. The released RNA was then quantified using SYBR Gold assay, and fluorescence was measured as described under 5.7.

7.10 *In vitro* cell viability

H1299 cells seeded in 96-well plates at a density of 6,000 cells per well were used to assess cytotoxicity. After incubation with PBAEs-siRNA polyplexes containing 20 pmol scrambled siRNA (siRNA negative control, siNC) ranging from N/P 3 to N/P 20 for 48 h, 10 µL of the Cell Counting Kit-8 (CCK-8, Sigma) reagent was added to develop color for 3-4 h. The optical density (OD) was measured on a Tecan plate reader at 450 nm and cell viability was calculated by dividing the values of groups treated with polyplexes by that obtained with the untreated group.

7.11 *In vitro* cellular uptake

H1299 cells were seeded in 24-well plates at a density of 15,000 cells per well and incubated with PBAEs-siRNA polyplexes containing 50 pmol of siRNA with N/P ratios of 3 to 10, where 20% of the siRNA was Alexa Fluor 647-labeled. Free siRNA and Lipofectamine 2000 containing equal amounts of siRNA were used as controls. After 24 h of incubation, cells were divided equally. Half of the cells were measured directly with an Attune NxT flow cytometer (ThermoFisher Scientific, Waltham, MA USA), and the other half were pre-mixed with 0.4% Trypan blue solution and measured comparably.

A549 cells were seeded in 96-well plates at a density of 6,000 cells per well and incubated with the same PBAEs-siRNA polyplexes containing 20 pmol of siRNA with an N/P ratio of 10. After 24 h of incubation, the cells were assessed on an Attune NxT flow cytometer (ThermoFisher Scientific).

7.12 *In vitro* endosomal escape

Hela-Gal8-mRuby3 cells were kindly provided by the lab of Professor Ernst Wagner (LMU Munich, Germany). Hela-Gal8-mRuby3 cells were seeded in the 8-well chamber slide (ibidi, Gräfelfing, Germany) at a density of 10,000 cells per well, and then incubated for 4 h with different PBAEs-siRNA polyplexes containing 40 pmol of siRNA (20% of which was Alexa Fluor 647-labeled). After incubation, the supernatant was discarded, and the chambers were rinsed with PBS for three times. The cells were first fixed with a 4% PFA solution at RT for 20 min and then stained with 0.5 µg/mL of DAPI solution for 8 min. After rinsing the chambers with PBS for at least three times, the cells were imaged using a SP8 inverted confocal laser scanning microscope (Leica Camera, Wetzlar, Germany) equipped with a 63X objective. The fluorescent dots of Gal8-mRuby3 were quantified using the Fuji plug-in of Image J.

7.13 *In vitro* eGFP knockdown

Protein knockdown experiments were conducted using H1299 cells stably expressing enhanced green fluorescent protein (eGFP). Polyplexes were formulated with siRNA targeting eGFP mRNA or scrambled siRNA with the same length. H1299/eGFP cells were seeded in 96-well plates at a density of 6,000 cells per well and then incubated with polyplexes containing 20 pmol siGFP or 20 pmol siNC for 48 h. Lipofectamine 2000 was used as a positive control, while free siRNA served as a negative control. After incubation, the cells were collected to perform the FACS analysis (Attune NxT Flow Cytometer, ThermoFisher Scientific). The eGFP knockdown efficiency was calculated by dividing the Median Fluorescence Intensity (MFI) of siRNA-treated group by that of the respective siNC-treated group.

7.14 *In vitro* EGFR knockdown

An EGFR knockdown experiment was conducted in A549 cells using polyplexes formulated with EGFR siRNA. Per well, 6,000 A549 cells were seeded in 96-well plates and treated with polyplexes containing either 20 pmol of EGFR siRNA or 20 pmol of scrambled siRNA (siNC) at an N/P ratio of 10 for 48 h. Following incubation, the cells were collected and stained with Vio® R667 anti-human EGFR antibody for 10 min. After washing twice using PBS, the cells were analyzed using a flow cytometer (Attune NxT) to assess EGFR expression.

7.15 Mucus penetration and uptake study

Air Liquid Interface (ALI) experiments were conducted utilizing Calu-3 cell culture. Specifically, Calu-3 cells were seeded at a density of 250,000 cells per well onto uncoated Transwell® polyester cell culture inserts (6.5 mm, 0.4 µm pore size) and were maintained in culture for three days until confluent. On day 4, the apical medium was removed to establish ALI conditions, and the medium in the basolateral chamber was replaced with 300 µl of PneumaCult™ ALI medium (STEMcell Technology, Vancouver, Canada). The medium was replaced every three days until the transepithelial electrical resistance (TEER) values stably reached 300 Ω*cm^2 when monitoring with an EVOM epithelial volt/Ω meter (World Precision Instruments, Sarasota, USA). Polyplexes and Lipofectamine 2000, each containing 100 pmol of siRNA, 20% of which was Alexa Fluor 647-labeled, were applied on top of Calu-3 monolayers without previous washing and incubated for 24 h. Free siRNA was employed as a negative control. Afterwards, the cells were stained with 100 µL of diluted Hoechst 33342 (for nuclear staining) and AF488-wheat germ agglutinin (for mucus staining) at 37°C for 20 min. Cells were then gently washed twice with PBS and mounted on glass slides using FluorSave™ reagent. Fluorescent images were immediately captured using a 40X objective on the SP8 inverted confocal laser scanning microscope (Leica Camera) and were processed using the Fuji plug-in of Image J.

7.16 *Ex vivo* activity in human precision-cut lung slices (hPCLS)

7.16.1 Human tissue, ethics statement and human precision-cut lung slices (hPCLS)

Human lung tissues were obtained from the University Hospital Großhadern of the Ludwig-Maximilian University (Munich, Germany) and the Asklepios Biobank of Lung Diseases (Gauting Germany). Participants provided written informed consent to participate in this study, in accordance with approval by the local ethics committee of the Ludwig Maximilian University Munich, Germany (Project 19–630). In brief, hPCLS were prepared from tumor-free peri-tumor tissue. The lung tissue was inflated with 3% agarose solution and then solidified at 4°C. The lung sections with a thickness of 500 µm were cut from the tissue blocks using a vibration microtome (HyraxV50) (Karl Zeiss AG, Oberkochen, Germany). hPCLS were cultured in DMEM F-12 medium supplemented with 0.1% FBS. Prior to experiments, hPCLS were cut into 4 mm diameter circular pieces using a biopsy puncher.

7.16.2 GAPDH gene silencing in hPCLS

Each well containing three punches of hPCLS in a 24-well plate was treated with different formulations containing either 100 pmol of siGAPDH or 100 pmol of siNC. Lipofectamine 2000 was included as a positive control and free siGAPDH as a negative control. After 48 h of incubation, the tissue punches were submerged in 1 mL TRIzol within lysing matrix D tubes and homogenized using a FastPrep 24 Tissue Lyzer (M.P. Biomedicals, Irvine, CA, USA). Subsequently, 200 µL of chloroform was added to each homogenized sample and mixed vigorously. The samples were then centrifuged at 11,000 g for 15 min at 4°C, after which the aqueous phase containing RNA was transferred to a new 1.5 mL Eppendorf tube. To precipitate the RNA, 500 µL of isopropanol was added and mixed thoroughly. After 10 min incubation at RT, the samples were centrifuged at 11,000 g for 10 min. The supernatant was discarded, and the RNA pellet was washed with 1 mL of ice-cold 75% ethanol, followed by centrifugation at 7,500 g for 5 min at 4°C. The supernatant was discarded again, and the RNA pellet was resuspended in 30 µL of RNase-free water. The extracted RNA was then processed for cDNA synthesis using a high-capacity cDNA synthesis kit (Applied Biosystems). Synthesized cDNA was diluted and subjected to quantitative PCR (qPCR) using SYBR™ Green PCR Master Mix (ThermoFisher Scientific), with Hs_GAPDH_2_SG primers specific for human GAPDH (Qiagen, Valencia, CA, US). Hs_ACTB_2_SG primers for human β-actin (Qiagen) were used as the normalization control.

7.17 *In vivo* distribution of polyplexes after pulmonary delivery

All animal experiments were conducted according to the German law of animal protection and approved by the Government of Upper Bavaria (ROB-55.2-2532.Vet_0220-171) and the Committee for Animal Experimentation of the Ludwig Maximilian University Munich, Germany.

Eight-week-old female BALB/c mice were intratracheally instilled with polyplexes containing 1 nmol of Alexa Fluor 647 labeled siRNA under ketamine/xylazine anesthesia. The control group received free Alexa Fluor 647-siRNA. After 24 h, mice were sacrificed with an overdose of ketamine/xylazine anesthesia, and organs including the heart, lung, liver, spleen and kidneys were harvested for imaging. Fluorescence was measured at an excitation wavelength of 635 nm and an emission wavelength of 668 nm using an IVIS Lumina III (PerkinElmer, Shelton, CT, USA). After imaging, the lungs were homogenized to obtain single-cell suspensions, using the Mouse Lung Dissociation Kit (Miltenyi Biotec, Germany) according to the manufacturer's protocol. The lung cells were first incubated with PBS solution containing Zombie UV™ and later stained with FITC anti-mouse CD45, BUV395 anti-mouse CD3, Vioblue anti-mouse CD4, APC-Cyanine7 anti-mouse CD8, PE-Cyanine7 anti-mouse F4/80, BUV605 anti-mouse CD11c, BV785 anti-mouse CD326, PE/Dazzle™594 anti-mouse CD170 and PerCP/Cyanine5.5 anti-mouse CD19 for 30 min at 4°C. The stained cells were measured using a Cytek® Aurora (San Diego, California, USA) implemented with autofluorescence extraction for the detection of cellular uptake.

7.18 Distribution of polyplexes in the lung

Eight-week-old female BALB/c mice were intratracheally instilled with polyplexes containing 1 nmol of pHrodo red-labeled siRNA under ketamine/xylazine anesthesia. After 24 h, the mice were sacrificed with an overdose of ketamine/xylazine anesthesia, and the lungs were harvested after lung perfusion. The lungs were then immersed in 4% PFA solution overnight. After PFA fixation, the lung tissues were embedded in paraffin and sliced into lung sections with thickness of 4 µm. The obtained slices were deparaffinized by incubating in xylene, followed by a series of ethanol dilutions. After hydration, the slices were stained with 0.5 µg/mL DAPI solution for nuclear visualization and imaged using a 10X objective on an SP8 inverted confocal laser scanning microscope (Leica Camera).

7.19 *In vivo* transfection evaluation of polyplexes

7.19.1 Safety evaluation

Eight-week-old female BALB/c mice were intratracheally instilled with different formulations containing 1 nmol of siGAPDH or 1 nmol of siNC, including PBAEs-siRNA and PEI-siRNA polyplexes. Control groups received either free siGAPDH or buffer only. After 24 h, the mice were sacrificed, and their lungs were first perfused with 10 mL of saline. Bronchoalveolar lavage fluid (BALF) was collected in a PBS/2mM EDTA buffer containing protease inhibitor cocktail (cOmplete™). The BALF was centrifuged at 500 g for 5 min at 4°C, and the supernatant was used to measure the concentration of pro-inflammatory cytokines using the LEGENDplex™ Mouse Cytokine Panel 2 kit (Biolegend, San Diego, California, USA). The lungs were harvested, with one lobe fixed in 4% PFA overnight and then embedded in paraffin for histological analysis via H&E staining, while the remaining tissue was stored in 1 mL of RNA-later solution for further analysis.

7.19.2 *In vivo* GAPDH gene silence efficacy of polyplexes

The lungs stored in RNA-later solution were transferred to lysing matrix D tubes and homogenized using a FastPrep 24 Tissue Lyzer (M.P. Biomedicals). RNA extraction was performed following the TRIzol-chloroform method as previously described under 5.16.2. The extracted RNA was then processed for cDNA synthesis using a high-capacity cDNA synthesis kit (Applied Biosystems). The synthesized cDNA was diluted and subjected to qPCR using SYBR™ Green PCR Master Mix (ThermoFisher Scientific) with Mm_GAPDH_3_SG primers (Qiagen) for GAPDH. Mm_ACTB_2_SG primer sspecific for mouse β-actin were used as the normalization control.

7.20 Statistical analysis

All data were expressed as means ± standard deviation (SD). All statistical analyses were performed using one-way analysis of variance (ANOVA) in GraphPad Prism or Student's t-test when specifically stated. Levels of significant differences were expressed as follows, *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

8 Acknowledgements

These studies were funded by the European Research Council (ERC-2022-COG-101088587). Min Jiang is funded by the China Scholarship Council. We gratefully acknowledge the provision of human biomaterial (non-fibrotic fresh tissue from non-CLD patient) and clinical data from the CPC-M bioArchive and its partners at the Asklepios Biobank Gauting, the LMU Hospital and the Ludwig-Maximilians-Universität München. We thank the patients and their families for their support. We additionally acknowledge ethris GmbH for whole body imaging resources and LMU's BMC flow cytometry core facility for expert advice and technical support. Figures were partially generated using biorender.com.

9 Conflict of Interest

Olivia Merkel and Benjamin Winkeljann are co-founders of RNhale GmbH. Olivia Merkel is a Scientific Advisory Board Member of Coriolis Pharma, AMW, and Corden Pharma as well as a consultant for PARI Pharma, AbbVie Deutschland, and Boehringer-Ingelheim International.

10 Supplementary Information

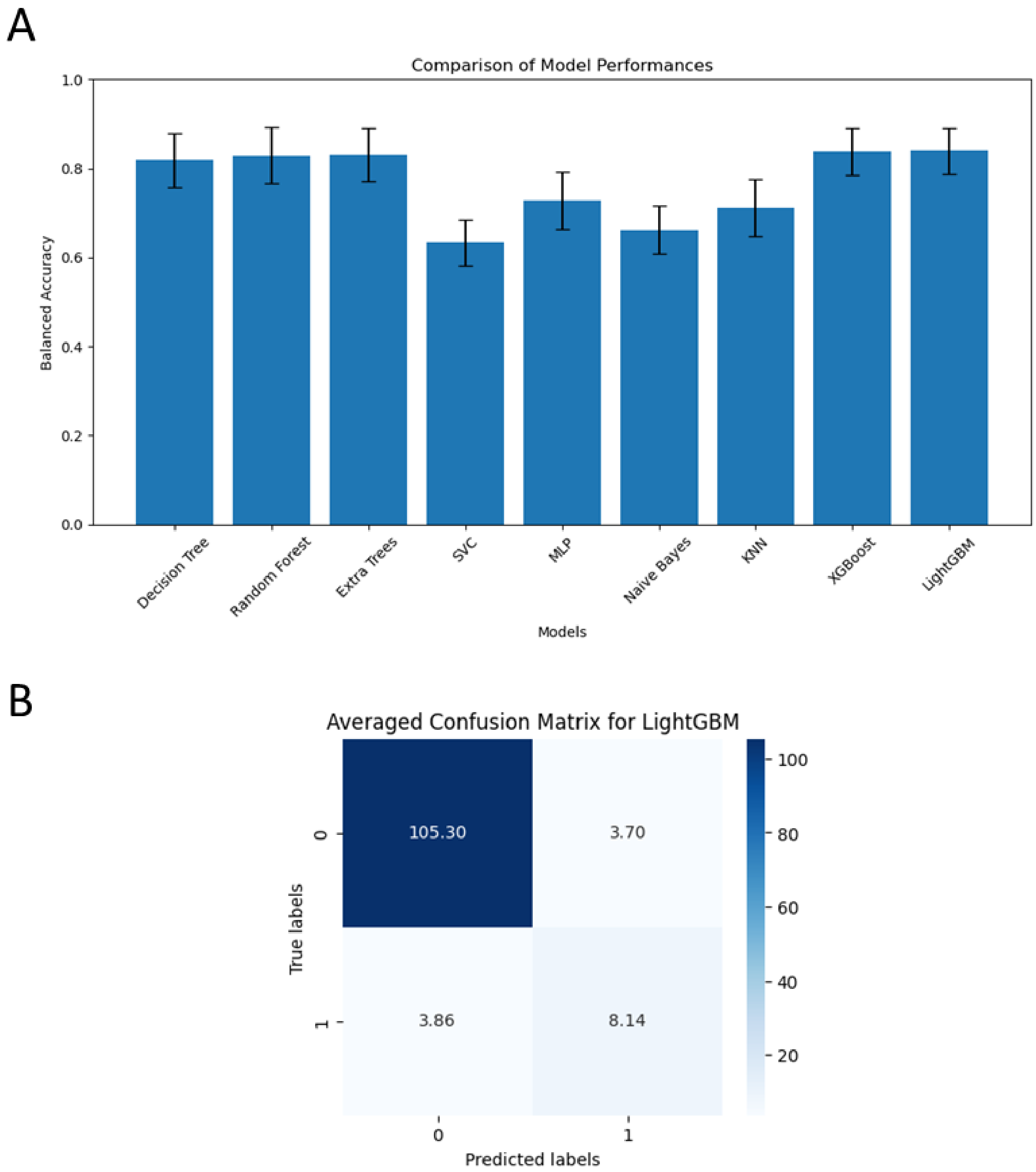


Figure IV.S1: Selection of Machine Learning Algorithm A) Comparison of different algorithms B) Averaged Confusion Matrix for default LightGBMClassifier

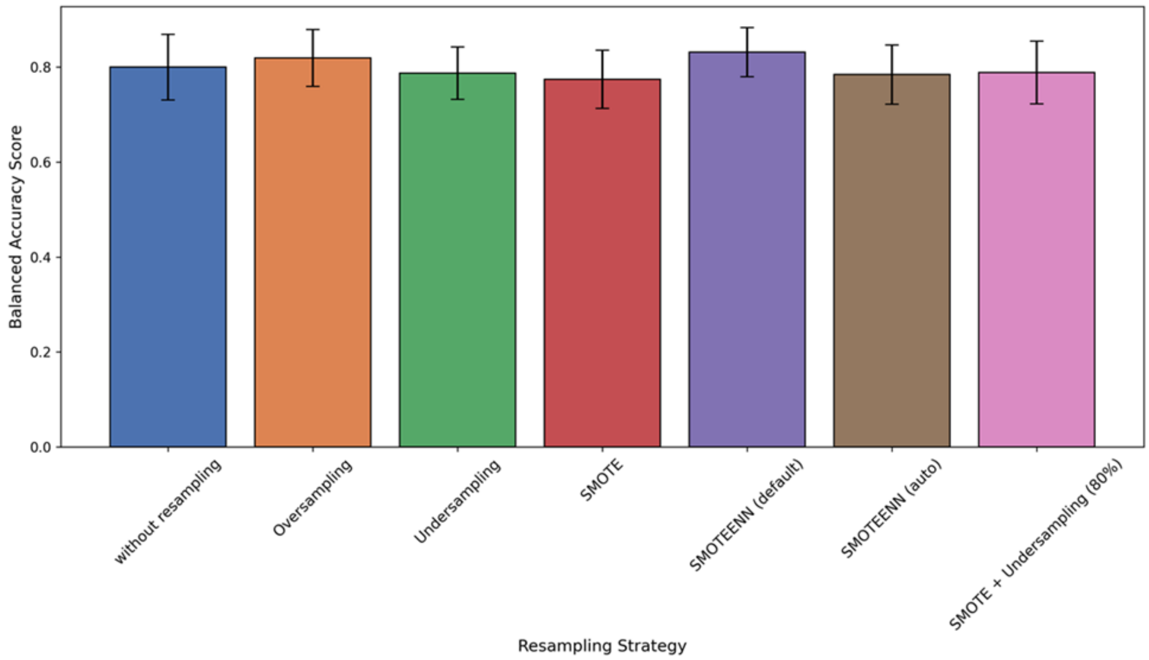


Figure IV.S2: Comparison of different resampling strategies to handle the unbalanced dataset.

Hyperparameter tuning using hyperopt:

100 evaluations were performed using the mean of 10 train test splits with replacement as objective. The hyperparameters are the following {'colsample_bytree': 0.9755088786798269, 'learning_rate': 0.1827587842746705, 'max_depth': 8, 'n_estimators': 465, 'num_leaves': 85, 'reg_lambda': 0.8324249896997891, 'subsample': 0.9331718683905172}

Figure IV.S3: Hyperparameter code for LGBMClassifier.

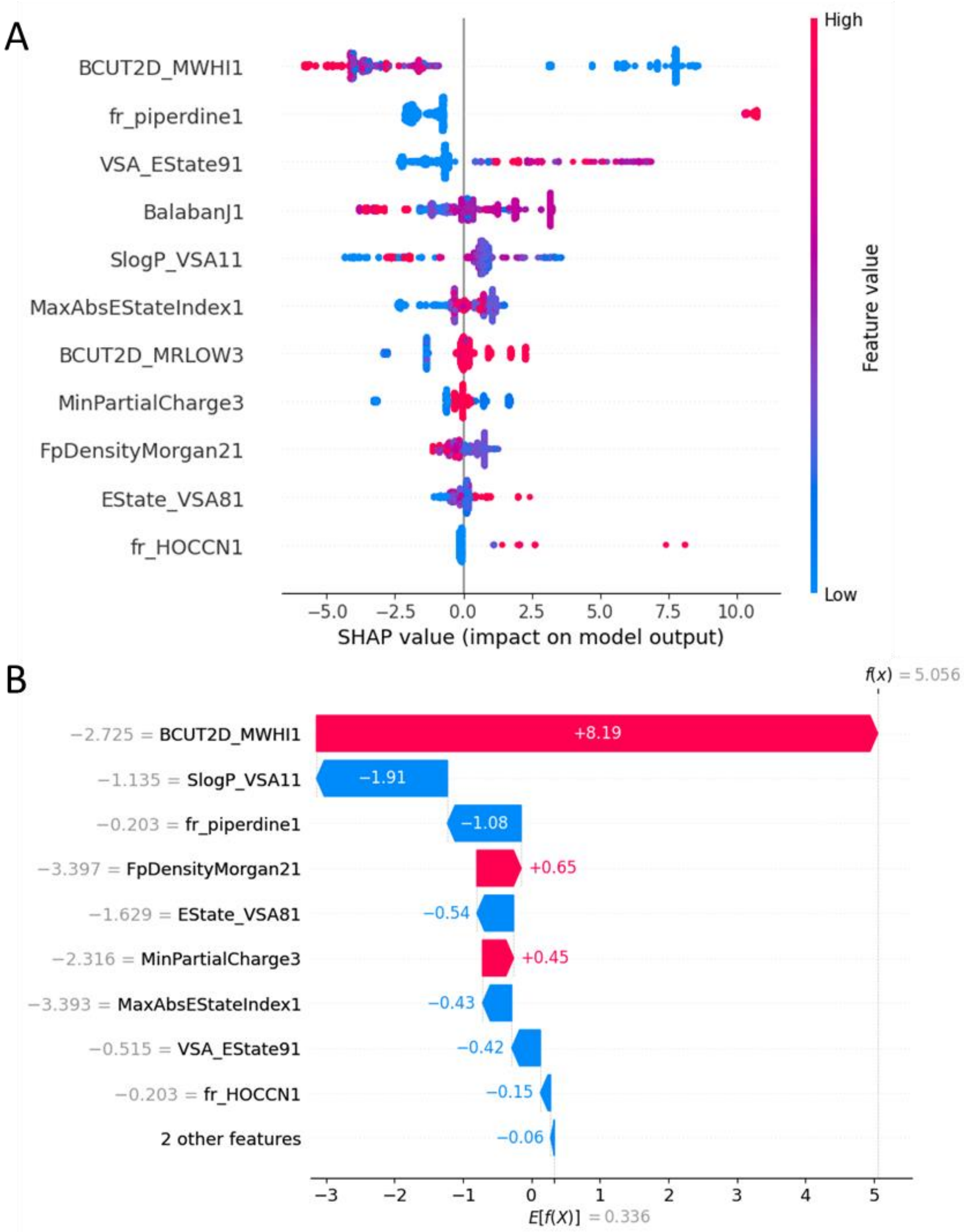


Figure IV.S4: SHAP values of the optimized model on the whole dataset (A) and for the predicted polymer as single point prediction (B).

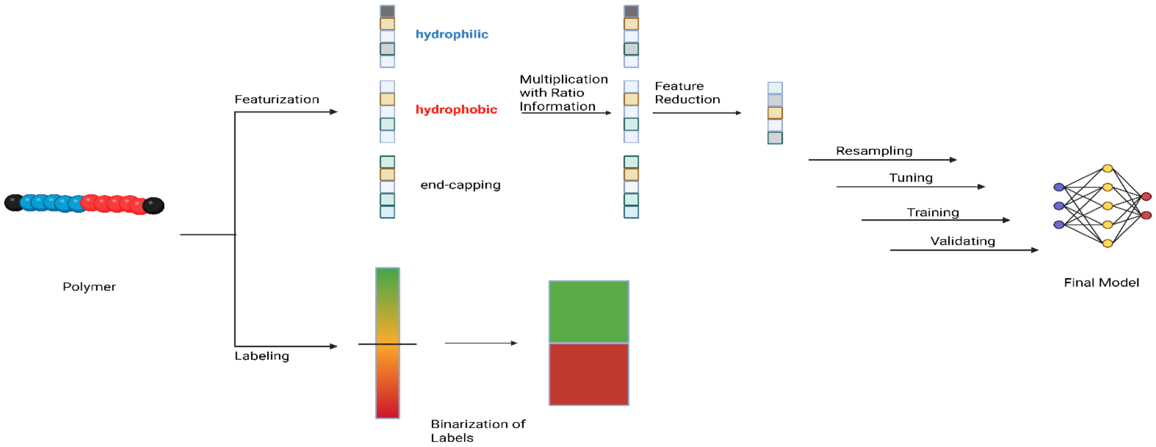
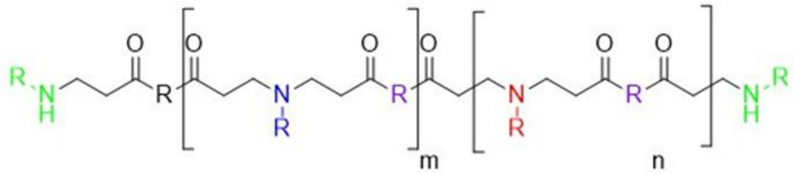


Figure IV.S5: Overview of the machine learning workflow



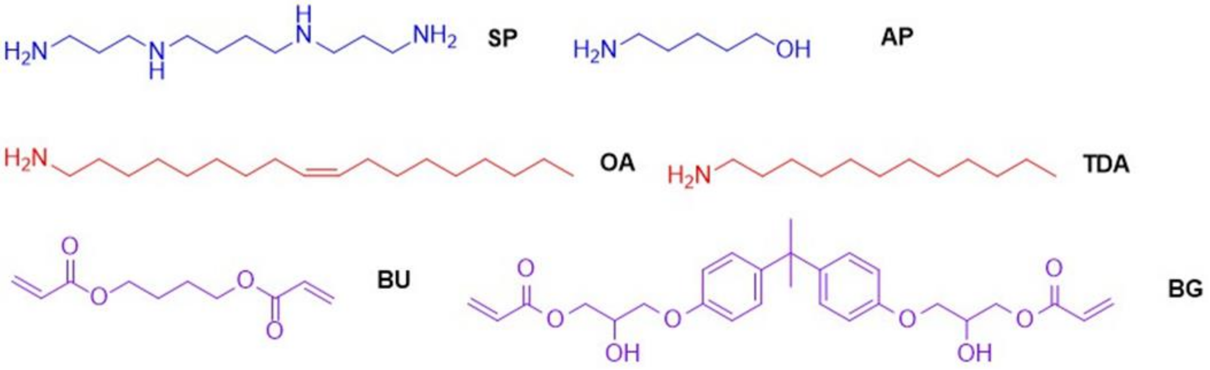


Figure IV.S6: Monomers used to design the Validation Set. SP, spermine; AP, 4-Amino-1-butanol; OA, oleylamine; TDA, tetradecylamine; Bu, 1,4-butanediol diacrylate; BG, bisphenol A glycerolate.

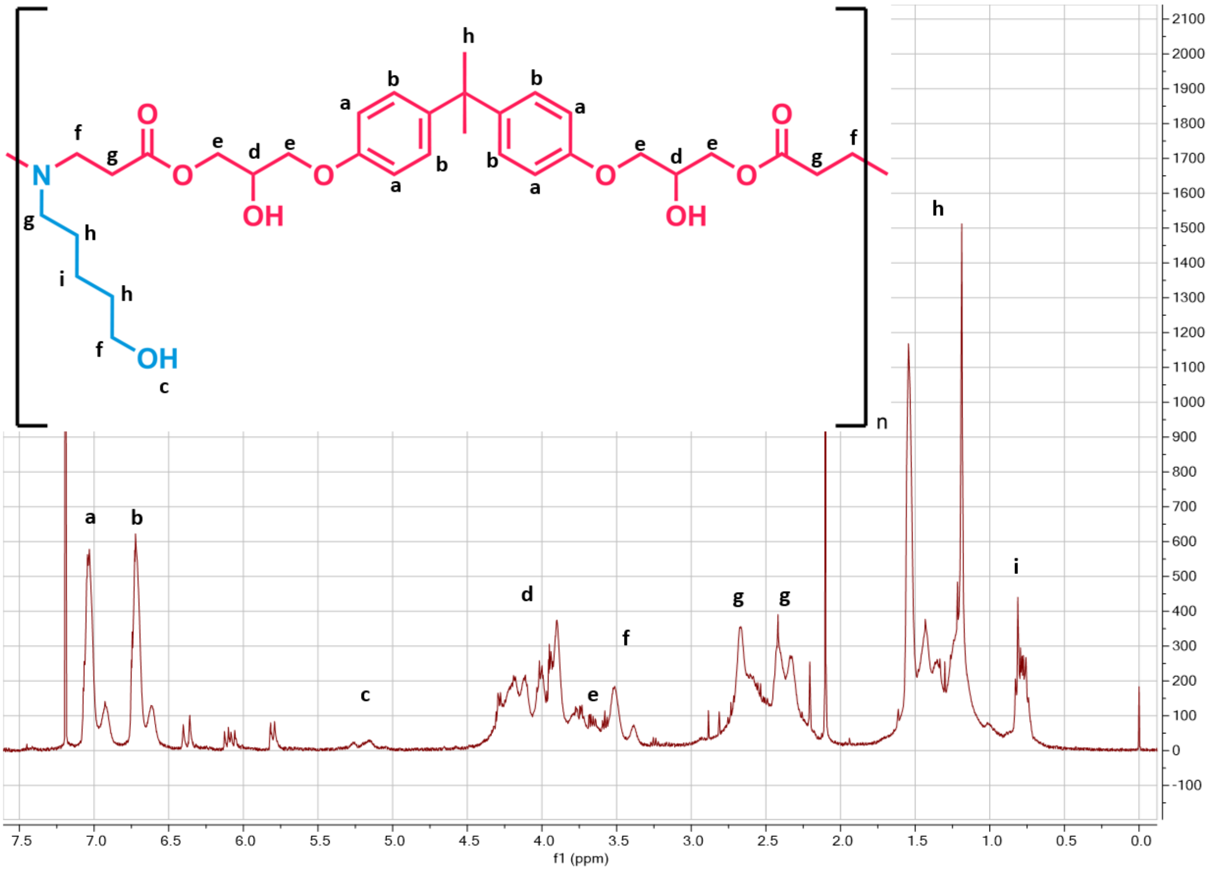


Figure IV.S7: 1H-NMR measurement of validation polymer AP-BG.

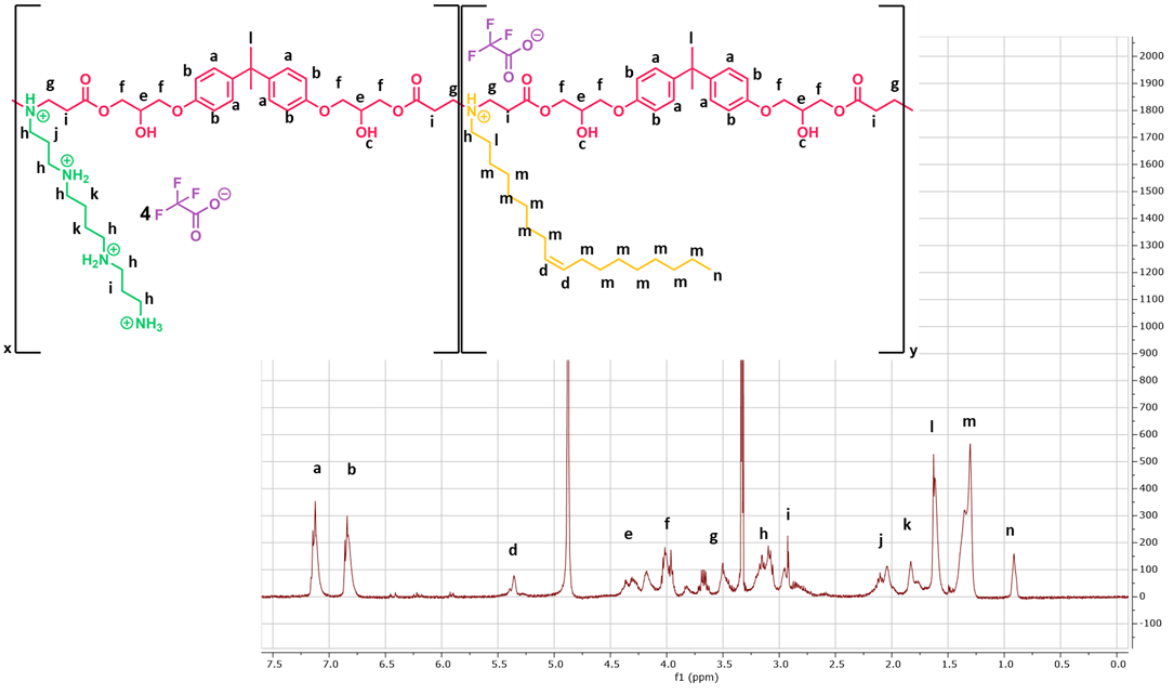


Figure IV.S8: 1H-NMR measurement of validation polymer SP-OA-BG

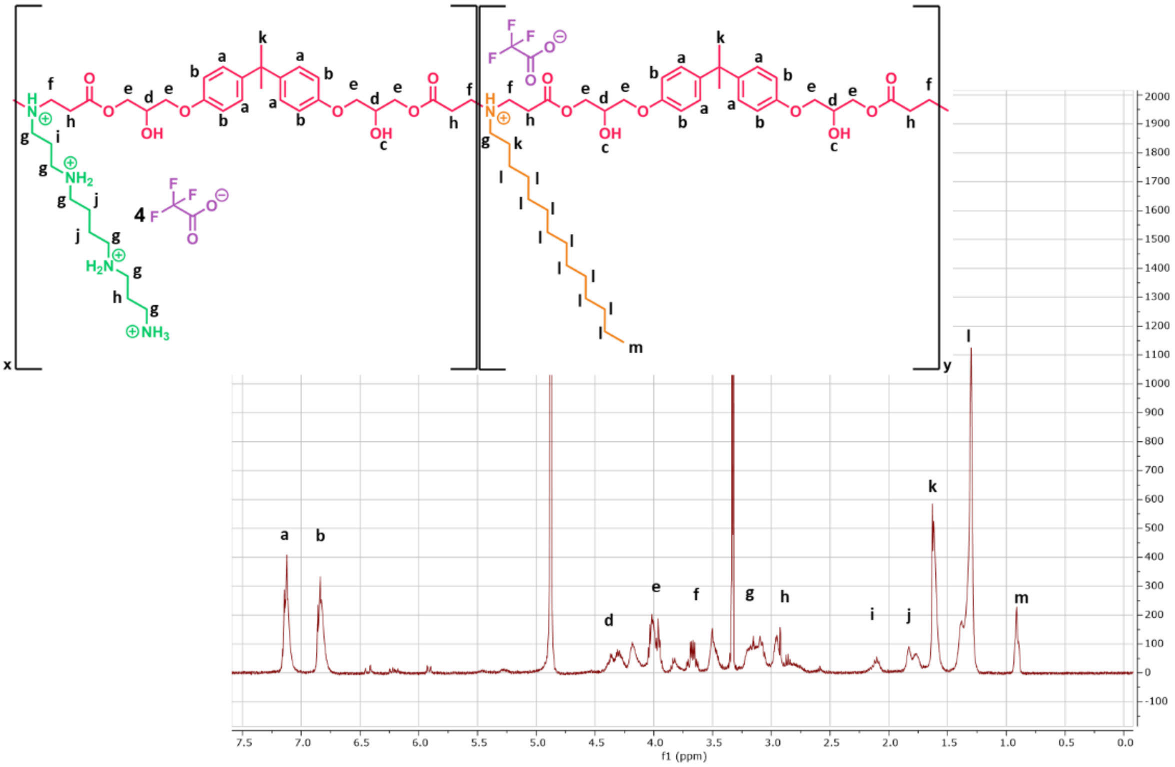


Figure IV.S9: 1H-NMR measurement of validation polymer SP-TDA-BG

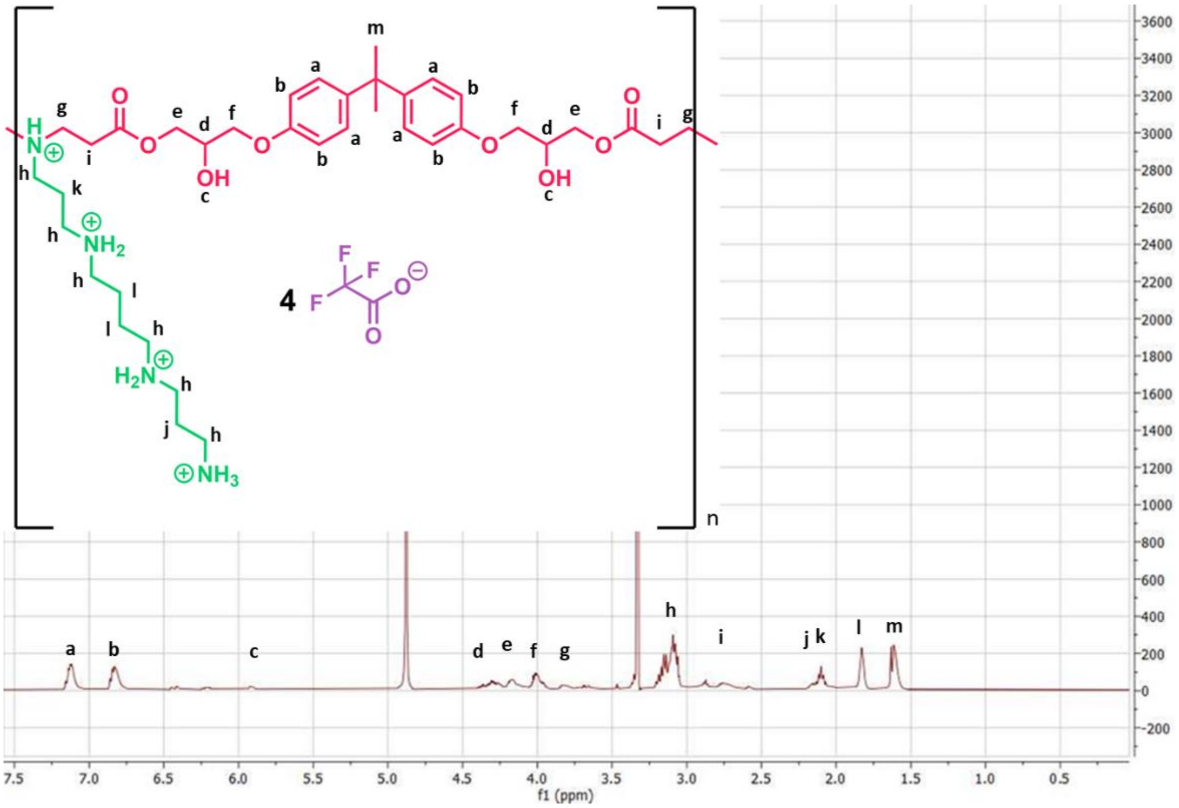


Figure IV.S10: 1H-NMR measurement of validation polymer SP-BG

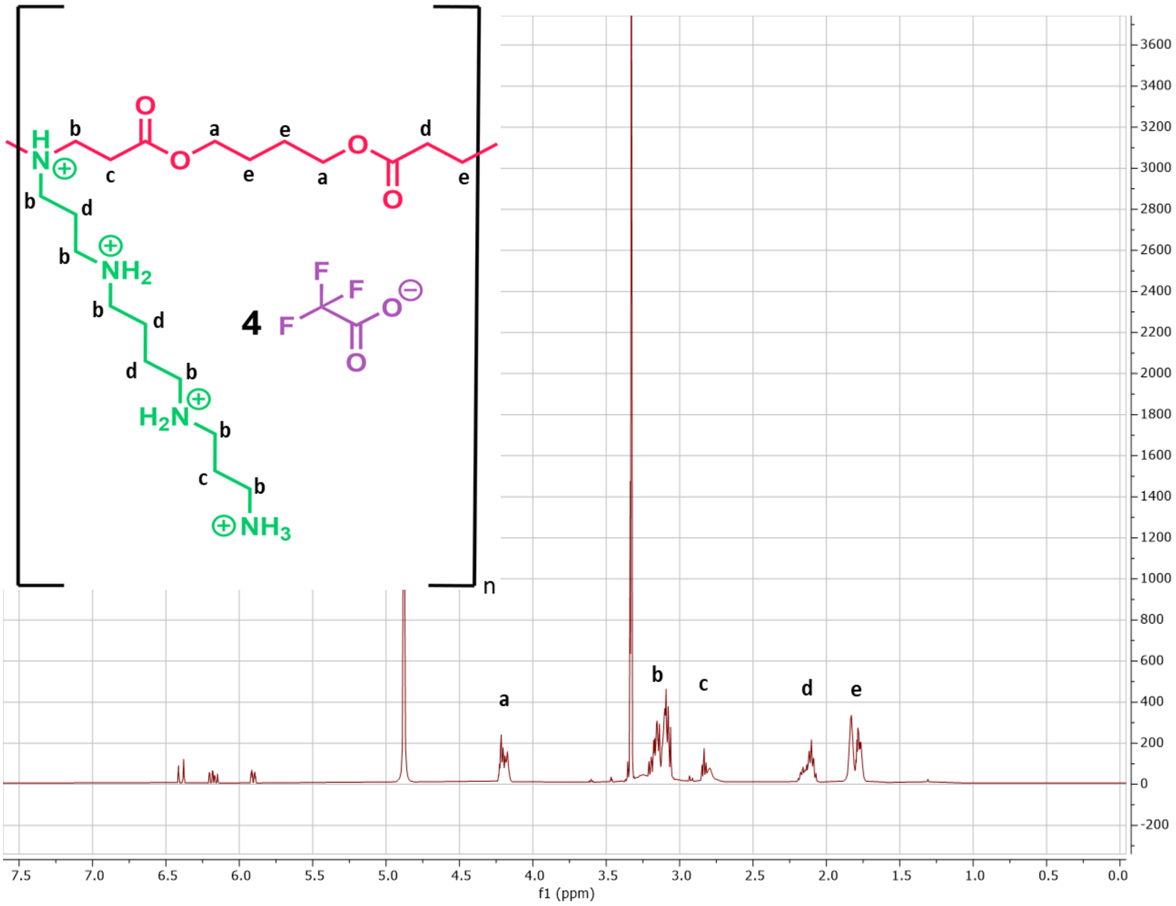


Figure IV.S11: 1H-NMR measurement of validation polymer SP-BU

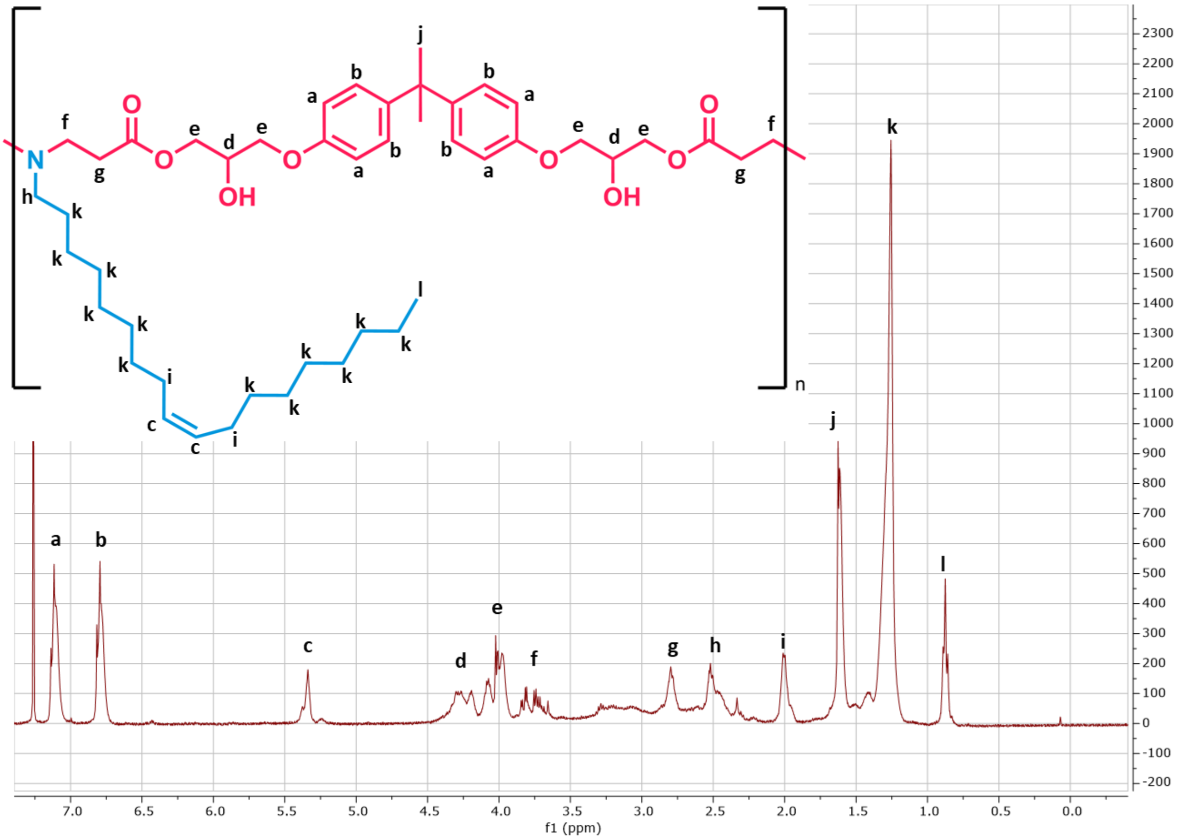


Figure IV.S12: 1H-NMR measurement of validation polymer OA-BG.

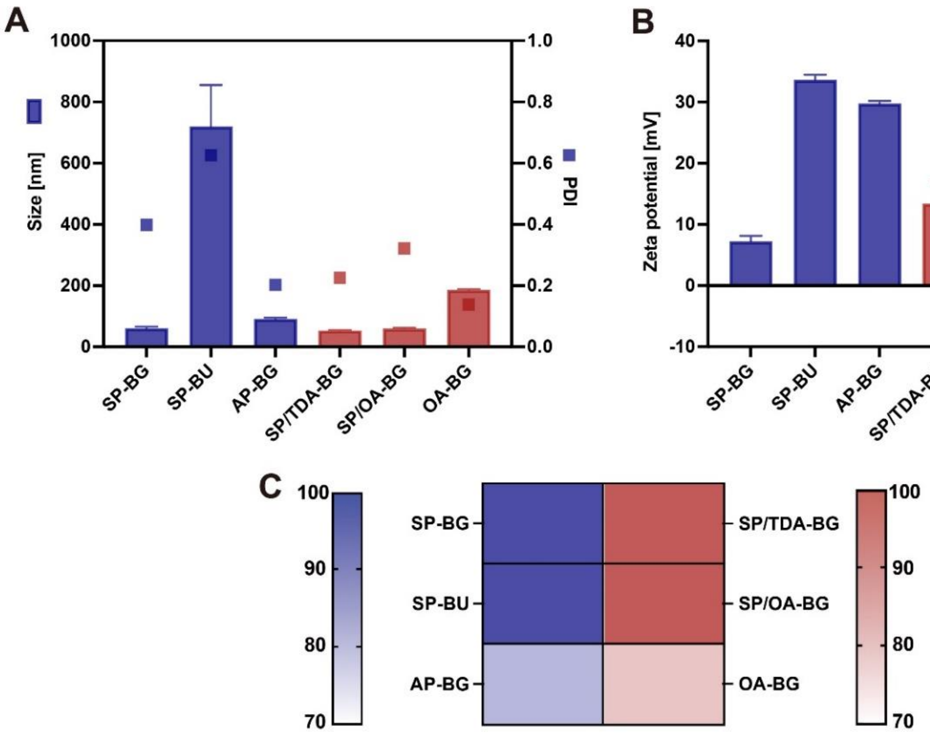


Figure IV.S13: Characterization of siRNA PBAEs polyplexes (A) Hydrodynamic diameter (represented by bar graph), polydispersity (represented by symbol) and (B) Zeta potential of siRNA-loaded polyplexes formulated at N/P ratio of 10. (C) siRNA encapsulation efficiency in the polyplexes formulated at N/P 10 with different polymers.

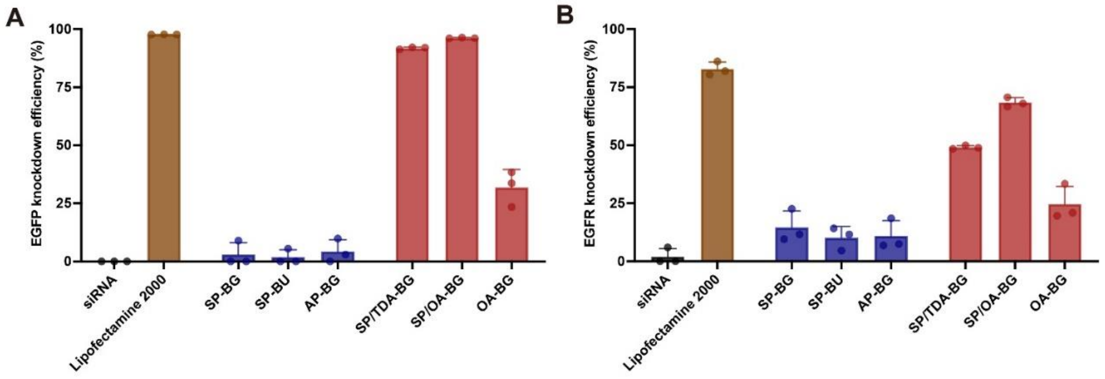


Figure IV.S14: *In Vitro* gene silencing efficiency**.** (A) Enhanced green fluorescent protein (eGFP) knockdown efficiency of siRNA polyplexes formulated at an N/P ratio of 10 in H1299/eGFP cells. (B) Epidermal growth factor receptor (EGFR) knockdown efficiency of siRNA polyplexes in A549 cells.

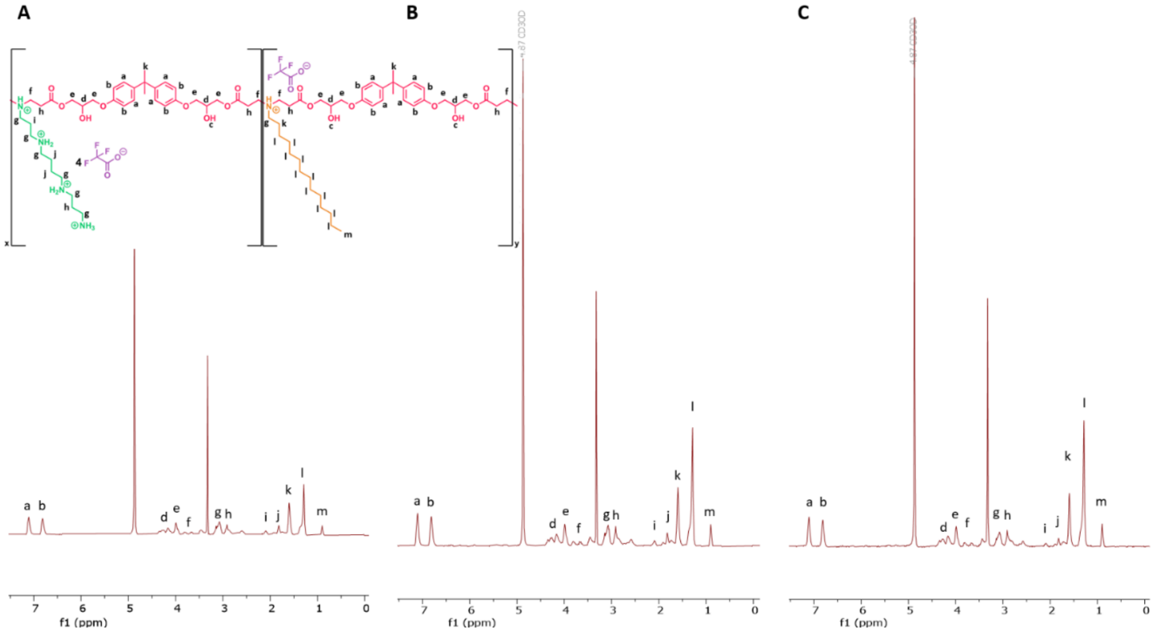


Figure IV.S15: 1H-NMR of the synthesized structures A) SP0.6/0.4TDA B) SP0.5/0.5TDA C) SP0.4/TDA0.6.

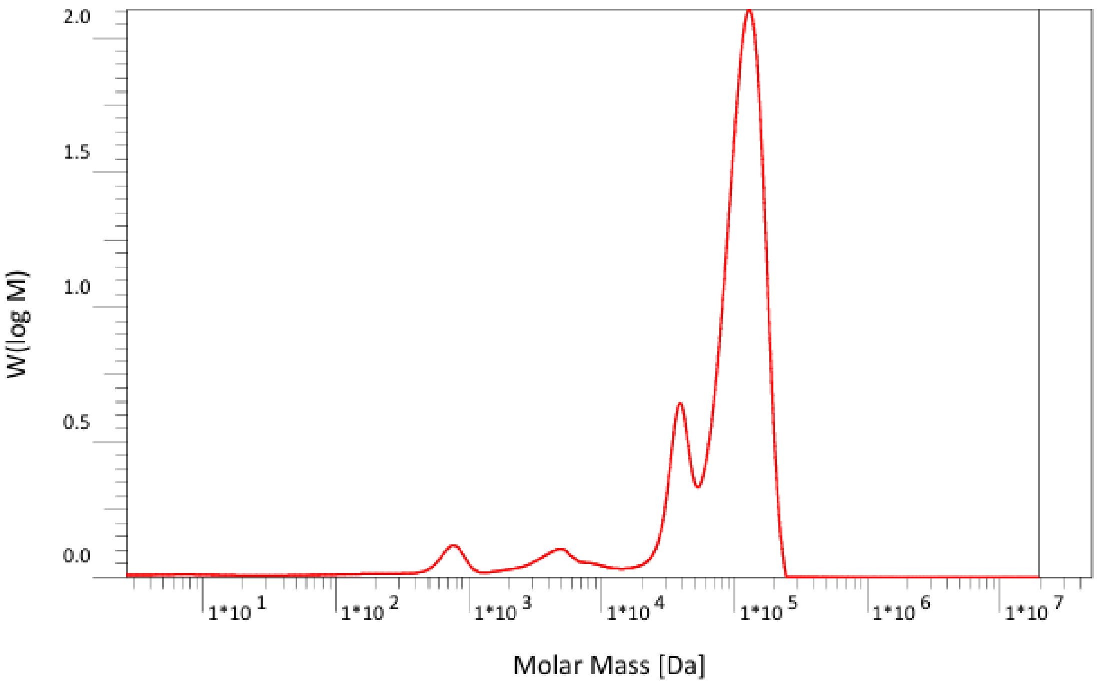


Figure IV.S16: GPC measurement of SP0.5/TDA0.5 which was tested *in vivo*.

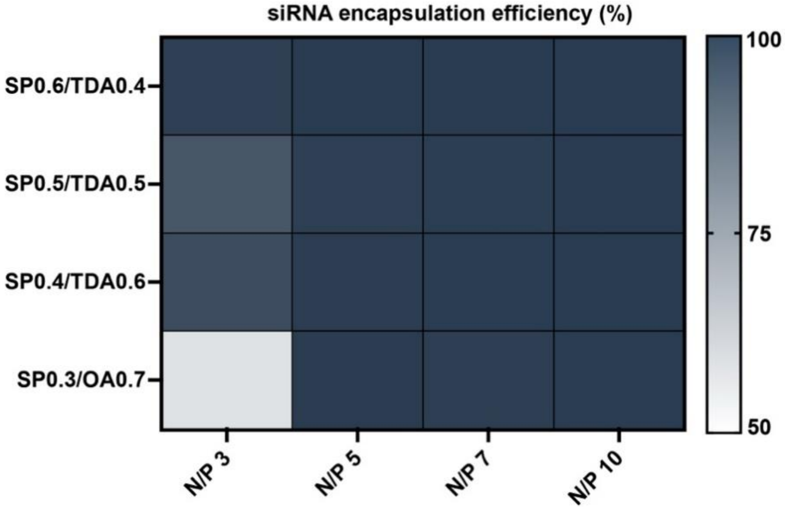


Figure IV.S17: siRNA encapsulation efficiency in the polyplexes prepared at different N/P ratios.

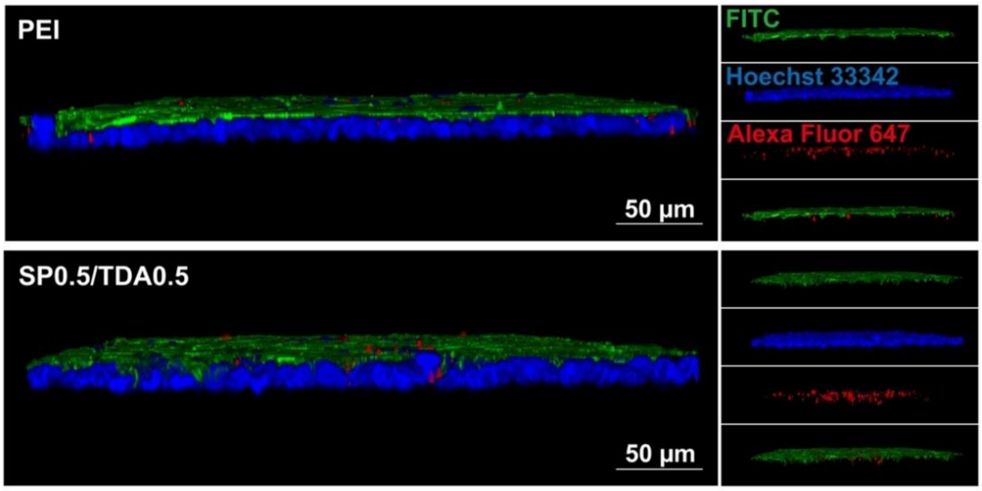


Figure IV.S18: Mucus penetration assay of siRNA-loaded PEI 25kDa polyplexes and PBAE SP0.5/TDA0.5 polyplexes in air-liquid interface (ALI) culture of Calu-3 cells. Scale bar, 50 µm.

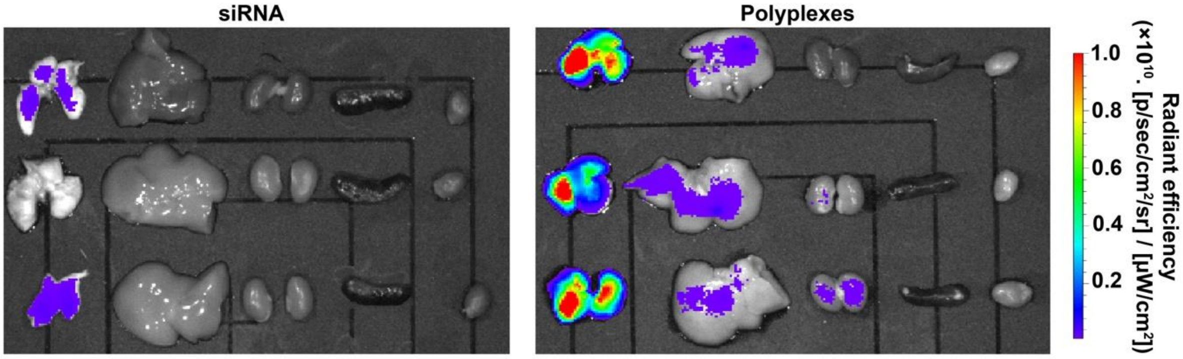


Figure IV.S19: Organ distribution after intratracheal instillation of free Alexa Fluor 647-labeled siRNA or siRNA-loaded polyplexes.

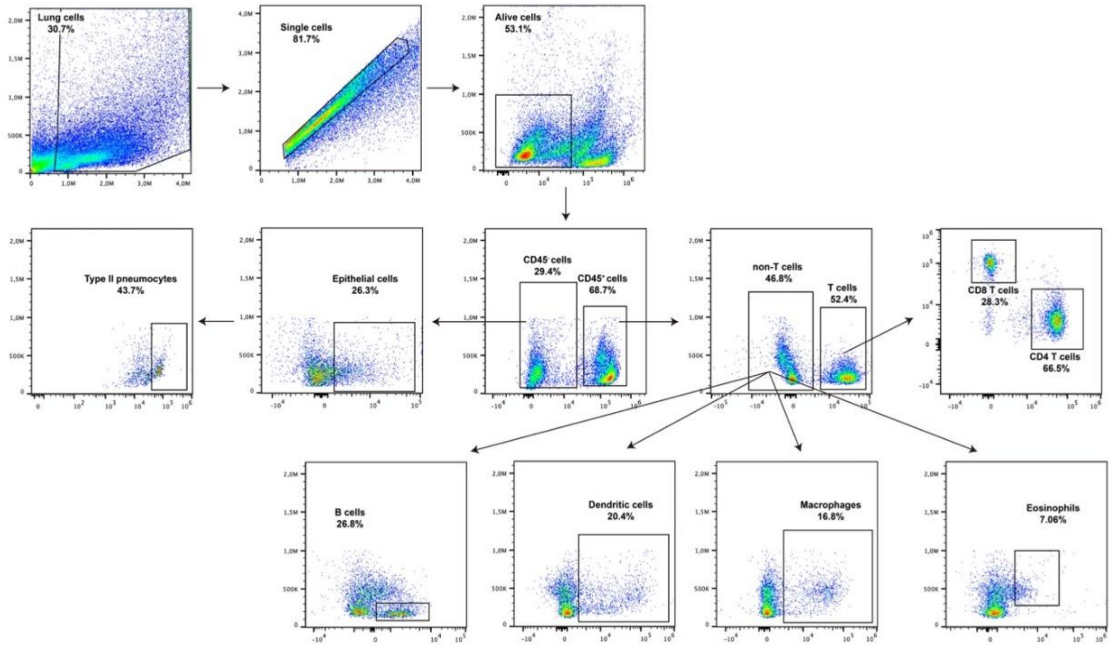
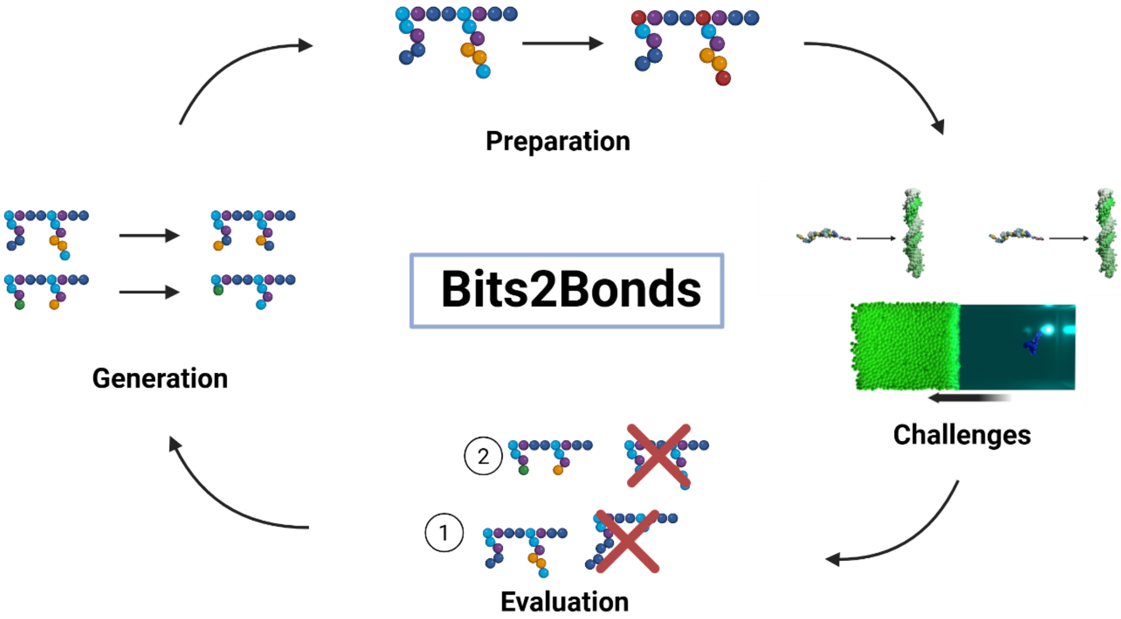


Figure IV.S20: Gating strategy of different cell types in the lung.

Chapter V - From Bits to Bonds - High throughput virtual screening of RNA nanocarriers using a combinatorial approach of Machine Learning and Molecular Dynamics

1 Graphical Abstract



2 Abstract

The implementation of high throughput methods for fuelling the design of effective nanocarriers for RNA delivery remains challenging. Traditional experimental screening is resource-intensive, while purely computational approaches face limitations, such as data scarcity for machine learning models and the high computational cost of molecular dynamics simulations. This work introduces a high-throughput virtual screening platform, "Bits2Bonds," integrating coarse-grained Molecular Dynamics (CG-MD) simulations with machine learning-driven optimization to design novel poly(β-amino ester) (PBAE) carriers for therapeutic siRNA delivery. The platform evaluates virtual polymers using MD-based "challenges" that simulate key hurdles in nucleic acid delivery such as membrane- and siRNA interaction (association/dissociation). The computational framework was calibrated and validated against experimental data, including synthesis and characterization of four distinct PBAEs, logP measurements, siRNA encapsulation assays, and cell culture knockdown experiments. This integrated approach provides a powerful tool for the *de novo* design and rapid virtual screening of optimized polymeric siRNA delivery systems.

Keywords: polyplex, Poly(beta)aminoesters, Martini 3, siRNA, nucleic acid, nanocarrier

3 Introduction

The field of RNA therapeutics has exploded in recent years, capturing the attention of researchers, pharmaceutical companies, capital providers, and the public alike. This surge in interest was ignited by milestones such as the 2018 approval of Patisiran, an RNA interference (RNAi)-based drug, and further propelled by the rapid deployment of mRNA vaccines against SARS-CoV-2[6,29,112]. This success highlights the potential of specific RNA modalities including small interfering RNA (siRNA), which holds immense promise for silencing disease-causing genes and treating previously "undruggable" targets. As of 2024, 20 RNA-based drugs are approved for clinical use, with hundreds more in development, underscoring the therapeutic potential of this class of molecules[141].

As a compelling alternative to LNPs,[118] polymeric cationic carrier systems provide advantages in tunability, complexity, and potential scalability. However, the design of functional yet safe polymeric nanocarriers remains a persistent challenge, partly due to an unclear or high toxicity of established carriers such as polyethylenimine (PEI). Hence, poly(β-amino esters) (PBAEs) have become a leading alternative.[26,81,183].

The search for improved polymeric drug delivery systems has traditionally relied on high-throughput screening (HTS) of polymer libraries[80,181]. This experimental approach, while valuable, is resource-intensive and limited by the chemical diversity of available compounds. The rise of computational power and sophisticated algorithms has enabled a powerful complementary approach: *virtual* high-throughput screening (vHTS). In vHTS, vast libraries of *virtual* molecules are rapidly assessed for their target binding, significantly accelerating the early stages of drug discovery, which has become standard practice in small molecule drug research[184–186].

However, vHTS has not been widely explored in the design of polymeric nanocarriers for RNA delivery, representing a significant gap in the field. While computational methods are used to study specific polymer-RNA interactions[155,187,188]or predict and optimize nanocarrier from data[43,127,128], a comprehensive, *de novo* virtual screening approach to identify novel, optimized polymeric carriers could be a big step forward. A primary limitation of solely data-driven methods, such as Machine Learning (ML), in this domain is the scarcity of high-quality datasets with comparable experimental conditions, annotation standards, and sufficient sample sizes, which are essential for building robust and generalizable models. Conversely, purely physics-based methods such as MD simulations are computationally highly demanding especially when using the established All-Atom approaches, which limits their use in high-throughput scenarios. Recent studies have demonstrated that integrating data-driven and physics-based approaches can not only accelerate the screening process but also provide deeper insights into underlying physical phenomena, facilitating a more systematic utilization of data[147,189]. However, realizing the full potential of these integrated approaches presents several challenges, including managing computational complexity[190], ensuring comparability between *in silico* and *in vitro* results[155,187], and establishing a virtual high-throughput framework for the discovery and optimization of PBAE based carrier systems.

This work addresses this critical need by developing and implementing a novel computational platform for the virtual screening of polyplex-forming polymers for siRNA delivery. We propose a novel approach utilizing MD-based virtual challenges to simulate the obstacles a molecule must overcome, coupled with an underlying optimization algorithm to iteratively identify high-performing structures. To enable high-throughput screening, we employed the Martini 3 force field and a simplified surrogate model of the polymers. Additionally, the optimization process was warm-started using a biased neural network trained via few-step reinforcement learning. Furthermore, we calibrated and validated the computational method to bridge the gap between *in silico* screening and experimental validation. To the best of our knowledge, this approach represents the first attempt to systematically optimize virtual polymer structures for enhanced formation of stable and effective siRNA delivery complexes.

4 Materials and Methods

The overall workflow was carried out by treating molecules as learnable Q Networks, where each output node is a probability of sampling a certain molecular fragment. When treating the process as a deterministic approach, one can see the neural network as a blueprint to build up a molecule. We first initialized the network using a reinforcement learning approach, where the model was trained to minimize the distance to a target molecule encoded as RDkit Descriptors. Subsequently, we carried out MD simulations to rate the performance of the molecules in challenging situations that are key for efficient RNA delivery. To optimize the RNA carrier molecules, we used a simple Genetic Algorithm, where random noise was added to the network weights to enable the construction of new molecules.

We ran this loop through multiple epochs, to optimize of performance score coming from the MD challenges. The simulations were validated and the calculation of the Performance Score was calibrated against wet lab experiments. The whole process is represented in Figure V.1A.

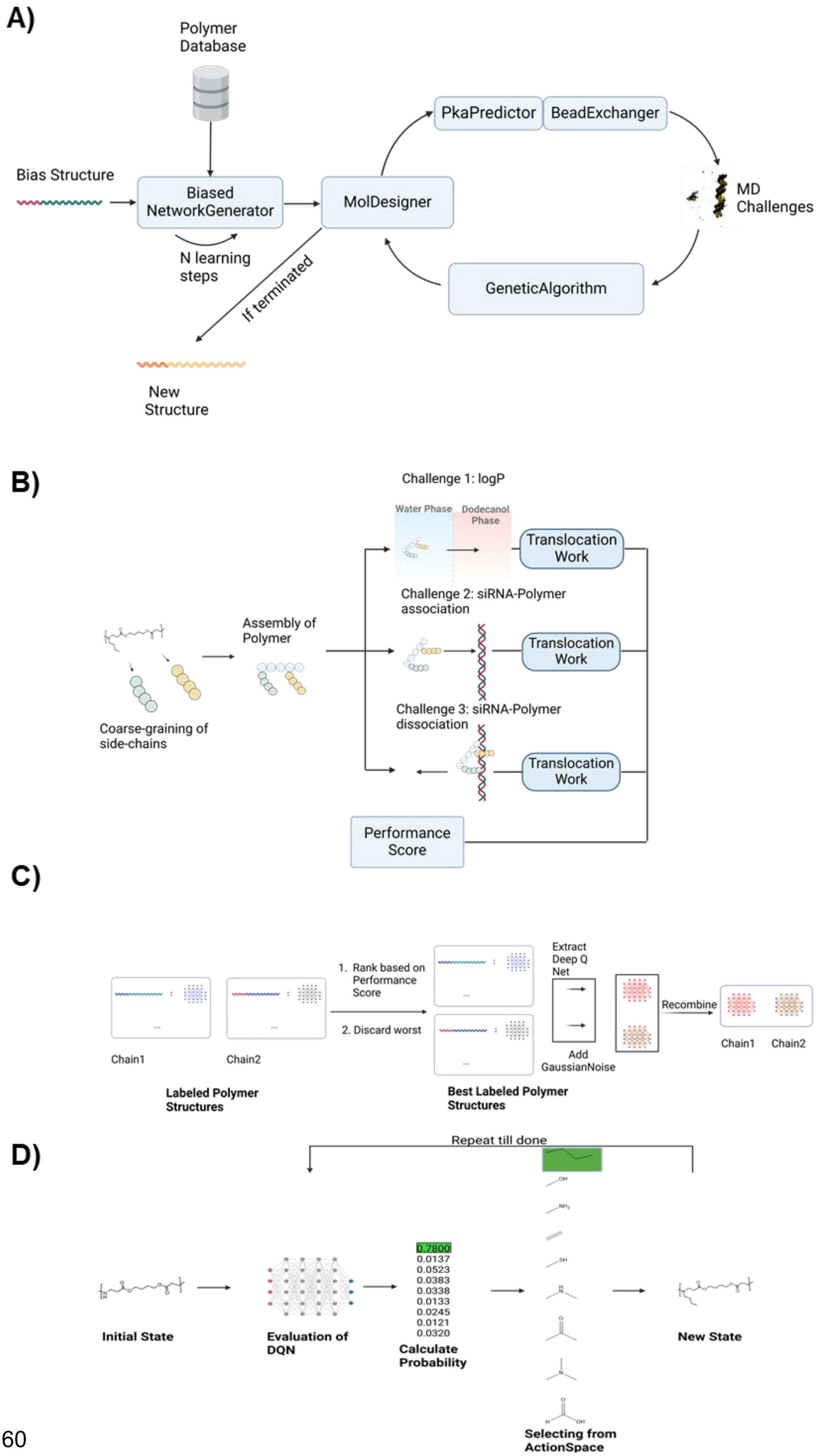

A)
Polymer Database
Bias Structure
Biased NetworkGenerator
N learning steps
MolDesigner
PkaPredictor
BeadExchanger
MD Challenges
GeneticAlgorithm
If terminated
New Structure
B)
Challenge 1: logP
Water Phase
Dodecanol Phase
Translocation Work
Challenge 2: siRNA-Polymer association
Translocation Work
Challenge 3: siRNA-Polymer dissociation
Translocation Work
Coarse-graining of side-chains
Assembly of Polymer
Performance Score
C)
Chain1
Chain2
Labeled Polymer Structures
1. Rank based on Performance Score
2. Discard worst
Best Labeled Polymer Structures
Extract Deep Q Net
Add GaussianNoise
Recombine
Chain1
Chain2
D)
Repeat till done
Initial State
Evaluation of DQN
0.7800
0.0137
0.0523
0.0383
0.0338
0.0133
0.0245
0.0121
0.0320
Calculate Probability
OH
NH2
SH
N
O
N
OH
Selecting from ActionSpace
New State

Figure V.1: Architecture of Bits2Bonds A) General overview of the software, B) Overview of the MD challenges applied, C) Overview of the Genetic Algorithm, D) Overview of the MolDesigner.

4.1 Biased Network Generation

To allow the Main Loop a warm start, a biased Q Network was constructed using a basic Reinforcement Learning approach. As a template we used an established PBAE structure[191]The reward function was designed as the cosine similarity (eq.V.1) where A represents the state vector and B the template vector based on their top 20 RDkit descriptors, which were evaluated in previous work[191]. The available action space was designed to fit common building blocks in polymer design and at the same time match available bead types in the Martini 3 force field. State representation of the molecules is selected to be a Morgan Fingerprint encoded as 2048 Bits. The network was trained using a MlpPolicy and we treated the number of timesteps as well as the number of actions as hyperparameters and observed their influence on the predicted molecules later (see Results).

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad \textbf{(eq.V.1)}$$

4.2 MolDesigner

The key element of the code is the MolDesigner, which takes the Networks as argument. Based on the network prediction, a molecule is assembled, taking the selected backbone and the available actions into account. In parallel, a bead information is designed that converts the molecular structure into a Martini3 representation. In general, MolDesigner is scalable in terms of representations that may run in parallel. The algorithm receives information from the Genetic Algorithm later on as well (Figure V.1A and V.1C).

4.3 pKa Predictor

For the MD Simulation Challenges, the molecules need to be assigned with pH dependent charges. To this end, we implemented a Graph Convolutional Network (GCN) approach to estimate the pKa values similar to that described in *Pan et al.*[192]. We further used the STONED Algorithm[193] to create 21,000 different possible side chains that were randomly charged to also allow the model to learn how to further treat a charged molecule. For pKa labelling, EPIK[194] was utilized. To allow separate protonation and deprotonation, we trained two separate models. Details about training, model architecture and model weights can be found in the Supplementary Information (Table V.S1, Figure V.S1).

4.4 Bead Exchanger

The BeadExchanger executes the information from the pKa Predictor. Beads are exchanged according to their pKa values. The return of the pKa Predictor is a list of pKa values for every protonable or deprotonable structure. The algorithm then iterates over the list as well as the respective beads and calculates the probability of being protonated using the Henderson-Hasselbalch equation:

$$pH = pK_a + \log \frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \qquad \textbf{(eq.V.2)}$$

Given a certain threshold. the model exchanges the bead in a deterministic manner. This is necessary to allow the Genetic Algorithm a comparable decision making and to stabilize the optimization. More detailed information can be found in Figure V.S2.

4.5 Genetic Algorithm

As an optimization function, we used a simple Genetic Algorithm approach. We would like to note that other algorithms could further improve the optimization process using prior knowledge from previous simulations. However, the focus of our work was to establish a system using a straightforward approach that can be improved and adapted if necessary.

To further optimize the molecules generated by Reinforcement Learning, using a scoring function from the MD challenges, we applied a Genetic Algorithm to Q-networks to manipulate the policy so that a slightly new policy was received[195]. We ranked the policies based on their performance in the MD challenges and treated the first molecule as elite, which is keeping its structure conserved. In this way, not only can the best current solution be retained, but the optimization progress can also be tracked across iterations. The mutation is carried out by adding random noise to the network weights (eq.V.3) where w’(i,j) are the updated weights on position i,j and w(i,j) are the original weights. To keep control over the mutation process, we implemented a mutation strength parameter as a hyperparameter μ. We investigated the influence of μ, which is a scalar to the GaussianNoise. We also allowed the parent molecules to switch side chains to allow additional variation in modification by introducing another binary hyperparameter. The Genetic Algorithm then returns the policies back to the MolDesigner that builds up new structures and bead models.

$$w'_{i,j} = w_{i,j} + \mu \cdot \mathcal{N}(0,1) \qquad \textbf{(eq.V.3)}$$

4.6 Molecular Dynamics

The steps described in the next section were performed fully automatically for every polymer investigated:

4.7 Creating of topology file

CG topology (.itp) and coordinate (.gro) files compatible with the Martini 3 force field were generated for GROMACS simulations using an automated Python script. This script utilized several inputs: a pandas DataFrame containing the CG polymer model definition (including residue types for lipophilic/hydrophilic chains at pH 8 and pH 4) generated by the MolDesigner module, a template .itp and .gro files representing the polymer backbone, and a separate file containing necessary bond parameters.

Topology file generation involved modifying the backbone template. New bead definitions were inserted into the [atoms] section, specifying atom type, residue number/name, charge group, and charge (e.g., +1 for SQ2p beads, 0 otherwise). The [bonds] section was subsequently populated using bond parameters (lengths and force constants) sourced from a parameter file, whose values were obtained according to the Martini 3 molecule parameterization guidelines for small molecules. Additionally, specific structural bonds linking anchor beads (residue numbers 14 and 28) to the first bead of their respective side chains were included with predefined parameters.

Coordinate files were generated by building upon the backbone template coordinates via algorithmically placing side chain bead coordinates relative to the corresponding backbone anchor points. The resulting .itp and .gro files provided the complete CG polymer description required for subsequent GROMACS simulations.

4.8 logP Challenge

The partitioning behaviour of the synthesized polymer between aqueous and hexadecane phases was analysed by MD simulations. A biphasic system comprising an aqueous layer and an organic layer (3500 hexadecane molecules) was constructed using GROMACS (2024.3) patched with PLUMED. The polymer was initially placed in the aqueous phase at a predefined position, followed by system solvation, charge neutralization, and energy minimization. The system was then equilibrated under NPT conditions, maintaining 298 K via the V-rescale thermostat and 1 bar pressure using the Parrinello-Rahman barostat.

To probe the energetics of transfer, Steered Molecular Dynamics (SMD) simulations were employed. Using PLUMED, a moving harmonic restraint of 100 KJ/mol /nm was applied to the polymer's centre of mass to guide its translocation across the water-hexadecane interface along the x-axis (normal to the interface) with a 10 fs timestep over 125,000 steps.

The work performed on the polymer during the SMD simulation was calculated by numerically integrating the force recorded by PLUMED along the x-coordinate, employing the trapezoidal rule. This calculated work profile provides an estimate of the energetic cost associated with moving the polymer between the two phases and into the hexadecane phase, testing its hydrophobicity.

4.9 siRNA association Challenge

The interaction between the synthesized polymers and siRNA was investigated using Steered SMD simulations performed in GROMACS, utilizing the PLUMED plugin. The initial polymer structure was placed within a pre-equilibrated simulation box that already contained the siRNA molecule, whose structure was obtained from a previous study[13,14]. The system underwent energy minimization first in vacuum to resolve steric clashes, followed by solvation and subsequent energy minimization in the presence of solvent. Equilibration was then carried out under NPT conditions (298 K, 1 bar) using the Berendsen thermostat and barostat to achieve a stable starting configuration for the SMD phase.

In the SMD simulations, designed to probe the polymer-siRNA interaction, a moving harmonic restraint was applied to the centre of mass of the polymer, analogous to the procedure in Section 2.2.1. The polymer was pulled along a defined reaction coordinate, oriented relative to the main siRNA axis. The work performed during this steered process was calculated by integrating the applied force along the displacement coordinate using data output by PLUMED. This yielded a work profile, providing a quantitative assessment of the polymer-siRNA interaction strength along the specified pathway.

4.10 Synthetic Accessibility Filtering

To incorporate synthetic feasibility, we computed the Synthetic Accessibility (SA) Score[196] for each candidate side chain. We integrated SA into the ranking by applying a penalty function: candidates with SA > 5 were penalized in the composite performance score, ensuring that highly complex substituents are deprioritized. We selected the threshold SA ≤ 5 to reflect moderate synthetic tractability.

4.11 Polymer Synthesis

In this study, four distinct PBAEs, considered reference polymers, with varying side chains were synthesized. The first polymer, designated AP, was derived from 5-aminopentan-1-ol as the sole side chain. The second and third polymers, OA/SP and TDA/SP, were synthesized by incorporating a 1:1 molar ratio of spermine (SP) with either oleylamine (OA) or tetradecylamine (TDA), respectively. The fourth polymer, SP, contained spermine as its only side chain.

For the synthesis, the respective reagents were dissolved in DMF. The reaction mixtures, contained in sealed vials, were stirred at 90 °C for 48 hours. Subsequently, the solvent was evaporated from the mixtures in petri dishes at room temperature over 48 hours. Polymers OA/SP and TDA/SP were deprotected by dissolving them in dichloromethane (DCM) and subsequently adding trifluoroacetic acid (TFA) (using 20 mL of DCM and 1 mL of TFA per

100 mg of polymer). These mixtures were stirred for 2 hours at room temperature, after which the solutions were evaporated at room temperature for 72 hours. The resulting solids were purified by precipitation from diethyl ether three times, followed by centrifugation (1250 x *g*, 2 min). Notably, polymer AP did not precipitate in diethyl ether; consequently, pentane was employed for its purification. Finally, all purified polymers were air-dried under a fume hood and then further dried in a vacuum oven at 40 °C for 48 hours to ensure complete removal of residual solvent. Structures and the molar side chain ratios of OA/SP and TDA/SP were analysed by 1H-NMR.

4.12 Nanoparticle Formulation

The preparation of PBAE-siRNA polyplexes involved an initial step of adjusting polymer stock solutions to various target concentrations using diethyl pyrocarbonate (DEPC)-treated water. Following this step, an equivalent volume of eGFP siRNA, previously brought to a specific concentration in 10 mM HEPES buffer (pH 5.4), was combined with the diluted polymer. These mixtures were then maintained at RT for a 30-minute period to allow for the self-assembly of siRNA-loaded polyplexes, achieving a range of polymer-to-RNA ratios, or so-called N/P ratios.

The N/P ratio, which quantifies the molar relationship between the protonable amine groups (N) of the polymer and the phosphate groups (P) of the siRNA, was a key parameter in determining the necessary polymer mass. This mass was ascertained using the following relationship:

m (polymer in pg) = n siRNA (pmol) x N/P x number of nucleotides siRNA x M protonable unit (g/mol)

Within this calculation, the number of nucleotides was considered to be 52 for the asymmetric 25/27mer siRNA used in this study. The molar mass of the protonable unit for each specific polymer was obtained by dividing the molar mass of its fundamental repeating unit by the quantity of protonable amines present in that unit.

4.13 Size and Zeta Potential Measurement

Size and zeta potential measurements were performed using a Malvern Zetasizer Ultra (Malvern Instruments, U.K.) via DLS and PALS, respectively using a pH 5.4 10 mM HEPES buffer as dispersant.

4.14 Modified SYBR Gold Assay

Determination of encapsulation was measured using a modified SYBR Gold assay. Nanoparticle solutions with various N/P ratios (50 pmol siRNA/well) were prepared at pH 5.4 and 7.4 in 10 mM HEPES buffer. After adding diluted SYBR Gold dye (8X), a 10-minute incubation in the dark was carried out. Fluorescence emission was measured using a Tecan Spark Plate Reader (TECAN, Männedorf, Switzerland) with 485 nm as excitation wavelength and 535 nm as emission wavelength. Encapsulation efficiency (EE) is the ability of the polymer to encapsulate RNA and was calculated based on the free siRNA in the sample. Note that the percent encapsulation was normalized to the amount of polymer in order to allow a fair comparison with the challenge scores, which were determined for a single molecule each. A more detailed calculation is provided in the Supplementary Information (Calculation S1). Briefly, the measured values at each N/P ratio were normalized to the fluorescence signal of 50 pmol free siRNA, multiplied by the siRNA-to-polymer molar ratio in the respective sample, and averaged across all tested N/P ratios to obtain the final EE value.

4.15 logP-experiments

For the log P assay, a calibration curve for each polymer between 0.05 mg/ml and 1.5 mg/ml in octanol was first created. Fluorescence emission was measured using a Tecan Spark Plate Reader (TECAN, Männedorf, Switzerland) at 384 nm excitation and 450 nm emission wavelengths. For all samples, a 1 mg/ml octanol solution was prepared and subsequently 100 µL of filtered 10 mM pH 5.4 HEPES buffer was added. Samples were incubated using an orbital shaker (24 hours at 250 rpm). Using the calibration curve, the polymer concentrations in the two phases were analysed and logP values were calculated.

4.16 In vitro eGFP Knockdown

Gene Knockdown experiments were conducted using H1299 cells stably expressing enhanced green fluorescent protein (eGFP). Nanoparticles were formulated with siRNA targeting eGFP mRNA or scrambled siRNA with the same length. H1299/eGFP cells were seeded in 96-well plates at a density of 6,000 cells per well and then incubated with polyplexes containing 20 pmol siGFP or 20 pmol of a negative control RNA (siNC) for 48 h. Lipofectamine 2000 was used as a positive control, while free siRNA served as a negative control. After incubation, the cells were collected by trypsinization to perform Flow Cytometer analysis of eGFP expression (Attune NxT Flow Cytometer, ThermoFisher Scientific). The eGFP knockdown efficiency was calculated by dividing the Median Fluorescence Intensity (MFI) of the siRNA-treated groups by that of the respective siNC-treated group.

5 Results and Discussion

5.1 Synthesis of Polymers and Nanoparticles

To validate the applicability of our software in practical experimental workflows, we synthesized eight distinct PBAEs. NMR spectroscopy confirmed the expected monomeric ratios (Figures V.S3-V.S10). The polymers were selected to represent a broader range of amphiphilic properties. AP-BG (Figures V.S3 and V.S11), OA-BU (Figures V.S9 and V.S17) and OA-BG (Figures V.S10 and V.S12) were chosen as representatives for hydrophobic polymers due to their low amine content, which limits protonation and consequently polarity. OA/SP-BG (Figures V.S4 and V.S12), TDA/SP-BG (Figures V.S5 and V.S13) and OA/SP-BU (Figures V.S8 and V.S16) exhibit a more balanced amphiphilic character that has been shown to be favourable for effective gene knockdown both *in vitro* and *in vivo*[191]. SP-BG (Figures V.S6 and V.S14) and SP-BU (Figures V.S7 and V.S15) were selected for their significant hydrophilicity, a characteristic typically associated with reduced *in vitro* knockdown efficacy[9]. When formulated with siRNA, all polymers, except the OA polymers, formed well suited particles with a hydrodynamic size < 100 nm (Figure V.S19) and a PDI < 0.2 (Figure V.S20) at higher N/P ratios.

5.2 Validation of the Hydrophobic Interface Challenge with logP data

Hydrophobicity influences RNA delivery with polymeric nanocarriers[197] because the nanoparticles must overcome barriers of amphiphilic membranes.

One example for such a barrier is the endosomal membrane, which the carrier system has to overcome, to escape the endosome and successfully deliver the cargo into the cytosol.

An established hypothesis for the enhanced endosomal escape of amphiphilic nanocarriers, is the interaction with phospholipids within the endosomal membrane[31].

To this end, we conducted a simulation that investigates the work required for a carrier system to move through a hydrophilic-lipophilic interface at low pH[28] and validated the results against experimental logP data (Figure V.2A). The hydrophilic SP-BG and SP-BU showed logP values of -1 and -0.6, while the amphiphilic OA/SP-BG, OA/SP-BU and TDA/SP-BG were more balanced with logP near zero. The hydrophobic OA polymers showed the highest logP of 3, with almost all sample in the octanol phase. (Figure V.2A, bars). Correspondingly, OA polymers required the lowest work to be pulled through the hydrophobic part of the biphasic system and SP polymers the highest (Figure V.2A, line). Furthermore, the medium logP values determined for OA/SP-BG, OA/SP-BU and TDA/SP-BU were consistent with the simulation results.

5.3 Validation of the siRNA Challenges with Encapsulation Efficiency data

An important criterion for a successful nanoparticulate siRNA delivery is the encapsulation and protection of cargo and at the same time cargo release into the cytosol to allow the formation of the RISC complex[198,199]. Correspondingly, we introduced two challenges where we measured the interaction of polymer and siRNA at pH 5.4 to mimic the formulation conditions, and at pH 7.4 to model neutral environments such as the cytosol[200].

The experimentally determined EE values (Figure V.2, bars) reflected the simulation results (Figure V.2, lines), showing the same trends at pH 5.4 (Figure V.2B) and pH 7.4 (Figure V.2C). Due to the high amine density, the SP polymers showed high EE at both pH values, with 1.24×10^{-2} and 1.09×10^{-2} encapsulated siRNA, respectively for SP-BG and a similar trend for SP-BU. In contrast, the EE of OA-BG was only 2.93×10^{-4} at pH 5.4 and 7.21×10^{-4} at neutral pH(). Furthermore, AP and the OA polymers showed a positive work requirement for both challenges, indicating no measurable RNA-polymer interaction in both experiment and simulation.

The three amphiphilic polymers showed balanced EE at both pH values (OA/SP-BG: 6.62 × 10^{-3} and 6.09 × 10^{-3}; TDA/SP-BG: 8.36 × 10^{-3} and 7.04 × 10^{-3}; OA/SP-BU: 3.76 x 10^{-3} and 3.05 x 10^{-3}), reflected by challenge values of −36.45 kJ/mol, −45.37 kJ/mol and -59.77 kJ/mol for OA/SP-BG, TDA/SP-BG and OA/SP-BU, respectively, at pH 4, and −47.67 kJ/mol, −30.48 kJ/mol and -41.57 kJ/mol, respectively, at pH 8.

5.4 Calibration and Fitting

Following the successful synthesis and formulation of distinct polymer-siRNA nanoparticles (Section 3.1), their functional efficacy was evaluated through cell culture experiments. Distinct performance levels consistent with the polymers' designed characteristics were revealed: OA/SP-BG and TDA/SP-BG with balanced amphiphilic character, demonstrated high GFP knockdown efficiency while the highly hydrophobic polymers AP-BG, OA-BG, and OA-BU as well as the significantly hydrophilic polymers SP-BG and SP-BU showed negligible activity, as anticipated. Interestingly, the amphiphilic SP/OA-BU showed only a small knockdown of 20.9% (Figure V.S13).

To convert the raw outputs of our MD simulations into a single quantitative predictor of polymer performance, we combined the three challenge outputs into a composite scoring function. We then calibrated this scoring function to the siRNA knockdown data, so that higher scores correspond to greater knock-down efficiency. This scoring function was designed to reward performance that closely matches the ideal target values observed in successful polymers (OA/SP-BG, TDA/SP-BG), while penalizing substantial deviations. To naturally capture the optimal amphiphilic behaviour of PBAEs, we selected a multi-dimensional Gaussian distribution as the basis of our scoring function. Specifically, it comprises a three-dimensional Gaussian reward component—centred on predetermined optimal values for the MD readouts (see eq.V.4). The centres and widths (sigmas) of the Gaussian component, which represent the target profile derived from the experimental winners, were held constant.

Calibration was performed using data from eight polymers spanning diverse side-chain chemistries and a range of hydrophobicity/cationic density. We estimated the amplitude by non-linear least-squares (scipy.optimize.curve_fit, default settings), minimizing the discrepancy between the scoring function output and a target metric defined as the Euclidean distance of each polymer's MD performance vector from an ideal reference profile. To improve agreement between simulation and experiment, the pH settings used by the BeadExchanger were adjusted to match the experimental buffer conditions, reducing systematic bias in predicted protonation states. The resulting fitted performance function provides a continuous score based on the three MD readouts.

$$S_{reward(x,y,z)} = 37.3917 * exp[-((x + 40)^2 / 1800 + (y + 40)^2 / 1250 + (z - 155)^2 / 450)] \qquad \textbf{(eq. V.4)}$$

Here, x, y, and z are the performance metrics for siRNA association (pH 4), siRNA dissociation (pH 8), and membrane interaction and $S_{reward(x,y,z)}$ is the performance score. The Gaussian reward function uses an amplitude of 37.3917 and is centred at x=-40, y=-33, and z=159. The spread of the reward is determined by standard deviations of 30 (for x), 30 (for y), and 30 (for z) in each respective dimension. The fit based on eight polymers provides a solid groundwork for mapping MD readouts to experimental knockdown; nonetheless, the sample size remains modest. We therefore explicitly acknowledge this limitation and plan to expand the calibration set and endpoints in subsequent iterations as additional data become available.

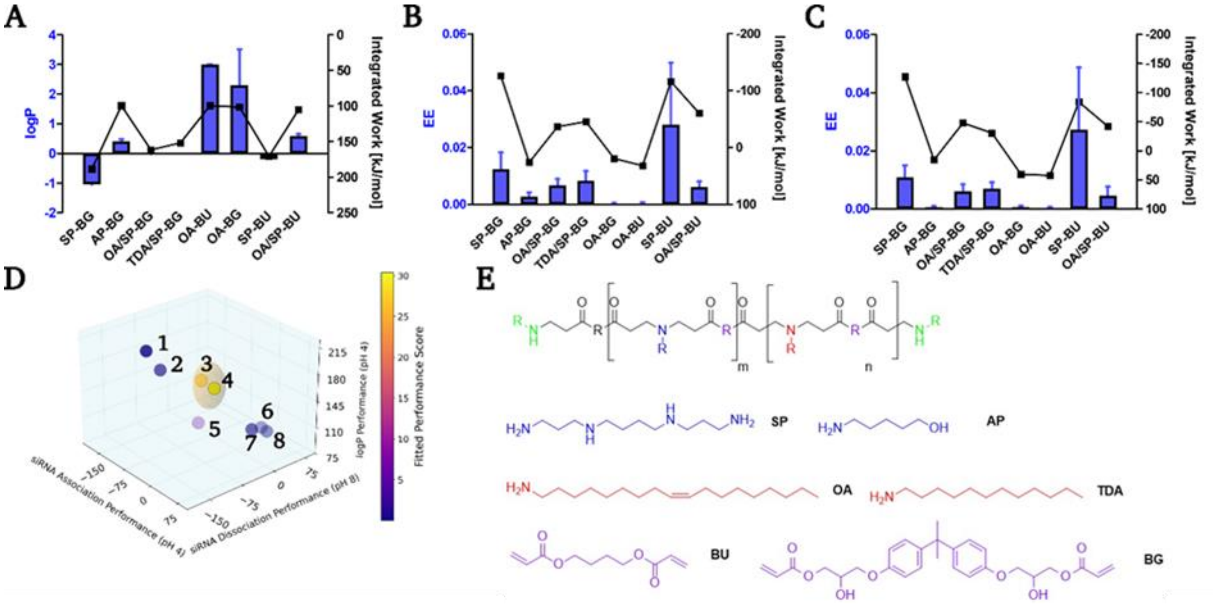


Figure V.2: Validation of the Model Approach A) Results of logP experiments (bar) together with the logP Challenge (line), B) Results of SYBR Gold Experiments at pH 5.4 (bar) together with the Association Challenge (line), C) Results of SYBR Gold Experiments at pH 7.4 (bar) together with the Dissociation Challenge (line), Each experiment was conducted 3 times and the mean and the standard deviation are reported here. D) Location of the eight reference polymers in the 3D performance space (siRNA association pH 4 vs. siRNA dissociation pH 8 vs. membrane interaction), coloured by their fitted performance score. 1:SP-BG 2:SP-BU 3 OA/SP-BG 4 TDA/SP-BG 5: OA/SP-BU 6:OA-BG 7:OA-BU 8: AP-BG. E) Components and corresponding nomenclature used for the synthesis of validation and calibration polymers.

5.5 Assessment of generated structures

5.5.1 Assessment of Cutoff value impact on generated structures

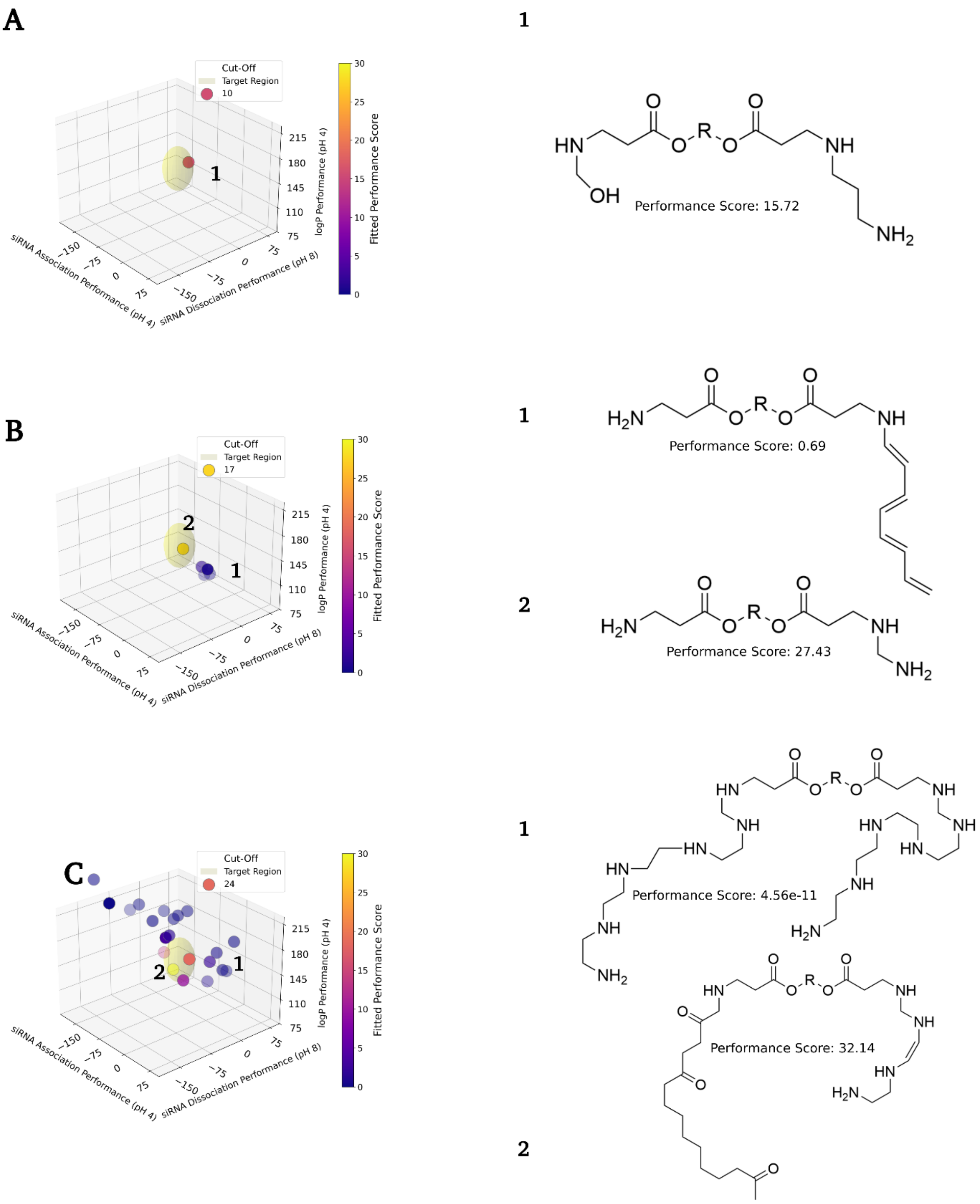


Figure V.3: Predicted polymer performance landscape for cutoff values. (A), (B), and (C) show results for Performance Score cutoff 10,17, and 24. Left: 3D performance space (siRNA association at pH 4 vs. siRNA dissociation at pH 8 vs. logP Performance) for iteratively generated structures. Right: Computed polymer structures and predicted Performance Score.

In our hyperparameter optimization, we first evaluated the impact of the performance score cutoff, here 10, 17, and 24, which determines the minimum performance score (calculated using the scoring function calibrated in Section 3.4) a generated polymer must achieve for the optimization process to potentially terminate or be considered successful. The choice of cutoff significantly influences the nature of the polymers generated. A higher cutoff, such as 24 (Figure V.3C), demands greater performance, potentially driving the optimization towards structures with high chemical similarity to the best-performing calibration polymers (Figure V.2D). Conversely, cutoff 10 (Figure V.3A) imposes a less stringent requirement, allowing the algorithm to accept structures that might be more distinct from the initial high-performers. To balance rigorous performance criteria with exploration of novel chemistries, we used a cutoff of 17 during hyperparameter tuning and 24 for the production run.

5.5.2 Assessment of Mutation Strength impact on generated structures

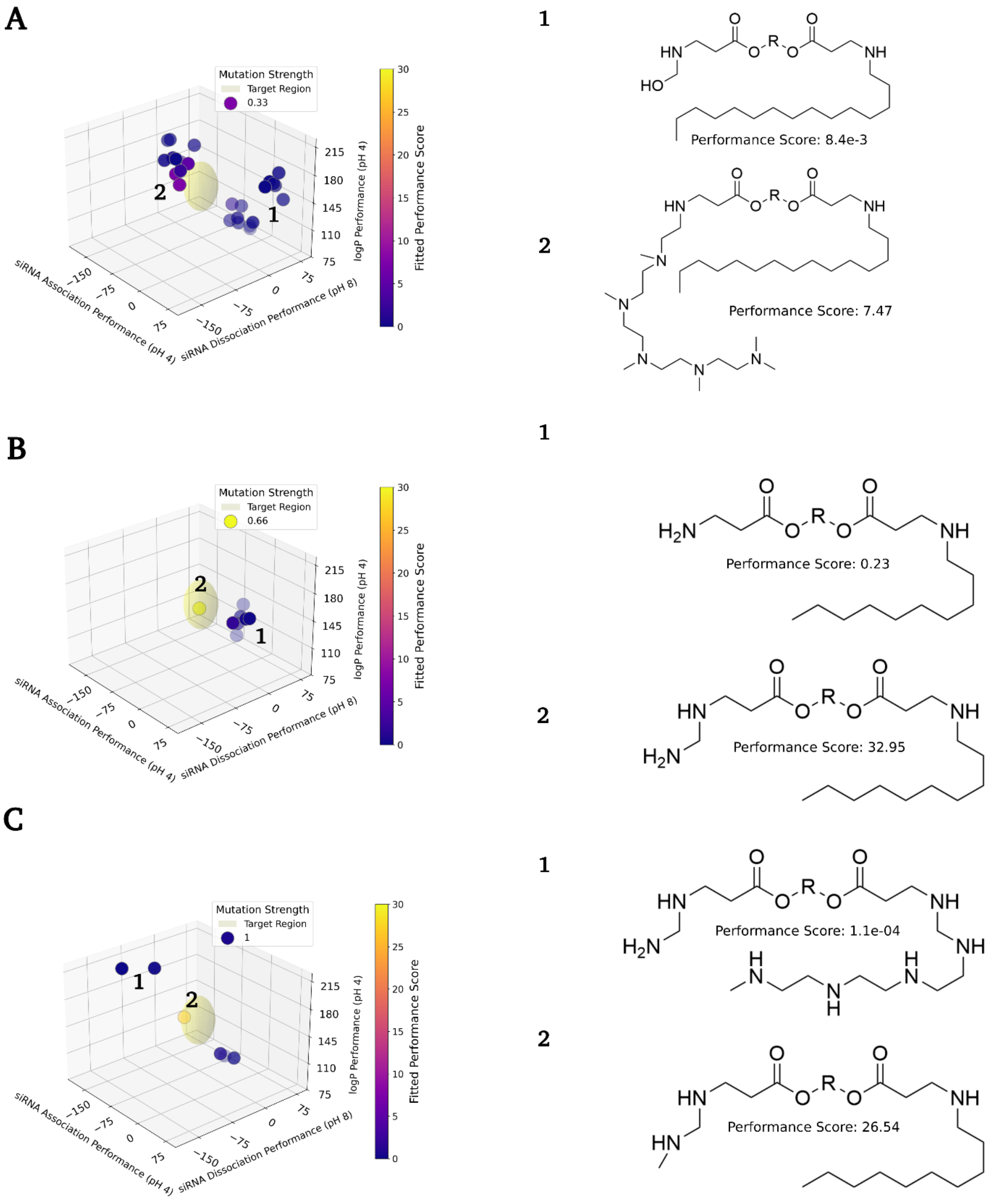


Figure V.4: Predicted polymer performance landscape for different mutation strengths. (A), (B), and (C) show results for mutation strengths 0.33 , 0.66,and 1 , respectively. Left: 3D performance space (siRNA association at pH 4 vs. siRNA dissociation at pH 8 vs. logP Performance) for iteratively generated structures. Right: Computed polymer structures and predicted Performance Score.

To assess our ML-MD combination effectiveness in exploring the chemical space for optimal polymers, we systematically varied the mutation strength parameter, which governs the extent of structural modifications during polymer generation, influencing the diversity of candidates produced. As expected, the lowest mutation strength (0.33) confined the generated polymers to a limited region within the multi-objective performance space (defined by pH 4 association, pH 8 dissociation, and membrane interaction metrics), clustering results closely together. This lack of dispersion, visualized in the performance space plot (Figure V.4A), indicated insufficient exploration beyond initial or similar structures. The maximum performance score of structures generated at the lower mutation strength (0.33) was only 7.47. Increasing the mutation strength to 0.66 enabled broader exploration across the performance space by allowing the generation of more diverse chemical motifs, such as those incorporating amine groups and alkyl side chains (Figure V.4B). This, in turn, yielded polymers with generally higher performance scores, exemplified by one candidate reaching 32.95 after eight episodes. The highest mutation strength tested of 1 rendered a polymer surpassing the set performance score threshold of 17 (Figure V.4C) in six episodes. This high-performing structure combined two key chemical motifs: a hydrophilic amine side chain and a substantial aliphatic hydrophobic side chain, likely contributing to its favourable predicted properties. These results demonstrate that a sufficiently high mutation strength (1 in this study) is crucial for escaping local optima and identifying high-performing candidates within a polymer design challenge.

5.5.3 Assessment of Mol Designer Stepsize Impact on generated Structures

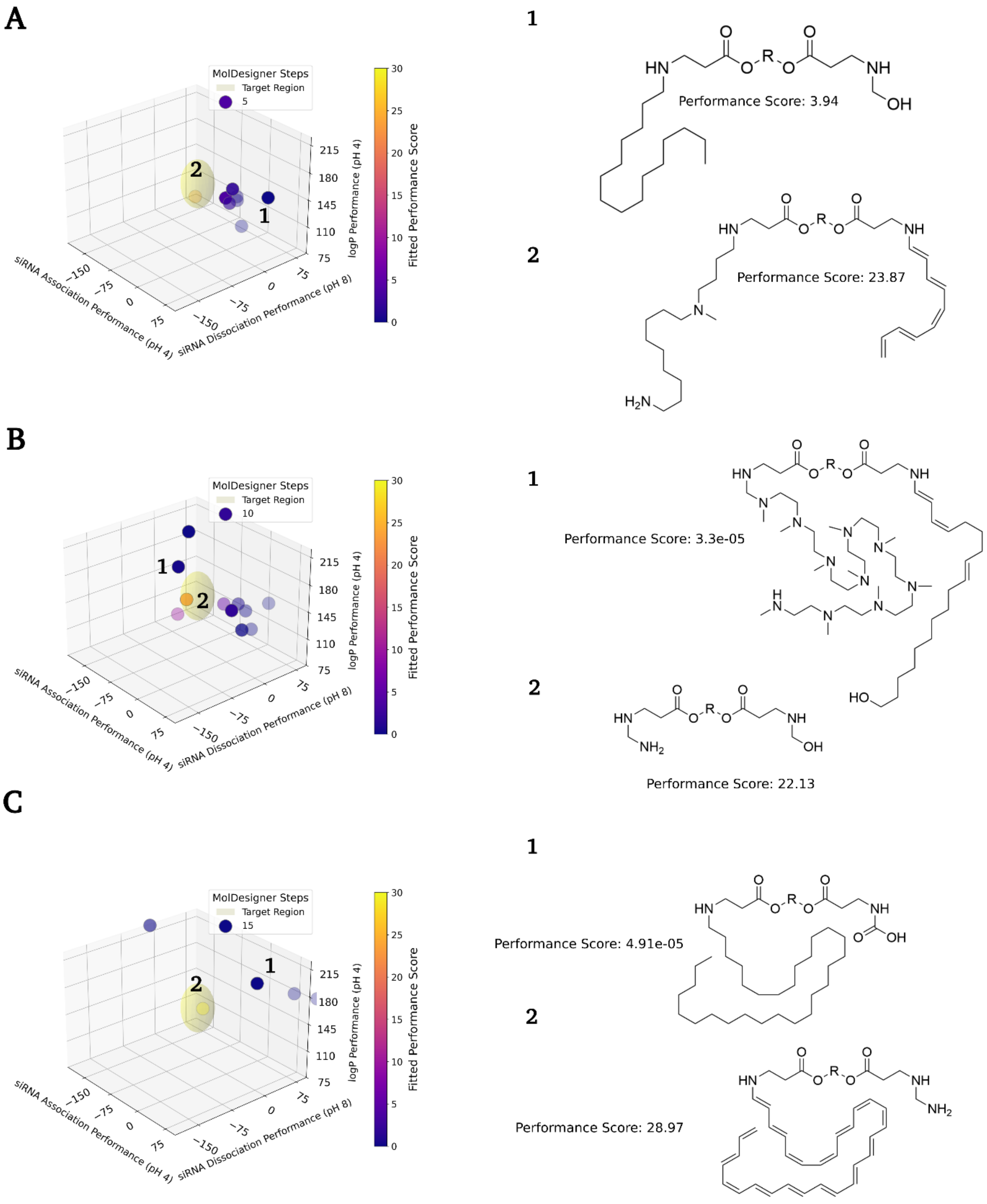


Figure V.5: Predicted polymer performance landscape for different MolDesigner Steps**.** (A), (B), and (C) show results from Stepsize 5 (A), 10 (B), and 15 (C), respectively. Left: 3D performance space (siRNA association at pH 4 vs. siRNA dissociation at pH 8 vs. logP Performance) for iteratively generated structures Right: Computed polymer structures and predicted Performance Score.

Following the optimization of mutation strength, the influence of side chain complexity was assessed by varying the 'Mol Designer steps' parameter. This parameter governs the iterative process of side chain construction within the polymer generation algorithm by determining the maximum number of monomer additions allowed per side chain. Keeping the cutoff and the mutation strength fixed at the optimal values of 17 and 1, respectively, simulations were conducted using Mol Designer step values of 5, 10, and 15. These settings correspond roughly to maximum side chain lengths of approximately 10-15, 20-30, and 30-40 heavy atoms.

The results indicated that simulations employing 5, 10 or 15 Mol Designer steps successfully identified polymers surpassing the performance score threshold of 17 (Figure V.5A-C). This indicates that the number of MolDesigner steps can be flexibly adjusted to achieve the desired polymer size. However, to remain consistent with typical side chain lengths reported in the literature[81,86] (10–15 heavy atoms), we fixed the number of MolDesigner steps to five for the production runs.

Based on these findings, hyperparameters of performance score cutoff = 24, mutation strength = 1 and Mol Designer steps = 5 were selected for subsequent polymer generation efforts. This combination should allow to effectively explore the chemical space for high-performing candidates while imposing constraints on side chain length, aiming to enhance the synthetic feasibility of predicted top-performing polymers.

5.5.4 Production Runs of New Structures

To assess both optimizer choice and practical robustness, we benchmarked the GA against a random-design baseline under a matched evaluation budget, keeping the evaluation pipeline identical (MD challenges, SA filtering, pH-dependent bead mapping) to isolate the optimizer's effect. Random sampling was markedly less efficient, requiring 23 ± 11 episodes on average to reach a performance score of 24 versus 5.5 ± 2.5 episodes for the GA (Figure

V.S23). Complementing this baseline, we executed three independent production runs in parallel using the RL warm start. The number of episodes to identify the first high-performing candidate varied by run – 3, 13, and 8 episodes, respectively (Figure V.6A–C) – indicating some sensitivity to the initial seed. Yet all runs converged within a modest number of iterations. The discovered candidates consistently contained amine functionalities, often short diaminals rather than the long polyamine chains typical of PBAEs, and their secondary chains included aliphatic alcohols and a polyunsaturated alkene chain. Together, these results show that while optimization time is stochastic and influenced by the warm start, the workflow reliably discovers diverse, novel polymers while the GA provides clear sample-efficiency advantages over random search under the same budget.

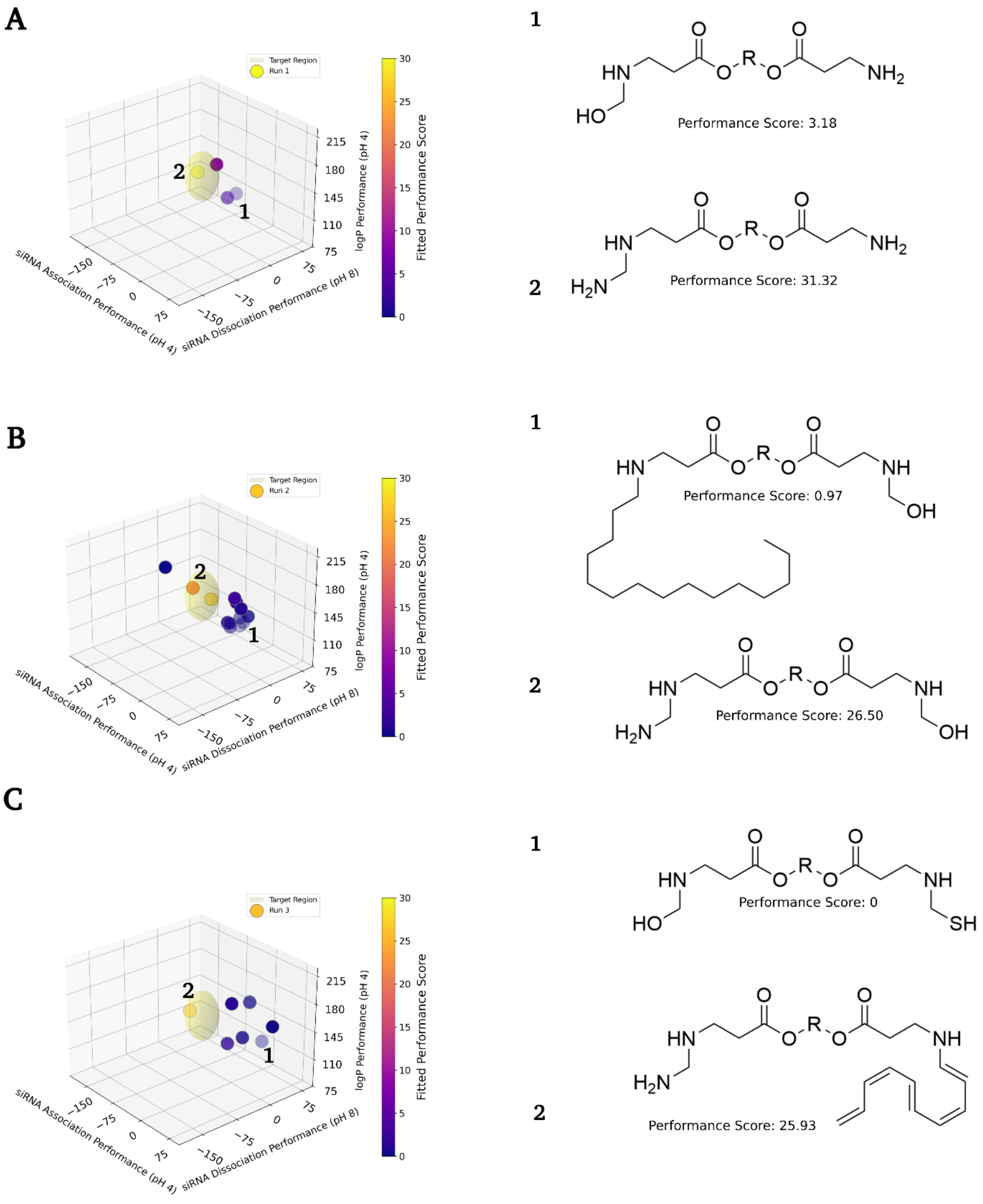


Figure V.6: Predicted polymer performance landscape in production. (A), (B), and (C) show triplicate runs of the software. Left: 3D scatter plots mapping siRNA association. Right: Computed polymer structures and predicted Performance Score.

6 Conclusion and Outlook

In this study, we have demonstrated, for the first time, the efficacy of a well-designed *in silico* pipeline approach based on a combination of ML and MD for identifying novel polymeric delivery systems. To this end, we introduced physico-chemical challenges as an innovative way for mimicking real-world hurdles of carrier systems. This pipeline effectively integrates MD simulations with an underlying optimization algorithm. We emphasize that this framework is broadly applicable to diverse delivery challenges, with PBAEs and siRNA serving as a representative model in our study. Furthermore, the developed software package possesses significant modularity. Key components, such as the polymer backbone scaffold, could be exchanged to represent different PBAEs and also other types of polymers. Additionally, constraints can be applied to restrict mutations to specific chemical moieties, and parameters governing side chain complexity (e.g., 'Mol Designer steps') can be adjusted, allowing for flexible adaptation to diverse polymer design challenges

We acknowledge that further optimizations beyond the scope of this initial version of Bits2Bonds are possible and necessary. Specifically, while we have considered the impact of individual monomers, factors such as molecular weight and monomer ratio, which also influence polymer properties, were deferred to future investigations. This decision was driven by the increased computational complexity associated with these parameters, which would compromise our objective of acceptable computational effort. However, since we observed, that the polymerization is kinetically trapped[129], we assume that using small compositional surrogates is an effective approximation and enables high-throughput exploration[201]. Furthermore, the synthesizability of the proposed polymers is not guaranteed. While automated synthesizability assessments are an active area of research[196,202] with some progress for small molecules[203], they remain a significant challenge for novel carrier systems like those explored here. Incorporating the SA score as a filtering criterion enabled the exclusion of synthetically inaccessible structures. The synthesis and subsequent computational as well as experimental optimization of the identified lead candidates are currently underway.

7 Data and code availability statement:

All code can be found at GitHub https://github.com/felixsie19/Bits2Bonds. All experimental data will be shared upon request.

8 Acknowledgements

The authors gratefully acknowledge the Leibniz Supercomputing Centre for funding this project by providing computing time on its Linux Cluster. This work was supported by ERC-2022-CoG-101088587 to O.M.M. The authors acknowledge the use of BioRender (BioRender.com) for the creation of Figures presented in this work.

9 LLM Statement

During the preparation of this work the authors used Gemini 2.5 to improve language conciseness. After using this tool/service, the authors reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

10 Supplementary Information

Table S1: Architecture of the CGNN and the selected hyperparameters for training.

Node Features	AtomType, hydrogen_donors, hydrogen_acceptors, Hybridization, Valence, Aromaticity, Ringsize, Charge
Layer1	Conv(30, 1024) + ReLU + BatchNorm
Layer2	Conv(1024. 512) + ReLU + BatchNorm
Layer3	Conv(512, 256) + ReLU + BatchNorm
Layer4	Conv(256, 512) + ReLU + BatchNorm
Layer5	Conv(512, 1024) + ReLU + BatchNorm
Layer6	Linear(1024, 128) + ReLU
Layer7	Linear(128, 16) + ReLU
Layer8	Linear(16, 1) + ReLU
optimizer	Adam
Initial lr	0.01
scheduler	ReduceLROnPlateau(mode =”min”,factor=0.7,patience=10,min_lr=0.001)
epochs	150

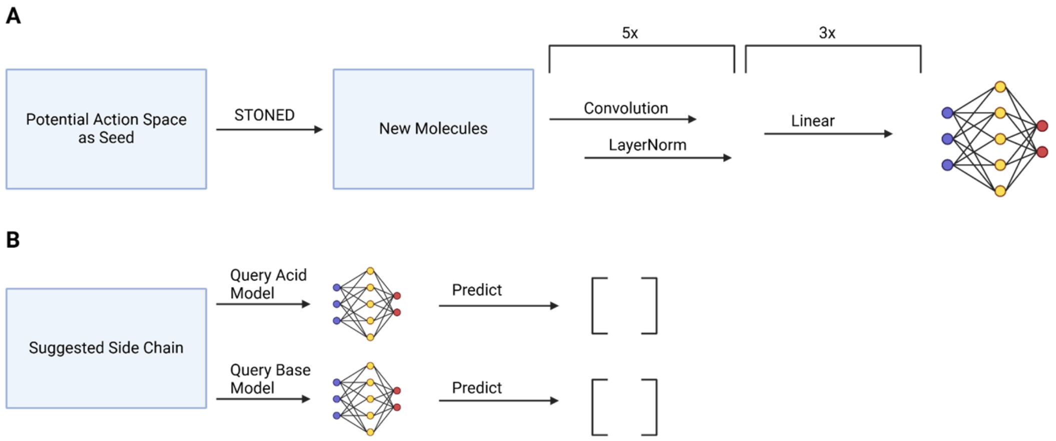


Figure V.S1: PkaPred Model A) Overview about the structure prediction process training B) Inference of pkaPred within the bead exchanger module.

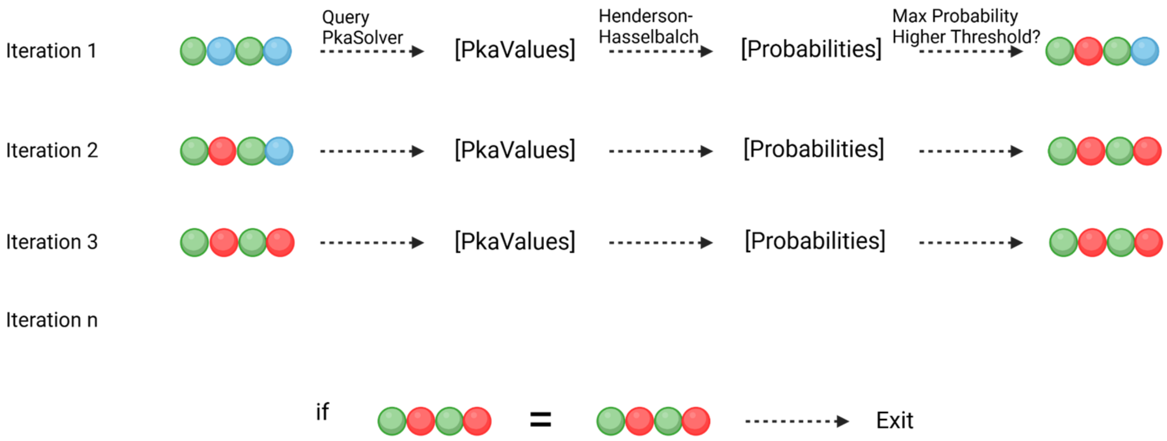


Figure V.S2: BeadExchanger. After setting the pH to a certain value, the BeadExchanger is queried and iterates over the beads as long as the probability of protonation is over a selected threshold.

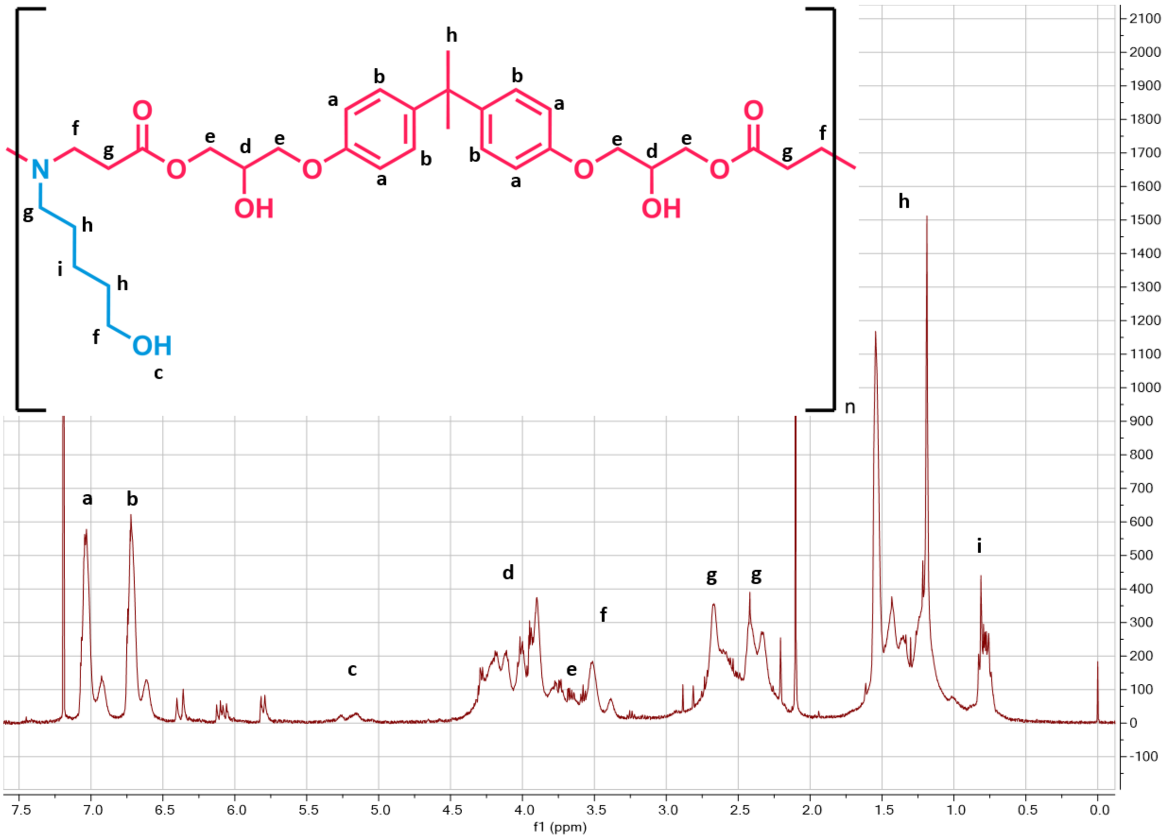


Figure V.S3: ^{1}H-NMR of AP-BG Polymer.

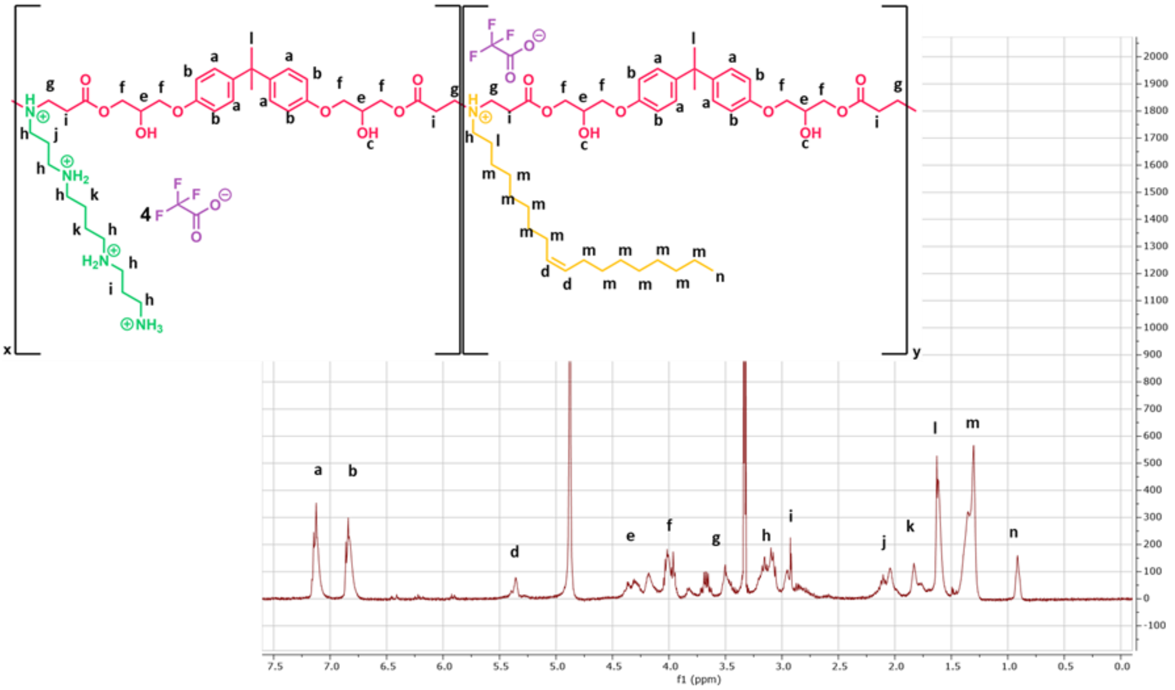


Figure V.S4: ^{1}H-NMR of OA/SP-BG Polymer.

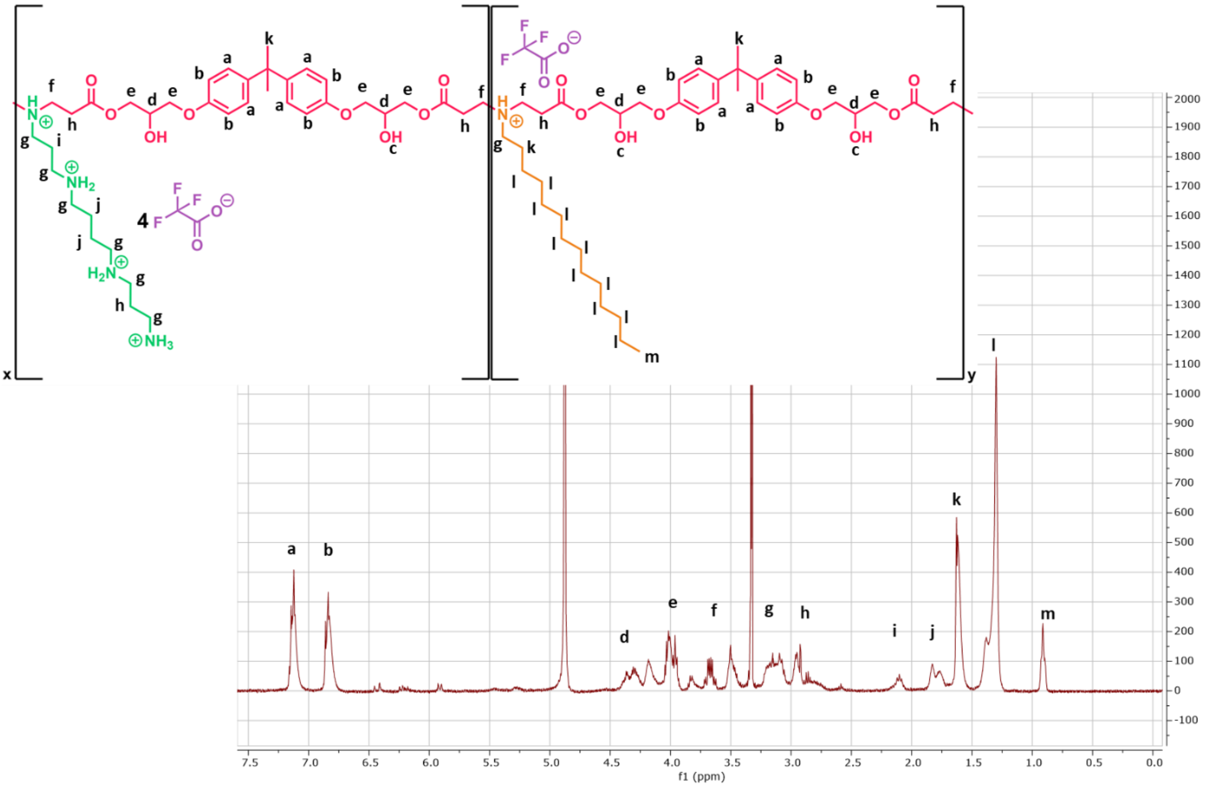


Figure V.S5: 1H-NMR of TDA/SP-BG Polymer.

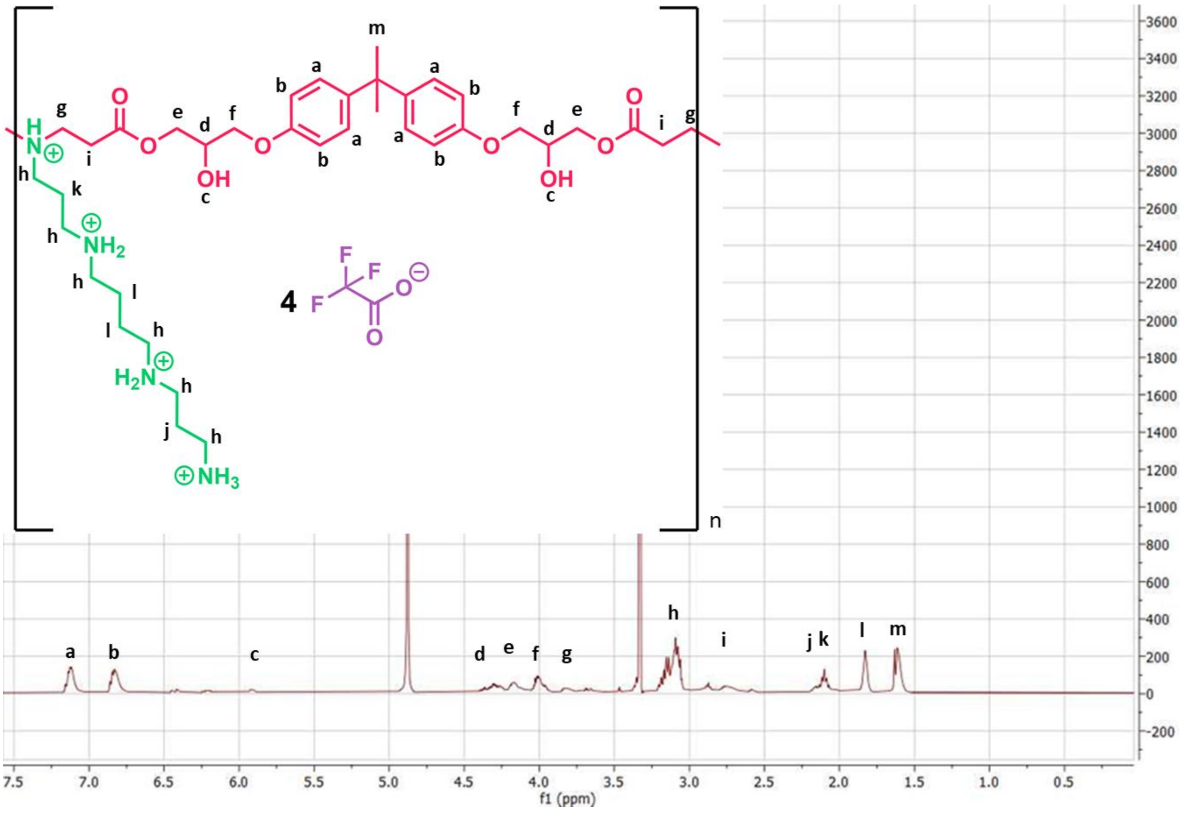


Figure V.S6: 1H-NMR of SP-BG Polymer.

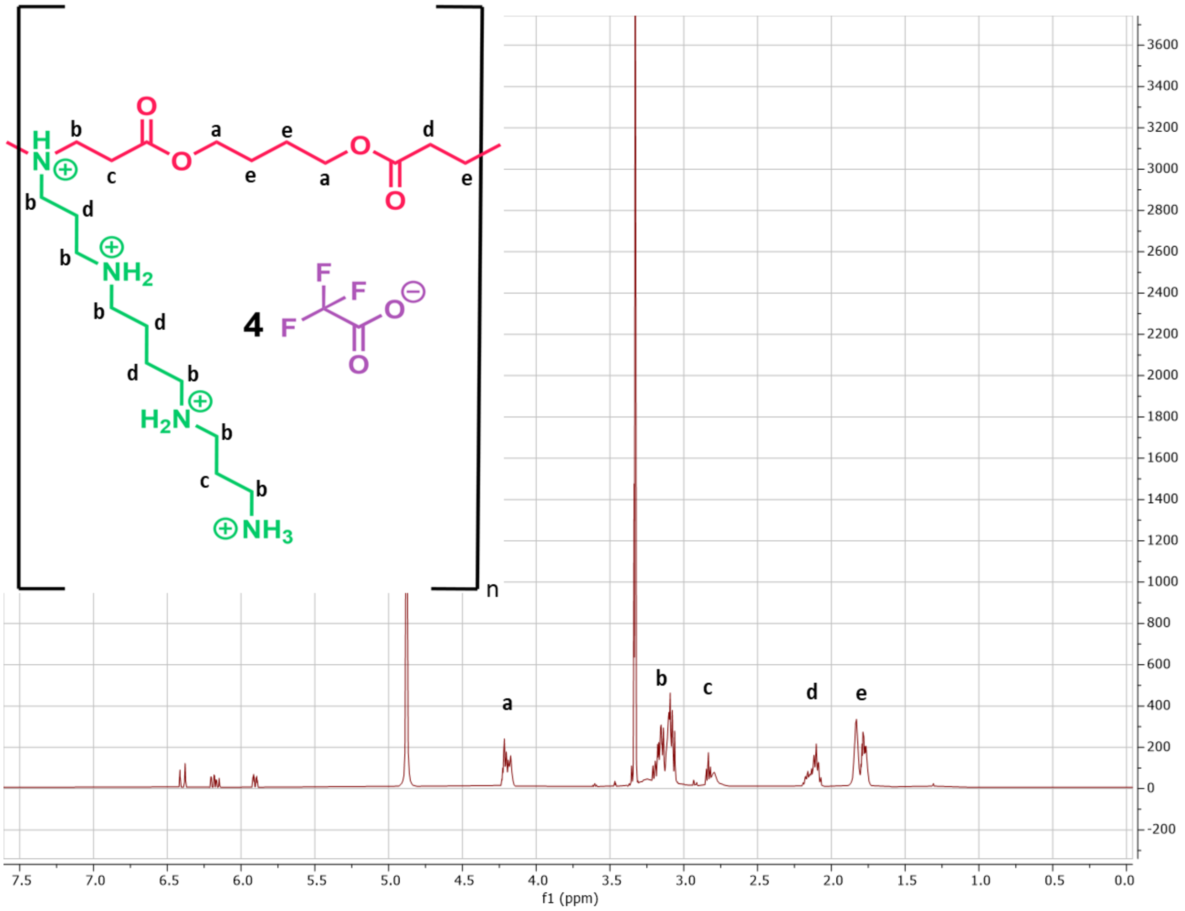


Figure V.S7: ^{1}H-NMR of SP-BU Polymer.

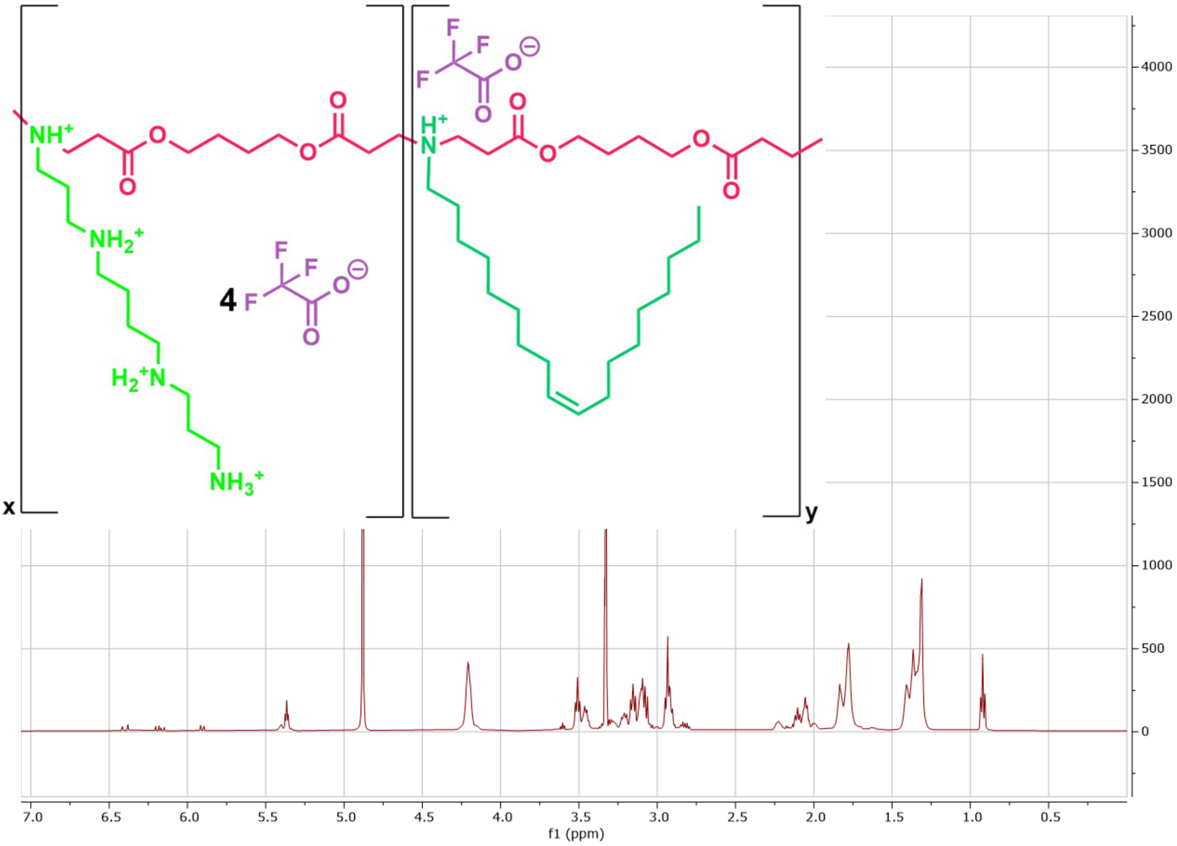


Figure V.S8: ^{1}H-NMR of OA/SP-BU Polymer.

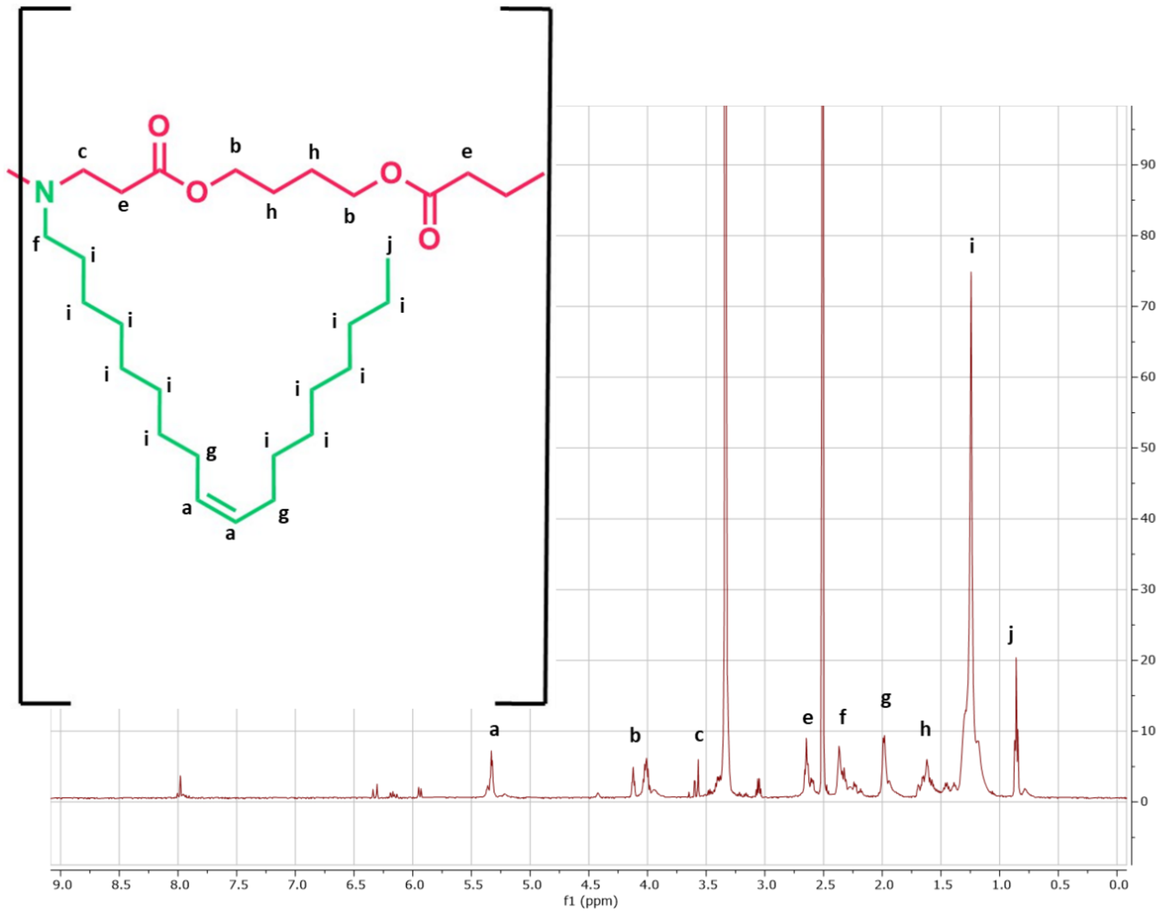


Figure V.S9: ^{1}H-NMR of OA-BU Polymer.

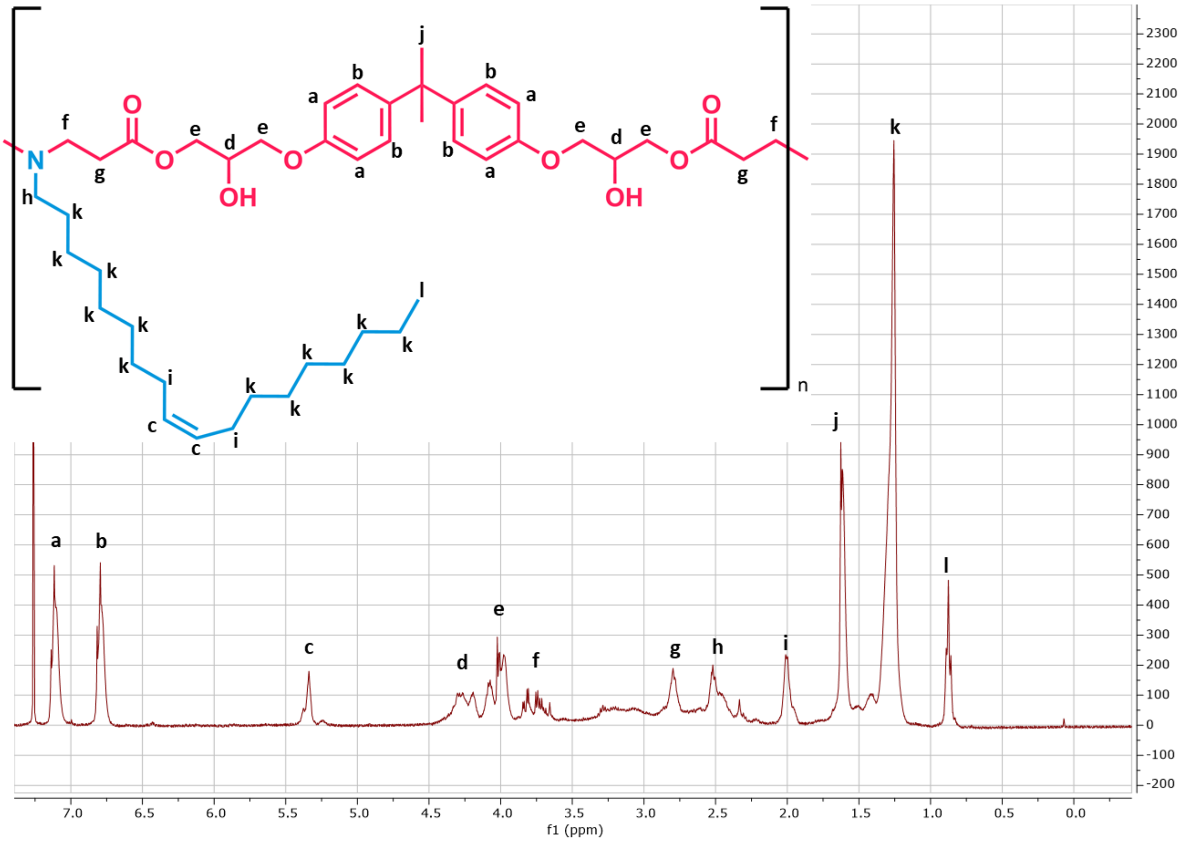


Figure V.S10: ^{1}H-NMR of OA-BG Polymer.

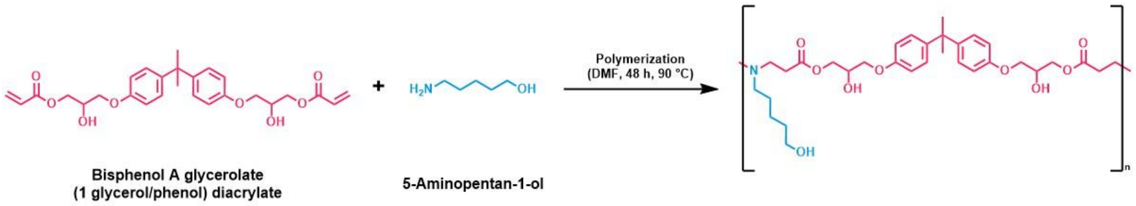


Figure V.S11: Synthesis route of AP-BG Polymer.

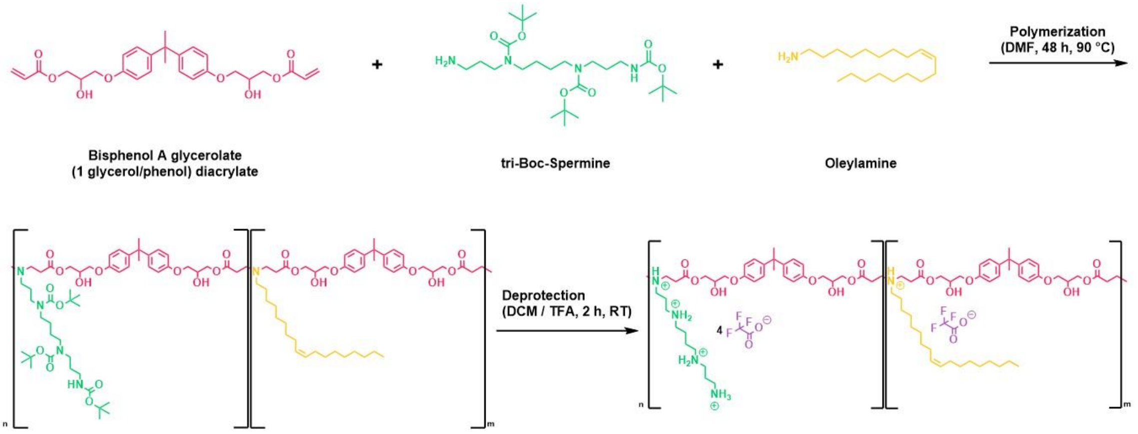


Figure V.S12: Synthesis route of OA/SP-BG Polymer.

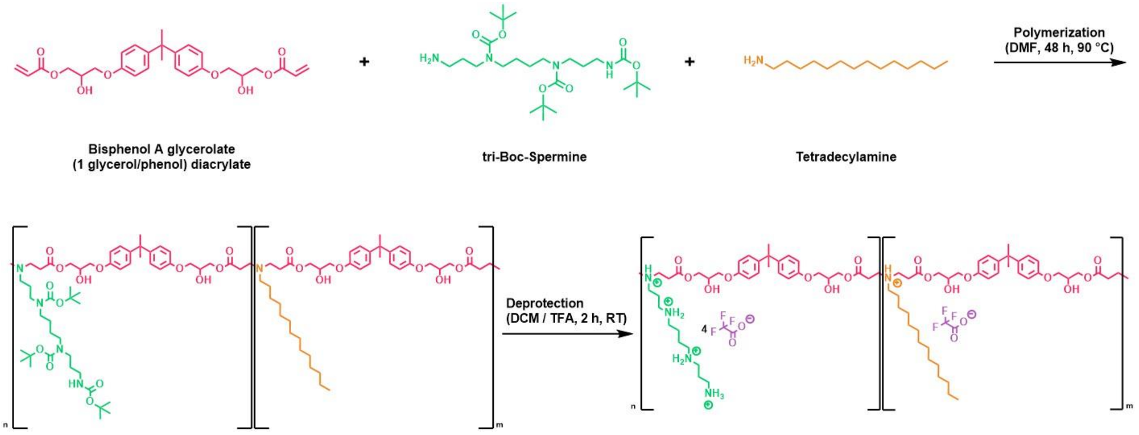


Figure V.S13: Synthesis route of TDA/SP-BG Polymer.

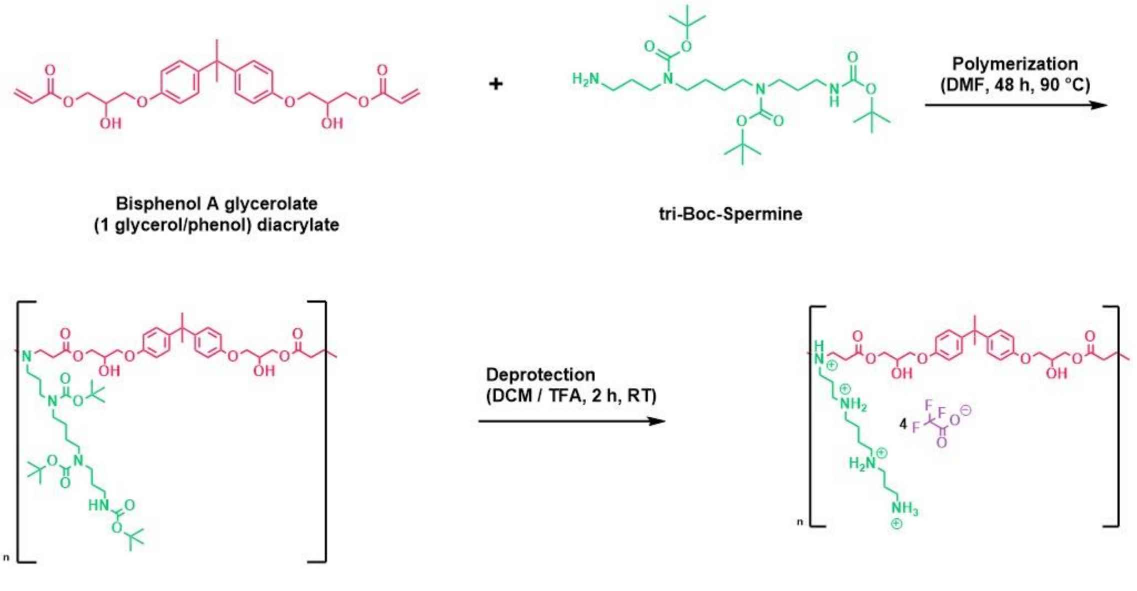


Figure V.S14: Synthesis route of SP-BG Polymer.

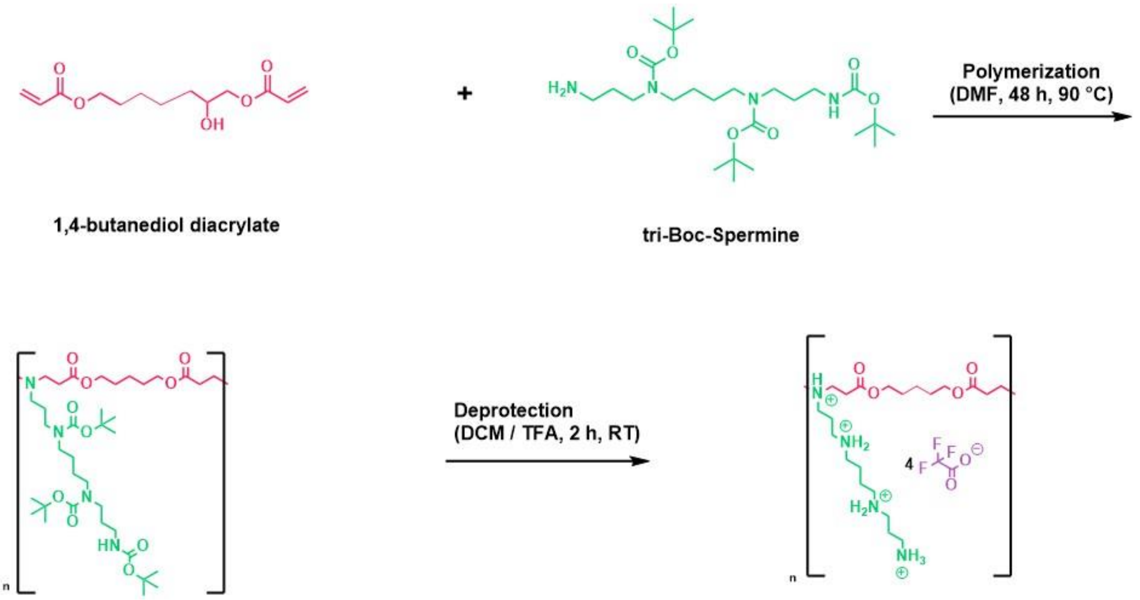


Figure V.S15: Synthesis route of SP-BU Polymer.

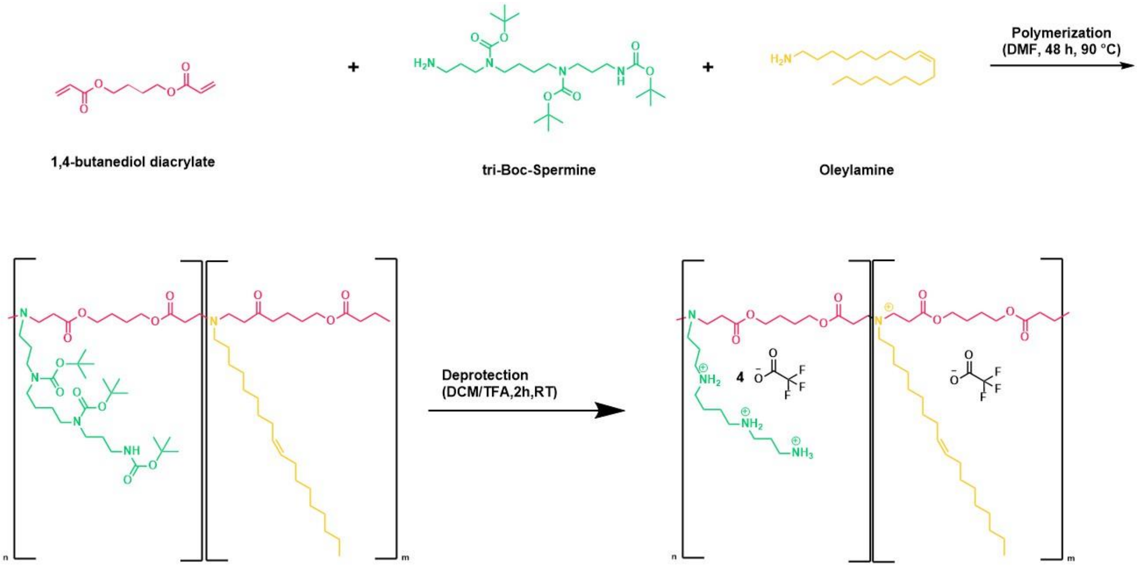


Figure V.S16: Synthesis route of OA/SP-BU Polymer.

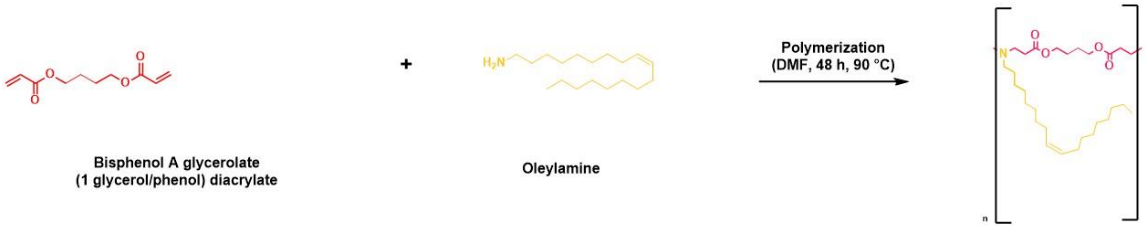


Figure V.S17: Synthesis route of OA -BU Polymer.

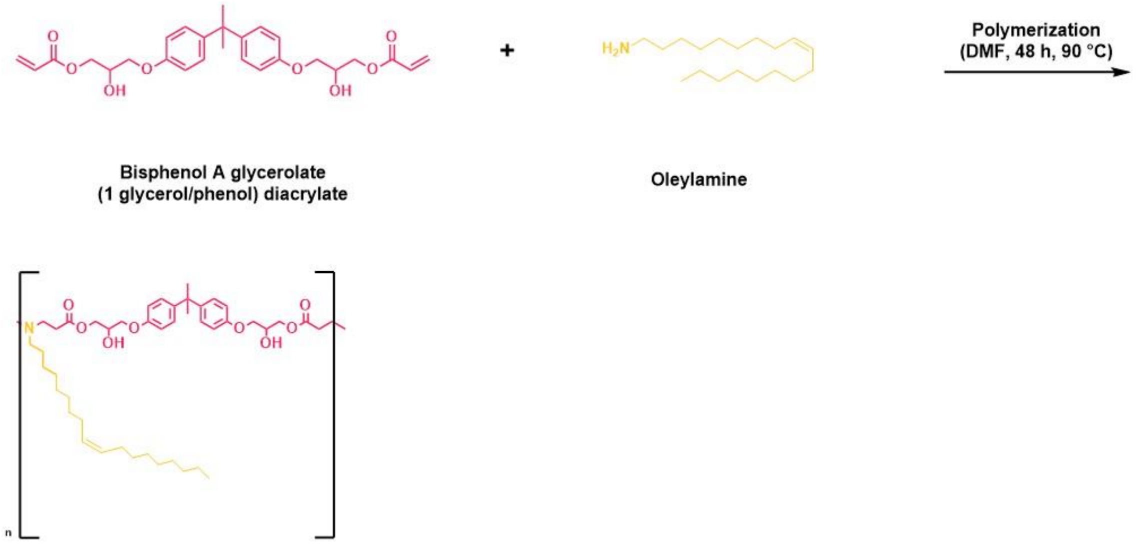


Figure V.S18: Synthesis route of OA -BG Polymer.

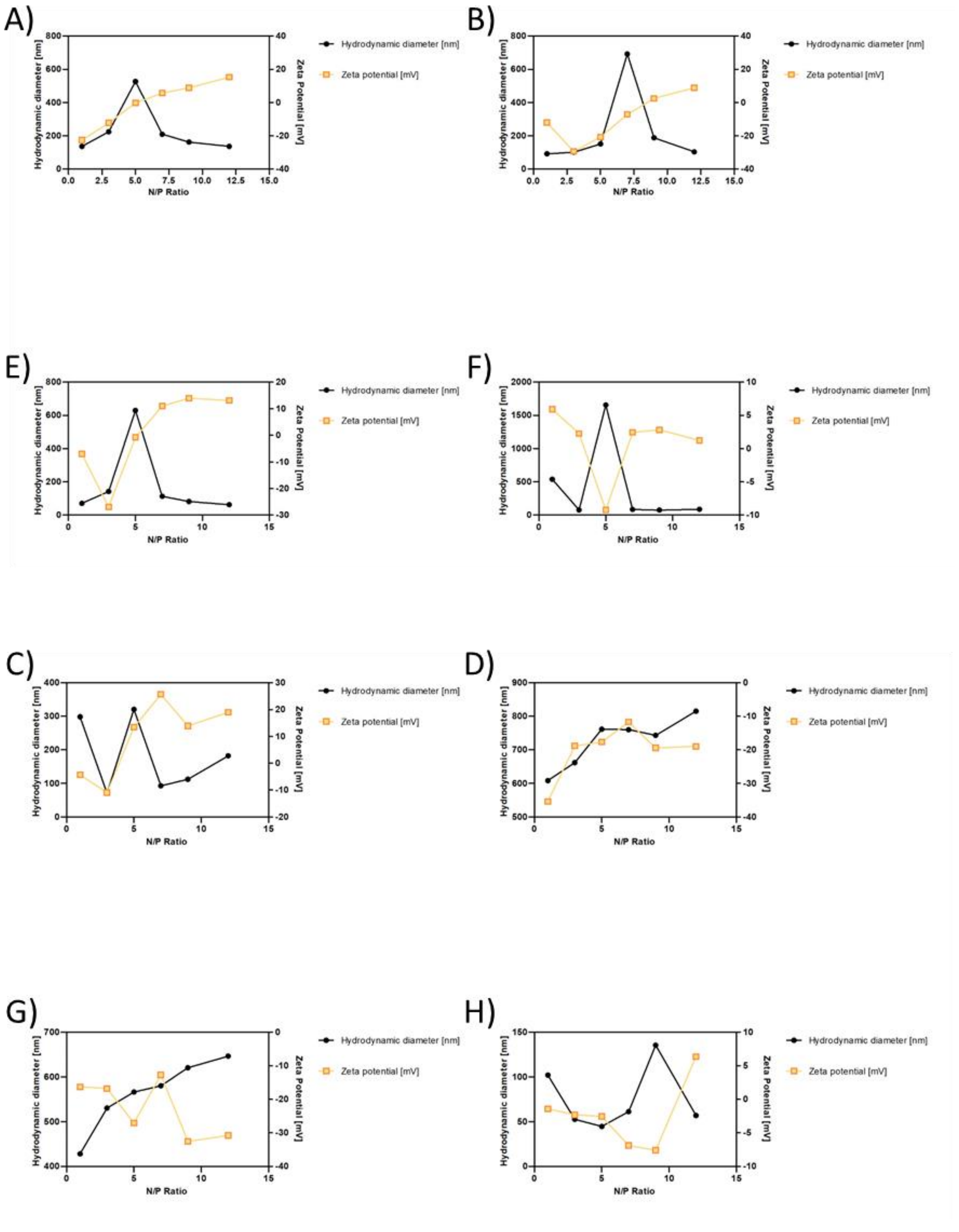


Figure V.S19: Size and Zeta potential of nanoparticles at different N/P ratios. A) AP-BG B) SP-OA-BG C) SP-TDA-BG D) SP-BG E) SP-BU F) OA-BG G) OA-BU H) OA-SP-BU

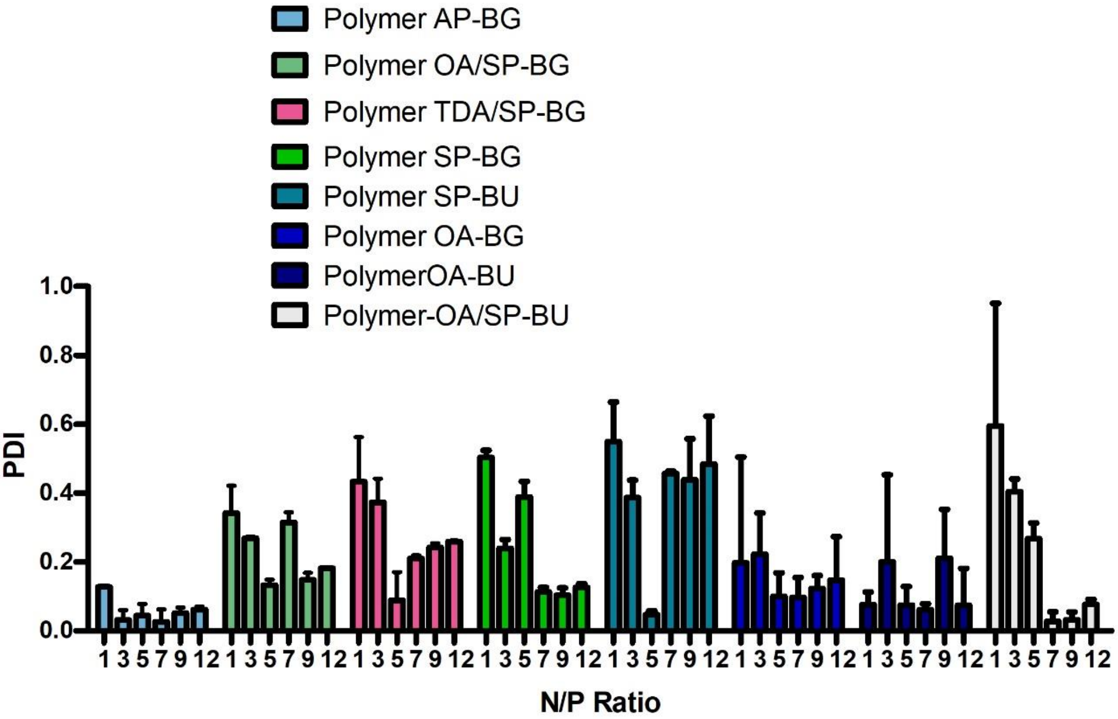


Figure V.S20: PDI of nanoparticles formulated with different polymers.

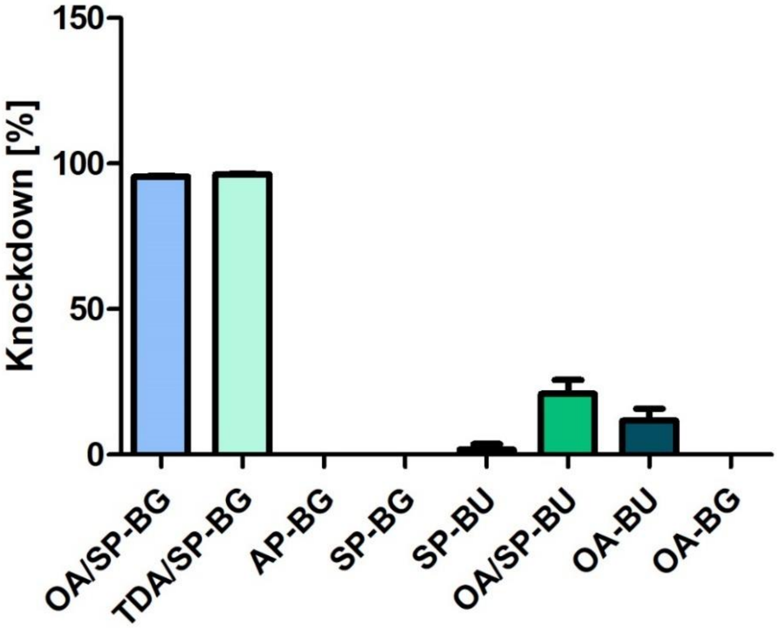


Figure V.S21: Knockdown results for the different calibration polymers on H1299 eGFP cells.

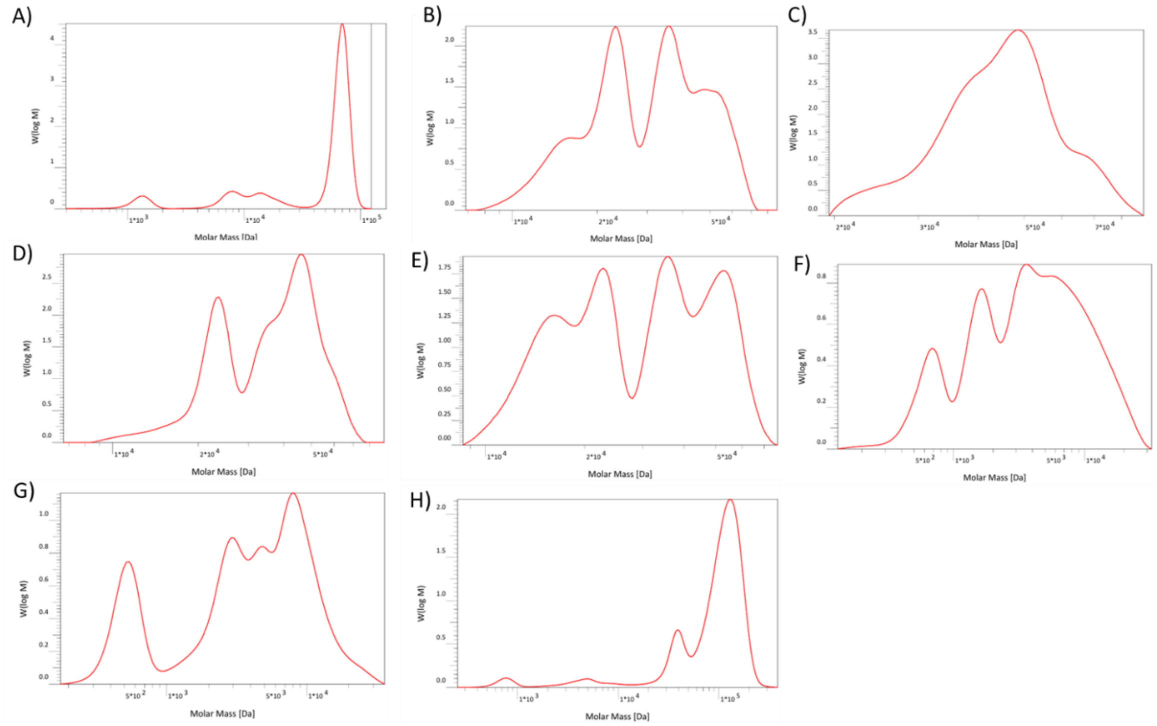


Figure V.S22: GPC-Traces of Calibration Polymers. Measurements were performed at 40°C in 0.1 M sodium chloride solution supplemented with 0.3% formic acid. A) OA/SP-BU B) OA-HP-BG C) OA/SP-BG D) SP-BU E) OA-HP-BU F) AP-BG G) SP-BD H) OA/SP-BG

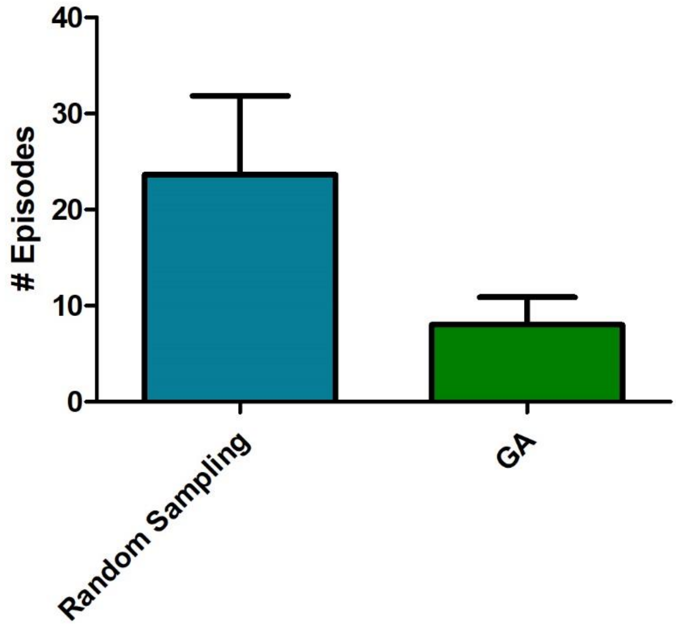


Figure V.S23: Comparison of Genetic Algorithm optimization vs. Random Sampling of Beads. '# Episodes' denotes the number of episodes required to reach a performance score >24.

Table V.S2: adjusted from Zimmermann et al, doi: 10.1016/j.jconrel.2022.09.021. Sequences of siRNAs used in the study. Nt = nucleotides; GFP = green fluorescence protein; NC = negative control; GAPDH = housekeeping gene GAPDH; A = Adenine; C = Cytosine; G = Guanine; U = Uracil; T = Thymine; p = phosphate residue; lower case bold letters = 2´-deoxyribonucleotides; capital letters = ribonucleotides; underlined capital letters = 2´-O-methylribonucleotides.

Name	Sense strand (5'-3')	Antisense strand (3'-5')	Length (nt)	
			Sense	Antisense
siGFP	pACCCUGAAGUUCAUCUGCACCAC**cg**	ACUGGGACUUCAAGUAGACGUGGUGGC	**25**	**27**
siNC	pCGUUAAUCGCGUAUAAUACGCGU**at**	CAGCAAUUAGCGCAUAUUAUGCGCAUAp	**25**	**27**

Table V.S3: Computational performance overview of the Bits2Bonds pipeline

Parameter	Description	Value
Hardware	GPU used for all production runs	NVIDIA RTX 3080 Ti
Parallelization	Number of simultaneous software instances	3
Mean wall-clock time per polymer evaluation	Full pipeline: MolDesigner → pKa / BeadExchanger → MD “challenges” → scoring	22 min 6 s ± 4 min 32 s
Throughput (polymer evaluations)	Completed full evaluations per hour	7.69
Throughput (side-chain screens)	Approximate rate based on parallelized side-chain sampling	15.38

Supplementary Calculation: Encapsulation-Efficiency (EE) Determination.

The encapsulation efficiency was determined for each polymer over a series of N/P ratios using 50 pmol siRNA per formulation. To compare the experimental results with molecular-dynamics simulations, we calculated an EE value that is not normalised to the amount of nitrogen per mole. The procedure is outlined below and illustrated with polymer SP as an example.

1. Definitions

Symbol	Meaning
EEN/P	Encapsulated siRNA (pmol) measured at a given N/P ratio
Amount of P	Total phosphate amount in the formulation (pmol)
N per RU	Nitrogen atoms per stochastic repeating unit (SRU)
Amount of SRU	Total SRUs present in the formulation (pmol)
EEN/P-value	Normalised encapsulation efficiency at a specific N/P ratio

EE-value	Overall polymer-specific encapsulation efficiency

2. Step-by-Step Calculation

1. Phosphate amount: Amount of P = Amount of siRNA × 52
2. Total SRUs: Amount of SRU = (N/P × Amount of P) / (N per RU)
3. EE(N/P) -value: EEN/P-value = (Amount of siRNA / Amount of SRU) × EE(N/P)
4. Overall EE-value: EE-value = $(1 / |Z|)\ \Sigma\ EE(N/P), \quad Z = \{1, 3, 5, 7, 9, 12\}$

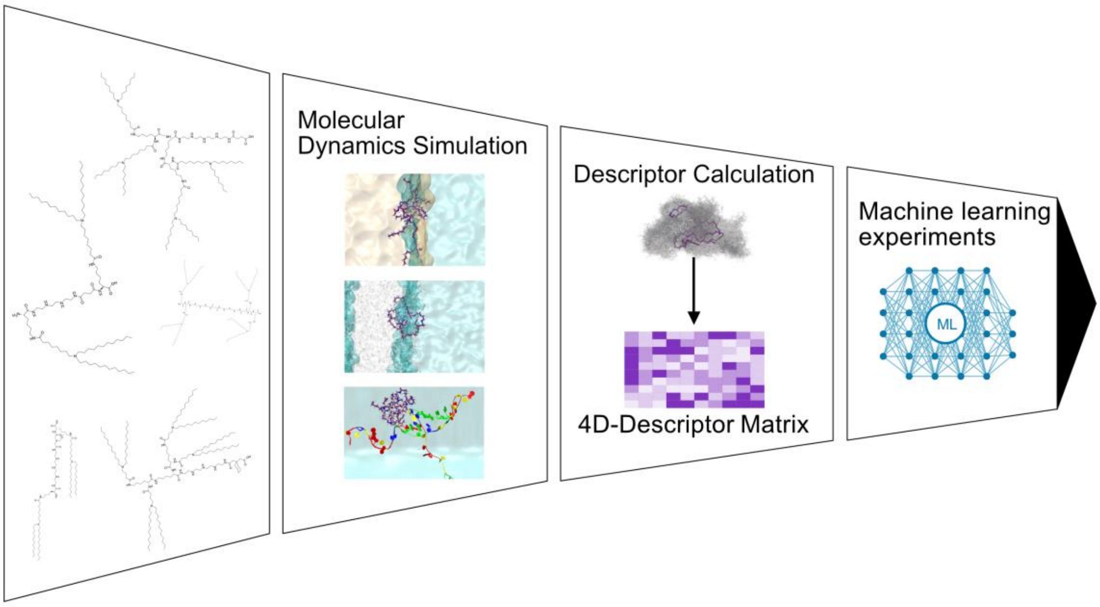
3. Worked Example (Polymer SP)

Input / Step	Value
N/P	1
EEN/P	2.989×10^{-1}
N per RU	4
Amount of P	50 pmol × 52 = 2.600×10^{3} pmol
Amount of SRU	$(1 \times 2.600 \times 10^{3}) / 4 = 6.50 \times 10^{2}$ pmol
EEN/P-value	$(50 / 650) \times 2.989 \times 10^{-1} = 2.299 \times 10^{-2}$

The above calculation is repeated for each N/P ratio in Z. The six resulting EEN/P-values are then averaged: EE-value_SP = $(1 / 6)\ \Sigma\ EE(N/P) = 1.244 \times 10^{-2}$.

Chapter VI - Capturing Molecular Motion by Integrating MD-Derived Descriptors into Predictive Machine Learning Models for RNA delivery

1 Graphical Abstract



2 Abstract

Drug-delivery vehicle performance is notoriously difficult to predict because successful transfection emerges from a multistep, tightly coupled process. Consequently, structure-transfection models built on static 2D/3D descriptors often generalize poorly, particularly in the presence of transfection cliffs and when extrapolating to chemically distinct carriers. Here, we use lipo-xenopeptides (LAX), sequence-defined, single-component amphiphiles with tunable pH responsiveness, as a representative case study to develop and benchmark a dynamics-aware prediction strategy for nucleic-acid delivery materials. We introduce a physics-informed machine-learning framework that integrates atomistic molecular dynamics (MD) with frame-resolved molecular descriptors to model transfection efficiency, termed 4D quantitative structure-transfection relationships (4D-QSTR). We performed all-atom MD simulations for a diverse library of 52 LAX carriers under physiologically relevant protonation ensembles at pH 5.0 and 7.4, across three environments representing key delivery challenges: behavior in a water-octanol interface as well as RNA- and membrane interactions. From each trajectory, we computed 3D RDKit descriptors per frame, summarized dynamics using time-windowed means and variances, and then applied probability-weighted aggregation across the three most populated charge microstates. Across multiple ML models and evaluation settings, 4D-QSTR features derived from equilibrated and full-trajectory windows improved rank-based prediction in challenging regimes, including chemically diverse splits and transfection-cliff scenarios and in several conditions outperformed static 2D/3D baselines. Beyond prediction, frame-wise analysis with rolling mean aggregation identified time-localized trajectory segments that maximized model performance, enabling mechanistic interrogation of carrier transitions at interfaces, within membranes, and near RNA. Together, our results indicate that dynamic, ensemble-aware descriptors capture delivery-relevant molecular behavior missed by static representations and establish a generalizable MD-ML workflow to support more explainable, closed-loop discovery and optimization of sequence-defined nucleic-acid delivery materials.

Keywords: 4D-QSTR, molecular dynamics, nucleic acid delivery, machine learning

3 Introduction

mRNA therapeutics have gained significant traction in recent years, with multiple approved mRNA vaccines[5,6,204] on the market. These advances enable prevention and, increasingly, treatment of diseases that were difficult to address before. Naked mRNA is rapidly degraded and shows limited cellular uptake, which is why efficient delivery systems are essential. Lipid nanoparticles (LNPs) are the current clinical standard and have transformed the field, yet they also pose challenges such as strict cold-chain storage[205], complications with repeated dosing[206], and notable manufacturing variability.[41]

Lipo-xenopeptides (LAX) offer a promising single-component alternative.[207] They offer strong nucleic-acid condensation, strong membrane interactions required for cellular entry, and endosomal escape in just one carrier molecule, uniting the key advantages of which commonly only one or another are described for polyplex or lipid systems, respectively.[31,208] Their defined sequence-based structure can be built with precise control via solid-phase synthesis, which supports rapid design-make-test-learn cycles and reproducible quality.

One of their specific advantages is a pronounced shift in logD between neutral and acidic conditions, which confers pH-responsive behavior that promotes extracellular stability and endosomal release.[31] Molecular properties are readily tunable through precise control of molecular weight and the use of exchangeable building blocks. Early studies demonstrated efficient RNA delivery across multiple cell lines in vitro and in vivo, positioning these materials as a promising carrier class.[209] Yet molecular carriers often exhibit complex structure-activity relationships, where improving one step of the delivery pathway can compromise another.[153,210] Achieving the right balance across condensation, protection, cellular uptake, endosomal escape, and cargo release remains difficult.

Machine learning (ML) has a long record of predicting structure-activity relationships for small molecules[211,212] and is now accelerating drug delivery research through mixture optimization[60], process optimization[213], and carrier discovery[191].

Currently, machine learning models in molecular design are frequently combined with 2D or 3D molecular descriptors[214,215] that capture structural or physicochemical properties. While these approaches offer the advantage of fast computation and are therefore widely used in material discovery[216], they often lack detailed information about the underlying molecular system. As a result, such descriptor-based models tend to identify molecules with similar performance profiles, but may underfit more complex structure-function relationships, particularly when the patterns in the training data differ from those in the test set.[154,217]

Molecular dynamics (MD) provide an ideal strategy for generating structured data suitable for machine learning frameworks, as they offer highly controlled environments for comparing molecular behavior. In the context of nucleic-acid delivery, simulations have primarily been employed to elucidate structural organization within lipid nanoparticles[218], characterize lipid/polymer RNA interactions[219–222], and to investigate endosomal escape mechanisms[223–226]. These studies typically rely on extensive and computationally demanding all-atom (AA)-, or coarse-grained simulations, the latter of which trade atomistic resolution for computational efficiency and may thereby obscure structure activity relationships. In small molecule discovery, integrated MD-ML approaches have been successfully applied to predict physicochemical properties using simulations[227,189] and docking data[228]. Riniker et al introduced a compact AA based framework employing integrated MD fingerprints to encode molecular descriptors for predicting free energy differences.[227] MD readouts were used by Chew et al as label for comparing different formulation encoding methods[229], while another work focused on the use of physics-informed descriptors from MD simulations.[230]

A promising different approach is the use of 4D-QSAR, originally proposed by Hopfinger et al.[231] and subsequently refined by several groups.[232–234] Particularly interesting is the work of Ash and Fourches[235], who computed WHIM[68] descriptors, a family of 3D molecular descriptors, across molecular-dynamics frames to test whether frame-wise fluctuations capture mechanistic signals underlying activity differences. They reported encouraging performance in low-data regimes and evidence that incorporating dynamics can mitigate activity-cliff effects.

We posit that extending this frame-aware descriptor strategy to drug-delivery materials discovery, especially for nucleic-acid carriers, could be especially impactful for three reasons:
first, the problem is intrinsically data-limited and 4D QSAR is especially powerful in solving low-data problems. Second, molecule-efficacy relationships often lack simple, interpretable SAR, complicating purely rational design. Therefore, encoding dynamic behavior may outperform static encodings, particularly when extrapolating to chemically distinct materials. And third, cliff-like transfection efficacy behavior has also been reported for nanocarrier material[153,210], and the fine-grained temporal/ensemble information from frame-resolved descriptors may help attenuate such effects. Systematic evaluation of dynamic, frame-aware descriptor strategies for nucleic-acid delivery remains unexplored, highlighting the need for rigorous benchmarking.

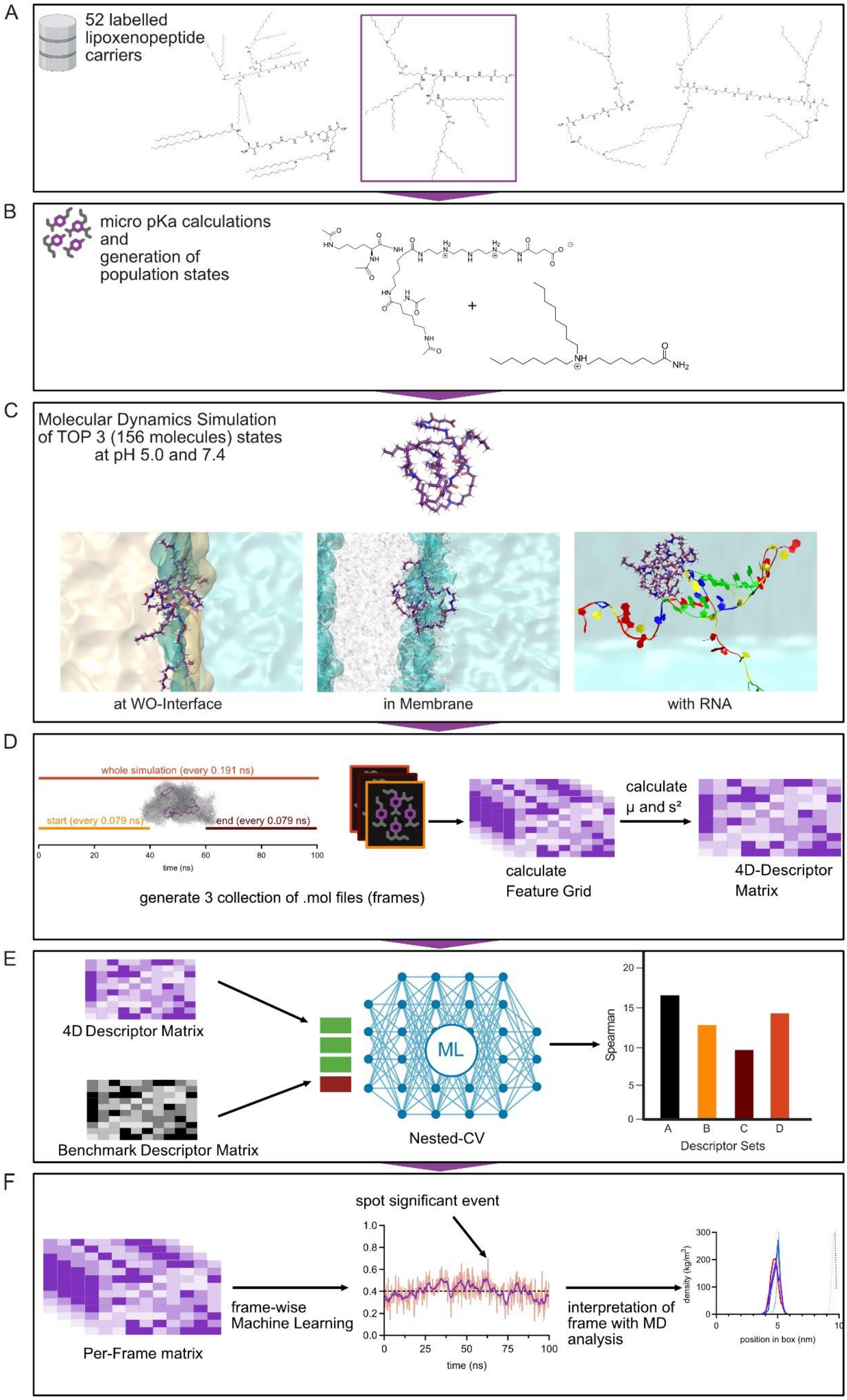
By expanding computational frameworks originally developed for small-molecule design, we advanced this direction toward nucleic-acid delivery by performing AA-MD simulations of a diverse library of lipo-xenopeptides (52) as a case study under physiologically relevant conditions, pH 5.0 (endosomal) and pH 7.4 (blood or cytosol). Each pH state was explored in three different environments relevant for RNA delivery efficiency: at the water-octanol (WO-) interface, in proximity to RNA and within a 1-palmitoyl-2-oleoxl-phosphatidylcholin (POPC) bilayer membrane. From the resulting trajectories, we extracted representative frames to calculate molecular descriptors, which were subsequently evaluated through a machine learning framework to assess their predictive power for transfection efficiency. We introduce an approach that integrates time-aware dynamic descriptors into drug-delivery prediction to test whether data-driven models can better capture the delivery process. We refer to this framework as 4D QSTR (quantitative structure–transfection relationship). Beyond LAXs, this integrative workflow establishes a potentially generalizable strategy for data-driven optimization of not only peptide-lipid hybrids but also other nucleic acid delivery systems.

4 Results and Discussion

4.1 Molecular Dynamics Simulation for 4D-Descriptor generation.

A case study comprised of 52 structures, known as LAX[31,210] (Scheme VI.1A, B and Table VI.S1), was chosen to evaluate the integration of MD-derived descriptors into a machine learning framework for predicting transfection efficiency. Leveraging the versatility of MD-simulations, two physiologically relevant pH conditions representative of RNA delivery conditions were examined: pH 5, representing the endosomal milieu and pH 7.4, mimicking neutral conditions within the body such as in the blood stream or cytosolic environment. Because ionizable lipids, polymers or LAXs can reversibly be protonated depending on pH, their charge states vary depending on the environment. To assign accurate protonation states at both pH levels, micro pKa values for all protonable groups and population distributions were calculated using Schrödinger's Epik suite[236] (Scheme VI.1B). In contrast to DFT calculations using Schrödinger's Jaguar,[237] , which were not feasible here due to the large molecular size (the smallest LAX contains >235 atoms), we employed Epik for micro-pKa estimation. Epik, which supports molecules up to 200 atoms accommodates most ionizable lipids but not the large LAXs molecules. To address this shortcoming, the carriers were fragmented before pKa prediction (Scheme VI.S1). Based on these distributions, the three most populated states at each pH were recombined while preserving stereochemistry and charge assignments and then subsequently used for MD simulations, yielding 156 structures (three per carrier) for each pH condition.The pH-dependent simulations were performed in different environments representative of those encountered by RNA carrier systems during their lifetime (Scheme VI.1C). These included WO-interfaces to probe carrier behavior at an interface and during early stages of self-assembly, as well as lipid bilayer systems (POPC model membrane) mimicking cell or endosomal membranes. In addition, carrier-RNA complexes were simulated to assess carrier behavior during formulation and stability in the presence of RNA. Resulting in 3 different environments, that pose as challenge for the carriers. Each system was simulated for 100 ns, a duration chosen based on preliminary 150 ns runs of carrier 1611 (state_2_1), which confirmed no change in structural stability after 100 ns (Figure VI.S1). This justified reducing simulation time to minimize cost across the 936 total simulations performed. To capture dynamic behavior, 3D structural states of the carrier molecules (referred to as frames) were extracted from the trajectories at defined time intervals using GROMACS tools (Scheme VI.1D). Frames were collected in different time windows: initially between 0-40 ns to capture initial rearrangements (referred to as start of simulation), at 60-100 ns to represent the equilibrated state, and over the full 100 ns trajectory (referred to as whole simulation), to generate time resolved data for the calculation of molecular descriptors used to construct the 4D descriptor set for the machine learning model. To encode molecular trajectories, we computed an extensive set of descriptors for every simulation frame. Unlike prior work[235] that targets a narrow subset of features, we deliberately broadened the descriptor panel to capture richer dynamical information. For each descriptor, we summarized temporal behavior by the frame-wise mean, reflecting the central tendency of atomic or molecular properties across the trajectory, and the frame-wise variance, quantifying the magnitude of temporal fluctuations (Scheme VI.1D).To ensure a physically meaningful representation across protonation/charge microstates, we further applied an ensemble-weighted aggregation over the three most probable charged conformers. Specifically, descriptor means were combined using the conformer occurrence probabilities as weights, thereby aligning the final representation with the underlying Boltzmann-like population and reducing bias from rare or non-representative states. This encoding integrates both time-averaged dynamics (mean/variance across frames) and chemical realism (probability-weighted conformer ensemble), providing a compact yet expressive feature set for downstream modeling

Scheme VI.1: Schematic illustration of the computational workflow used for integrating molecular dynamics-derived descriptors in a machine learning model. A) Generation of 52 three-dimensional (.mol) structures from two-dimensional (2D) inputs. B) Fragmentation of molecules at defined structural points to generate 3D .mol files for suitable pKa predictions using Schrödingers Epik, followed by calculation of micro pKa values for all protonable groups at pH 5.0 and 7.4. Carriers were subsequently recombined, while preserving calculated charge and protonation assignment, and stereochemistry. C) Molecular dynamics simulation of structures in three distinct environments, illustrated by snapshots taken after 100ns from the three different setups of carrier 1621 D) Extraction of trajectory frames as .mol files representing different stages of the 100 ns simulation. Consecutive calculation of RDkit descriptors for each dropped frame, followed by computation of weighted mean and standard deviation. Resulting in a 4D QSTR descriptor set for different simulations stages and weighted mean 4D QSTR descriptors per frame. E) Comparing 4-QSTR descriptors across simulation stages with 2D and 3D RDkit baselines using different machine learning models F) Identification of significant events using frame wise weighted 4D descriptors during the simulation.


A
52 labelled lipoxenopeptide carriers
B
micro pKa calculations and generation of population states
+
C
Molecular Dynamics Simulation of TOP 3 (156 molecules) states at pH 5.0 and 7.4
at WO-Interface
in Membrane
with RNA
D
whole simulation (every 0.191 ns)
start (every 0.079 ns)
end (every 0.079 ns)
0
20
40
60
80
100
time (ns)
generate 3 collection of .mol files (frames)
calculate Feature Grid
calculate μ and s²
4D-Descriptor Matrix
E
4D Descriptor Matrix
Benchmark Descriptor Matrix
ML
Nested-CV
Spearman
20
15
10
5
0
A
B
C
D
Descriptor Sets
F
Per-Frame matrix
frame-wise Machine Learning
spot significant event
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
time (ns)
interpretation of frame with MD analysis
density (kg/m³)
300
200
100
0
0
5
10
position in box (nm)

4.2 Deterministic vs Weighted Approach in 4D-QSAR Calculations.

To demonstrate that our approach to compute weighted molecular descriptors outperforms the conventional deterministic method using only the top state, we performed a simple experiment where descriptor vectors for the top state were compared to a weighted vector over the top three states. This analysis was conducted across all simulations, now called challenges, and pH conditions. We quantified agreement using Spearman's correlation coefficient between experimental values and predictions from an ExtraTrees model, which is recognized for strong out-of-the-box performance on small datasets with high-dimensional features. The results showed that all weighted vectors had higher Spearman values than the deterministic calculation (Figure VI.1). Therefore, for further experiments, we choose to use the weighted calculation of our MD derived descriptors.

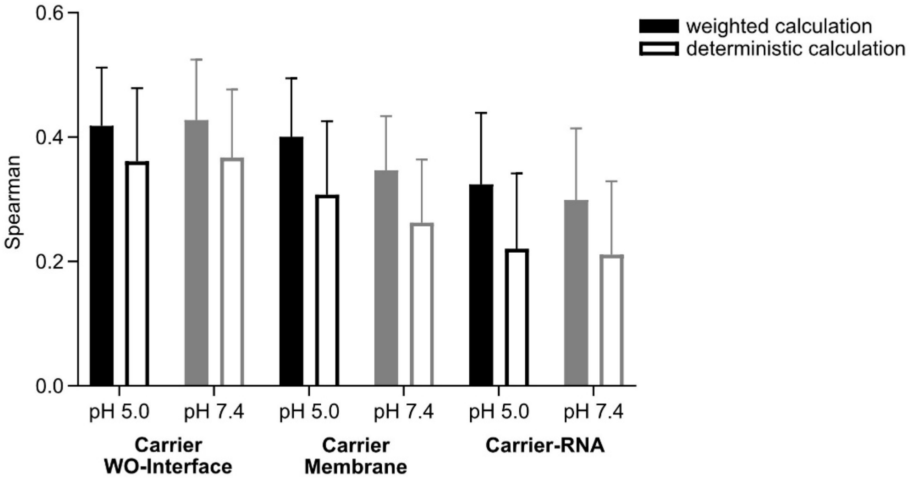


Figure VI.**1:** Comparison of the performance of a weighted vs. a deterministically calculated 4D descriptor set with Spearman´s ρ. Means were calculated from frame-wise performance (mean ± SD).

4.3 Comparison of 4D-QSAR Descriptors from Different Simulation Segments with 2D and 3D Benchmarks

Following the calculation of the descriptor matrices, they were evaluated within a standardized ML pipeline, screening multiple algorithms and selecting the best-performing model per feature set. As baselines, we included 2D RDKit descriptors (conventional benchmark) and 3D RDKit descriptors (less information-rich 3D reference) to contextualize potential gains from our MD-derived representations (Scheme VI.1E). Model comparison used 5-fold cross-validation. Because molecular discovery frequently requires extrapolation to chemically distinct structures, we added a chemically diverse split that maximizes train-test dissimilarity in chemical space.

A persistent challenge in molecular ML is the presence of cliffs, where, in this specific case, minor structural changes lead to large differences in transfection outcomes. We hypothesized that a 4D-QSTR-style encoding of dynamics would outperform conventional baselines, particularly in extrapolation and cliff scenarios, consistent with prior observations in the literature.[4,5]

The 2D RDKit baseline achieved higher Spearman correlations (Figure VI.2) than any of the more information-rich feature sets, including the 3D baseline, under standard cross-validation (CV). This aligns with the well-known strength of 2D encoding when train-test similarity is high. [239,240] Simpler representations can resist overfitting and avoid incorporating simulation noise. Moreover, most datasets, like this one, were historically generated through iterative optimization of closely related 2D scaffolds, which inherently favors descriptors that capture 2D structural variation. This design bias likely contributes to the consistently strong performance of 2D-derived features. However, in more challenging settings, performance dropped: for the similarity-constrained split (Chem div) the 2D baseline reached 0.418, and for cliff prediction it reached 0.450. The 3D baseline performed similarly in these settings, achieving 0.394 and 0.452, respectively. Overall, the early stages of the simulations generally showed limited predictive power, whereas descriptors derived from equilibrium and full-trajectory windows yielded substantially better results for cliff prediction. For example, for the membrane system at pH 5.0, the equilibrium window achieved a Spearman correlation above 0.6, and for the WO interface combining both pH conditions, the equilibrium Spearman correlation reached 0.667. Interestingly, while the membrane performed well at pH 5.0 but not at pH 7.4, the WO interface simulation showed the opposite, namely that performance is strong at neutral pH but not at acidic pH.

For the chemical diversity split, WO pH 7.4 and membrane pH 5.0 again outperformed the baselines over the full simulation, with Spearman correlations of 0.576 and 0.540, respectively. The strongest performance in the chemical diversity setting was obtained when concatenating all descriptor vectors and both pH conditions into a single representation, where the full-trajectory model reached a Spearman correlation of 0.636. In contrast, combining only the start-window descriptors provided essentially no ranking ability (Spearman = 0.006), further underscoring how poorly informed the initial simulation frames are. The consistently strong performance of descriptors aggregated over all vectors and the entire simulation suggests that capturing the full temporal and contextual information may be critical for achieving robust extrapolation in complex processes such as drug delivery.

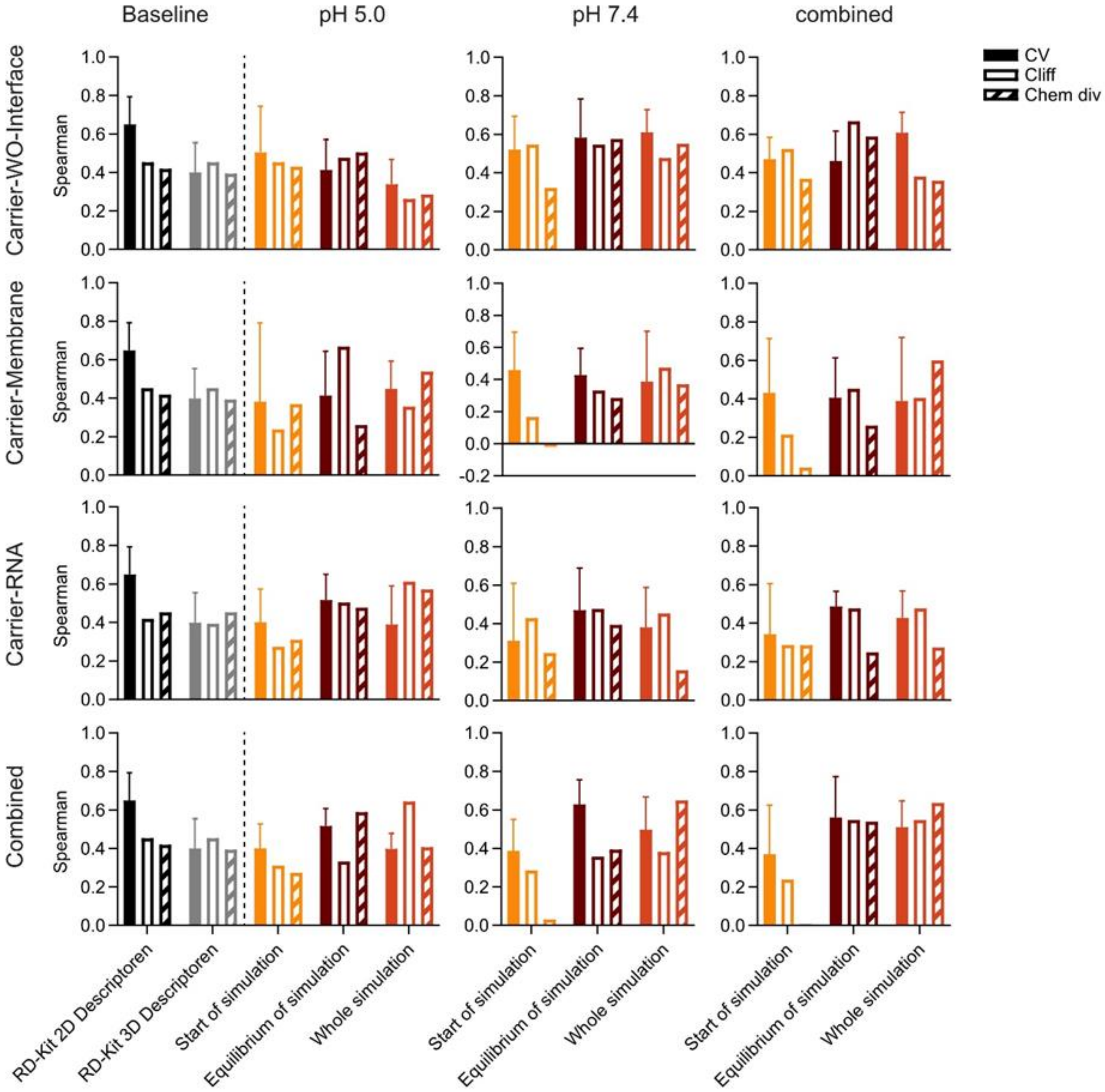


Figure VI.**2:** Performance of 4D-QSTR descriptors compared with 2D RDKit and 3D RDKit descriptors as baselines for different simulations and pH levels, evaluated at different parts of the simulation—start (0–40 ns), equilibrium (60–100 ns), and whole (0–100 ns)—as well as a combination of the pH levels per simulation and a combination of the pH levels across all simulations. Three different tests: cross-validation (CV), cliffs, and chemical diversity (chem_div).

We note that the comparatively poor performance observed for the RNA and membrane simulations at pH 7.4 may stem from a modeling choice made to keep the simulations lightweight where we omitted intermolecular (molecule‑molecule) interactions. This simplification is defensible because the resulting, unbiased single-molecule representations are subsequently mapped to experimental biological data, which can reintroduce contextual information during model training. Nonetheless, excluding collective effects can remove relevant complexity and thereby obscure aspects of molecular behavior.

In particular, microenvironment-dependent protonation may differ between isolated molecules (as estimated by our EPIK-based calculations) and molecules embedded in micellar or nanoparticulate assemblies, where local dielectric properties, ionic strength, and neighbor interactions can shift apparent pKa and may result in different molecular behavior. [31,241] These aggregate-level effects, absent in single-molecule trajectories, could plausibly contribute to the lowered predictive performance at pH 7.4 when looking at RNA and membrane challenges.

For the restricted (most diverse) splits, robust analyses (e.g., multiple randomizations of the same constraint) were not feasible due to strict splitting conditions. Results should therefore be interpreted as single point estimates of achievable performance under the chosen split. We consistently observed that full-trajectory setups capture substantially more information, which in turn leads to markedly stronger predictive performance.

Motivated by these findings, we wondered if our approach can be used to spot significant events that mainly drive performance and therefore potentially derive findings for the mechanisms of nucleic acid delivery.

4.4 Frame-wise 4D-QSTR Descriptor calculation for Identifying Key Time Points in MD-Simulations.

Since the frame-wise means computed for different simulation segments exhibited distinct predictive performance (Figure VI.2), we asked whether per-frame ML performance could help identify salient molecular behaviors in the simulated environment (Scheme VI.1F). This is nontrivial, as meaningful events need not occur at the same absolute timepoint for every molecule. To accommodate temporal misalignment, we also computed a rolling mean over 11 frames, assuming a broad enough timeframe to aggregate information without losing too much individual information and used the trajectory-wide mean as a baseline. Deviations of the rolling mean from this baseline were then used to flag significant events, enabling detection of transient behaviors that may drive model performance without requiring strict synchronization across molecules.

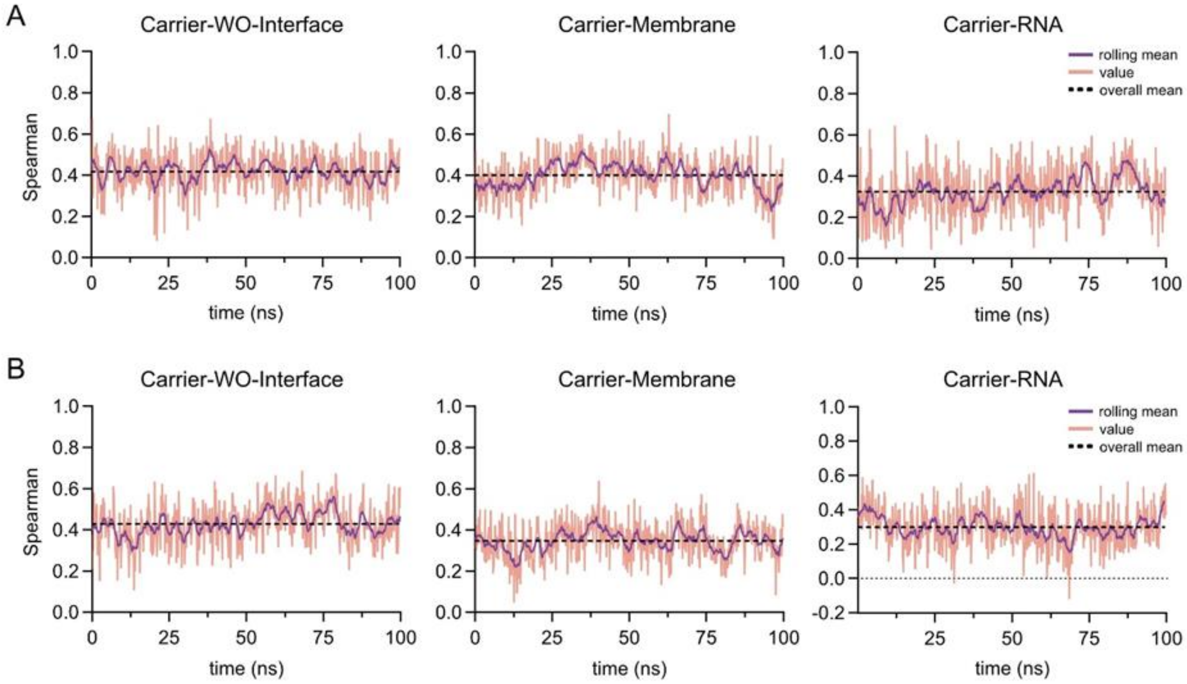


Figure VI.3: Performance of 4D-QSTR descriptors over the 100 ns simulation time. (A) pH 5.0; (B) pH 7.4. Shown are the 11-frame rolling mean and the overall mean of the prediction across all frames. One frame corresponds to 0.19125 ns. 523 frames were used for descriptor calculation.

Figure VI.3 shows that significant events were detected across all challenges. In concordance with low performance for RNA at pH 7.4 in the descriptor screening (Figure VI.2), we observed the lowest overall mean Spearman correlation and even excursions below zero for this condition. This suggests that the corresponding trajectories carry limited information about downstream transfection efficiency. Mechanistically, this is plausible: RNA‑carrier interactions are typically most pronounced under acidic conditions, whereas neutral pH favors disassembly with relatively modest, less informative variation across carriers. An additional factor may be the absence of explicit intermolecular (material‑material) interactions in our simulations, as discussed above, which could further attenuate the signal at pH 7.4.

To probe the link between dynamics and predictivity, we further examined individual molecules at the timepoint of peak Spearman correlation (Table VI.1) to assess whether distinctive conformational or interaction patterns emerged at these peaks. This molecule-level inspection provides qualitative context for the model's most informative windows and guides hypotheses for follow-up simulations.

Table VI.1: Overview of frames and timepoints per simulation environment with the highest Spearman value and the carriers that showed the lowest error over the whole trajectory for each simulation

Environment	pH	frame	time (ns)	Carrier with lowest error	Carrier with second lowest error	Carrier with third lowest error
WO-Interface	5	202	38.6325	1762	1867	1868
RNA	5	64	12.2400	1862	1869	1613
Membrane	5	329	62.9213	1868	1762	1869
WO-Interface	7	356	68.0850	1762	1858	1755
RNA	7	298	56.9925	1862	1613	1869
Membrane	7	210	40.1625	1755	1862	1762

Subsequently, the three carriers with the lowest overall prediction error were extracted, selecting carriers that correlate well with overall predictions and allow potential explainability (Table VI.1). To investigate possible key events at the distinct timepoints (Table VI.1), the trajectories of carrier 1762 were analyzed at both pH levels (Figure VI.4). Carrier 1762 was selected as representative system for WO-interface and membrane simulations, as it consistently appeared among the top three performers in the interfacial simulations (Table VI.1). Mean square displacement (MSD) was evaluated and plotted (Figure VI.4A, E). Because MSD reflects the spatial movement of the carrier during the simulation, it provides a noise-reduced measure for interpreting dynamic transitions identified with the frame-wise prediction analysis (Figure VI.3, Table VI.1).

Figure VI.4A shows the MSD of 1762 at the WO-interface for pH 5.0 and pH 7.4. At both pH values, the carrier reached a plateau in displacement, indicating that reduced mobility (interfacial pinning) at the interface occurs for both carriers. This plateau appeared earlier at pH 5.0, suggesting that the carrier becomes immobilized sooner due to higher electrostatic interactions with water. This observation agrees with the time of the highest Spearman value occurring sooner for this pH. At pH 7.4, the event occurred later in time but followed the same trend. These differences align with the expected interplay of electrostatic interactions with the aqueous phase and lipophilic interactions with octanol, which are characteristic of the LAX carriers Although the difference between the two pH values is modest, the smaller MSD at pH 5.0 indicates slightly stronger interfacial confinement under acidic conditions.

To further visualize this interfacial pinning, density distributions were analyzed at the frame of maximum model performance and ±one frame (Δt = 0.191 ns) for both interfacial environments (Figure VI.4B, D). A small but distinct shift of the carrier toward the interface was evident at the key frame, where the distribution also showed the sharpest and highest peak at both pH levels. After this frame, the carrier remained in closer vicinity to the aqueous phase for both pH values, suggesting stabilization at the interface. At pH 7.4 this shift occurred later and was less pronounced, yet the peak sharpened similarly. These observations suggest the possibility of conformers with more information for the model than at other timepoints.

This behavior was supported by the calculation of the area under the curve (AUC) for each frame between 4.2 nm and 6.2 nm (roughly the water-octanol interface area) and for the whole box (Figure VI.4C), revealing the proportion of the carrier residing in this interface. This value increased notably for both pH values, followed by a subsequent incline. The carrier with the higher total charge (pH 5.0) was almost residing to 100% in the interfacial area.

Another factor contributing to the carrier's interaction with the aqueous phase is hydrogen bonding. The number of hydrogen bonds (H-bonds) between 1762 and water (Figure VI.4D) showed a local minimum preceding the key frame, followed by an increase above the simulation average. This pattern could indicate a structural transition, in which carrier 1762 adopts a conformation with enhanced interfacial interactions, likely one of the configurations carrying the highest predictive relevance within the dataset.

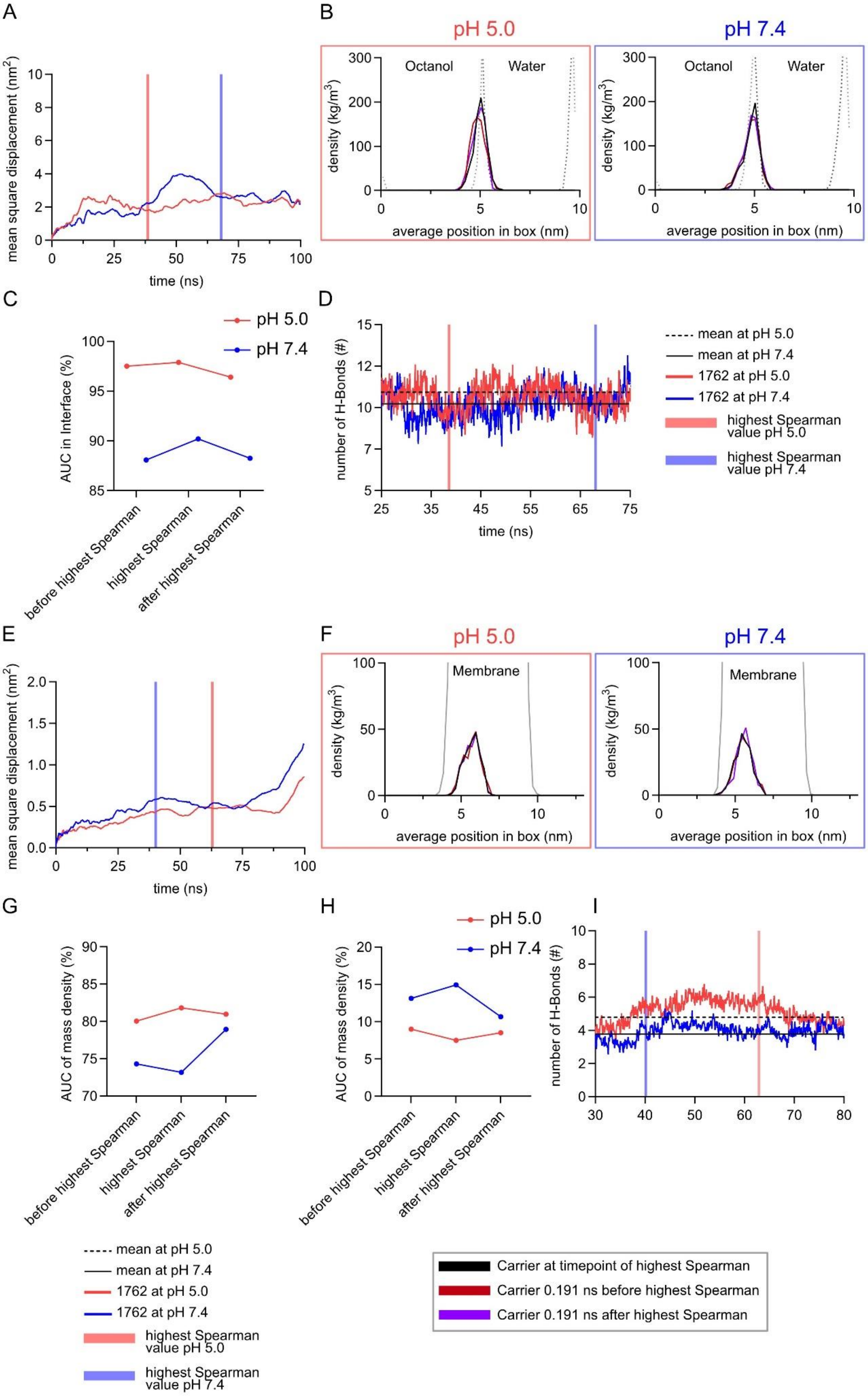

A
mean square displacement (nm^2)
time (ns)
B
pH 5.0
pH 7.4
Octanol
Water
density (kg/m^3)
average position in box (nm)
C
pH 5.0
pH 7.4
AUC in Interface (%)
before highest Spearman
highest Spearman
after highest Spearman
D
number of H-Bonds (#)
time (ns)
mean at pH 5.0
mean at pH 7.4
1762 at pH 5.0
1762 at pH 7.4
highest Spearman value pH 5.0
highest Spearman value pH 7.4
E
mean square displacement (nm^2)
time (ns)
F
pH 5.0
pH 7.4
Membrane
density (kg/m^3)
average position in box (nm)
G
AUC of mass density (%)
before highest Spearman
highest Spearman
after highest Spearman
H
pH 5.0
pH 7.4
AUC of mass density (%)
before highest Spearman
highest Spearman
after highest Spearman
I
number of H-Bonds (#)
mean at pH 5.0
mean at pH 7.4
1762 at pH 5.0
1762 at pH 7.4
highest Spearman value pH 5.0
highest Spearman value pH 7.4
Carrier at timepoint of highest Spearman
Carrier 0.191 ns before highest Spearman
Carrier 0.191 ns after highest Spearman

Figure VI.4: MD trajectory analysis of carrier 1762, one of the systems with the lowest overall prediction errors in interfacial simulations, shown for pH 5.0 (red) and pH 7.4 (blue). (A) Mean-square displacement (MSD) of carrier 1762 at the water–octanol (WO) interface. (B) Mass-density profiles at the WO-interface at the key frame of maximal model performance and at times ±0.191 ns relative to that frame. (C) Percentage of the area under the curve (AUC) between 4.2 and 6.2 nm representing carrier enrichment at the WO-interface. (D) Number of hydrogen bonds (H-bonds) between carrier 1762 and water over time, including the overall simulation mean. (E) MSD of carrier 1762 embedded in a POPC membrane at both pH values. (F) Membrane-spanning density profiles of carrier 1762 at the key frame and at times ±0.191 ns relative to that frame. (G) AUC between 5 and 7 nm quantifying carrier distribution within the membrane leaflet interior.
(H) AUC at the membrane–water interface (0 nm – 5 nm), indicating transient changes in interfacial localization. MSD, H-bonds depict the weighted mean of analysis outputs across simulation time, and weighted mean density profiles were computed from coordinate snapshots (.gro) at the indicated frames for carrier 1762 (n = 1). (I) Number of hydrogen bonds (H-bonds) between carrier 1762 and water over time, including the overall simulation mean.

Carriers embedded in a POPC membrane naturally exhibit more restricted motion than in water or octanol; consequently, the MSD values are smaller (Figure VI.4E). At pH 7.4, carrier movement increases initially and then reaches a plateau around the key frame, suggesting that the carrier has reached a membrane region where it gets trapped. Possibly due to interactions with both lipid headgroups of POPC and the aqueous phase. The overall MSD is larger than for the pH 5.0 simulation, likely reflecting weaker interactions for the less charged carrier. At pH 5.0 a similar confinement event is observed but occurs later in the trajectory.

The corresponding mass density distribution of the carrier 1762 reveals a noticeable shift toward the center of the membrane at pH 5.0, indicating potential local perturbation of the membrane. (Figure VI.4F). In Contrast, during the pH 7 simulation, the carrier gradually migrates toward the membrane‑water contact area after the key frame it moves back. (Figure VI.4F) As in the WO-interface simulations, the AUC analysis of the mass density distribution further highlights these changes. When evaluating the area spanning from bilayer midplane (7nm) to the midpoint of a single leaflet (5nm) (Figure VI.4G), a marginal increase in carrier density is observed at pH 5.0, which diminishes after 0.191 ns. Conversely, for the membrane-water region (Figure VI.4H), the AUC decreases over the same interval, suggesting reduced carrier occupancy at the boundary. At pH 7.4, the opposite trend is observed. The carrier density decreases within the leaflet interior and increases at the interface, consistent with enhanced interfacial location. After the key frame the carrier moves back towards the "starting" frame (frame before the key frame) suggesting that this movement provides us, like for the WO-interface with conformers with the highest predictive information. This behavior aligns with the notion that increased lipophilic interactions at neutral pH strengthens the carrier's association with the membrane.

Hydrogen-bond analysis revealed fewer overall H-bonds between carrier 1762 and the POPC membrane at pH 7.4 compared to pH 5.0. At the lower pH, the number of H-bonds decreased after the key frame, consistent with structural rearrangements occurring in the membrane under these conditions. At neutral pH, the number of H-bonds increased after the point of maximal model performance and then reached a plateau.

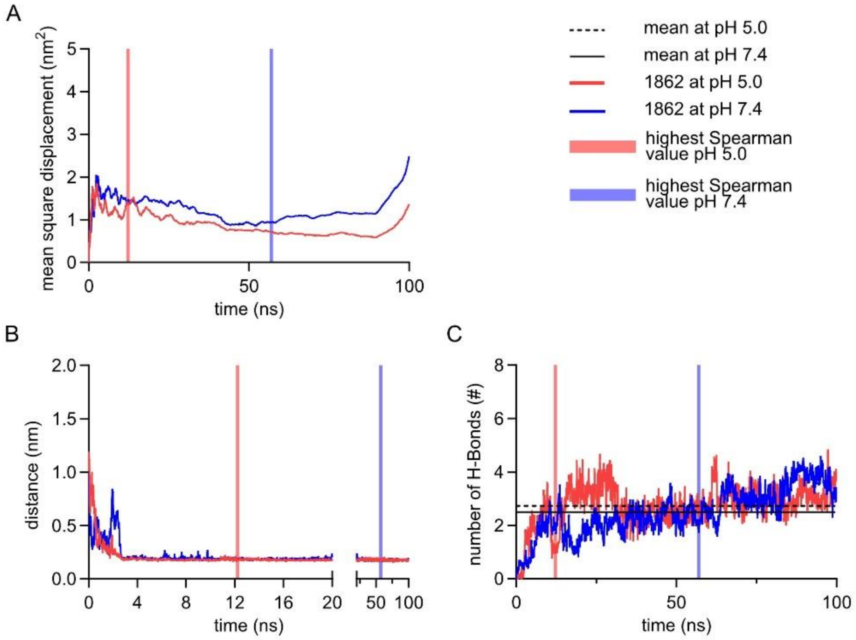


Figure VI.5: MD trajectory analysis of RNA simulations for the 1862 the carrier with the lowest error over the whole simulation. (A+B) Weighted mean distance between carrier and RNA. (C) Weighted rolling mean (11-frame window) of the number of hydrogen bonds (H-bonds) between carrier and RNA. All means are weighted by the relative state occurrence within the population at the respective pH.

In both pH conditions for Carrier-RNA interactions, the MSD for 1862, the carrier with the lowest overall error for this environment, shows pronounced motion that is not attributable to free diffusion; after binding to RNA (Figure VI.5A), a decrease in MSD is visible in both curves (Figure VI.4E), followed by a plateau. For both pH values, the frames with the highest Spearman's coefficient occur close to the onset of this plateau phase. Carriers in the vicinity of RNA, especially when simulated as single molecules, initially rely mostly on electrostatic interactions[242] with the negatively charged phosphate groups of RNA.

The observed differences are evident in the simulation analyses for 1862 (Figure VI.5B). At pH 7.4, carriers required significantly longer to approach RNA compared to pH 5.0. This trend is further supported by the hydrogen bond analysis (Figure VI.5C): while the overall mean number of hydrogen bonds differs only marginally between the two conditions, a distinct increase occurs earlier at pH 5.0 (before 25 ns) than at pH 7.4 (after 50 ns) in concordance with the timepoints of the highest Spearman value. These temporal differences likely account for the shift in time points of highest Spearman correlation. It could also provide a potential explanation for the marked disparity in predicted performance between the two pH levels. As noted above, EPIK-predicted protonation states were used for pH 7.4 and pH 5.0; however, lipophilic microenvironments can alter effective charge states[31], and particularly at pH 7.4, the predicted states may not accurately capture the true speciation relevant for RNA binding. In contrast, carrier-membrane and WO-interface simulations at pH 7.4 yielded comparable performance. Notably, the simple phase model (WO-interface) performed remarkably well, in line with experimental data[31], highlighting the predictive value of simplified models for assessing carrier transfection efficiency.

Across all three simulation environments, the frames with the highest Spearman correlations consistently aligned with major structural transitions in the MD trajectories, highlighting the ability of the learned representations to capture physically meaningful states.

5 Conclusions

In our study on integrating MD simulations into a ML framework for predicting the transfection efficiency of LAX, we show that MD-derived descriptors can meaningfully predict performance, especially in settings where conventional featurization struggles. We further demonstrate how MD and ML can be combined to enhance explainability by proposing a workflow that highlights time-localized, mechanistically relevant events along the delivery pathway.

As this is, to our knowledge, the first demonstration of such observations in this context, several limitations warrant attention. First, the absence of explicit molecule-molecule interactions may discard information that could improve predictive power, for example influence on the micro pKa values, or phase-behavior. Future work needs to carefully balance computational cost and information gain. Second, broader data coverage is essential. Expanding beyond a single case study to additional carriers, cell lines and cargo and incorporate additional simulations that further reflect the delivery process.

Looking ahead, we envision a closed-loop platform that integrates MD-informed descriptors with multi-objective ML optimization coupled to automated synthesis and formulation. Such a system could prioritize informative experiments via active learning, map structure–activity with time-resolved attributions and iterate rapidly toward candidates with improved performance and tolerability. Together, these advances would move the field toward self-driving discovery for nanocarriers and accelerate the development of lipo-xenopeptide-based delivery systems for new applications.

6 Materials and Methods

6.1 Micro pKa determination with EPIK

Starting from two-dimensional structures of lipid tails and headgroups (PCD) (Scheme VI.S1 for division convention), these were embedded as 3D-mol files and converted into Schrödinger input files for micro-pKa calculations using Schrödinger's (version 2025-1) Epik software (version 7.1).[236] pKa values were determined using a pH threshold of 1, upper and lower charge level windows of + 10 and – 10, and a maximum of 10 population states per molecule at the pH. A report for each structure was generated. Afterwards 3D mol files with correct stereochemistry of the different protonated populations were built with RDkit[243] and an inhouse script. The top three states per pH value and carrier, i.e. the carrier population with the highest percentages at this pH, were selected. This resulted in 153 structures per pH value, which were subsequently used for parametrization and molecular dynamics simulation as described below.

6.2 Parametrization

3D Mol files from the previous step were changed with Open Babel to mol2 files. The molecules were then parametrized with AmberTools23.[244] Partial charges were calculated via the Gasteiger method, to reduce computational cost, GAFF2 was used as a forcefield for the carriers, as well as for octanol. Since molecules were compared to each other, the minor loss in accuracy associated with the Gasteiger charge model was considered negligible. These parameters were then processed with Parmchk2 and topologies and PDB files were obtained using the tleap program. Finally, the PDB and topology files were converted into GROMACS input files using the Python library ParmEd and an in-house script.

6.3 Molecular Dynamics Simulation (MDS)

All molecular dynamics (MD) simulations were performed with GROMACS 2022.3[245–247] at 298.15 K and 1 atm. Temperature and pressure control were applied using combinations of the Nose–Hoover, V-rescale, C-rescale, Berendsen, and Parrinello–Rahman algorithms as specified for each system. Energy minimization was conducted stepwise for 50,000 steps, followed by short NVT and NPT equilibrations and 100 ns production runs. Electrostatics were treated with the Particle Mesh Ewald (PME) method and cutoff distances of 1.2 nm (or 0.9 nm for membrane systems). Lennard–Jones interactions used a force-switch scheme between 1.0 – 1.2 nm (or 0.9 nm cutoff for membrane systems). Dispersion corrections for energy and pressure were disabled unless stated otherwise. LINCS constraints were applied to all bonds involving hydrogen atoms, center-of-mass motion removal was disabled, and all integrations used the leap-frog algorithm. To minimize storage demands, water-molecules were excluded from the trajectory (xtc) output file for the Carrier-RNA and Carrier-Membrane simulations.

6.3.1 Carrier in vacuum

To assess carrier behavior in the absence of solvent, additional vacuum simulations were performed under NVT conditions. The same electrostatics and non-bonded settings were applied as in solvated systems, while center-of-mass motion was removed every 2 ps. Each system was equilibrated before a 5 ns production run (*dt = 2 fs*).

6.3.2 Carrier in Water–Octanol Interface

The water–octanol interface was prepared following the GROMACS tutorial[248] for biphasic systems. Carriers were placed in a 5 × 5 × 10 nm box, solvated with a pre-equilibrated 10 ns octanol layer, and subsequently with TIP3P water. Systems were neutralized according to their total charge. Equilibration consisted of 0.2 ns NVT (Nose–Hoover thermostat) during which carriers were pulled into the octanol layer with constant force, followed by 1.5 ns NPT (C-rescale barostat) with positional restraints on the carriers. Production simulations of 100 ns NPT (V-rescale thermostat, Parrinello–Rahman barostat) were then performed.

6.3.3 Carrier-RNA Systems

Model mRNA was built using Schrödinger's Maestro Suite, and a short fragment (40 bp, Table VI.S3) was parametrized with AmberTools23 using the Amber nucleic acid force field.

Files were converted to GROMACS input with ParmEd. RNA was positioned in the center of a 10 nm cubic box and restrained throughout the simulation. Carriers were placed at (2.5, 2.5, 7.5 nm) to ensure identical starting conditions. After solvation and neutralization, systems underwent energy minimization, 0.2 ns NVT, and NPT equilibration (with restraints), followed by 100 ns NPT production using the same parameters as for the water–octanol systems, but with restraints on the RNA for equal conditions.

6.3.4 Carrier– POPC Membrane Systems

A POPC bilayer containing 100 lipids per leaflet was built using CHARMM-GUI membrane Builder[249] and parametrized with the Lipid21 force field. The membrane was pre-equilibrated for 100 ns before carrier insertion. Systems were assembled in an 8.2 × 8.2 × 14 nm box, solvated, neutralized, and adjusted to 0.15 M NaCl. After energy minimization, a 125 ps NVT equilibration (V-rescale thermostat) was followed by a 125 ps NVT run in which carriers were pulled into the membrane center (Nose–Hoover thermostat). Subsequent 1.25 ns NPT (Berendsen barostat, grouped V-rescale thermostats) and 10 ns NPT (C-rescale barostat) equilibration phases allowed membrane relaxation. Production runs of 100 ns NPT followed with PME electrostatics and dispersion correction enabled.

6.3.5 Analysis

All simulations were run for 100 ns. After completion, carriers were centered and three segments from the trajectories were selected for frame extraction: the initial phase (0–40 ns, frames), the equilibrated phase (60–100 ns, frames), and the entire simulation (0–100 ns, frames). For each segment, .gro files were saved at defined time intervals. The resulting .gro files were converted to mol format using Open Babel, and molecular descriptors were calculated with RDKit via an in-house Python script. To conserve disk space, water molecules were excluded from trajectory outputs in RNA and membrane simulations. All simulations were performed once per system; subsequent statistical analysis was based on frame sampling as described above. For visualization of trajectory snapshots VMD 2[250] was used. Mean Square Displacement (MSD) was calculated via GROMACS, then weighed, based on the percentage of occurrence at this pH and the mean calculated via an inhouse Python script. Density distributions were calculated per frame and state via GROMACs and then the weighted mean was calculated as well. GraphPad Prism (GraphPad Software, La Jolla, USA, v. 10.6.1) was used to calculate overall mean of hydrogen bonds and area under the curve for density distributions. Affinity Designer 218.2 (version 2.5.7, Serif Ltd., West Bridgford, UK), and GraphPad Prism were used for visualization.

6.4 Carrier Encoding and Feature Generation

To featurize the molecular dynamics (MD) trajectories, each carrier was represented using a comprehensive set of molecular descriptors. To ensure that spatial and geometric information was preserved, we employed descriptor families that explicitly encode 3D molecular structure, including WHIM, GETAWAY, and related geometrical indices. Descriptor computation was performed using the RDKit cheminformatics library (version 2024.9.1).

For each frame of every trajectory, a total of molecular 984 descriptors were calculated. Subsequently, the mean and standard deviation across all frames were determined, yielding 1968 aggregated features per molecule. Descriptor calculation was performed independently for all simulation types (WO-Interface, RNA, membrane) and at two pH values (5 and 7.4).

To account for conformational variability and protonation effects, the three most probable protonation states were extracted for each pH condition. Descriptors of these states were weighted according to their relative population probabilities, resulting in a weighted descriptor vector per pH value. For benchmarking, a deterministic encoding using only the most likely state was also evaluated.

To capture both environment- and pH-dependent behavior, combined descriptor vectors were constructed: pH-combined vectors, concatenating pH 5.0 and 7.4 features. Simulation-combined vectors, merging descriptors across all simulation types to evaluate whether integrated environmental information improves model performance.

As a computationally lighter baseline, 2D molecular descriptors were also generated for each carrier. Furthermore, to benchmark trajectory-derived features against static molecular representations, 3D descriptors of the unprotonated, energy-minimized vacuum structure were computed using RDKit (see Section Molecular Dynamics Simulation (MDS)).

Descriptor sets were prepared for three temporal segments of each trajectory: start phase, Equilibrium phase, whole simulation.

This multi-scale encoding strategy was designed to capture both transient and equilibrium structural features relevant to carrier performance.

6.5 Model Selection and Feature Comparison

A model zoo comprising 13 different regression algorithms (Table S2) was used to systematically benchmark predictive performance across descriptor types. Prior to training, all feature sets were standardized using a StandardScaler (scikit-learn v.1.6.1). Transfection efficiency labels were log-transformed to normalize their distribution. Models were evaluated under three complementary data-splitting strategies: 5-fold cross-validation (CV) for generalization performance, chemical diversity splits to test extrapolation to novel chemotypes, and transfection-cliff splits to probe model sensitivity to steep response changes. Given that relative performance trends are often more informative than absolute numeric agreement, spearman rank correlation between predicted and experimental values was used as the primary evaluation metric.

6.6 Single-Frame Descriptor Analysis

To investigate temporal patterns within the trajectories, frame-wise descriptor sets were generated for every simulation frame and protonation state. For each frame, both weighted and deterministic protonation encodings were computed, and models were evaluated using 5-fold CV. For every simulation type, pH value, and frame index, the Spearman correlation coefficient between predicted and experimental activities was calculated. To smoothen short-term fluctuations, a rolling mean over 11 consecutive frames was applied. Local maxima in these smoothened correlation profiles were interpreted as potential dynamically relevant events.

7 Conflicts of interest

O.M. is a consultant for PARI Pharma GmbH, Boehringer-Ingelheim International, and AbbVie Deutschland GmbH on unrelated projects. O.M. is advisory board member for Coriolis Pharma GmbH, Corden Pharma GmbH, and AMW GmbH. O.M., and B.W., have equity interests in RNhale GmbH.

8 Acknowledgments

O.M. gratefully acknowledges funding from the DFG (Project 516634310) and ERC (ERC-2022-COG-101088587).

This work was supported through the Leibniz Supercomputing Centre in the framework of the project *AVOCADO2* to B.W.

9 Statement for the use of LLMs

During the preparation of this manuscript the authors used ChatGPT to improve readability and language. The text was reviewed afterwards, and the authors take full responsibility for the content of the publication.

10 Supplementary Information

Materials. CleanCap FLuc mRNA (5moU) was obtained from TriLink BioTechnologies, San Diego, CA, USA. Murine neuroblastoma cell line Neuro2a (N2a) was purchased from the American Type Culture Collection, ATCC, Manassas, VA, USA. The human cervix carcinoma cell line (HeLa) was obtained from the German Collection of Microorganisms and Cell Cultures GmbH, DSMZ, Braunschweig, Germany. The human embryonic kidney cells (HEK-293T)…. Dulbecco's Modified Eagle's Medium (DMEM) low glucose with sodium bicarbonate, sodium pyruvate and L-glutamine, and fetal bovine serum (FBS) were purchased from Sigma-Aldrich, St. Louis, MO, USA. Penicillin-streptomycin (10,000 U/mL; 10 mg/mL) and trypsin/EDTA 10× were purchased from PAN-Biotech GmbH, Aidenbach, Germany. Luciferase Cell Culture Lysis 5× reagent and beetle luciferin sodium salt were obtained from Promega, Madison, WI, USA, and ATP from Roche Diagnostics, Mannheim, Germany. Coenzyme A trilithium salt, DL-dithiothreitol, and glycylglycine were purchased from Sigma-Aldrich, St. Louis, MO, USA. 3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2H-tetrazolium bromide (MTT) was purchased from Carl Roth, Karlsruhe, Germany. Dimethyl sulfoxide (DMSO) was obtained from Fisherscientific, Loughborough, UK. D(+)-Glucose 1-hydrate was purchased from Applichem, Darmstadt, Germany. Ethylenediaminetetraacetic acid (EDTA) disodium salt dihydrate was purchased from Merck, Darmstadt, Germany. 4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES) was purchased from BIOMOL GmbH, Hamburg, Germany.

Particle formation. The FLuc-mRNA was diluted in HBG (20 mmol/L of HEPES, 5% (w/v) glucose, pH 7.4) to a concentration of 25 µg/mL. LAF-XP carriers were diluted in purified water to appropriate concentrations for calculated N/P (nitrogen/phosphate) ratios under consideration of all primary, secondary, and tertiary amines except the tertiary amine within the N-(trifluoroethyl)iminodiacetyl (TFE-IDA). (Table VI.S1) Particles were formed by mixing equal volumes of mRNA dilution and LAF dilution *via* rapid pipetting, followed by 40 min incubation at RT yielding a final mRNA concentration of 12.5 µg/mL.

Cell culture. The murine neuroblastoma cell line Neuro2A (N2a), the human cervix carcinoma cell line (HeLa) and the human embryonic kidney cells (HEK-293T) were cultured in Dulbecco's Modified Eagle's Medium (DMEM)-low glucose (1 g/L glucose) containing L-glutamine, sodium bicarbonate and sodium pyruvate supplemented with 10% FBS, 100 U/mL of penicillin, and 100 µg/mL of streptomycin. Cells were cultured at 37 °C and 5% CO_2 at a relative humidity of 95%.

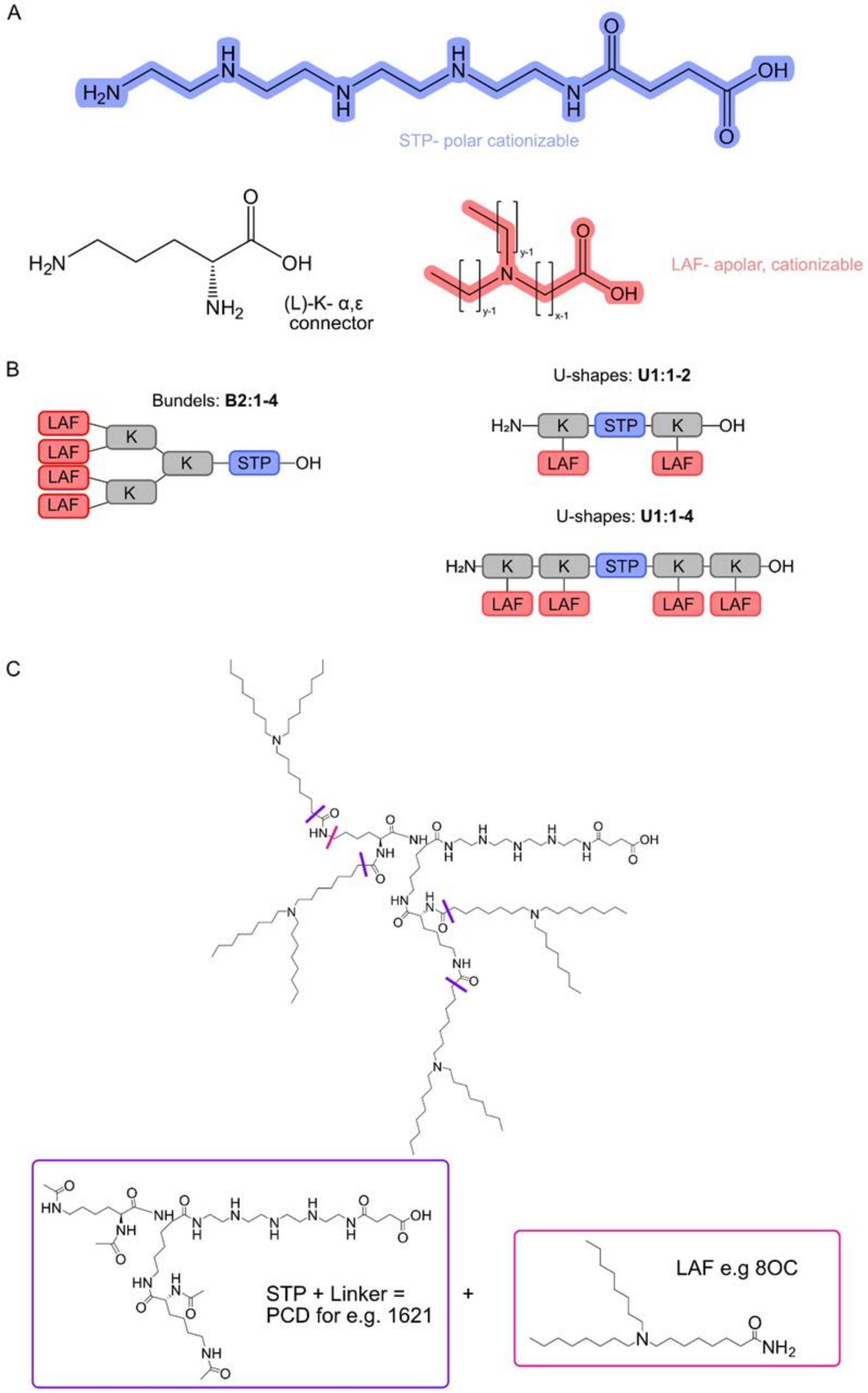
Luciferase expression assay. One day prior to transfection, 10,000 cell/well in case of N2a and HEK-293T cells and 5000 cells/well in case of HeLa cells were seeded in 96-well plates. Shortly before the transfection, cell culture medium was replaced by 99 µL fresh medium supplemented with 10% FBS. LAF-XP polyplexes were formed as described above (12.5 µg/mL mRNA-FLuc) and transfected at a dose of 12.5 ng mRNA-FLuc per well. HBG buffer (1 µL) was used as negative control. After incubation at 37 °C for 24 h, the medium was removed, cells were lysed with 100 µL of cell culture 0.5× lysis buffer, and frozen at −80 °C overnight. Prior to measurement, plates were thawed (1 h, RT, 25 rpm) on a rocking shaker. Cell lysates were 1:100 diluted in PBS and mixed thoroughly. Luciferase activity in 35 µL of diluted lysate was measured with a Centro LB 960 microplate luminometer (Berthold Technologies, Bad Wildbad, Germany) after addition of 100 µL LAR buffer (20 mmol/L glycylglycine, 1 mmol/L $MgCl_2$, 0.1 mmol/L EDTA, 3.3 mmol/L dithiothreitol, 0.55 mmol/L adenosine 5′-triphosphate, 0.27 mmol/L coenzyme A, pH 8.0 - 8.5) supplemented with 5% (v/v) of a mixture of 10 mmol/L luciferin-sodium and 29 mmol/L glycylglycine with a measurement duration of 10 s. Transfection efficiency was calculated as relative light units (RLU) per seeded number of cells per well after background subtraction (*i.e.*, RLU values of HBG-treated cells). Experiments were carried out in triplicates.

Table VI.S1: Amines in lipo-xenopeptides. As well as used N/P, molecular weight

ID-number	MW free base	MW HCl salt (36,5 g/mol per HCl)	N/P	counted amines/ oligomer
1611	1501.42	1720.42	18	6
1613	2585.29	2840.79	24	7
1621	2136.44	2391.94	24	7
1719	2984.89	3386.39	12	11
1730	2407.85	2772.85	12	10
1745	2536.03	2937.53	12	11
1746	1277.02	1496.02	18	6
1752	2360.92	2616.42	24	7
1753	2809.79	3065.29	24	7
1754	2473.15	2728.65	24	7
1755	2697.57	2953.07	24	7
1758	2760.46	3161.96	12	11
1759	3209.33	3610.83	12	11
1760	2872.68	3274.18	12	11
1761	3097.11	3498.61	12	11
1762	2360.92	2616.42	24	7
1763	1389.24	1608.24	18	6
1764	1613.67	1832.67	18	6
1765	1445.35	1664.35	18	6
1766	1557.56	1776.56	18	6
1791	2369.01	2624.51	24	7
1792	2603.06	2858.56	24	7
1793	2333.1	2588.60	24	7
1794	2531.23	2786.73	24	7
1813	2094.41	2349.91	24	7
1814	2318.84	2574.34	24	7
1816	2928.79	3330.29	12	11
1821	1968.48	2187.48	18	6
1822	1896.66	2115.66	18	6

1823	2640.21	3005.21	12	10
1824	2874.26	3239.26	12	10
1825	2604.03	2969.03	12	10
1826	2802.43	3167.43	12	10
1827	1473.4	1692.40	18	6
1840	1543.54	1762.54	18	6
1841	1583.6	1802.60	18	6
1842	1517.45	1736.45	18	6
1843	1555.55	1774.55	18	6
1844	1458.39	1640.89	18	5
1845	1500.47	1682.97	18	5
1858	3069.06	3470.56	12	11
1859	3149.19	3550.69	12	11
1860	3016.89	3418.39	12	11
1861	3093.08	3494.58	12	11
1862	2898.76	3227.26	12	9
1863	2982.92	3311.42	12	9
1864	2403	2658.50	24	7
1865	2443.07	2698.57	24	7
1867	2415.01	2670.51	24	7
1868	2317.85	2536.85	24	6
1869	2359.93	2578.93	24	6
1888	1598.49	1817.49	18	6
1909	2457.96	2713.46	24	7

Scheme VI.S1: Carrier structure and fragmentation for pKa calculation. A) Building blocks for all carriers simulated. B) Different topologies found in the dataset build out of the building blocks. C) Fragmentation for pKa calculation shown on 1621.


A
STP- polar cationizable
(L)-K- α,ε connector
LAF- apolar, cationizable
B
Bundels: B2:1-4
U-shapes: U1:1-2
U-shapes: U1:1-4
LAF
K
STP
H₂N
OH
C
STP + Linker = PCD for e.g. 1621
+
LAF e.g 8OC

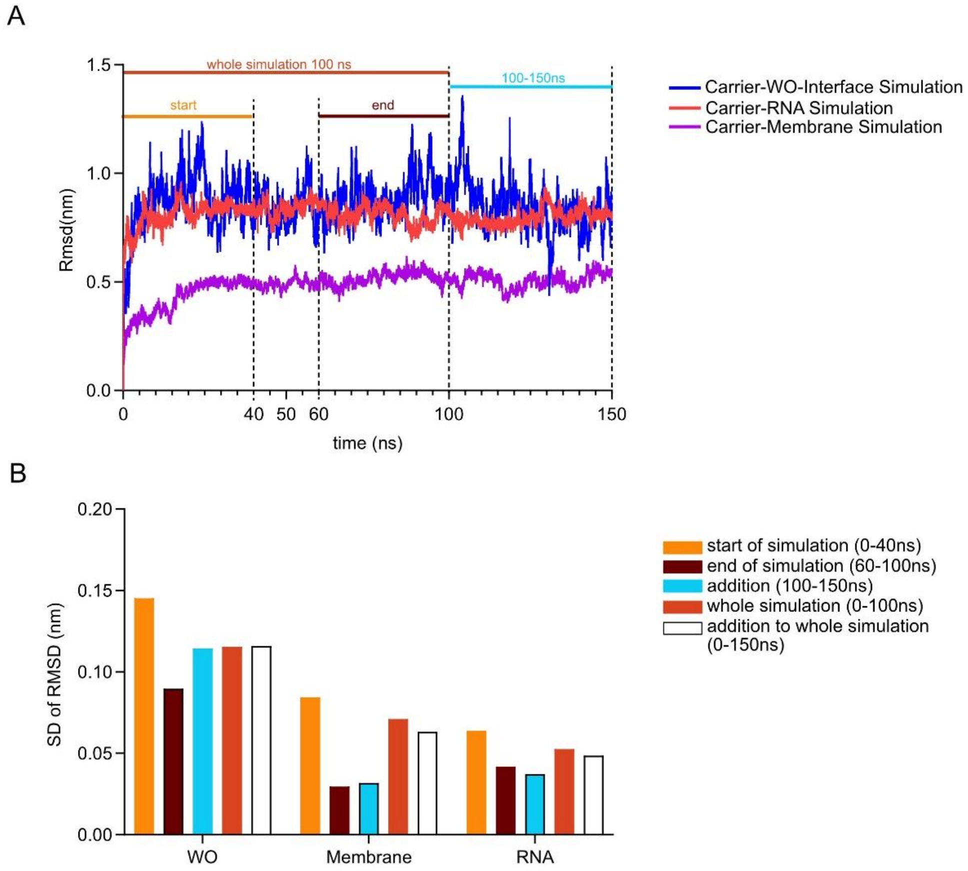


Figure VI.S1: Root Mean Square displacement (RMSD) of 1611 at pH 5 in different simulation setups over 150ns. (A) Shows RMSD over time for the three simulation environments while also showing the different parts of interest of the simulation (B) Standard deviation of the RMSD in the different time intervals. (n = 1)

Table VI.S2: Models used in the model zoo with respective hyperparameter and the python library they were imported from

Model Name	Key Hyperparameters	Library
DummyMean	strategy = "mean"	scikit-learn (sklearn.dummy)
Linear	default parameters	scikit-learn (sklearn.linear_model)
Ridge	alpha = 1.0, random_state = 42	scikit-learn (sklearn.linear_model)

Lasso	alpha = 1e-3, max_iter = 5000, random_state = 42	scikit-learn (sklearn.linear_model)
ElasticNet	alpha = 1e-3, l1_ratio = 0.5, max_iter = 50000, random_state = 42	scikit-learn (sklearn.linear_model)
KNN	n_neighbors = 7	scikit-learn (sklearn.neighbors)
SVR	C = 10.0, gamma = "scale", epsilon = 0.1	scikit-learn (sklearn.svm)
RandomForest	n_estimators = 300, n_jobs = -1, random_state = 42	scikit-learn (sklearn.ensemble)
ExtraTrees	n_estimators = 400, n_jobs = -1, random_state = 42	scikit-learn (sklearn.ensemble)
GradientBoosting	default parameters, random_state = 42	scikit-learn (sklearn.ensemble)
MLP	hidden_layer_sizes = (256,128), activation = 'relu', alpha = 1e-4, learning_rate_init = 1e-3, max_iter = 500	scikit-learn (sklearn.neural_network)
XGB	n_estimators = 600, max_depth = 6, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 0.8, reg_lambda = 1.0, tree_method = 'hist', random_state = 42	xgboost

LGBM	n_estimators = 1000,
	num_leaves = 63,
	learning_rate = 0.05,
	subsample = 0.8,
	colsample_bytree = 0.8,
	reg_lambda = 1.0,
	random_state = 42
		lightgbm

Table VI.S3.- Sequenz of the mRNA part containing 40 Bases

5´*GACGGCAACAUCCUGGGGCACAAGCUGGAGUACAACUACA*3´

Chapter VII - Meta-Learning as a Promising Strategy for Lipid Nanoparticle Optimization and Ionizable Lipid Discovery

1 Abstract

The rapid expansion of LNP based RNA therapeutics has created an urgent need for predictive tools that can accelerate the design of formulations and novel lipid compounds. However, formulation development remains challenging due to complex, multistep delivery mechanisms and the scarcity of high-quality experimental data. Conventional machine-learning approaches often struggle to extrapolate to new chemical scaffolds, cargos, and cell types. Here, we explore few-shot meta learning (FSL) as a strategy to overcome data scarcity in early-stage LNP development. Using a recently published dataset on lipid-based delivery systems, we created chemically, and contextually coherent meta-learning tasks based on data provenance and formulation conditions. Several FSL algorithms were benchmarked against supervised baselines using both Morgan fingerprints and graph-based encodings. To emulate challenging extrapolation, all siRNA-related data were withheld during meta-training and used solely for testing. Model-agnostic meta-learning (MAML) substantially outperformed conventional supervised and transfer-learning baselines, achieving an average R^2 of 0.38 ± 0.049 for siRNA delivery, compared with near-zero performance for non-meta models. In a retrospective active-learning simulation, meta-trained models identified high-performing candidates within the first acquisition rounds, achieving markedly higher hit rates and enrichment factors than random forest and random selection baselines. To validate these findings experimentally, we synthesized 15 new ionizable lipids and generated in vitro transfection data across multiple cell lines and RNA cargos. Despite the very small dataset, MAML achieved superior predictive performance to RF across all settings, including Pearson correlations up to 0.63 for siRNA delivery. Together, these results demonstrate that FSL provides a powerful and generalizable framework for guiding formulation design in data-limited environments, enabling faster and more informed exploration of the RNA delivery design space.

Keywords: Lipids, Meta Learning, Few-Shot Learning, Machine Learning, Lipid Nanoparticle

2 Main

The field of RNA therapeutics has expanded rapidly in recent years, transforming from a niche research area into a cornerstone of modern drug development. Pioneering approvals such as the mRNA vaccines Comirnaty (Pfizer-BioNTech) and Spikevax (Moderna), and siRNA drugs including Onpattro, Givlaari, and Oxlumo (Alnylam Pharmaceuticals), have demonstrated the therapeutic potential of RNA across infectious, genetic, and metabolic diseases[251,252].

A major contributor to this progress is the advancement of lipid nanoparticle (LNP) technology, which protects fragile RNA from degradation and enables efficient delivery to target tissues. Building on the clinical success of LNP mRNA vaccines during the COVID-19 pandemic[5,6], LNPs have become the leading non-viral platform for mRNA delivery, with ongoing expansion to other RNA modalities. Compared with viral vectors, whose translation can be limited by immunogenicity, toxicity, manufacturing complexity, and payload constraints[14,253], LNPs offer synthetic tunability, favorable biocompatibility, and scalable production via microfluidic-mixing methods[254,255]. An LNP typically comprises four to five lipid components: an ionizable lipid, phospholipid, cholesterol, and PEG-lipid, occasionally supplemented by a targeting lipid[30]. Each component serves a distinct physicochemical function, but the ionizable lipid plays the dominant role in RNA encapsulation, endosomal escape, and delivery efficiency[33]. Over the past decade, thousands of ionizable lipid structures have been synthesized and screened, and high-throughput (HT) formulation and testing platforms have been developed to accelerate discovery[256–258]. Nevertheless, LNP optimization remains a high-dimensional, multi-parameter problem, where optimal performance depends not only on chemical composition but also on RNA type[259], target tissue[42] and mixing conditions[255].

The wide spread use of artificial intelligence (AI) and machine learning (ML) provides powerful strategies to navigate this complex formulation landscape. ML models have successfully been applied to predict encapsulation efficiency[260], particle size[261], cell selectivity[262], and in vitro transfection performance from experimental data[128,263]. Existing approaches generally fall into two categories: (i) high-throughput screening (HTS) + ML integration, where large, well-controlled datasets are used for model training[92,256,264]. Although robust, these approaches are often resource- and material-intensive. Or (ii) data aggregation from literature, where information from multiple studies is merged to expand the accessible chemical space. While inexpensive, this strategy introduces heterogeneous data quality, experimental bias, and domain noise, which can compromise generalizability[265,266].

Moreover, most conventional ML models are limited by their inability to extrapolate to new chemical scaffolds, tissue types, or RNA cargo scenarios that inherently suffer from data scarcity[154,240,267].

Transfer learning (TL) has been proposed to reuse prior knowledge from related tasks, e.g., leveraging models trained on large molecular datasets to fine-tune predictions for specific targets[71,72]. However, in molecular ML applications, TL frequently faces complex case-to-case differences, making its implementation quite cumbersome[268].

An alternative paradigm, few-shot learning (FSL), directly addresses data scarcity by teaching models to learn new tasks from only a few labeled examples[269]. Rather than focusing on single prediction tasks, FSL trains on distributions of tasks, enabling rapid adaptation to novel conditions. FSL has already demonstrated promise in drug discovery, where it has improved small-molecule activity prediction[270], drug-target interaction modeling[271], and ADMET property estimation[272]. Despite these advances, no studies to date have explored FSL for drug-delivery optimization, even though formulation research often faces the same low-data challenges.

In this proof-of-concept study, we explore the feasibility of few-shot learning for the early-stage development of novel ionizable lipids. Specifically, we: (i) benchmark multiple FSL algorithms and molecular featurization strategies to simulate extrapolation to unseen cargos; (ii) investigate a meta-trained model within a retrospective active-learning framework, assessing whether it can guide formulation decisions for a held-out RNA cargo; and (iii) validate experimentally by synthesizing a library of 15 ionizable lipids and testing their performance across multiple cell types and RNA cargos.

We hypothesize that few-shot learning can leverage shared latent representations of molecular and formulation descriptors to generalize across different LNP compositions and biological contexts, thereby accelerating the design-make-test-learn cycle for RNA therapeutics. Here, we provide a reproducible framework for AI-guided formulation design in data-scarce regimes.

We employed a recently published dataset on lipid-based delivery systems for model development and evaluation[264]. In the corresponding study, the authors used historical data to design novel lipids and reported promising outcomes. However, we argue that the design of entirely new formulations for previously unseen cargo types or cell lines is difficult when relying on datasets that do not include such variations. Nevertheless, such datasets can still be valuable for enabling a model to meta-learn transferable knowledge, allowing it to rapidly extract useful information from related challenges even when only limited new data is available.

To demonstrate this concept, we divided the datasets into tasks in two steps. First, the data were split according to their source to ensure that each task contained data originating from the same source, thereby avoiding potential biases introduced by merging data from different origins. Second, the data were partitioned by a defined criterion to ensure that each task comprised only comparable data within itself. Additionally, a label-based binning-splitting approach was used to ensure the same label distribution in support and query set (Figure VII.1A). Subsequently, we compared several few-shot learning (FSL) algorithms with classical supervised learning models (Figure VII.1B). Because molecular featurization has a major influence on model performance in both conventional and meta-learning settings, we evaluated two different molecular encodings: a bit-vector representation based on Morgan fingerprints, and a learnable graph-based embedding similar to the one used in the original dataset's publication. To emulate a challenging extrapolation task, all siRNA-related data were excluded from the training set and reserved as a holdout test set. Model robustness was further assessed using a random seed strategy that redistributed the data into different support-query splits.

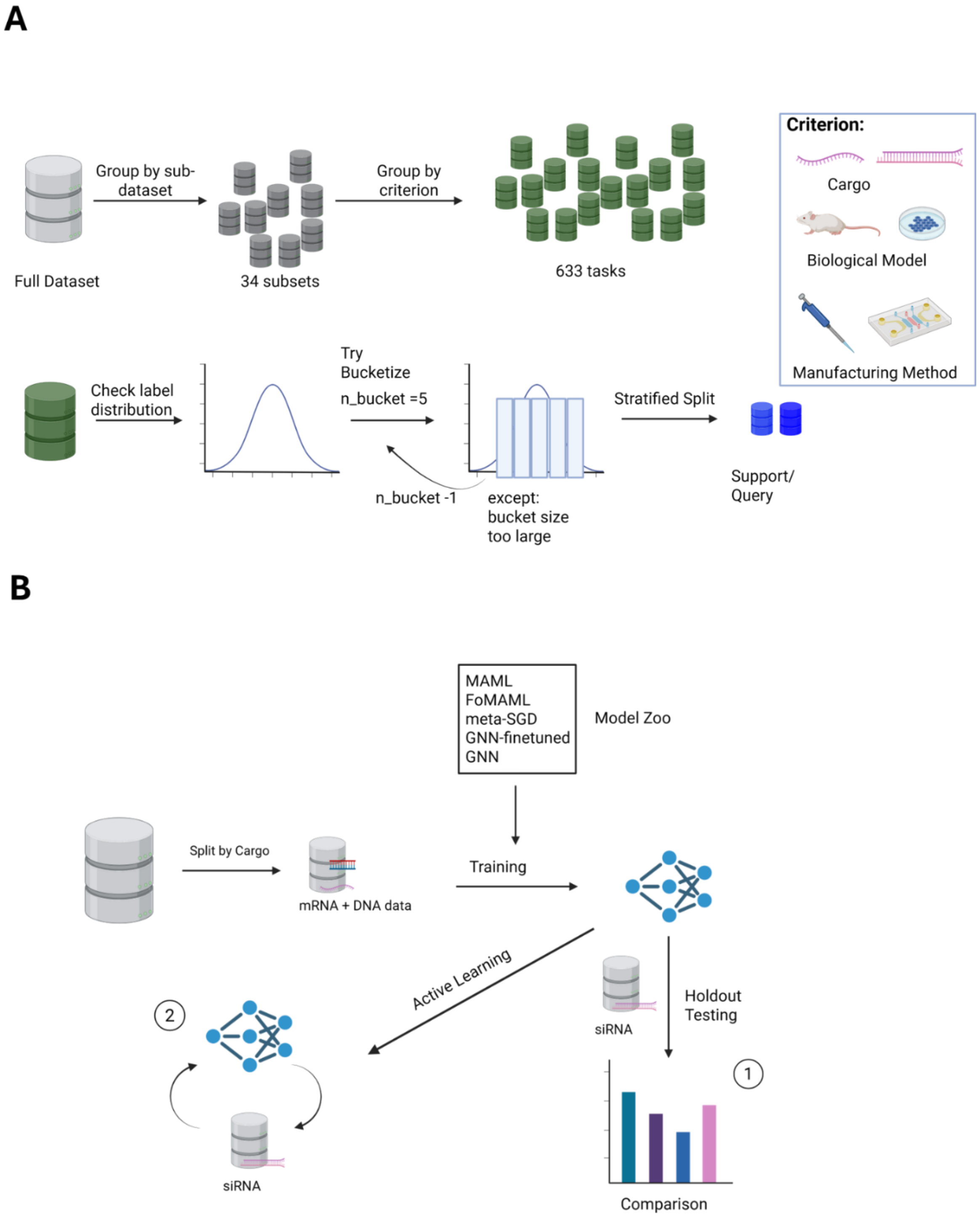


Figure VII.1: A) Schematic overview of data preparation. Full data was grouped into source-related subsets and further grouped into tasks suitable for meta-learning. Tasks were discretized for proper support-query splits. B) Overview over the experiments used to test the potential of cargo holdout by 1) model screening and 2) active learning.

As expected, the supervised baseline models trained only on the meta-training data without further fine-tuning showed no meaningful generalization for the unseen cargo, resulting in R^2 values close to zero (Figure VII.2A). After fine-tuning on the support set, the performance improved slightly, reaching mean R^2 values of 0.046 ± 0.028 for the fingerprint representation and 0.078 ± 0.021 for the graph-based encoding. In contrast, all few-shot models achieved substantially higher performance. The model based on model-agnostic meta-learning (MAML) performed best, yielding an average R^2 of 0.38 ± 0.049 for the fingerprint features. Interestingly, the graph-based version performed slightly worse, with R^2 values of 0.28 ± 0.16 for MAML and 0.29 ± 0.15 for first-order MAML (FoMAML). This result was unexpected, as graph-based encodings are generally assumed to be more expressive[273]. A plausible explanation is that the higher parameterization and complexity of the graph models led to reduced robustness in this low-data regime, which is also reflected in the larger performance variance.

In molecular discovery workflows, algorithms are often used to prioritize new candidates for experimental testing. This strategy, known as active learning (AL), leverages model predictions to guide data acquisition[60,274–276]. To investigate whether few-shot learning could also be beneficial in this context, we applied the meta-trained MAML model from the previous experiment in an AL-like simulation. The holdout siRNA dataset was again used but restricted to samples matching the criterion "Whitehead_siRNA_whitehead_lipidoid_generic_cell_nan_in_vitro" to simulate realistic laboratory conditions, where data typically originate from one research group and one cell line. To keep the setup straightforward, the top 5 % of samples were defined as hits, and a greedy acquisition strategy was applied, where the model iteratively selected the candidate with the highest predicted performance. Although batch selection is common in practice, we opted for batch-1 acquisition here to control for batch effects and to evaluate the intrinsic ranking ability of each method on identical, incremental updates. For benchmarking, we compared the results to random selection and to a random forest (RF) model, which is often used in small-data and active learning settings[277,278]. All molecules were encoded as Morgan fingerprints, as this representation had shown superior stability in our previous experiments. The initial training set was constructed using a centroid-based sampling approach to ensure broad coverage of the feature space.

Across 100 simulated acquisition iterations, the MAML model consistently demonstrated strong early hit detection, identifying four of the top five hits within the first few iterations (Figure VII.2B). The RF model also detected one hit early on but failed to discover additional high-performing candidates, while random selection eventually surpassed RF performance. This suggests that the RF model struggled to capture the biological context underlying the structure-activity relationships. Overall, the MAML-based strategy identified 36 out of 59 possible hits, compared with only 3 hits for RF and 5 hits for random selection. To further assess the ability of meta-learned models to guide early formulation optimization, we compared the performance of the MAML model to the RF baseline across several active-learning-related metrics sampled at iteration 5 to monitor early discovery (Figure VII.2C). Overall, MAML clearly outperformed the RF baseline in all evaluated criteria. The hit rate (Hit@k) and enrichment factor (EF@k) of MAML were four times higher than those of RF, indicating a substantially improved capability to identify high-performing formulations among the top-ranked candidates. Similarly, the best-so-far@k score was higher for MAML (4.18 vs. 2.91), confirming that the meta-learned model more consistently selected top-yielding formulations during the iterative search. In addition, MAML achieved a markedly lower simple regret (0.36 vs. 1.63), demonstrating faster convergence towards optimal formulations. The corresponding yield@k further supports this trend, with MAML producing roughly an order-of-magnitude higher mean yield compared to the RF baseline (3.17 vs. 0.18). These results highlight the advantage of few-shot meta-learning in guiding candidate selection under limited-data conditions, especially in the early stages.

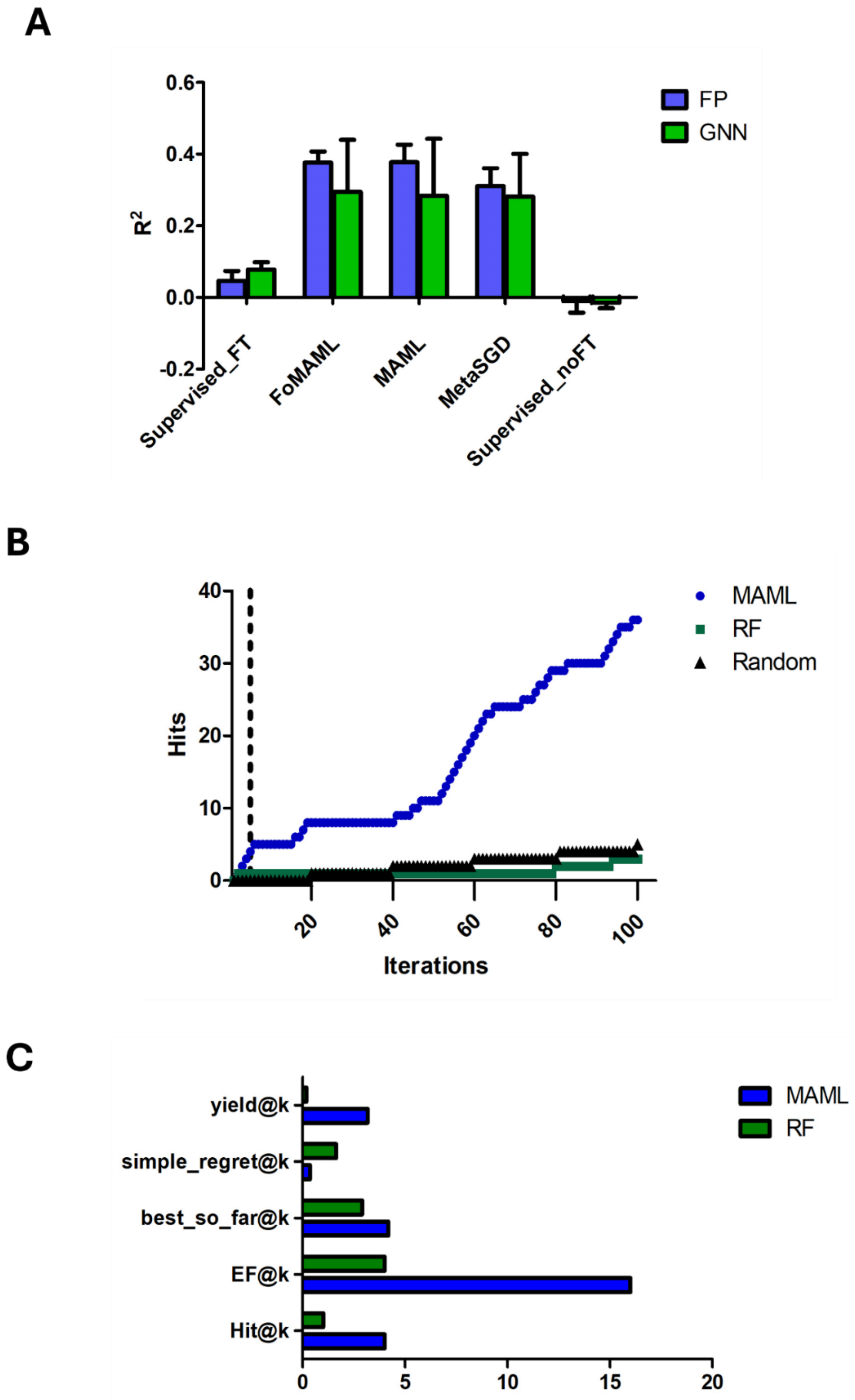


Figure VII.2: Results of the siRNA holdout experiments A) Model comparison of different few-shot algorithms (FoMAML, MAML, mSGD), supervised ANN (supervised_noFT) and a transfer learning ANN (supervised_FT) ranked by R2 value. B) Active Learning of the MAML meta model compared to RandomForrest (RF) and Random Picking baselines. C) Active Learning comparison of MAML and RF at round 5.

To validate our meta-learning approach, we synthesized a small library of 15 new ionizable lipids (Supplementary Table VII.1) and formulated corresponding LNPs for siRNA and mRNA delivery. Acrylates were first obtained by esterification of the respective alcohols with acryloyl chloride and subsequently reacted with polyamine head groups via a solvent-free aza-Michael reaction (Figure VII.3). Conversion rates were characterized via ^{1}H NMR (Figure VII.S4 – VII.S24) and final lipid structures were confirmed via MS ESI (Supplementary Table VII.1). The resulting crude lipids were formulated with cholesterol, 1,2-distearoyl-sn-glycero-3-phosphocholine, and DMG-PEG2000 in ethanol and mixed with siRNA or mRNA using a high-throughput microfluidic device. The resulting particles were dialyzed against PBS and characterized for size and polydispersity index (PDI) (Figure VII.S25).

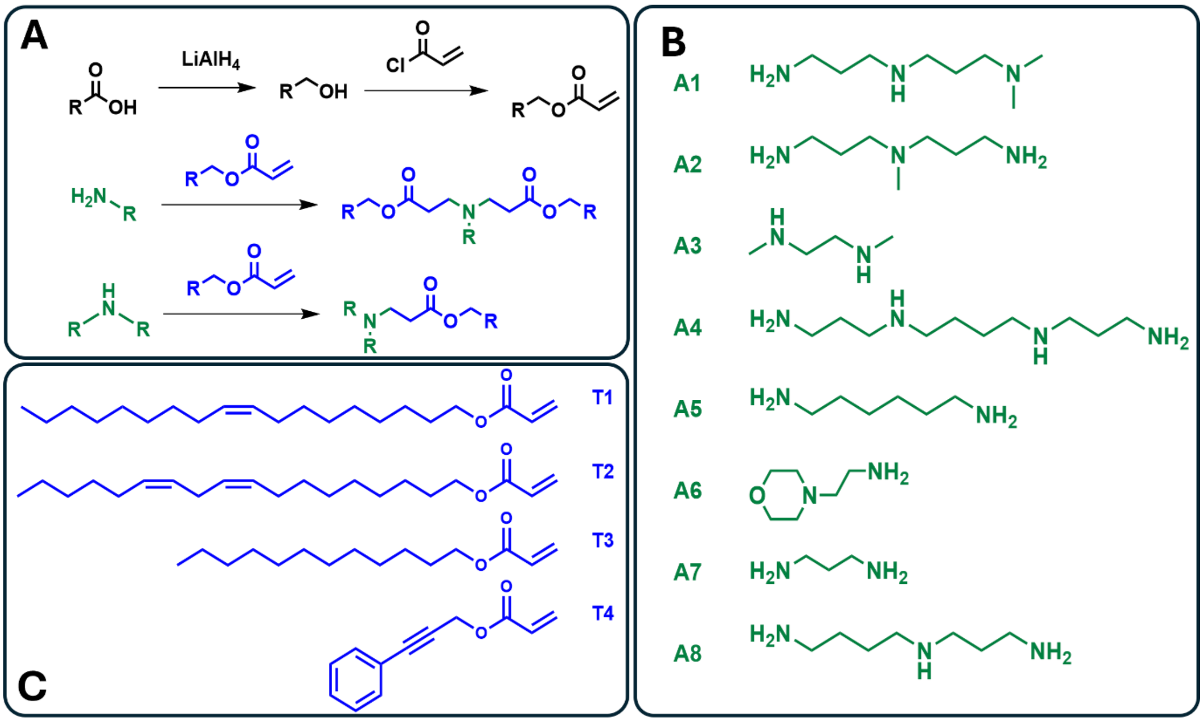


Figure VII.3: Synthesis route and lipid design of ionizable lipids. (A) Starting from either the carboxylic acid (reduced to the corresponding alcohol) or directly from the alcohol, the alcohol was esterified with acryloyl chloride; subsequent solvent-free aza-Michael addition furnished the final ionizable lipids. (B) Amine head groups and (C) alkyl tails employed in this study. Combination of various alkyl tails and amine head groups led to 15 chemical diverse ionizable lipids.

LNP transfection efficiency was evaluated in epithelial (H1299, A549, MDA-MB-231) and dendritic (DC2.4) cell lines, quantifying either Firefly luciferase knockdown (siRNA) (Figure VII.4A) or mRNA-mediated luciferase expression (Figure VII.4B). This dataset of labeled *in-vitro* data served as the basis to test and validate our model on own data. We note that for some treatments, the remaining expression exceeded 100%. For siRNA knockdown experiments, stable reporter cell lines are required, but their physiological state can influence apparent knockdown. Upon LNP treatment, some cells reduce proliferation and redirect metabolic resources toward processing the particles via the endo-lysosomal pathway, particularly in cases with high uptake but limited endosomal escape. We observed a similar trend where untreated cells proliferated freely and entered partial quiescence, while LNP-treated H1299-Luc cells showed enhanced endo-lysosomal trafficking and stress, which resulted in slower proliferation and failure to reach quiescence. Consequently, these cells displayed higher apparent luciferase “remaining expression” compared to the untreated reference.

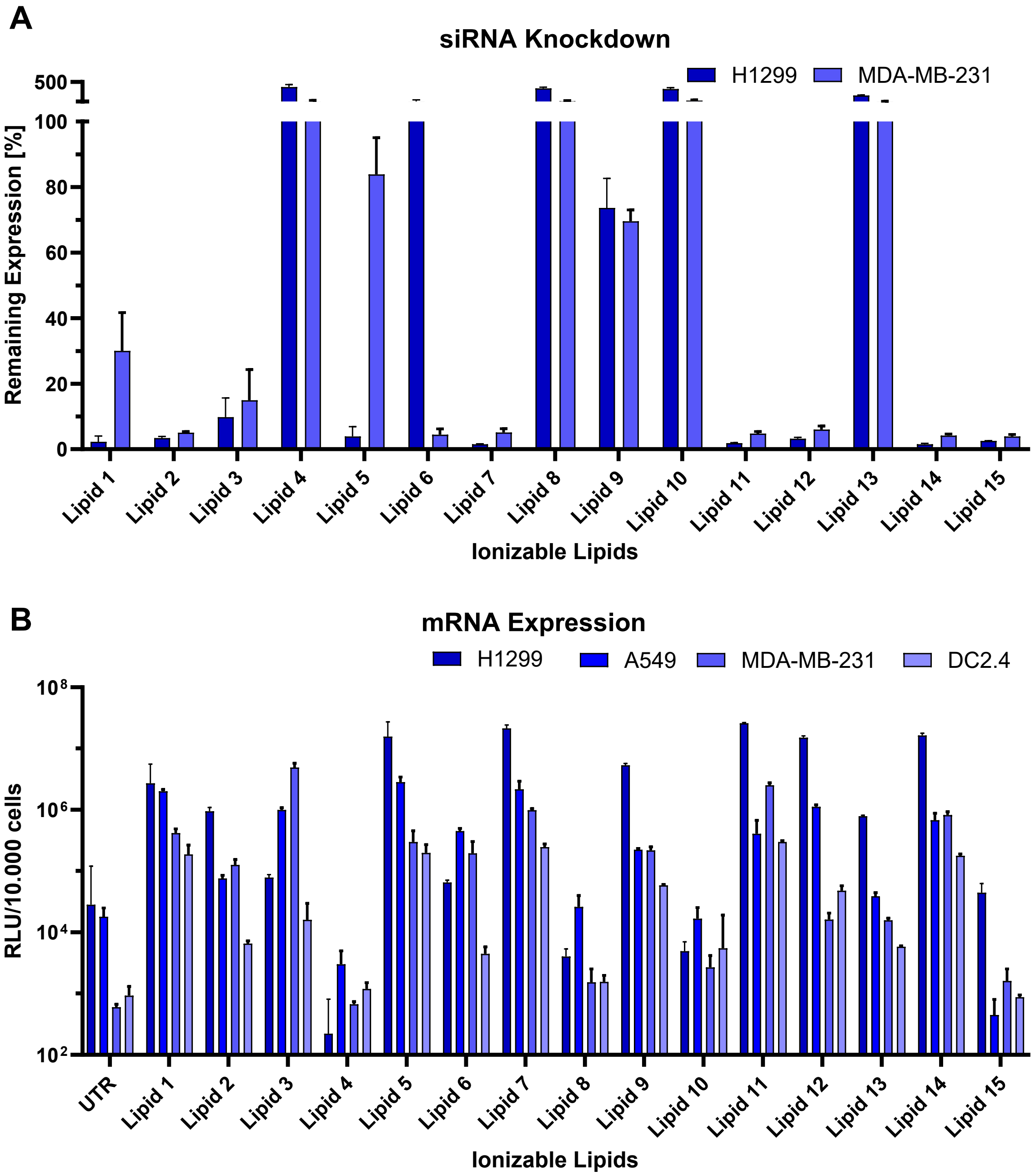


Figure VII.4: Transfection efficiency of LNPs formulated with ionizable lipids 1–15. (A) H1299 and MDA-MB-231 firefly-luciferase reporter cells were treated with LNPs loaded with luciferase-specific siRNA (50 pmol) for 48 h; knockdown is reported as remaining expression (%) relative to untreated controls (UTR, set to 100%). (B) H1299, A549, MDA-MB-231 and DC2.4 cells were treated with LNPs encapsulating firefly-luciferase mRNA (150 ng) for 24 h; luciferase activity (RLU/10,000 cells) is shown on a log10 scale, with UTR indicating untreated controls. Data are mean ± SD from technical replicates (n=3).

A five-fold cross-validation (5-CV) setup was used, with nine lipids serving as the support set, three as validation data for checkpoint selection, and three as test data (Figure VII.5A). This experiment aimed to assess whether few-shot learning provides an advantage over traditional models such as random forests in very low-data scenarios. Model performance was evaluated for mRNA transfection in A549, DC2.4, H1299, and MDA-MB-231 cells, and for siRNA transfection in H1299-FLuc and MDA-MB-231-FLuc cells.

Despite the small dataset, the MAML model achieved notable predictive power for siRNA delivery, with Pearson correlation coefficients of 0.63 for H1299 and 0.61 for MDA-MB-231 (Figure VII.5), clearly outperforming the RF baseline, which reached 0.27 and 0.45, respectively. For mRNA transfection, moderate correlations were obtained for DC2.4 (r = 0.35), H1299 (r = 0.37), and MDA-MB-231 (r = 0.36), while no positive correlation was observed for A549. Nonetheless, the MAML models consistently outperformed the RF baselines across all settings. These findings indicate that few-shot meta-learning can provide valuable predictive insights even in very-low-data environments, supporting its potential as a practical tool for early-stage lipid formulation design, where experimental data generation remains costly and time-consuming.

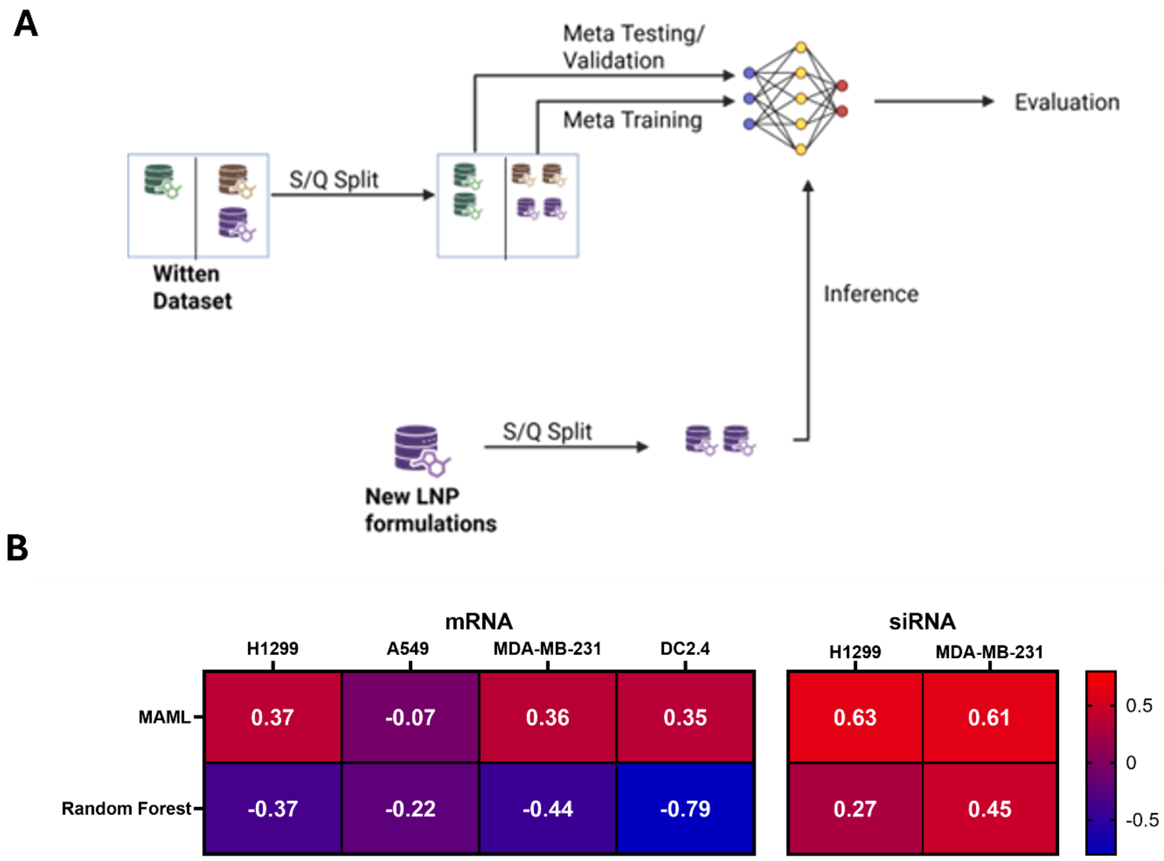


Figure VII.5: A) Split strategy to novel lipids using the meta trained model. B) Pearson r values of MAML vs RF for different cell lines and cargo.

Overall, our findings demonstrate the potential of meta-trained models as a promising strategy for early-stage formulation development. Guiding the discovery process in the right direction from the outset, can help researchers and institutions reduce costly late-stage failures while effectively leveraging historical data that are often difficult to integrate into conventional data-driven approaches.

Looking ahead, to extend the applicability of these models to clinically more relevant systems, in vivo validation will be necessary. The promise of meta-learning in formulation development lies in its ability to integrate information that would otherwise be difficult to combine, thereby substantially reducing development time and costs. Future work will expand the training corpus across additional datasets and refine the meta-learning strategy beyond MAML and MetaSGD, and by jointly optimizing formulation composition together with lipid-component discovery within a single framework.

3 Supplementary Information

3.1 Materials and Methods

3.1.1 Materials

3-Phenyl-2-propin-1-ol, oleic acid, spermidine, spermine, 4-(2-Aminoethyl)-morpholin, N,N-Dimethylethylendiamin, N,N-Dimethyldipropylenetriamine, lithium aluminium hydrid in hexanes (1M), citric acid monohydrate, sodium citrate dihydrate, sodium acetate, RPMI-1640 Medium, Dulbecco's Phosphate Buffered Saline (PBS), 2-mercaptoethanol, heat-inactivated Fetal Bovine Serum (FBS) and cholesterol were purchased from Sigma-Aldrich (Taufkirchen, Germany). Acryloylchlorid, 1,3-Diamino-propan, triethylamine, PBS 10X and all solvents were purchased from fisher scientific. Linoleic acid, 3,3'-Diamino-N-methyldipropylamine, 1-Dodecanol were purchased from TCI Chemicals (Germany). 1,6-Diaminohexane was purchased from Thermo Fisher Scientific. DMG-PEG 2000, 1,2-Distearoyl-sn-glycero-3-phosphocholine (DSPC) was purchased from Avanti. Fluc mRNA was purchased from Ribopro,. Silencer™ Firefly Luciferase (GL2 + GL3) siRNA and its scrambled negative control siRNA were purchased from Thermofisher (Waltham, Massachusetts, USA). If not otherwise specified, highly purified water (Arium® Pro Ultrapure Water System, Sartorius AG, Göttingen, Germany) was used for all the experiments.

3.1.2 Data Preparation

All computational work was carried out using python v3.11. The full dataset from Ref 16 was used. The data was initially grouped into subgroups by using the "split_name_for_normalization" column. The subgroups were split into 20 molecules large tasks by grouping by the one hot encoded criterion: "Delivery_target_dendritic", "Delivery_target_generic_cell","Delivery_target_liver", "Delivery_target_lung", "Delivery_target_lung_epithelium", "Delivery_target_macrophage", "Delivery_target_muscle", "Delivery_target_spleen", "Helper_lipid_ID_DOPE", "Helper_lipid_ID_DOTAP", "Helper_lipid_ID_DSPC", "Helper_lipid_ID_MDOA", "Helper_lipid_ID_None", "Route_of_administration_in_vitro", "Route_of_administration_intramuscular", "Route_of_administration_intratracheal", "Route_of_administration_intravenous","Batch_or_individual_or_barcoded_Barcoded", "Batch_or_individual_or_barcoded_Individual", "Cargo_type_mRNA", "Cargo_type_pDNA", "Cargo_type_siRNA", "Model_type_A549","Model_type_BDMC", "Model_type_BMDM","Model_type_HBEC_ALI", "Model_type_HEK293T","Model_type_HeLa","Model_type_IGROV1","Model_type_Mouse ","Model_type_RAW264p7"

To add ratio information, the lipid composition columns were transformed into floating numbers and added to the data. As target information, the "quantified_delivery" column was used, since it already represents a standard scaled label. Zero-variance tasks as well as duplicated were removed from the dataset. To allow a later graph encoding of the respective molecules, the SMILES code was added for each formulation point. The data was subsequently split into support and query (10/10). To ensure a comparable training, the label distribution was discretized, and the support-query split was stratified. A training set was created by removing all tasks that contain siRNA as cargo from the full set. The removed data was used for the holdout set (64 tasks) and the validation set (8 tasks).

3.1.3 Model Comparison

The Model Comparison experiment was performed by comparing different meta learning models (FoMAML, MAML, MetaSGD- all from learn2learn v 0.2.0) to basic supervised models (no finetuning and finetuning from torch v2.6.0). Featurization into fingerprints was performed using Morgan Fingerprints with r=4 and 2048 bits (using RDkit v2024.9.5). Graph encoding for the graph neural network featurization as well as the base GNN were used from chemprop v2.2.1. Message Passing and Mean Aggregation were applied prior to one hidden ReLU layer and one linear regression head. As fingerprint base model a basic pytorch model was used with two hidden ReLu layers and one regression head. As loss function MSE was selected. Data was subsequently loaded into a specialized DataLoader class (8 tasks per batch) and was tested over 10 different random seeds and the mean, and the standard deviation were calculated. All variables and hyperparameters were selected based on prior optimization and testing.

3.1.4 Active Learning

For the active learning experiment a greedy-active learning strategy was used where the datapoint with the highest predicted value was picked after every round. As retrospective dataset, the siRNA holdout set from the model comparison was used as well as the best model. As baseline a Random Forrest Model was selected and both models were compared against a random picking algorithm. To mimic the initial few-shot data available, 10 datapoints from the dataset were selected as starting points. The points were sampled based on a high-diversity sampling, that selected points that had high distances in Euclidean space spanned by fingerprints and ratio information. To obtain a realistic learning curve, 100 iterations with one sample pick were performed. Based on the obtained curve, several metrices were calculated: yield@k, simple_regret@k, best_so_far@k, EF@k and Hits@k with k being the number of iterations (here fixed at 5). Calculations and explanations of the metrices:

Let X denote the finite candidate set with size N and let f(x) be the objective measured experimentally, for example a transfection readout. During an active-learning run, the algorithm selects a sequence x_t for t = 1..k and yields observations y_t = f(x_t). Define the incumbent after t queries as b_t = max_{i ≤ t} y_i. All metrics are computed with respect to the same candidate pool X.

1) Hits@k

Let T be the set of top items, for example the highest-scoring fraction of X according to f. Hits@k is the count of selected items that belong to T across the first k iterations: Hits@k = Σ_{t=1..k} 1[x_t ∈ T].

2) Enrichment Factor (EF@k)

EF@k measures enrichment over random selection: EF@k = (Hits@k / k) / (|T| / N). EF@k greater than 1 indicates better-than-random retrieval of top candidates.

3) Best-so-far@k

The best outcome encountered up to iteration k: best_so_far@k = b_k = max_{t ≤ k} y_t.

4) Simple regret@k

Simple regret quantifies the gap to the global best available in the pool: simple_regret@k = y* − b_k, where y* = max_{x ∈ X} f(x).

5) Yield@k (cumulative normalized yield)

To make results comparable across datasets, we report the cumulative sum of min-max normalized outcomes. Define y_min = min_{x ∈ X} f(x) and y_max = max_{x ∈ X} f(x). For each iteration t, compute the normalized value ỹ_t = (y_t − y_min) / (y_max − y_min). Then yield@k = Σ_{t=1..k} ỹ_t.

3.1.5 Own Lipids Test

The test on the synthesized novel lipids was performed using a 5-fold CV approach where the lipids were split (9 train/3 val/3 test) for the MAML model and (12 train/3 test) for the RF baseline. The lipids were tested and the mean Pearson value was calculated based on the predicted values for the test points vs the experimental labels. The data was standardized using a StandardScaler. Featurization for the MAML model was performed using the GNN featurization method described in Section Model Comparison. RF featurization was performed using Morgan Fingerprints described in Section Model Comparison.

For detailed information about the experiments, we would like to refer the reader to https://github.com/felixsie19/FewShotLNPs.

3.1.6 Chemical synthesis

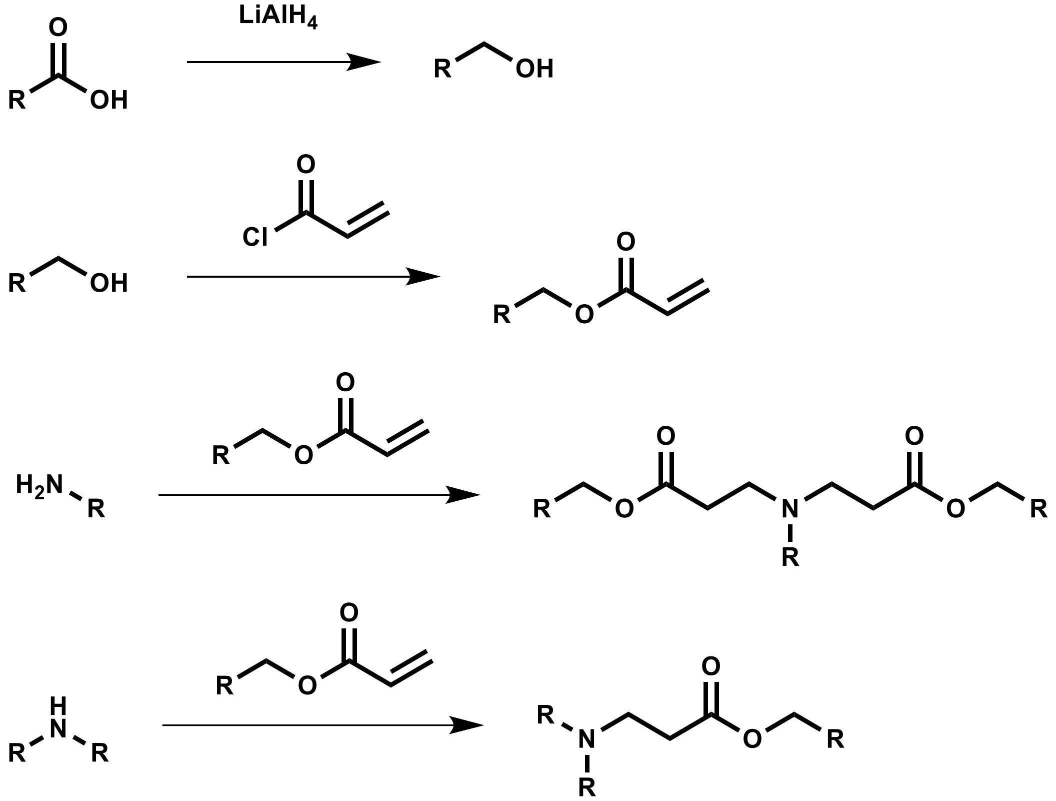


Figure VII.S1: Synthesis Scheme of Lipids.

Synthesis of oleyl alcohol and linoleyl alcohol

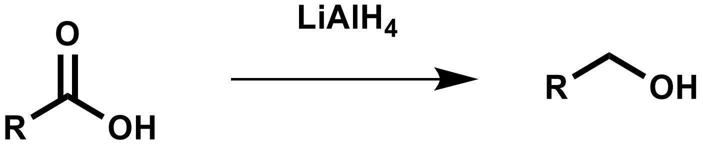


Figure VII.S2: Synthesis of oleyl alcohol and linoleyl alcohol.

Oleic acid or linoleic acid (8 mmo, 1 eq) were dissolved in 100 ml of anhydrous THF. Solution was cooled to 0°C and 1 M LiAlH4 (12 mmol, 1,5 eq) was added dropwise. After 30 min the ice bath was removed, and reaction was carried out at RT overnight. The reaction was quenched with water and 1 M NaOH and filtered through celite 545.

Synthesis of alkyl acrylates

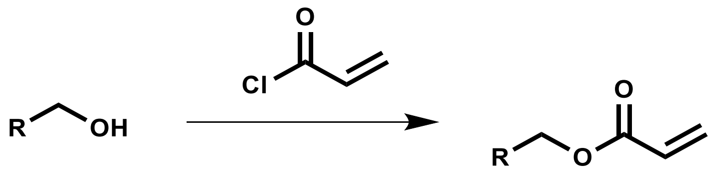


Figure VII.S3: Synthesis of alkyl acrylates.

Oleyl alcohol, linoleyl alcohol, dodecanol or 3-phenyl-2-propin-1-ol (1 mmol, 1 equiv.) were dissolved in 10 ml of anhydrous dichloromethane together with triethylamine (1,5 mmol, 1,5 equiv.). Acryloyl chloride (1.2 mmol, 1.2 equiv.) was dissolved in 20 ml of anhydrous CDCl2 and added dropwise to the reaction at 0 °C for 30 min. Afterwards the ice bath was removed and kept stirring at RT overnight. The mixture was diluted with CH2Cl2 and washed with brine twice and sat. H2CO3. The organic layer was dried over MgSO4, filtered, and concentrated in vacuo. The residue was purified by a CombiFlash PuriFlash Rf200i chromatography system (Teledyne ISCO) with gradient elution from cyclohexane/ ethylacetate to 100:0 to 0:100 cyclohexane/ethyl acetate.

Synthesis of ionizable lipids

Final lipids were synthesized through a solvent free aza-michael reaction of respective amines and acrylates. Acrylates and amines were added into vials and placed on a shaker at 250 rpm for > 120 h at RT. Lipids were used without further purification. Acrylates were added in excess: equivalents of acrylates were calculated by x = 2 eq for every primary amine + 1 eq for every secondary amine + 2 eq excess (x = 1 * for every N-H bond + 2). Conversion was monitored via ^{1}H NMR and final mass was confirmed by MS-ESI.

3.1.7 LNP formulation

For siRNA LNPs, Fluc siRNA was dissolved in 10 mM Citrate buffer pH = 4. Lipid were dissolved at 1 mM in EtOH with a molar ratio of (ionizable lipid/cholesterol/DSPC/DMG-PEG2000 50/38.5/10/1.5). LNPs were formulated with a high throughput microfluidics device (Sunscreen, Unchained Labs). Flow rate ratios were 3:1 (aqueous phase:organic phase), and total flow rate was 10,000 µl/min on the Sunny 100 X chip.

For mRNA LNPs, Fluc mRNA was dissolved in 10 mM Citrate buffer pH = 4. Lipids were dissolved at 3 mM in EtOH with a molar ratio of (ionizable lipid/cholesterol/DSPC/DMG-PEG2000 50/38.5/10/1.5). LNPs were formulated with a high throughput microfluidics device (Sunscreen, Unchained Labs). Flow rate ratios were 3:1 (aqueous phase:organic phase), and total flow rate was 10,000 µl/min on the Sunny 190 T chip.

After formulations obtained via microfluidics were dialyzed overnight against 1X PBS. Particle size and Polydispersity Index were measured by Dynamic Light Scattering (DLS) with a Wyatt DynaPro Plate Reader II.

3.1.8 Luciferase expression assay (mRNA)

To assess mRNA expression efficiency in submerged cell culture, A549, H1299, DC2.4, and MDA-MB-231 cells were seeded at a density of 10,000 cells per well in 200 µL medium in 96-well plates. A549 and H1299 were cultured in RPMI 1640 + 10% FBS; DC2.4 in RPMI 1640 + 10% FBS + 1% 2-β-mercaptoethanol; MDA-MB-231 in DMEM High Glucose + 10% FBS. After 24 h, medium was replaced with fresh medium, and cells were transfected with 150 ng mLuc-encapsulating LNPs. D-Lin-MC3-DMA served as a positive control and untreated cells served as blank. Following 24 h incubation at 37 °C and 5 % CO_2, medium was removed, and cells were lysed with 0.5× lysis buffer (100 µL per well) and incubated for 30 min at room temperature. Luciferase activity was measured on a Tecan Spark plate reader (TECAN, Männedorf, Switzerland). A 35 µL aliquot of cell lysate was read for 10 s after automatic addition of 100 µL LAR buffer (20 mM glycylglycine; 1 mM $MgCl_2$; 0.1 mM EDTA; 3.3 mM DTT; 0.55 mM ATP; 0.27 mM coenzyme A; pH 8–8.5) supplemented with 10% (v/v) of a mixture of 10 mM luciferin and 29 mM glycylglycine. Transfection efficiency was calculated and reported as relative light units (RLU) per well.

3.1.9 Luciferase knockdown assay (siRNA)

siRNA-mediated knockdown of firefly luciferase (Fluc) mRNA was assessed in H1299-PGK-eGFP-Luc and MDA-MB-231-Luc reporter cell lines. H1299-PGK-eGFP-Luc cells were seeded at 2,500 cells per well in 200 µL RPMI 1640 + 10% FBS; MDA-MB-231-Luc cells were seeded at 6,000 cells per well in 200 µL DMEM High Glucose + 10% FBS. After 24 h, medium was replaced with fresh medium, and cells were transfected with Fluc siRNA containing LNPs. Following 48 h incubation at 37 °C and 5 % CO2, luciferase activity was measured on a Tecan Spark plate reader (TECAN, Männedorf, Switzerland). A 35 µL aliquot of cell lysate was read for 10 s after automatic addition of 100 µL LAR buffer (20 mM glycylglycine; 1 mM $MgCl_2$; 0.1 mM EDTA; 3.3 mM DTT; 0.55 mM ATP; 0.27 mM coenzyme A; pH 8–8.5) supplemented with 10% (v/v) of a mixture of 10 mM luciferin and 29 mM glycylglycine. Untreated cells were set to 100 % firefly luciferase expression, and knockdown efficiency was calculated as the remaining expression.

Table VII.S1: Overview over components mixed for synthesis, the respective theoretical and actual masses as well as the lipid name.

Lipid No.	Full Lipid Code	Amine No.	Alkyltail No.	Calculated mass	Found	Lipid
Lipid 1	A1T1	A1	T1	1127,04	1127,03	L11
Lipid 2	A1T2	A1	T2	1120,99	1120,99	L10
Lipid 3	A1T3	A1	T3	880,80	880,80	L13
Lipid 4	A1T4	A1	T4	718,38	718,38	L40
Lipid 5	A2T1	A2	T1	1435,31	1435,31	L7
Lipid 6	A2T2	A2	T2	1427,24	1427,25	L6
Lipid 7	A2T3	A2	T3	1106,99	1106,99	L15
Lipid 8	A2T4	A2	T4	890,43	890,43	L34
Lipid 9	A3T2	A3	T2	729,64	729,64	L21
Lipid 10	A3T3	A3	T3	569,52	569,52	L18
Lipid 11	A4T2	A4	T2	2123,84	2123,84	L44
Lipid 12	A5T2	A5	T2	1398,22	1398,22	L45
Lipid 13	A6T2	A6	T2	771,65	771,65	L46
Lipid 14	A7T2	A7	T2	1356,17	1356,17	L47
Lipid 15	A8T2	A8	T2	1747,51	1747,51	L48

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.39 (dd, *J* = 17.3, 1.5 Hz, 2H), 6.12 (dd, *J* = 17.3, 10.4 Hz, 2H), 5.81 (dd, *J* = 10.4, 1.5 Hz, 2H), 5.41 – 5.28 (m, 10H), 4.15 (t, *J* = 6.7 Hz, 4H), 4.06 (q, *J* = 7.2 Hz, 6H), 2.86 (t, *J* = 6.6 Hz, 1H), 2.76 (dd, *J* = 7.8, 6.3 Hz, 6H), 2.67 – 2.52 (m, 1H), 2.54 – 2.35 (m, 12H), 2.21 (s, 8H), 2.01 (q, *J* = 6.6 Hz, 20H), 1.63 (ddt, *J* = 18.0, 11.1, 4.9 Hz, 18H), 1.28 (dd, *J* = 12.8, 5.0 Hz, 109H), 0.93 – 0.84 (m, 15H).

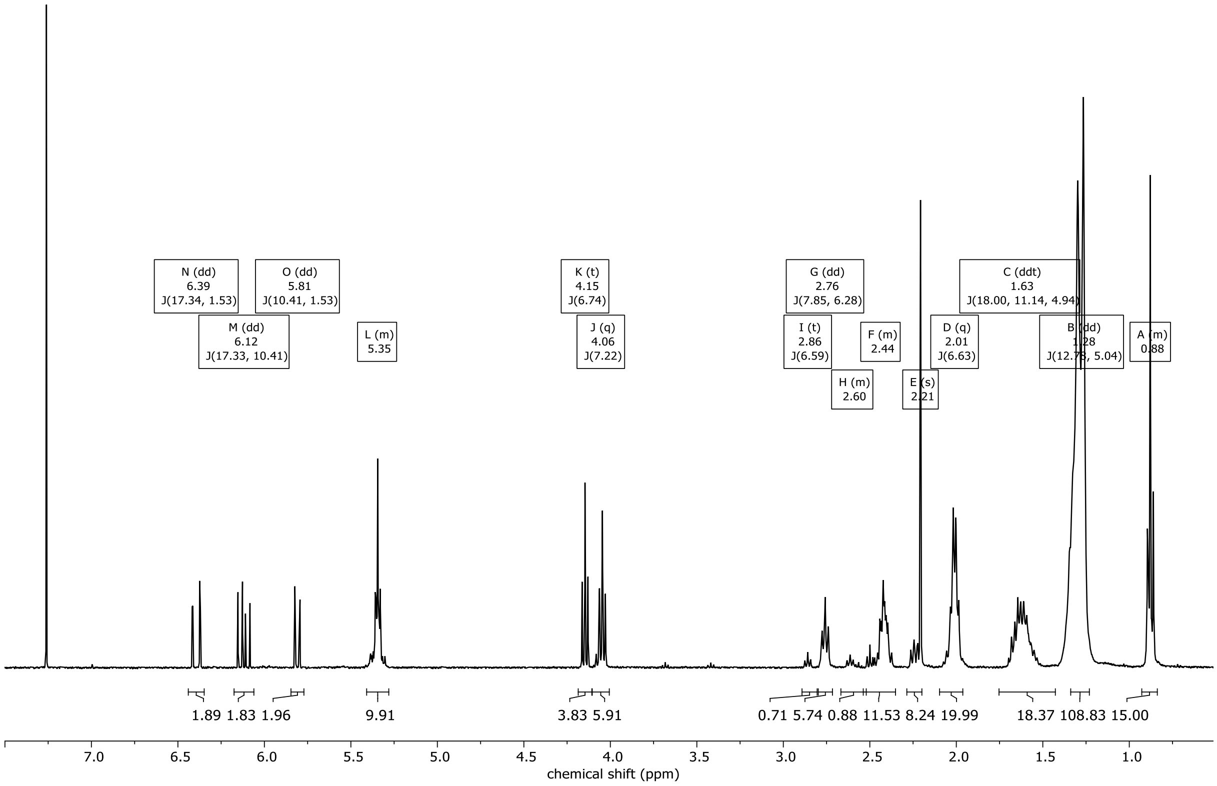


Figure VII.S4: 1H NMR of final Lipid1 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 5.37 – 5.20 (m, 4H), 3.99 (dt, *J* = 8.3, 6.8 Hz, 2H), 2.80 (t, *J* = 6.6 Hz, 1H), 2.70 (q, *J* = 5.4 Hz, 3H), 2.57 (t, *J* = 7.1 Hz, 1H), 2.44 (t, *J* = 6.6 Hz, 1H), 2.33 (dt, *J* = 25.7, 7.4 Hz, 3H), 2.12 (d, *J* = 3.7 Hz, 1H), 1.98 (q, *J* = 6.9 Hz, 4H), 1.56 (tt, *J* = 13.9, 7.1 Hz, 5H), 1.38 – 1.15 (m, 18H), 1.00 – 0.65 (m, 3H).

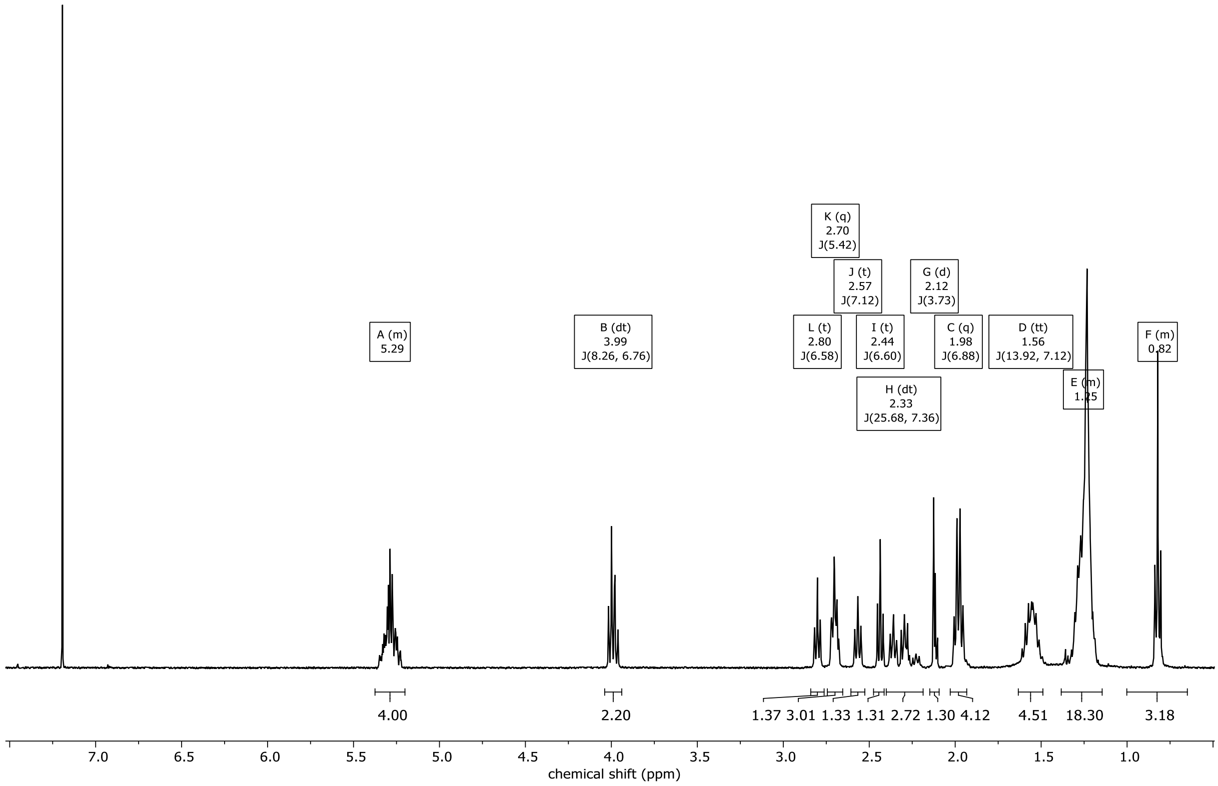


Figure VII.S5: 1H NMR of final Lipid2 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.39 (dd, *J* = 17.3, 1.5 Hz, 2H), 6.11 (dd, *J* = 17.3, 10.4 Hz, 2H), 5.80 (dd, *J* = 10.4, 1.5 Hz, 2H), 4.14 (t, *J* = 6.7 Hz, 4H), 4.10 – 4.00 (m, 6H), 2.85 (s, 0H), 2.76 (td, *J* = 7.4, 1.8 Hz, 6H), 2.54 – 2.36 (m, 12H), 2.35 – 2.17 (m, 10H), 1.72 – 1.49 (m, 16H), 1.38 – 1.18 (m, 102H), 0.91 – 0.83 (m, 15H).

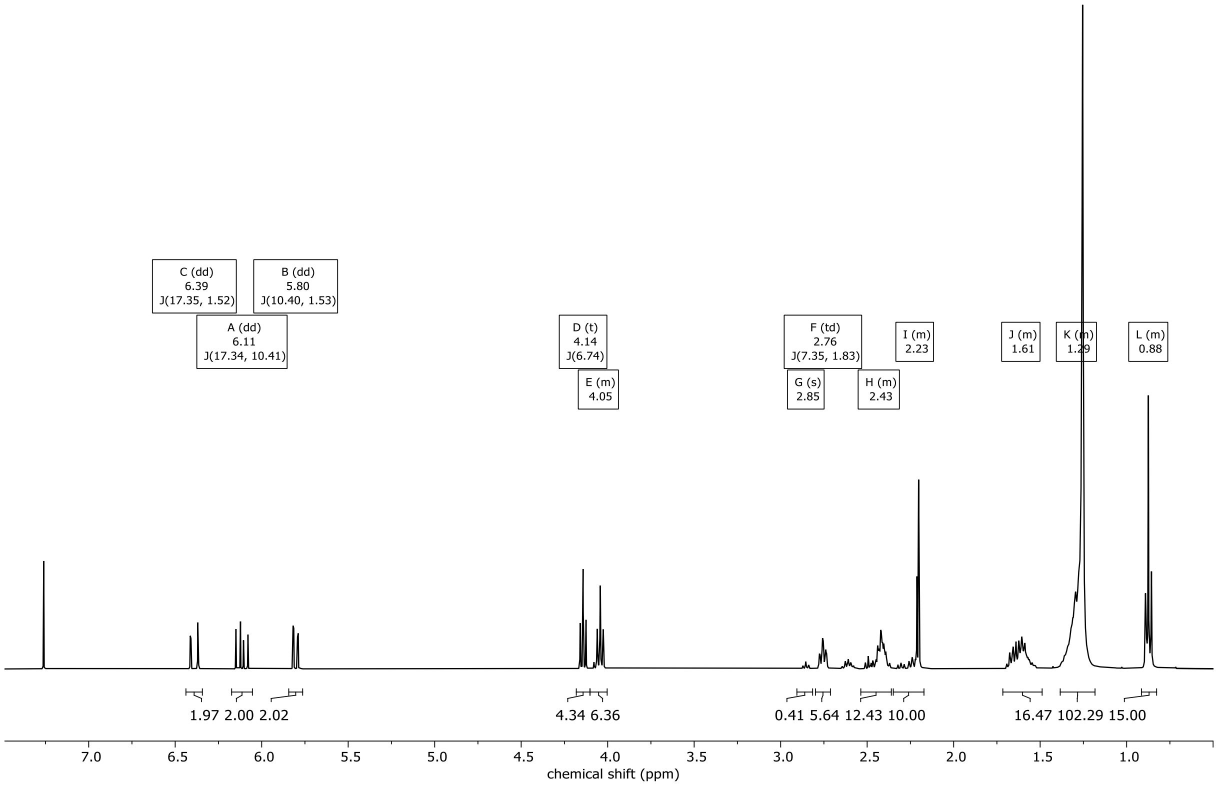


Figure VII.S6: 1H NMR of final Lipid3 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.49 – 7.42 (m, 8H), 7.35 – 7.28 (m, 12H), 6.55 – 6.43 (m, 0H), 6.19 (dd, *J* = 17.3, 10.5 Hz, 0H), 5.90 (dd, *J* = 10.5, 1.4 Hz, 0H), 4.90 (s, 7H), 2.84 – 2.74 (m, 7H), 2.56 – 2.47 (m, 7H), 2.46 – 2.37 (m, 7H), 2.20 (s, 8H).

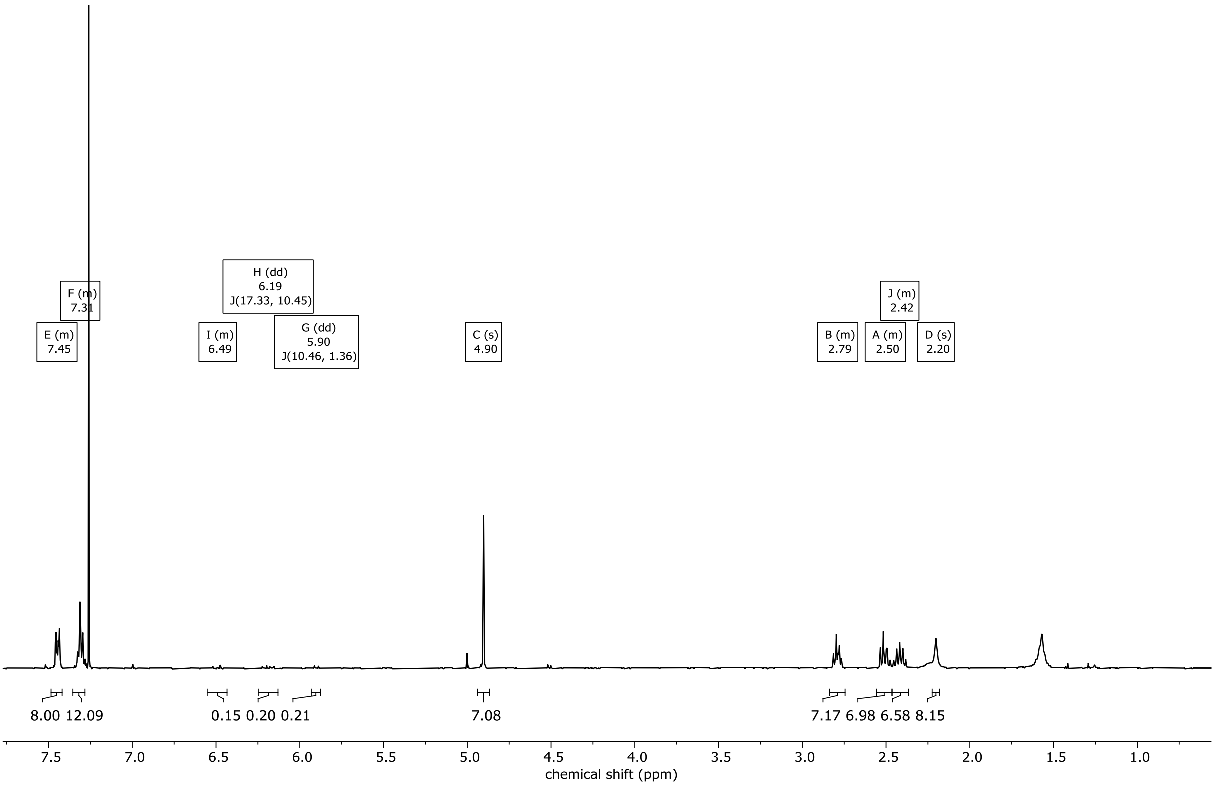


Figure VII.S7: 1H NMR of final Lipid4 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.39 (dd, *J* = 17.3, 1.5 Hz, 3H), 6.12 (dd, *J* = 17.3, 10.4 Hz, 3H), 5.81 (dd, *J* = 10.4, 1.5 Hz, 3H), 5.41 – 5.29 (m, 11H), 4.15 (t, *J* = 6.7 Hz, 5H), 4.05 (t, *J* = 7.0 Hz, 7H), 2.87 (t, *J* = 6.6 Hz, 2H), 2.76 (t, *J* = 7.4 Hz, 5H), 2.64 (t, *J* = 7.1 Hz, 2H), 2.54 – 2.31 (m, 12H), 2.22 – 2.15 (m, 3H), 1.72 – 1.56 (m, 19H), 1.29 (dd, *J* = 17.5, 7.2 Hz, 142H), 0.93 – 0.84 (m, 18H).

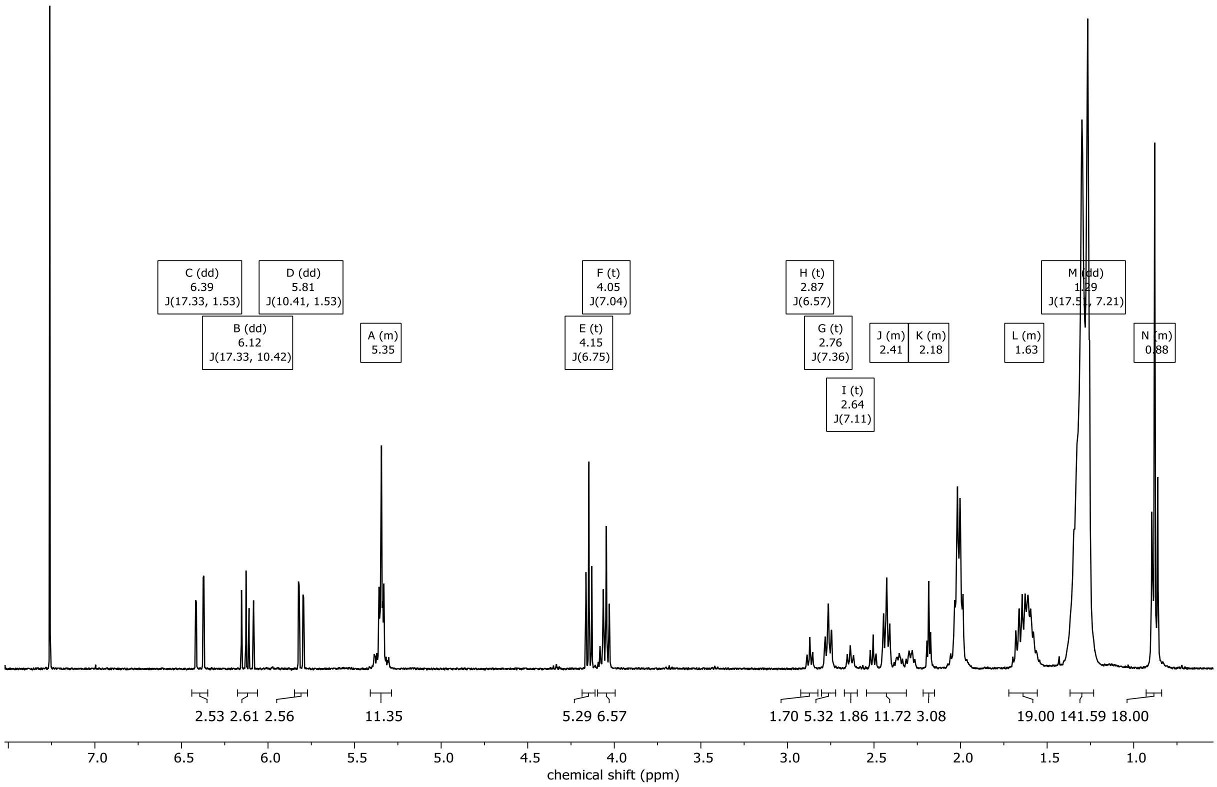


Figure VII.S8: 1H NMR of final Lipid5 crude

^{1}H NMR (400 MHz, CDCl$_3$) δ 5.44 – 5.27 (m, 4H), 4.06 (dt, *J* = 8.1, 6.8 Hz, 2H), 2.87 (q, *J* = 6.3 Hz, 1H), 2.77 (s, 1H), 2.73 – 2.56 (m, 1H), 2.54 – 2.39 (m, 3H), 2.27 – 2.15 (m, 4H), 2.05 (q, *J* = 6.9 Hz, 4H), 1.64 (dt, *J* = 25.1, 7.6 Hz, 8H), 1.45 – 1.21 (m, 19H), 0.93 – 0.85 (m, 3H).

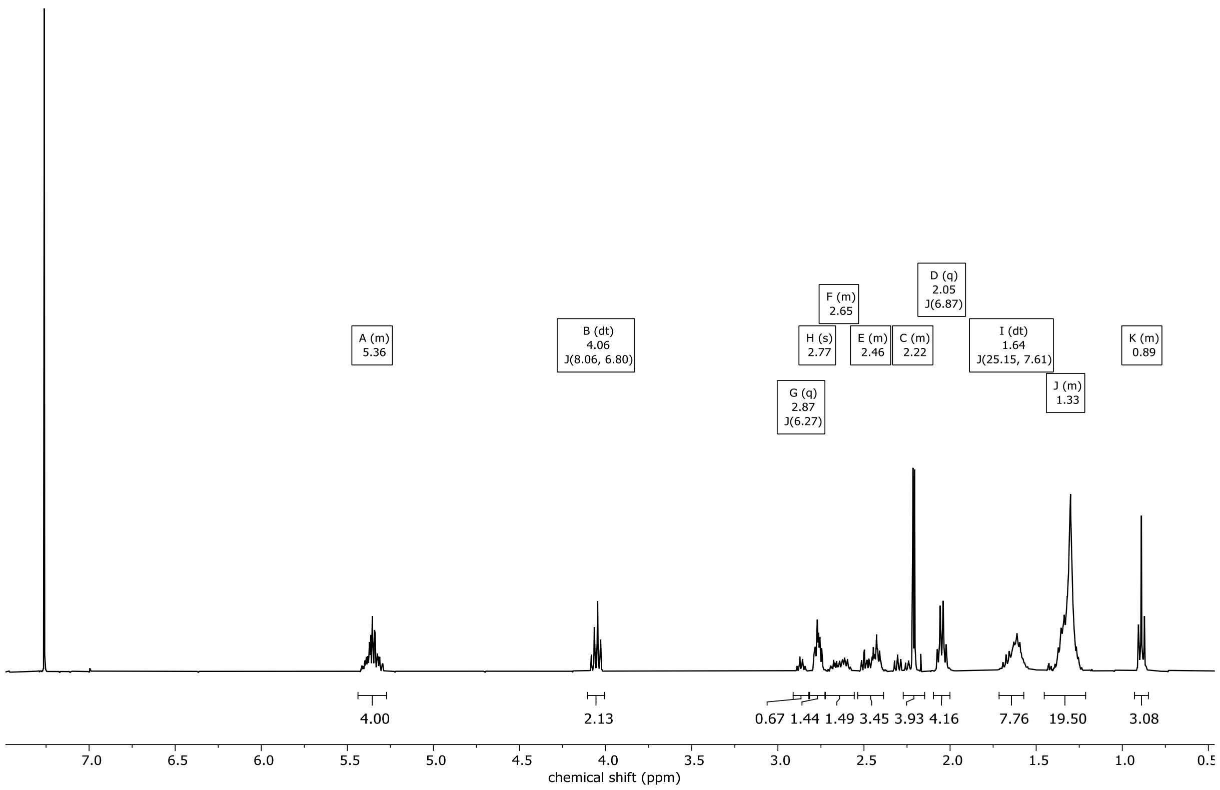


Figure VII.S9: 1H NMR of final Lipid6 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.39 (dd, *J* = 17.3, 1.5 Hz, 2H), 6.12 (dd, *J* = 17.3, 10.4 Hz, 2H), 5.81 (dd, *J* = 10.4, 1.5 Hz, 2H), 4.15 (t, *J* = 6.7 Hz, 4H), 4.06 (dt, *J* = 8.1, 6.7 Hz, 4H), 2.87 (s, 1H), 2.81 – 2.72 (m, 2H), 2.64 (s, 1H), 2.54 – 2.41 (m, 4H), 2.39 – 2.23 (m, 3H), 2.19 (d, *J* = 3.7 Hz, 2H), 1.72 – 1.52 (m, 11H), 1.39 – 1.25 (m, 77H), 0.92 – 0.83 (m, 12H).

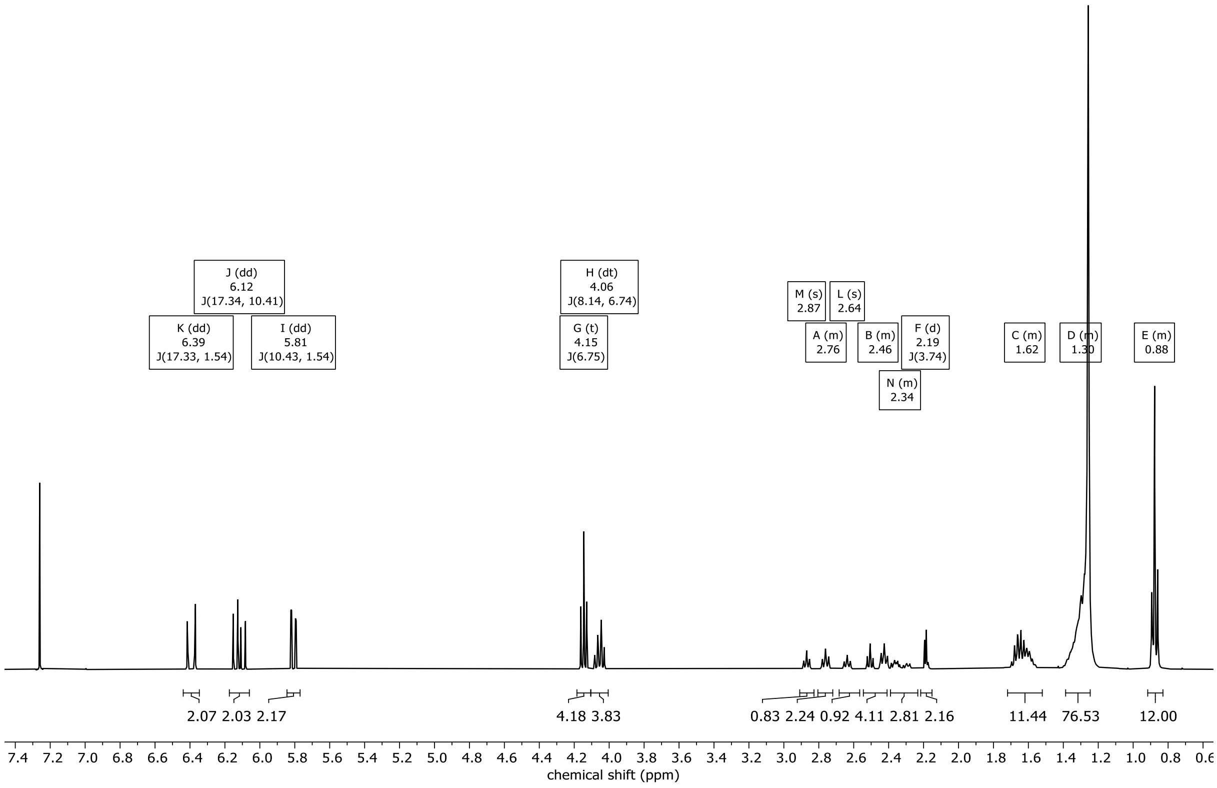


Figure VII.S10: 1H NMR of final Lipid7 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.48 – 7.41 (m, 8H), 7.36 – 7.27 (m, 12H), 4.92 – 4.88 (m, 8H), 2.84 – 2.76 (m, 8H), 2.61 – 2.38 (m, 12H), 2.27 (t, *J* = 7.3 Hz, 3H), 2.15 (d, *J* = 9.2 Hz, 3H).

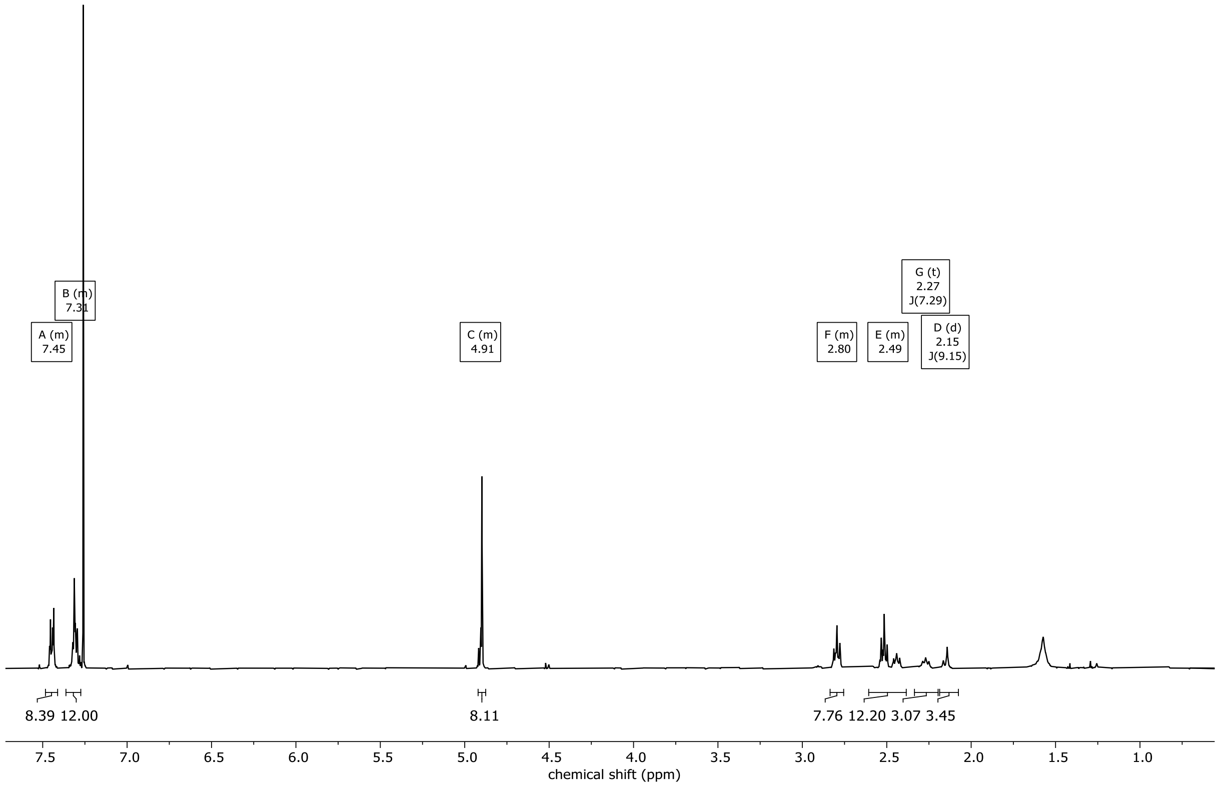


Figure VII.S11: 1H NMR of final Lipid8 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 5.44 – 5.27 (m, 4H), 4.06 (td, *J* = 6.8, 1.0 Hz, 2H), 2.81 – 2.61 (m, 5H), 2.54 – 2.42 (m, 5H), 2.24 (d, *J* = 6.7 Hz, 3H), 1.73 – 1.52 (m, 4H), 1.48 – 1.24 (m, 18H), 1.01 – 0.74 (m, 3H).

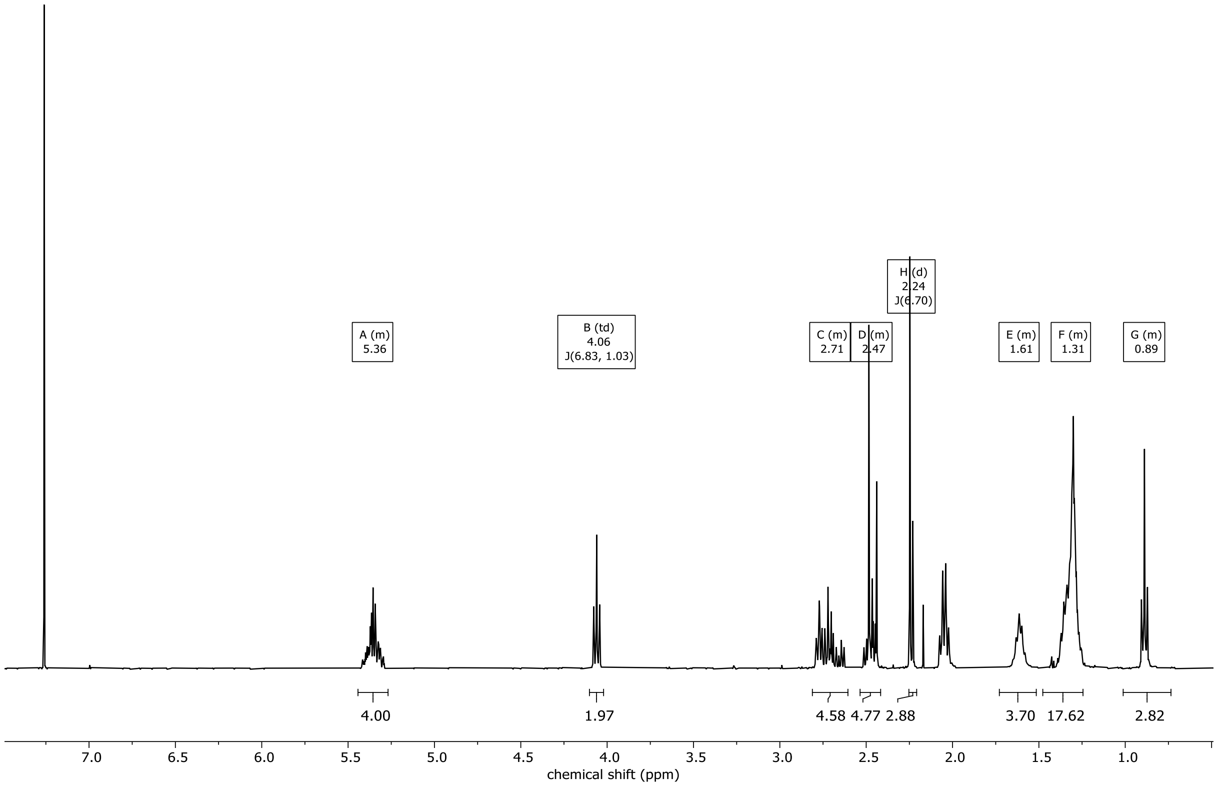


Figure VII.S12: 1H NMR of final Lipid9 crude

^{1}H NMR (400 MHz, CDCl$_3$) δ 6.39 (dd, *J* = 17.3, 1.5 Hz, 2H), 6.12 (dd, *J* = 17.3, 10.4 Hz, 2H), 5.81 (dd, *J* = 10.4, 1.5 Hz, 2H), 4.15 (s, 2H), 4.06 (s, 2H), 2.72 (t, *J* = 7.3 Hz, 4H), 2.51 – 2.43 (m, 8H), 2.25 (s, 6H), 1.72 – 1.55 (m, 9H), 1.34 – 1.25 (m, 79H), 0.92 – 0.84 (m, 12H).

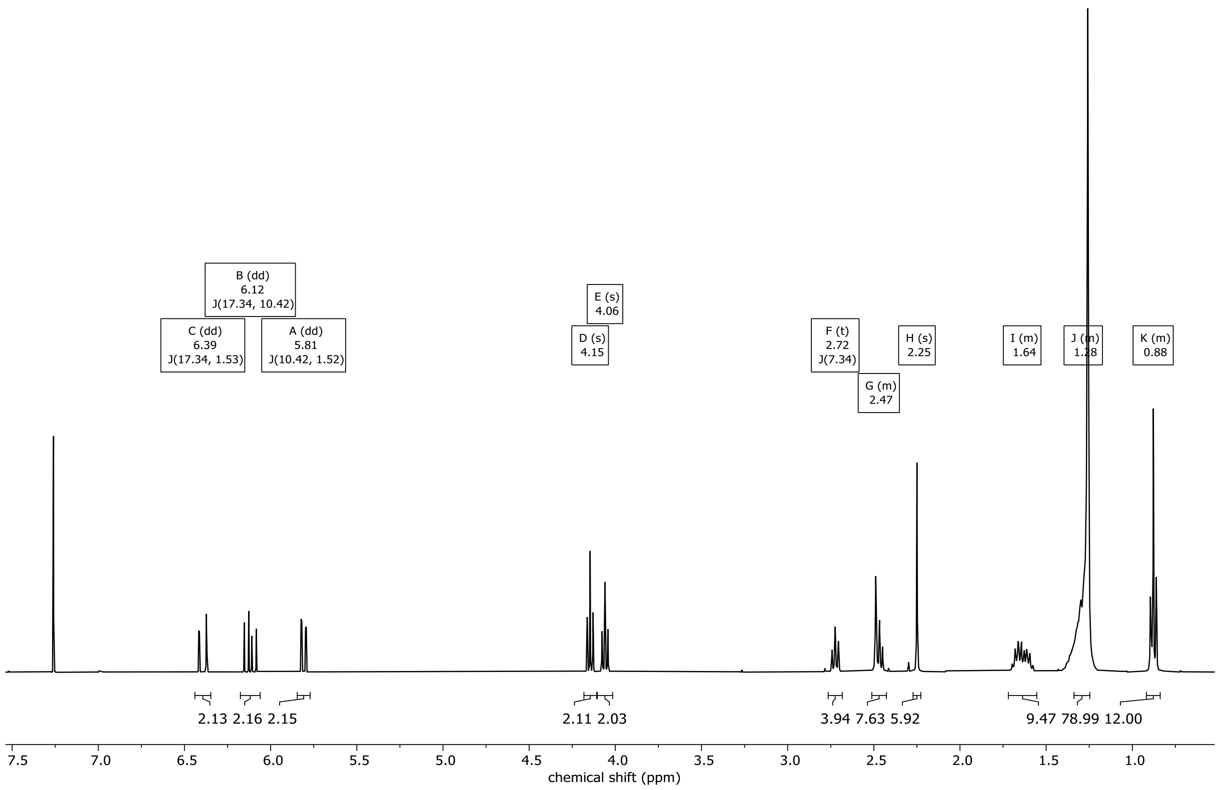


Figure VII.S13: 1H NMR of final Lipid10 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 5.44 – 5.27 (m, 4H), 4.06 (dt, *J* = 8.0, 6.8 Hz, 2H), 2.86 (t, *J* = 6.2 Hz, 1H), 2.77 (t, *J* = 6.3 Hz, 4H), 2.63 (dd, *J* = 25.6, 7.9 Hz, 2H), 2.45 (ddd, *J* = 29.4, 12.9, 6.1 Hz, 4H), 2.05 (q, *J* = 6.8 Hz, 4H), 1.60 (d, *J* = 7.7 Hz, 5H), 1.47 – 1.21 (m, 21H), 0.93 – 0.85 (m, 4H).

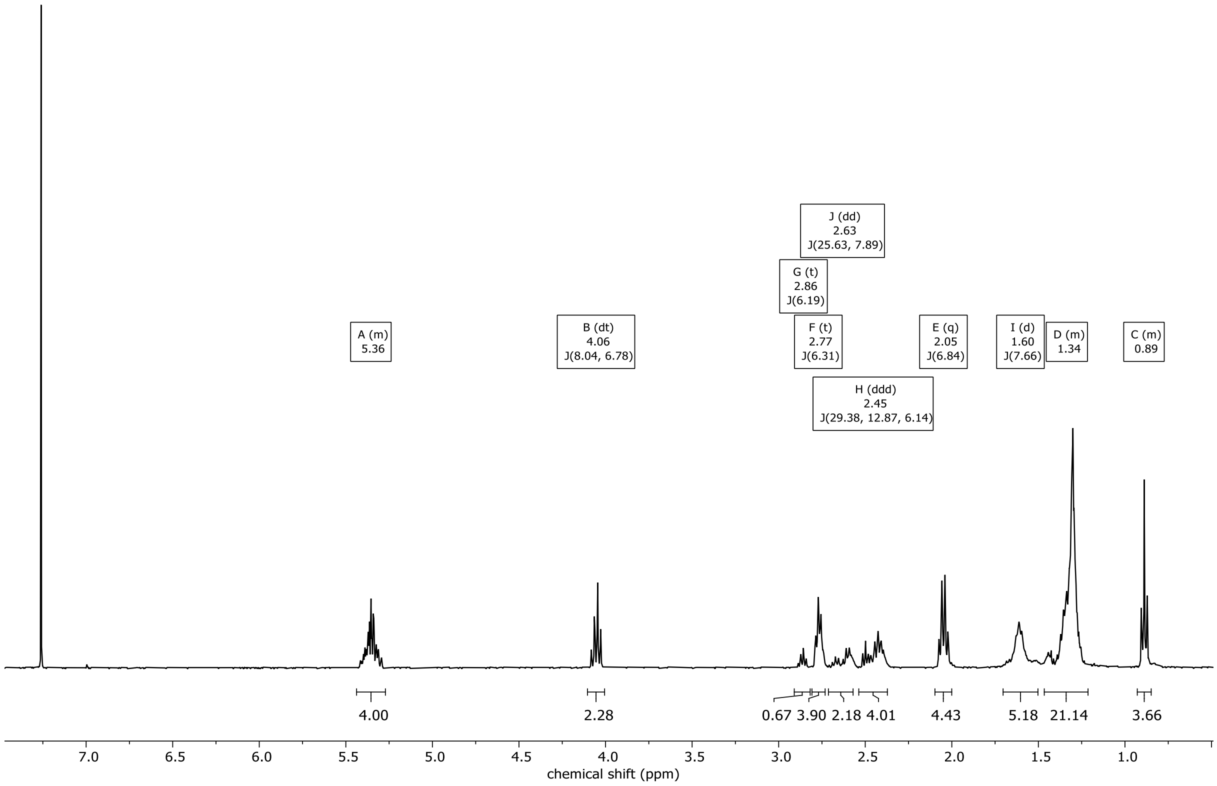


Figure VII.S14: 1H NMR of final Lipid11 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.40 (dd, *J* = 17.3, 1.5 Hz, 0H), 6.12 (dd, *J* = 17.3, 10.4 Hz, 0H), 5.81 (dd, *J* = 10.4, 1.5 Hz, 0H), 5.44 – 5.27 (m, 4H), 4.15 (t, *J* = 6.8 Hz, 1H), 4.06 (d, *J* = 8.7 Hz, 1H), 2.87 (t, *J* = 6.5 Hz, 1H), 2.76 (q, *J* = 6.0 Hz, 3H), 2.64 – 2.56 (m, 1H), 2.51 (t, *J* = 6.5 Hz, 1H), 2.46 – 2.35 (m, 1H), 2.05 (q, *J* = 6.9 Hz, 5H), 1.76 – 1.16 (m, 29H), 0.93 – 0.85 (m, 4H).

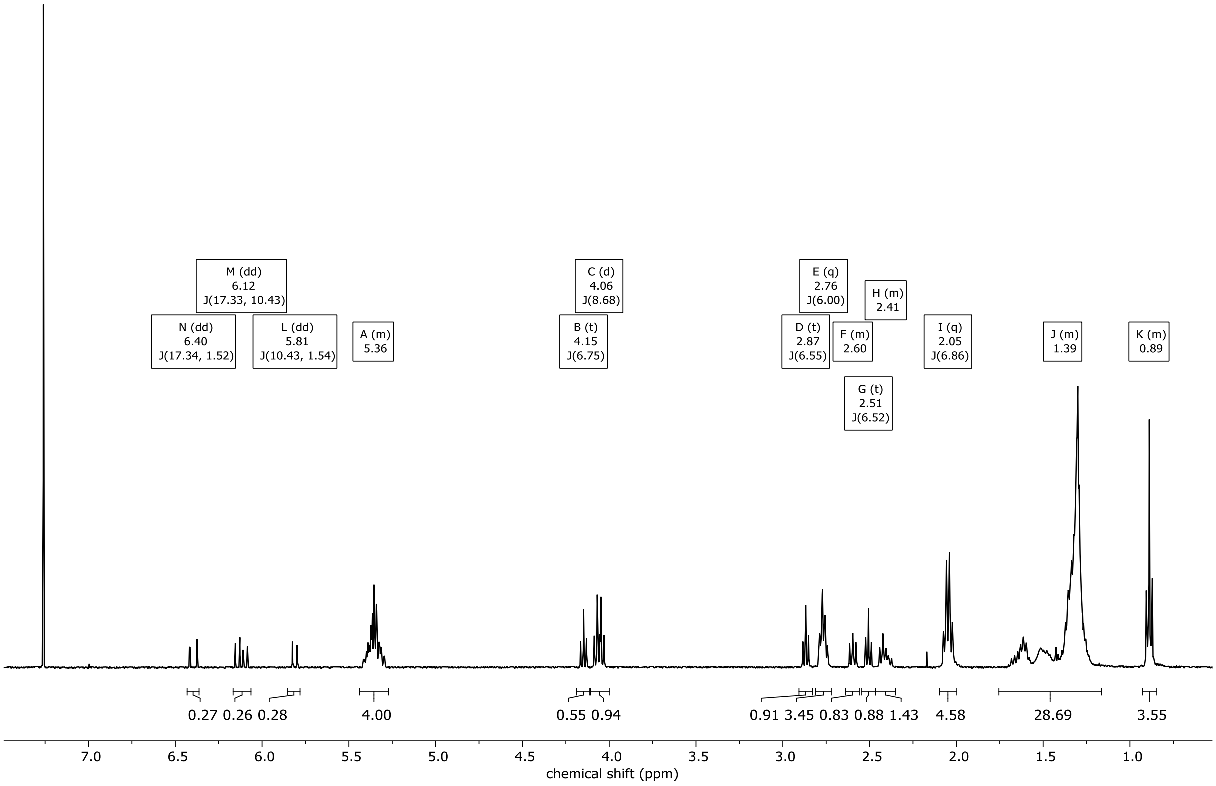


Figure VII.S15: 1H NMR of final Lipid12 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.39 (dd, *J* = 17.3, 1.5 Hz, 0H), 6.12 (dd, *J* = 17.3, 10.4 Hz, 0H), 5.81 (dd, *J* = 10.4, 1.5 Hz, 0H), 5.44 – 5.27 (m, 4H), 4.14 (t, *J* = 6.7 Hz, 0H), 4.06 (dt, *J* = 9.0, 6.8 Hz, 2H), 3.73 – 3.66 (m, 2H), 2.89 (t, *J* = 6.6 Hz, 1H), 2.84 – 2.68 (m, 3H), 2.05 (q, *J* = 6.9 Hz, 4H), 1.72 – 1.56 (m, 3H), 1.45 – 1.21 (m, 17H), 0.93 – 0.84 (m, 3H).

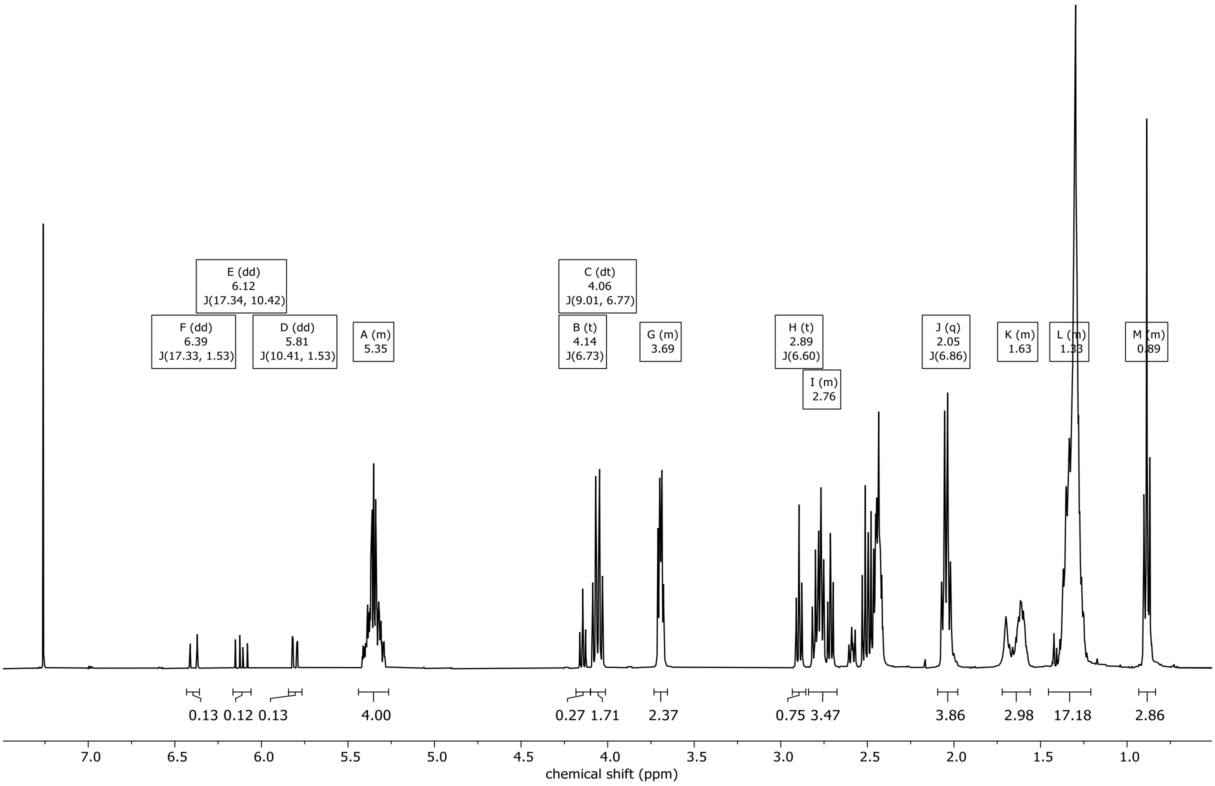


Figure VII.S16: 1H NMR of final Lipid13 crude

^{1}H NMR (400 MHz, CDCl$_3$) δ 5.37 – 5.20 (m, 4H), 3.99 (dt, *J* = 8.1, 6.8 Hz, 2H), 2.85 – 2.49 (m, 5H), 2.48 – 2.32 (m, 2H), 1.98 (q, *J* = 6.8 Hz, 4H), 1.65 – 1.49 (m, 3H), 1.39 – 1.14 (m, 17H), 0.86 – 0.78 (m, 3H).

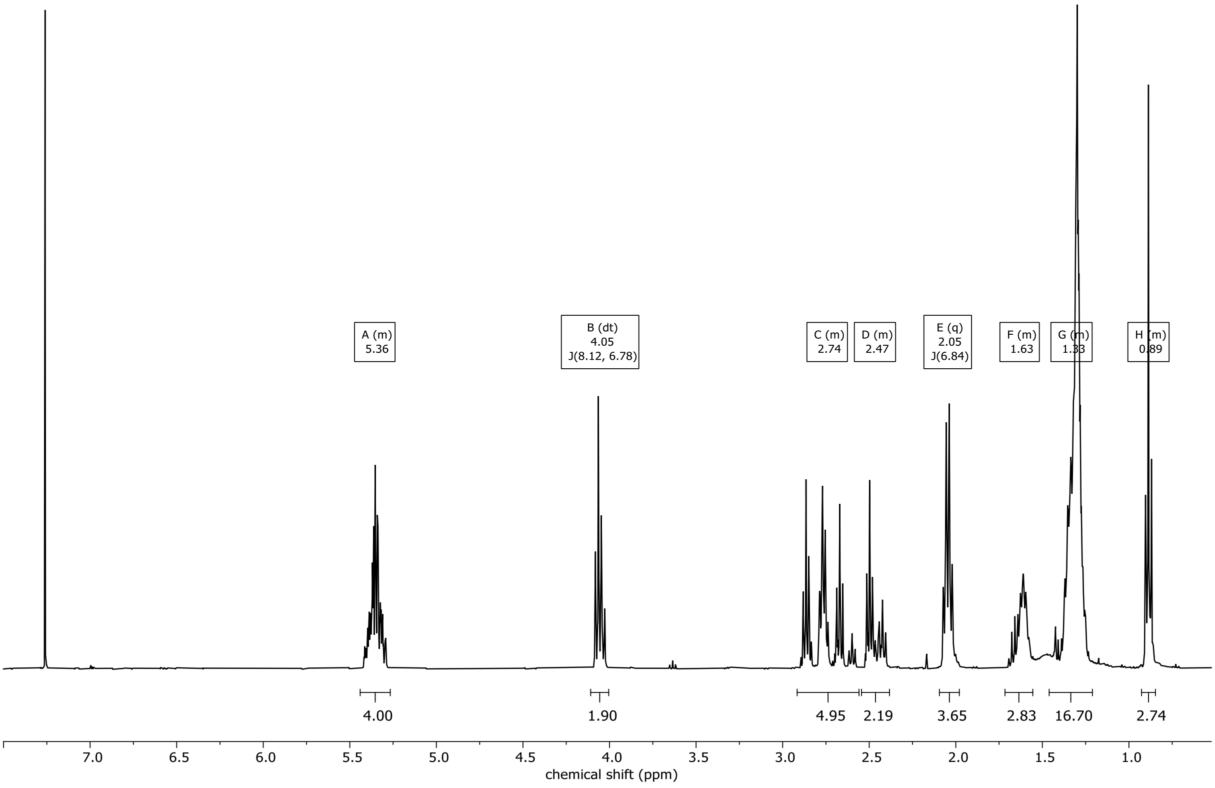


Figure VII.S17: 1H NMR of final Lipid14 crude

^{1}H NMR (400 MHz, CDCl$_3$) δ 5.37 – 5.20 (m, 4H), 3.99 (dt, *J* = 8.3, 6.8 Hz, 2H), 2.79 (dd, *J* = 5.9, 2.6 Hz, 1H), 2.75 – 2.65 (m, 4H), 2.60 – 2.48 (m, 0H), 2.48 – 2.27 (m, 3H), 1.98 (q, *J* = 6.9 Hz, 4H), 1.54 (p, *J* = 6.8 Hz, 3H), 1.50 – 1.14 (m, 19H), 0.89 – 0.77 (m, 3H).

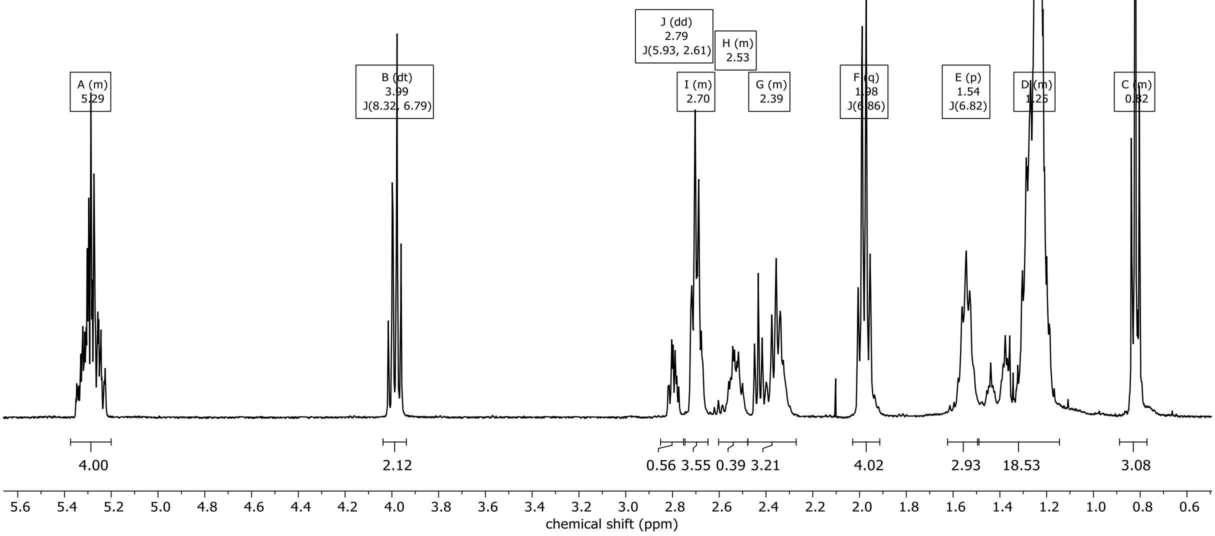


Figure VII.S18: 1H NMR of final Lipid15 crude

^{1}H NMR (400 MHz, $CDCl_3$) δ 5.40 – 5.27 (m, 2H), 3.61 (t, *J* = 6.7 Hz, 2H), 2.09 – 1.95 (m, 4H), 1.61 – 1.48 (m, 2H), 1.29 (ddt, *J* = 17.9, 14.5, 4.9 Hz, 22H), 0.96 – 0.81 (m, 3H).

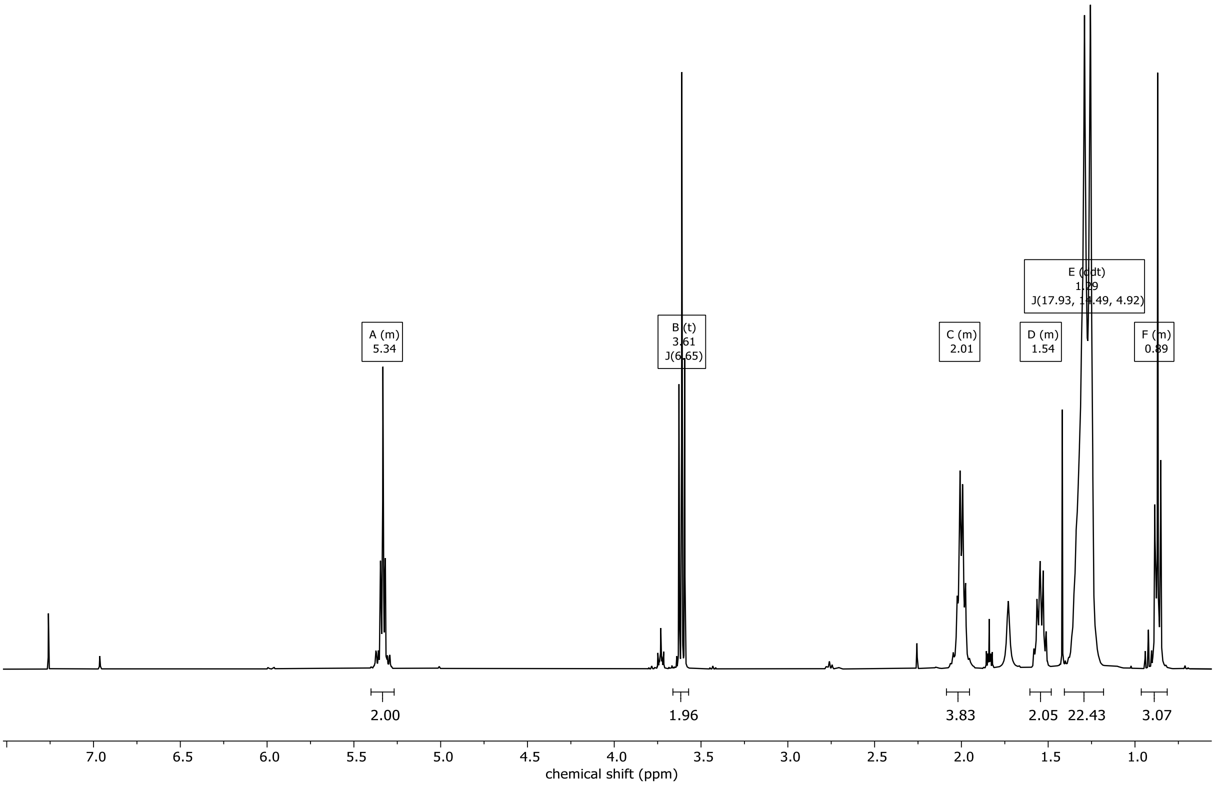


Figure VII.S19: 1H NMR of linoleyl alcohol

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.32 (dd, *J* = 17.3, 1.5 Hz, 1H), 6.04 (dd, *J* = 17.4, 10.4 Hz, 1H), 5.73 (dd, *J* = 10.4, 1.5 Hz, 1H), 5.37 – 5.20 (m, 4H), 4.08 (t, *J* = 6.7 Hz, 2H), 2.75 – 2.64 (m, 2H), 1.98 (q, *J* = 6.9 Hz, 4H), 1.65 – 1.53 (m, 2H), 1.35 – 1.14 (m, 16H), 0.88 – 0.77 (m, 3H).

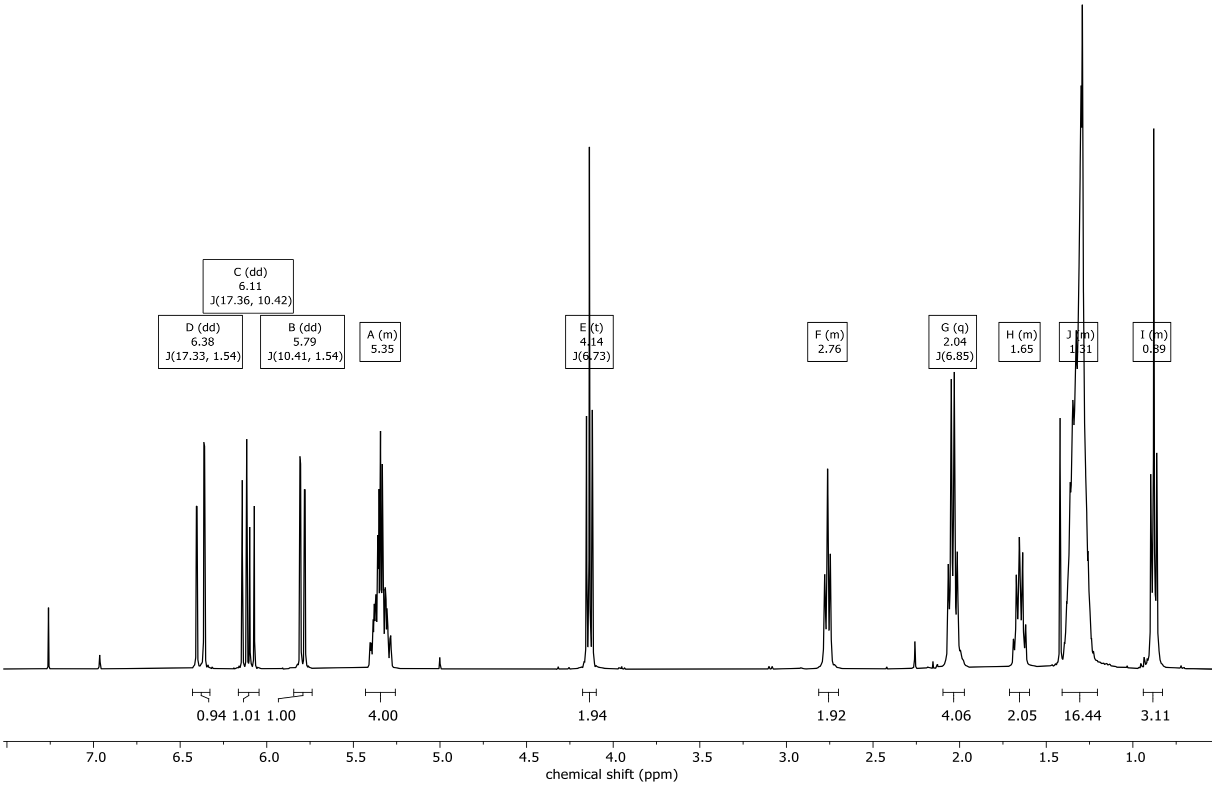


Figure VII.S20: 1H NMR of oleyl alcohol

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.38 (dd, *J* = 17.3, 1.6 Hz, 1H), 6.11 (dd, *J* = 17.3, 10.4 Hz, 1H), 5.79 (dd, *J* = 10.5, 1.6 Hz, 1H), 5.47 – 5.17 (m, 2H), 4.14 (t, *J* = 6.7 Hz, 2H), 2.00 (q, *J* = 6.6 Hz, 4H), 1.71 – 1.60 (m, 2H), 1.41 – 1.20 (m, 23H), 0.92 – 0.81 (m, 3H).

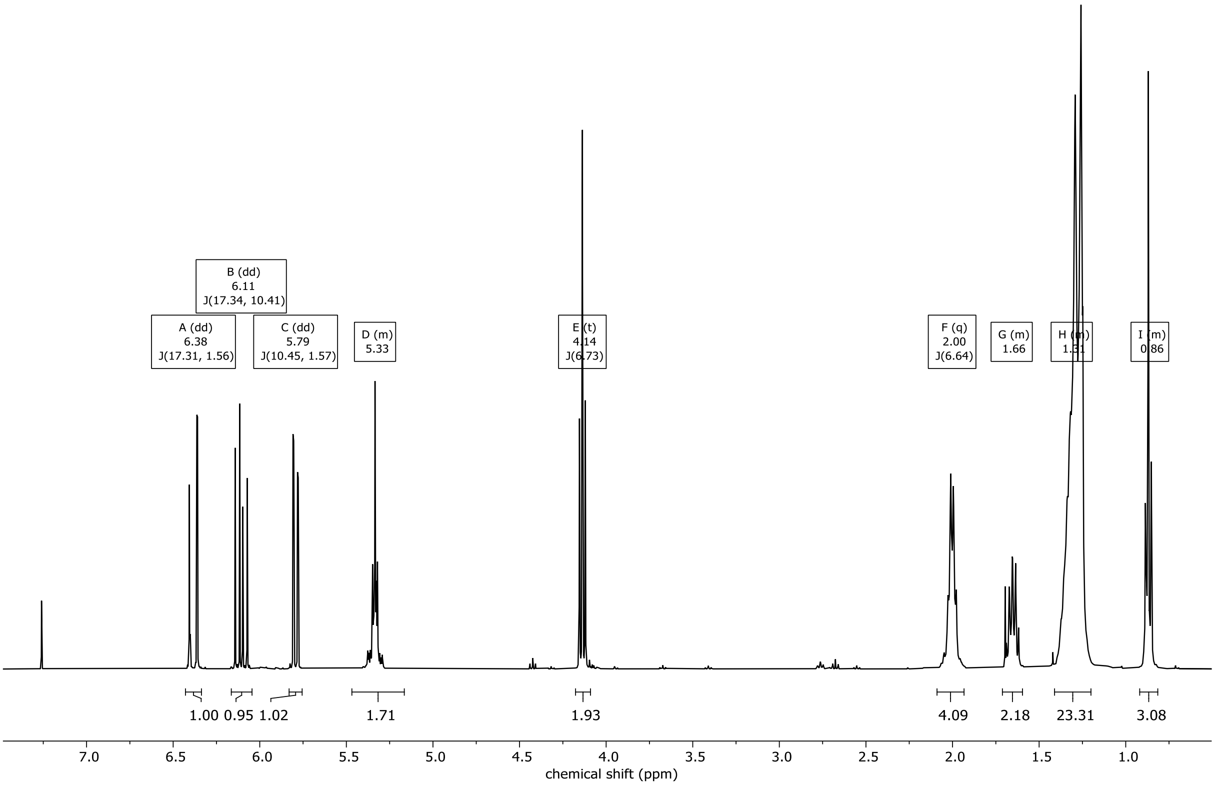


Figure VII.S21: 1H NMR of Oleyl acrylate

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.52 – 7.41 (m, 2H), 7.36 – 7.27 (m, 3H), 6.49 (dd, *J* = 17.3, 1.4 Hz, 1H), 6.18 (dd, *J* = 17.4, 10.4 Hz, 1H), 5.88 (dd, *J* = 10.4, 1.4 Hz, 1H), 5.00 (s, 2H).

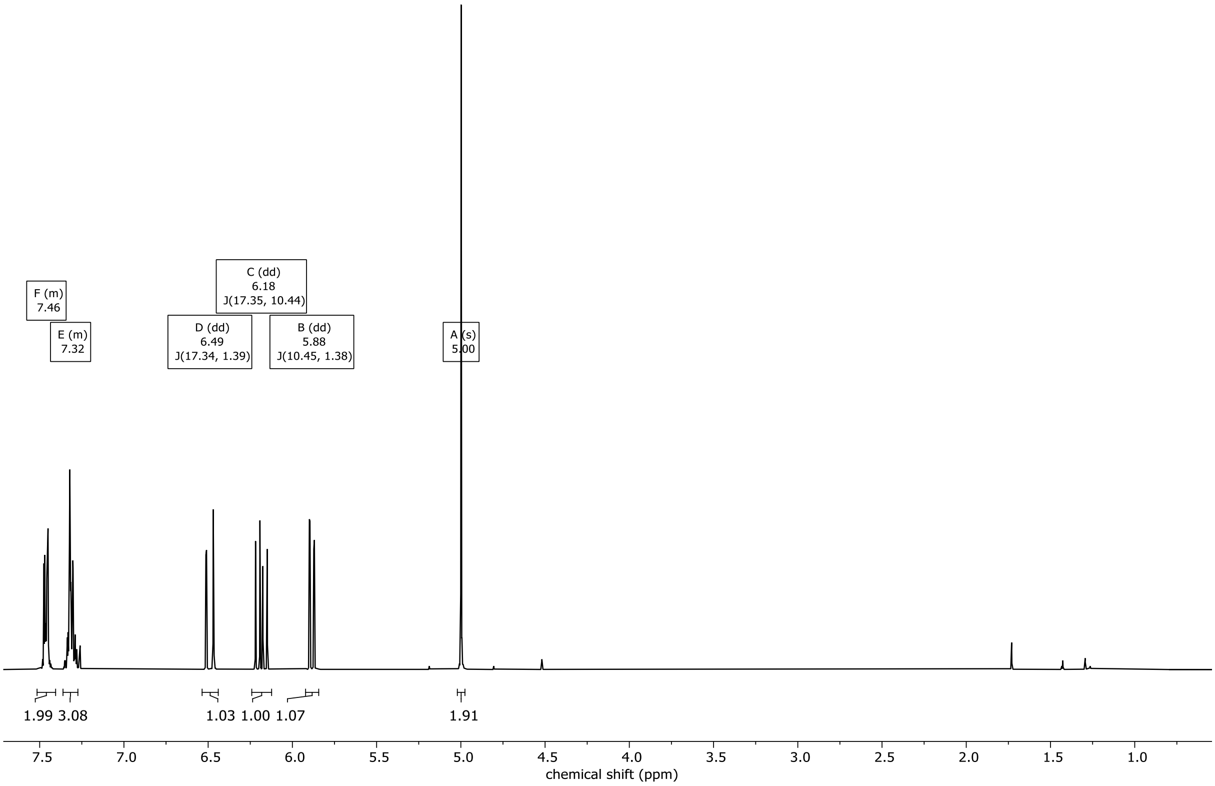


Figure VII.S22: 1H NMR of 3-phenylprop-2-yn-1-yl acrylate

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.38 (dd, *J* = 17.3, 1.5 Hz, 1H), 6.11 (dd, *J* = 17.3, 10.4 Hz, 1H), 5.84 – 5.75 (m, 1H), 4.14 (t, *J* = 6.8 Hz, 2H), 1.65 (dq, *J* = 8.0, 6.6 Hz, 2H), 1.33 – 1.20 (m, 18H), 0.94 – 0.81 (m, 3H).

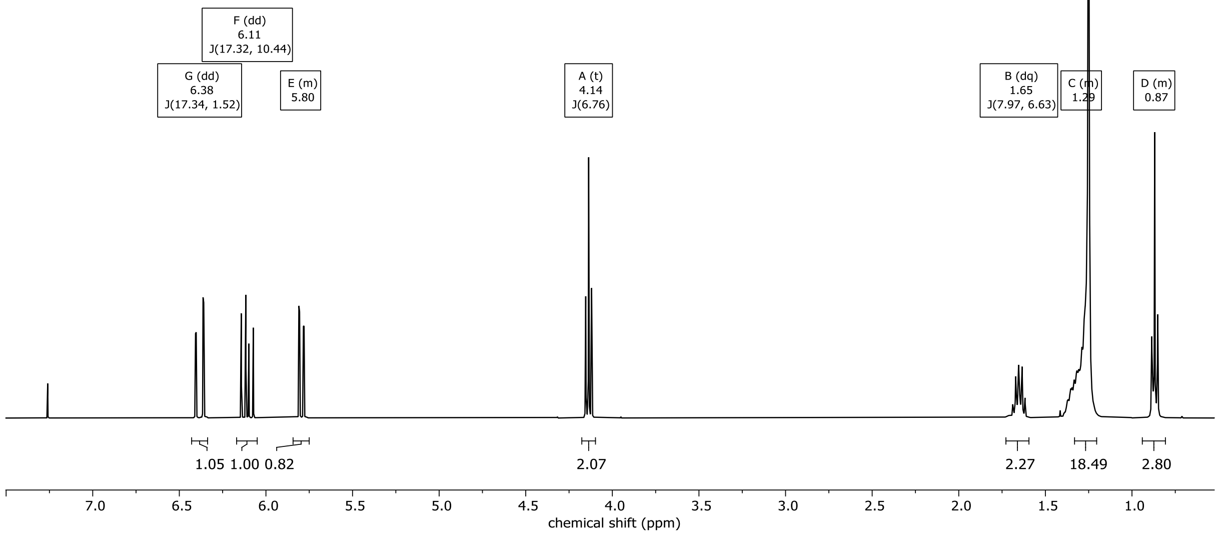


Figure VII.S23: 1H NMR of dodecyl acrylate

^{1}H NMR (400 MHz, $CDCl_3$) δ 5.44 – 5.27 (m, 4H), 3.64 (t, *J* = 6.6 Hz, 2H), 2.81 – 2.73 (m, 2H), 2.05 (q, *J* = 6.8 Hz, 4H), 1.62 – 1.51 (m, 2H), 1.42 – 1.22 (m, 16H), 0.93 – 0.84 (m, 3H).

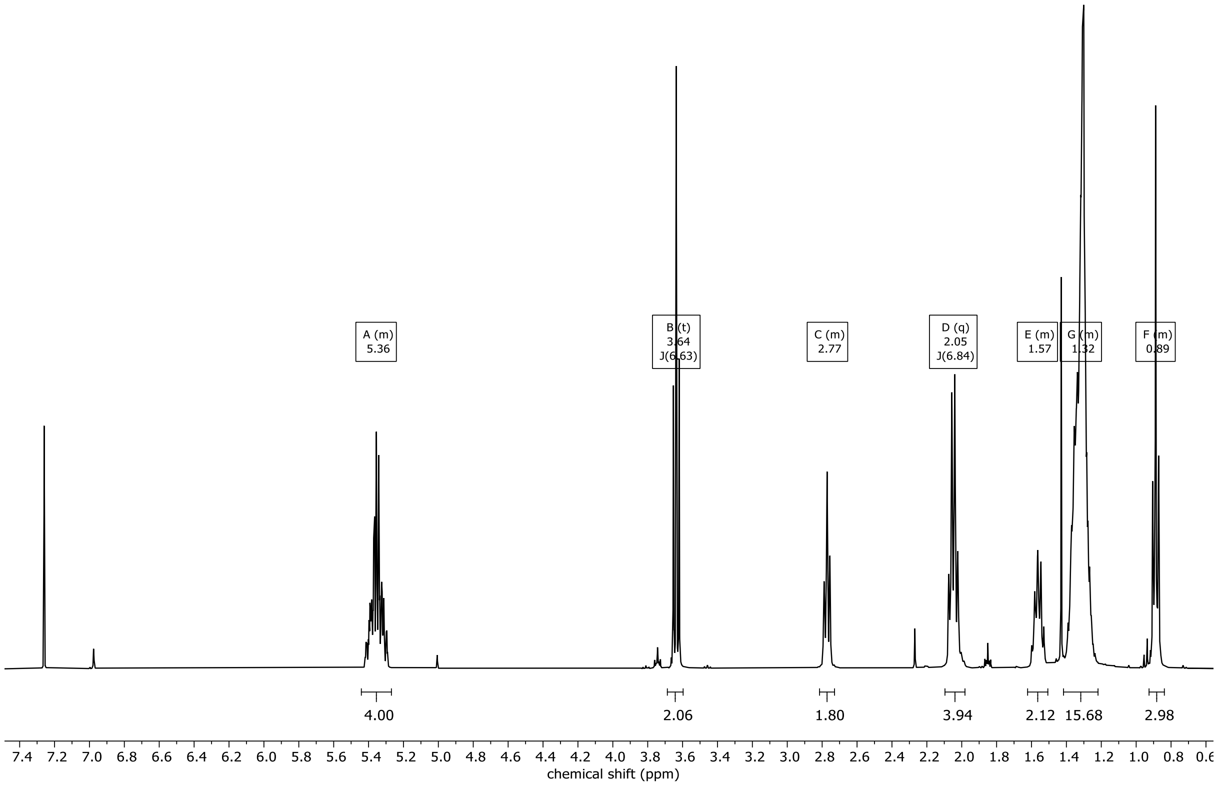


Figure VII.S24: 1H NMR of Linoleyl alcohol

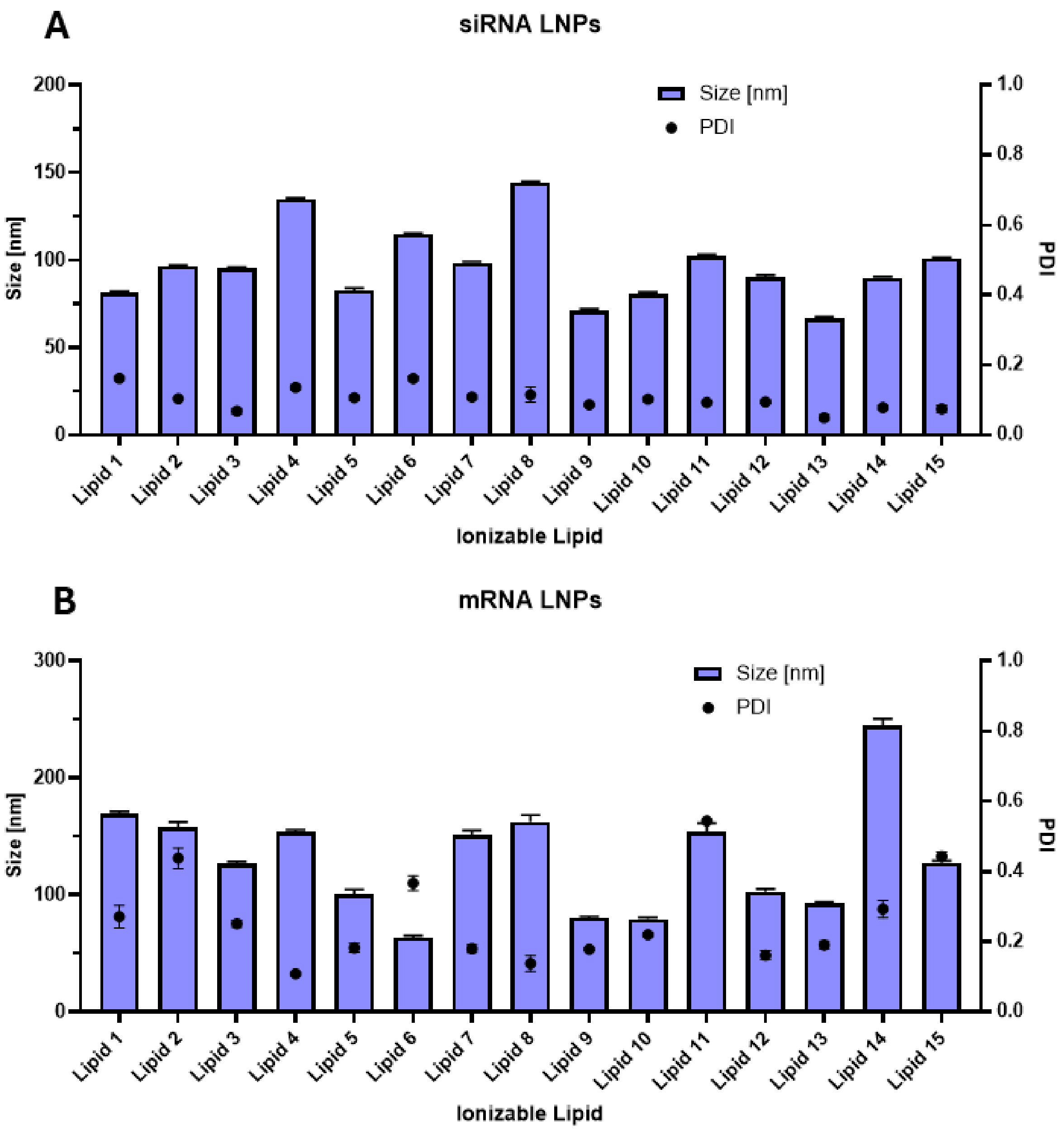


Figure VII.S25: Size and Polydispersity Index of siRNA LNPs (**A**) and mRNA LNPs (**B**) after dialysis and sterile filtration. Data is shown as mean + SD, n=3 technical replicates.

Chapter VIII - Summary and Perspective

The complexity and multidimensionality of drug delivery processes make them particularly amenable to data-driven approaches. Although the integration of such methods into drug delivery is still in its infancy, this thesis shows that different computational strategies can already address distinct problems along the modern formulation pipeline.

Chapter II demonstrated how the properties of polymeric nanocarriers can be optimized at the level of synthesis using Design of Experiments (DoE). This work underlined that experimental evidence should be generated through a structured experimental plan and subsequent statistical analysis, and that meaningful structure-property relationships can be established even for complex polymer systems. In addition, we showed that integrating Python-based workflows into the experimental process offers flexible ways to handle and analyse data more efficiently. This idea was further advanced in Chapter III, where we incorporated machine learning (ML) as a flexible extension once DoE became limiting. There, we showed how robust models can be applied to small, orthogonally designed datasets and used to identify synthesis parameters that optimize nanocarriers for in vitro transfection efficiency.

As outlined in the introduction, the quantity and quality of available data remain critical bottlenecks. Cheminformatics is considerably more mature in its use of ML than drug delivery, primarily because large, standardised and easily accessible datasets exist, which allows researchers to focus on model development rather than basic data assembly. In drug delivery, the situation is more complex. Multi-component formulations, multi-scale readouts, high experimental noise, fragmented datasets and the lack of standardised protocols all contribute to the challenge. In this thesis, literature-derived data were used to illustrate how such fragmented information can be merged into a more informative system that can guide the discovery of new polymeric carrier materials (Chapter IV). In this setting, ML enabled the prediction of carrier performance in five out of six cases in vitro, and the lead candidate also showed promising results in vivo.

For lipid nanoparticle (LNP) development, substantially more data and research are available, driven by the success of recently approved LNP-based products. However, when

LNP formulation design is cast as an ML problem, the optimisation space becomes highly complex, because multi-component systems result in a formulation space with effectively thousands of dimensions. In Chapter VII, we therefore investigated whether meta-learning approaches can help to avoid strong biases toward specific historical datasets and improve generalisation. The results showed that such methods can serve as powerful base models for active learning. In simulated optimisation tasks on an unseen cargo, meta-learning based models were able to identify early hits up to 17 times faster than a conventional baseline, effectively supporting both the discovery and optimisation of new formulations.

In the absence of large, standardised datasets, physics-informed systems offer an attractive alternative, since they provide more detailed mechanistic insight into molecular behaviour and can build robust datasets from scratch. This was demonstrated in Chapter V, where we developed a program that samples molecules, labels them in a high-throughput manner using molecular dynamics (MD) simulations, and optimises candidates in silico by combining simulation with AI-based optimisation and molecule preparation. The complete simulation workflow was validated and calibrated against wet-lab experiments to ensure sufficient realism. We identified several interesting and structurally novel candidates with limited similarity to previously known high-performing structures, which indicates that such approaches can promote true novelty in carrier design. Although this workflow still involves trade-offs between physical realism and computational speed, ongoing advances in hardware, for example massively parallel GPU execution, are likely to reduce simulation and optimisation times and will make such approaches increasingly practical.

A similar rationale applies to 4D-QSTR, introduced in Chapter VI. This framework aggregates time-resolved dynamic information from MD simulations into ML-usable descriptors and provides complementary signal on molecular behaviour, particularly for extrapolation and cliff-like tasks that are central challenges in early material discovery. Across several benchmarks, we observed that incorporating dynamic information improved predictive performance precisely in those regimes where conventional 2D and 3D descriptors tend to struggle. For some simulations, performance gains of up to 20 % relative to 2D/3D baselines were observed. At the same time, 2D descriptor baselines remained stronger for standard, random-split tasks. This led us to hypothesise that datasets generated by expert structure optimisation carry an implicit 2D human bias, which still favours simple fingerprints in familiar regions of chemical space. Overall, these findings suggest that the complexity of the prediction task must be reflected in the complexity of the information that is provided to the model, especially when the goal is to discover genuinely new carrier structures.

Taken together, the results presented in this thesis can be viewed as one of the early, systematic attempts to integrate data-driven approaches into drug-delivery workflows. We provide new insights into the synthesis of PBAEs, how their polymeric properties relate to their performance as nanoparticle systems, and how this knowledge can be used for rational design and optimisation. Furthermore, by demonstrating how literature-based data can be repurposed for carrier discovery and optimisation, we outline a path that is accessible to research groups worldwide that may not have access to high-throughput experimentation but can still benefit from data-driven guidance to save time and material. The physics-based approaches introduced here provide a starting point for other researchers to extend, adapt and improve such systems for different delivery challenges. All code used in this thesis is openly available on GitHub (https://github.com/felixsie19), which supports transparency and reuse.

Looking ahead, it is reasonable to expect that artificial intelligence will continue to develop rapidly and will be progressively integrated into drug delivery. Recent studies already report strong performance on difficult design problems and indicate a trend toward increasingly automated workflows for synthesis, formulation and testing. While such automation is essential for building robust models, it is equally important to remember that data-driven approaches are only as reliable as the underlying data. The rapid pace of technological development therefore needs to be accompanied by community-wide standards for manufacturing, characterisation and biological testing in order to ensure that data are comparable across laboratories. As shown in Chapter VII, certain methods can mitigate lab-to-lab variability and batch effects, but high-performing AI methods still rely predominantly on large, well-curated datasets.

Recent research in AI increasingly focuses on the concept of “world models” as a next step in model development. The underlying idea is to move beyond purely data-driven, static learning toward trial-and-error learning in explicitly modelled or simulated environments. Chapter V already illustrated how such ideas could look in the context of drug delivery by coupling MD simulations with ML-based optimisation. Future work may explore different strategies to use such models to accelerate the path from early discovery to clinical studies.

It will be essential that expertise in simulations, ML and their integration into experimental workflows is established early in scientific training, so that future generations of researchers develop an intuitive understanding of data-driven methods and can exploit them to address unresolved therapeutic challenges.

Bibliography

(1) *EU/3/19/2210 - orphan designation for treatment of beta thalassaemia intermedia and major | European Medicines Agency (EMA)*. https://www.ema.europa.eu/en/medicines/human/orphan-designations/eu-3-19-2210 (accessed 2025-09-10).

(2) Raal, F. J.; Kallend, D.; Ray, K. K.; Turner, T.; Koenig, W.; Wright, R. S.; Wijngaard, P. L. J.; Curcio, D.; Jaros, M. J.; Leiter, L. A.; Kastelein, J. J. P. Inclisiran for the Treatment of Heterozygous Familial Hypercholesterolemia. *N. Engl. J. Med.* **2020**, *382* (16), 1520–1530. https://doi.org/10.1056/NEJMoa1913805.

(3) Rosenberg, S. A.; Aebersold, P.; Cornetta, K.; Kasid, A.; Morgan, R. A.; Moen, R.; Karson, E. M.; Lotze, M. T.; Yang, J. C.; Topalian, S. L.; Merino, M. J.; Culver, K.; Miller, A. D.; Blaese, R. M.; Anderson, W. F. Gene Transfer into Humans — Immunotherapy of Patients with Advanced Melanoma, Using Tumor-Infiltrating Lymphocytes Modified by Retroviral Gene Transduction. *N. Engl. J. Med.* **1990**, *323* (9), 570–578. https://doi.org/10.1056/NEJM199008303230904.

(4) Friedmann, T.; Roblin, R. Gene Therapy for Human Genetic Disease? *Science* **1972**, *175* (4025), 949–955. https://doi.org/10.1126/science.175.4025.949.

(5) Baden, L. R.; Sahly, H. M. E.; Essink, B.; Kotloff, K.; Frey, S.; Novak, R.; Diemert, D.; Spector, S. A.; Rouphael, N.; Creech, C. B.; McGettigan, J.; Khetan, S.; Segall, N.; Solis, J.; Brosz, A.; Fierro, C.; Schwartz, H.; Neuzil, K.; Corey, L.; Gilbert, P.; Janes, H.; Follmann, D.; Marovich, M.; Mascola, J.; Polakowski, L.; Ledgerwood, J.; Graham, B. S.; Bennett, H.; Pajon, R.; Knightly, C.; Leav, B.; Deng, W.; Zhou, H.; Han, S.; Ivarsson, M.; Miller, J.; Zaks, T. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* **2021**, *384* (5), 403–416. https://doi.org/10.1056/NEJMoa2035389.

(6) Polack, F. P.; Thomas, S. J.; Kitchin, N.; Absalon, J.; Gurtman, A.; Lockhart, S.; Perez, J. L.; Marc, G. P.; Moreira, E. D.; Zerbini, C.; Bailey, R.; Swanson, K. A.; Roychoudhury, S.; Koury, K.; Li, P.; Kalina, W. V.; Cooper, D.; Frenck, R. W.; Hammitt, L. L.; Türeci, Ö.; Nell, H.; Schaefer, A.; Ünal, S.; Tresnan, D. B.; Mather, S.; Dormitzer, P. R.; Şahin, U.; Jansen, K. U.; Gruber, W. C. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N. Engl. J. Med.* **2020**, *383* (27), 2603–2615. https://doi.org/10.1056/NEJMoa2034577.

(7) Koeberl, D.; Schulze, A.; Sondheimer, N.; Lipshutz, G. S.; Geberhiwot, T.; Li, L.; Saini, R.; Luo, J.; Sikirica, V.; Jin, L.; Liang, M.; Leuchars, M.; Grunewald, S. Interim Analyses of a First-in-Human Phase 1/2 mRNA Trial for Propionic Acidaemia. *Nature* **2024**, *628* (8009), 872–877. https://doi.org/10.1038/s41586-024-07266-7.

(8) ModernaTX, Inc. *A Global, Phase 1&#x2F;2, Open-Label, Dose Optimization Study to Evaluate the Safety, Tolerability, Pharmacodynamics, and Pharmacokinetics of mRNA-3705 in Participants With Isolated Methylmalonic Acidemia Due to Methylmalonyl-CoA Mutase Deficiency*; Clinical trial registration NCT04899310; clinicaltrials.gov, 2025. https://clinicaltrials.gov/study/NCT04899310 (accessed 2025-09-10).

(9) Seker Yilmaz, B.; Gissen, P. Genetic Therapy Approaches for Ornithine Transcarbamylase Deficiency. *Biomedicines* **2023**, *11* (8), 2227. https://doi.org/10.3390/biomedicines11082227.

(10) Kulkarni, J. A.; Witzigmann, D.; Thomson, S. B.; Chen, S.; Leavitt, B. R.; Cullis, P. R.; van der Meel, R. The Current Landscape of Nucleic Acid Therapeutics. *Nat. Nanotechnol.* **2021**, *16* (6), 630–643. https://doi.org/10.1038/s41565-021-00898-0.

(11) Setten, R. L.; Rossi, J. J.; Han, S. The Current State and Future Directions of RNAi-Based Therapeutics. *Nat. Rev. Drug Discov.* **2019**, *18* (6), 421–446. https://doi.org/10.1038/s41573-019-0017-4.

(12) Balwani, M.; Sardh, E.; Ventura, P.; Peiró, P. A.; Rees, D. C.; Stölzel, U.; Bissell, D. M.; Bonkovsky, H. L.; Windyga, J.; Anderson, K. E.; Parker, C.; Silver, S. M.; Keel, S. B.; Wang, J.-D.; Stein, P. E.; Harper, P.; Vassiliou, D.; Wang, B.; Phillips, J.; Ivanova, A.; Langendonk, J. G.; Kauppinen, R.; Minder, E.; Horie, Y.; Penz, C.; Chen, J.; Liu, S.; Ko, J. J.; Sweetser, M. T.; Garg, P.; Vaishnaw, A.; Kim, J. B.; Simon, A. R.; Gouya, L. Phase 3 Trial of RNAi Therapeutic Givosiran for Acute Intermittent Porphyria. *N. Engl. J. Med.* **2020**, *382* (24), 2289–2301. https://doi.org/10.1056/NEJMoa1913147.

(13) Moran, N. First Gene Therapy Approved. *Nat. Biotechnol.* **2012**, *30* (12), 1153–1153. https://doi.org/10.1038/nbt1212-1153.

(14) Nidetz, N. F.; McGee, M. C.; Tse, L. V.; Li, C.; Cong, L.; Li, Y.; Huang, W. Adeno-Associated Viral Vector-Mediated Immune Responses: Understanding Barriers to Gene Delivery. *Pharmacol. Ther.* **2020**, *207*, 107453. https://doi.org/10.1016/j.pharmthera.2019.107453.

(15) Jiang, Z.; Dalby, P. A. Challenges in Scaling up AAV-Based Gene Therapy Manufacturing. *Trends Biotechnol.* **2023**, *41* (10), 1268–1281. https://doi.org/10.1016/j.tibtech.2023.04.002.

(16) Wang, J.-H.; Gessler, D. J.; Zhan, W.; Gallagher, T. L.; Gao, G. Adeno-Associated Virus as a Delivery Vector for Gene Therapy of Human Diseases. *Signal Transduct. Target. Ther.* **2024**, *9* (1), 78. https://doi.org/10.1038/s41392-024-01780-w.

(17) van den Berg, A. I. S.; Yun, C.-O.; Schiffelers, R. M.; Hennink, W. E. Polymeric Delivery Systems for Nucleic Acid Therapeutics: Approaching the Clinic. *J. Controlled Release* **2021**, *331*, 121–141. https://doi.org/10.1016/j.jconrel.2021.01.014.

(18) Graham, F. L.; van der Eb, A. J. A New Technique for the Assay of Infectivity of Human Adenovirus 5 DNA. *Virology* **1973**, *52* (2), 456–467. https://doi.org/10.1016/0042-6822(73)90341-3.

(19) Peng, L.; Wagner, E. Polymeric Carriers for Nucleic Acid Delivery: Current Designs and Future Directions. *Biomacromolecules* **2019**, *20* (10), 3613–3626. https://doi.org/10.1021/acs.biomac.9b00999.

(20) Boussif, O.; Lezoualc'h, F.; Zanta, M. A.; Mergny, M. D.; Scherman, D.; Demeneix, B.; Behr, J. P. A Versatile Vector for Gene and Oligonucleotide Transfer into Cells in Culture and in Vivo: Polyethylenimine. *Proc. Natl. Acad. Sci.* **1995**, *92* (16), 7297–7301. https://doi.org/10.1073/pnas.92.16.7297.

(21) Debus, H.; Baumhof, P.; Probst, J.; Kissel, T. Delivery of Messenger RNA Using Poly(Ethylene Imine)–Poly(Ethylene Glycol)-Copolymer Blends for Polyplex Formation: Biophysical Characterization and in Vitro Transfection Properties. *J. Controlled Release* **2010**, *148* (3), 334–343. https://doi.org/10.1016/j.jconrel.2010.09.007.

(22) Urban-Klein, B.; Werth, S.; Abuharbeid, S.; Czubayko, F.; Aigner, A. RNAi-Mediated Gene-Targeting through Systemic Application of Polyethylenimine (PEI)-Complexed siRNA in Vivo. *Gene Ther.* **2005**, *12* (5), 461–466. https://doi.org/10.1038/sj.gt.3302425.

(23) Moghimi, S. M.; Symonds, P.; Murray, J. C.; Hunter, A. C.; Debska, G.; Szewczyk, A. A Two-Stage Poly(Ethylenimine)-Mediated Cytotoxicity: Implications for Gene Transfer/Therapy. *Mol. Ther.* **2005**, *11* (6), 990–995. https://doi.org/10.1016/j.ymthe.2005.02.010.

(24) Fischer, D.; Bieber, T.; Li, Y.; Elsässer, H.-P.; Kissel, T. A Novel Non-Viral Vector for DNA Delivery Based on Low Molecular Weight, Branched Polyethylenimine: Effect of Molecular Weight on Transfection Efficiency and Cytotoxicity. *Pharm. Res.* **1999**, *16* (8), 1273–1279. https://doi.org/10.1023/A:1014861900478.

(25) Casper, J.; Schenk, S. H.; Parhizkar, E.; Detampel, P.; Dehshahri, A.; Huwyler, J. Polyethylenimine (PEI) in Gene Therapy: Current Status and Clinical Applications. *J. Controlled Release* **2023**, *362*, 667–691. https://doi.org/10.1016/j.jconrel.2023.09.001.

(26) Lynn, D. M.; Langer, R. Degradable Poly(β-Amino Esters): Synthesis, Characterization, and Self-Assembly with Plasmid DNA. *J. Am. Chem. Soc.* **2000**, *122* (44), 10761–10768. https://doi.org/10.1021/ja0015388.

(27) Green, J. J.; Langer, R.; Anderson, D. G. A Combinatorial Polymer Library Approach Yields Insight into Nonviral Gene Delivery. *Acc. Chem. Res.* **2008**, *41* (6), 749–759. https://doi.org/10.1021/ar7002336.

(28) Hua, S.; de Matos, M. B. C.; Metselaar, J. M.; Storm, G. Current Trends and Challenges in the Clinical Translation of Nanoparticulate Nanomedicines: Pathways for Translational Development and Commercialization. *Front. Pharmacol.* **2018**, *9*. https://doi.org/10.3389/fphar.2018.00790.

(29) Adams, D.; Gonzalez-Duarte, A.; O'Riordan, W. D.; Yang, C.-C.; Ueda, M.; Kristen, A. V.; Tournev, I.; Schmidt, H. H.; Coelho, T.; Berk, J. L.; Lin, K.-P.; Vita, G.; Attarian, S.; Planté-Bordeneuve, V.; Mezei, M. M.; Campistol, J. M.; Buades, J.; Brannagan, T. H.; Kim, B. J.; Oh, J.; Parman, Y.; Sekijima, Y.; Hawkins, P. N.; Solomon, S. D.; Polydefkis, M.; Dyck, P. J.; Gandhi, P. J.; Goyal, S.; Chen, J.; Strahs, A. L.; Nochur, S. V.; Sweetser, M. T.; Garg, P. P.; Vaishnaw, A. K.; Gollob, J. A.; Suhr, O. B. Patisiran, an RNAi Therapeutic, for Hereditary Transthyretin Amyloidosis. *N. Engl. J. Med.* **2018**, *379* (1), 11–21. https://doi.org/10.1056/NEJMoa1716153.

(30) Hald Albertsen, C.; Kulkarni, J. A.; Witzigmann, D.; Lind, M.; Petersson, K.; Simonsen, J. B. The Role of Lipid Components in Lipid Nanoparticles for Vaccines and Gene Therapy. *Adv. Drug Deliv. Rev.* **2022**, *188*, 114416. https://doi.org/10.1016/j.addr.2022.114416.

(31) Thalmayr, S.; Grau, M.; Peng, L.; Pöhmerer, J.; Wilk, U.; Folda, P.; Yazdi, M.; Weidinger, E.; Burghardt, T.; Höhn, M.; Wagner, E.; Berger, S. Molecular Chameleon Carriers for Nucleic Acid Delivery: The Sweet Spot between Lipoplexes and Polyplexes. *Adv. Mater.* **2023**, *35* (25), 2211105. https://doi.org/10.1002/adma.202211105.

(32) Jayaraman, M.; Ansell, S. M.; Mui, B. L.; Tam, Y. K.; Chen, J.; Du, X.; Butler, D.; Eltepu, L.; Matsuda, S.; Narayanannair, J. K.; Rajeev, K. G.; Hafez, I. M.; Akinc, A.; Maier, M. A.; Tracy, M. A.; Cullis, P. R.; Madden, T. D.; Manoharan, M.; Hope, M. J. Maximizing the Potency of siRNA Lipid Nanoparticles for Hepatic Gene Silencing In Vivo. *Angew. Chem. Int. Ed.* **2012**, *51* (34), 8529–8533. https://doi.org/10.1002/anie.201203263.

(33) Chatterjee, S.; Kon, E.; Sharma, P.; Peer, D. Endosomal Escape: A Bottleneck for LNP-Mediated Therapeutics. *Proc. Natl. Acad. Sci.* **2024**, *121* (11), e2307800120. https://doi.org/10.1073/pnas.2307800120.

(34) Hagedorn, L.; Jürgens, D. C.; Merkel, O. M.; Winkeljann, B. Endosomal Escape Mechanisms of Extracellular Vesicle-Based Drug Carriers: Lessons for Lipid Nanoparticle Design. *Extracell. Vesicles Circ. Nucleic Acids* **2024**, *5* (3), 344–357. https://doi.org/10.20517/evcna.2024.19.

(35) Jung, H. N.; Lee, S.-Y.; Lee, S.; Youn, H.; Im, H.-J. Lipid Nanoparticles for Delivery of RNA Therapeutics: Current Status and the Role of in Vivo Imaging. *Theranostics* **2022**, *12* (17), 7509–7531. https://doi.org/10.7150/thno.77259.

(36) Ai, L.; Li, Y.; Zhou, L.; Yao, W.; Zhang, H.; Hu, Z.; Han, J.; Wang, W.; Wu, J.; Xu, P.; Wang, R.; Li, Z.; Li, Z.; Wei, C.; Liang, J.; Chen, H.; Yang, Z.; Guo, M.; Huang, Z.; Wang, X.; Zhang, Z.; Xiang, W.; Sun, D.; Xu, L.; Huang, M.; Lv, B.; Peng, P.; Zhang, S.; Ji, X.; Luo, H.; Chen, N.; Chen, J.; Lan, K.; Hu, Y. Lyophilized mRNA-Lipid Nanoparticle Vaccines with Long-Term Stability and High Antigenicity against SARS-CoV-2. *Cell Discov.* **2023**, *9* (1), 9. https://doi.org/10.1038/s41421-022-00517-9.

(37) Messerian, K. O.; Zverev, A.; Kramarczyk, J. F.; Zydney, A. L. Characterization and Associated Pressure-Dependent Behavior of Deposits Formed during Sterile Filtration of mRNA-Lipid Nanoparticles. *J. Membr. Sci.* **2023**, *684*, 121896. https://doi.org/10.1016/j.memsci.2023.121896.

(38) Maeki, M.; Uno, S.; Niwa, A.; Okada, Y.; Tokeshi, M. Microfluidic Technologies and Devices for Lipid Nanoparticle-Based RNA Delivery. *J. Controlled Release* **2022**, *344*, 80–96. https://doi.org/10.1016/j.jconrel.2022.02.017.

(39) Agha, A.; Waheed, W.; Stiharu, I.; Nerguizian, V.; Destgeer, G.; Abu-Nada, E.; Alazzam, A. A Review on Microfluidic-Assisted Nanoparticle Synthesis, and Their Applications Using Multiscale Simulation Methods. *Discov. Nano* **2023**, *18* (1), 18. https://doi.org/10.1186/s11671-023-03792-x.

(40) Palanki, R.; L. Han, E.; M. Murray, A.; Maganti, R.; Tang, S.; L. Swingle, K.; Kim, D.; Yamagata, H.; C. Safford, H.; Mrksich, K.; H. Peranteau, W.; J. Mitchell, M. Optimized Microfluidic Formulation and Organic Excipients for Improved Lipid Nanoparticle Mediated Genome Editing. *Lab. Chip* **2024**, *24* (16), 3790–3801. https://doi.org/10.1039/D4LC00283K.

(41) Strelkova Petersen, D. M.; Chaudhary, N.; Arral, M. L.; Weiss, R. M.; Whitehead, K. A. The Mixing Method Used to Formulate Lipid Nanoparticles Affects mRNA Delivery Efficacy and Organ Tropism. *Eur. J. Pharm. Biopharm.* **2023**, *192*, 126–135. https://doi.org/10.1016/j.ejpb.2023.10.006.

(42) Dahlman, J. E.; Kauffman, K. J.; Xing, Y.; Shaw, T. E.; Mir, F. F.; Dlott, C. C.; Langer, R.; Anderson, D. G.; Wang, E. T. Barcoded Nanoparticles for High Throughput in Vivo Discovery of Targeted Therapeutics. *Proc. Natl. Acad. Sci.* **2017**, *114* (8), 2060–2065. https://doi.org/10.1073/pnas.1620874114.

(43) Kumar, R.; Le, N.; Oviedo, F.; Brown, M. E.; Reineke, T. M. Combinatorial Polycation Synthesis and Causal Machine Learning Reveal Divergent Polymer Design Rules for Effective pDNA and Ribonucleoprotein Delivery. ChemRxiv October 21, 2021. https://doi.org/10.26434/chemrxiv-2021-cdmpf.

(44) N. Politis, S.; Colombo, P.; Colombo, G.; M. Rekkas, D. Design of Experiments (DoE) in Pharmaceutical Development. *Drug Dev. Ind. Pharm.* **2017**, *43* (6), 889–901. https://doi.org/10.1080/03639045.2017.1291672.

(45) Fukuda, I. M.; Pinto, C. F. F.; Moreira, C. dos S.; Saviano, A. M.; Lourenço, F. R. Design of Experiments (DoE) Applied to Pharmaceutical and Analytical Quality by Design (QbD). *Braz. J. Pharm. Sci.* **2018**, *54*, e01006. https://doi.org/10.1590/s2175-97902018000001006.

(46) Yu, L. X.; Amidon, G.; Khan, M. A.; Hoag, S. W.; Polli, J.; Raju, G. K.; Woodcock, J. Understanding Pharmaceutical Quality by Design. *AAPS J.* **2014**, *16* (4), 771–783. https://doi.org/10.1208/s12248-014-9598-3.

(47) Wolpert, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Comput.* **1996**, *8* (7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341.

(48) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(49) 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*; KDD ’16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

(50) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.

(51) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2018; Vol. 31.

(52) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444. https://doi.org/10.1038/nature14539.

(53) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323* (6088), 533–536. https://doi.org/10.1038/323533a0.

(54) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2012; Vol. 25.

(55) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv February 22, 2017. https://doi.org/10.48550/arXiv.1609.02907.

(56) Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; Zhou, Y. Deep Learning Scaling Is Predictable, Empirically. arXiv December 1, 2017. https://doi.org/10.48550/arXiv.1712.00409.

(57) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15* (56), 1929–1958.

(58) Müller, R.; Kornblith, S.; Hinton, G. When Does Label Smoothing Help? arXiv June 10, 2020. https://doi.org/10.48550/arXiv.1906.02629.

(59) Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*; PMLR, 2015; pp 448–456.

(60) Ortiz-Perez, A.; Tilborg, D. van; Meel, R. van der; Grisoni, F.; Albertazzi, L. Machine Learning-Guided High Throughput Nanoparticle Design. *Digit. Discov.* **2024**, *3* (7), 1280–1291. https://doi.org/10.1039/D4DD00104D.

(61) Kusne, A. G.; Yu, H.; Wu, C.; Zhang, H.; Hattrick-Simpers, J.; DeCost, B.; Sarker, S.; Oses, C.; Toher, C.; Curtarolo, S.; Davydov, A. V.; Agarwal, R.; Bendersky, L. A.; Li, M.; Mehta, A.; Takeuchi, I. On-the-Fly Closed-Loop Materials Discovery via Bayesian Active Learning. *Nat. Commun.* **2020**, *11* (1), 5966. https://doi.org/10.1038/s41467-020-19597-w.

(62) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96. https://doi.org/10.1038/s41586-021-03213-y.

(63) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(64) Bonachéra, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and

Adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, *46* (6), 2457–2477. https://doi.org/10.1021/ci6002416.

(65) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(66) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717. https://doi.org/10.1021/jm000942e.

(67) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. In *3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Springer Netherlands: Dordrecht, 1998; pp 339–353. https://doi.org/10.1007/0-306-46857-3_18.

(68) Todeschini, R.; Gramatica, P. The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling. *SAR QSAR Environ. Res.* **1997**, *7* (1–4), 89–115. https://doi.org/10.1080/10629369708039126.

(69) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Message Passing Neural Networks. In *Machine Learning Meets Quantum Physics*; Schütt, K. T., Chmiela, S., von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K., Müller, K.-R., Eds.; Springer International Publishing: Cham, 2020; pp 199–214. https://doi.org/10.1007/978-3-030-40245-7_10.

(70) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. arXiv February 4, 2018. https://doi.org/10.48550/arXiv.1710.10903.

(71) Wang, Y.; Wang, J.; Cao, Z.; Farimani, A. B. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.* **2022**, *4* (3), 279–287. https://doi.org/10.1038/s42256-022-00447-x.

(72) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. ChemRxiv March 7, 2023. https://doi.org/10.26434/chemrxiv-2022-jjm0j-v4.

(73) Bai, S.; Zhang, F.; Torr, P. H. S. Hypergraph Convolution and Hypergraph Attention. arXiv October 10, 2020. https://doi.org/10.48550/arXiv.1901.08150.

(74) Chew, A. K.; Afzal, M. A. F.; Kaplan, Z.; Collins, E. M.; Gattani, S.; Misra, M.; Chandrasekaran, A.; Leswing, K.; Halls, M. D. Leveraging High-Throughput Molecular Simulations and Machine Learning for Formulation Design. June 18, 2024. https://doi.org/10.26434/chemrxiv-2024-4lff6.

(75) Zong, Y.; Lin, Y.; Wei, T.; Cheng, Q. Lipid Nanoparticle (LNP) Enables mRNA Delivery for Cancer Therapy. *Adv. Mater.* **2023**, *35* (51), 2303261. https://doi.org/10.1002/adma.202303261.

(76) Kon, E.; Elia, U.; Peer, D. Principles for Designing an Optimal mRNA Lipid Nanoparticle Vaccine. *Curr. Opin. Biotechnol.* **2022**, *73*, 329–336. https://doi.org/10.1016/j.copbio.2021.09.016.

(77) De, A.; Ko, Y. T. Why mRNA-Ionizable LNPs Formulations Are so Short-Lived: Causes and Way-Out. *Expert Opin. Drug Deliv.* **2023**, *20* (2), 175–187. https://doi.org/10.1080/17425247.2023.2162876.

(78) Schoenmaker, L.; Witzigmann, D.; Kulkarni, J. A.; Verbeke, R.; Kersten, G.; Jiskoot, W.; Crommelin, D. J. A. mRNA-Lipid Nanoparticle COVID-19 Vaccines: Structure and Stability. *Int. J. Pharm.* **2021**, *601*, 120586. https://doi.org/10.1016/j.ijpharm.2021.120586.

(79) Patel, A. K.; Kaczmarek, J. C.; Bose, S.; Kauffman, K. J.; Mir, F.; Heartlein, M. W.; DeRosa, F.; Langer, R.; Anderson, D. G. Inhaled Nanoformulated mRNA Polyplexes

for Protein Production in Lung Epithelium. *Adv. Mater.* **2019**, *31* (8), 1805116. https://doi.org/10.1002/adma.201805116.

(80) Rui, Y.; Wilson, D. R.; Tzeng, S. Y.; Yamagata, H. M.; Sudhakar, D.; Conge, M.; Berlinicke, C. A.; Zack, D. J.; Tuesca, A.; Green, J. J. High-Throughput and High-Content Bioassay Enables Tuning of Polyester Nanoparticles for Cellular Uptake, Endosomal Escape, and Systemic in Vivo Delivery of mRNA. *Sci. Adv.* **2022**, *8* (1), eabk2855. https://doi.org/10.1126/sciadv.abk2855.

(81) Anderson, D. G.; Akinc, A.; Hossain, N.; Langer, R. Structure/Property Studies of Polymeric Gene Delivery Using a Library of Poly(β-Amino Esters). *Mol. Ther.* **2005**, *11* (3), 426–434. https://doi.org/10.1016/j.ymthe.2004.11.015.

(82) Capasso Palmiero, U.; Kaczmarek, J. C.; Fenton, O. S.; Anderson, D. G. Poly(Β-amino Ester)- *Co* -poly(Caprolactone) Terpolymers as Nonviral Vectors for mRNA Delivery In Vitro and In Vivo. *Adv. Healthc. Mater.* **2018**, *7* (14), 1800249. https://doi.org/10.1002/adhm.201800249.

(83) Kozielski, K. L.; Ruiz-Valls, A.; Tzeng, S. Y.; Guerrero-Cázares, H.; Rui, Y.; Li, Y.; Vaughan, H. J.; Gionet-Gonzales, M.; Vantucci, C.; Kim, J.; Schiapparelli, P.; Al-Kharboosh, R.; Quiñones-Hinojosa, A.; Green, J. J. Cancer-Selective Nanoparticles for Combinatorial siRNA Delivery to Primary Human GBM in Vitro and in Vivo. *Biomaterials* **2019**, *209*, 79–87. https://doi.org/10.1016/j.biomaterials.2019.04.020.

(84) Shi, J.; Zhang, Y.; Ma, B.; Yong, H.; Che, D.; Pan, C.; He, W.; Zhou, D.; Li, M. Enhancing the Gene Transfection of Poly(β-Amino Ester)/DNA Polyplexes by Modular Manipulation of Amphiphilicity. *ACS Appl. Mater. Interfaces* **2023**, *15* (36), 42130–42138. https://doi.org/10.1021/acsami.3c03802.

(85) Li, Y.; He, Z.; A, S.; Wang, X.; Li, Z.; Johnson, M.; Foley, R.; Sáez, I. L.; Lyu, J.; Wang, W. Artificial Intelligence (AI)-Aided Structure Optimization for Enhanced Gene Delivery: The Effect of the Polymer Component Distribution (PCD). *ACS Appl. Mater. Interfaces* **2023**, *15* (30), 36667–36675. https://doi.org/10.1021/acsami.3c05010.

(86) Eltoukhy, A. A.; Siegwart, D. J.; Alabi, C. A.; Rajan, J. S.; Langer, R.; Anderson, D. G. Effect of Molecular Weight of Amine End-Modified Poly(β-Amino Ester)s on Gene Delivery Efficiency and Toxicity. *Biomaterials* **2012**, *33* (13), 3594–3603. https://doi.org/10.1016/j.biomaterials.2012.01.046.

(87) Lv, H.; Zhang, S.; Wang, B.; Cui, S.; Yan, J. Toxicity of Cationic Lipids and Cationic Polymers in Gene Delivery. *J. Controlled Release* **2006**, *114* (1), 100–109. https://doi.org/10.1016/j.jconrel.2006.04.014.

(88) Wolfert, M. A.; Dash, P. R.; Nazarova, O.; Oupicky, D.; Seymour, L. W.; Smart, S.; Strohalm, J.; Ulbrich, K. Polyelectrolyte Vectors for Gene Delivery: Influence of Cationic Polymer on Biophysical Properties of Complexes Formed with DNA. *Bioconjug. Chem.* **1999**, *10* (6), 993–1004. https://doi.org/10.1021/bc990025r.

(89) Design of Experiments. In *Applied Biostatistics for the Health Sciences*; John Wiley & Sons, Ltd, 2022; pp 508–541. https://doi.org/10.1002/9781119722717.ch11.

(90) Kumar, R. Materiomically Designed Polymeric Vehicles for Nucleic Acids: Quo Vadis? *ACS Appl. Bio Mater.* **2022**, *5* (6), 2507–2535. https://doi.org/10.1021/acsabm.2c00346.

(91) Patel, R. A.; Webb, M. A. Data-Driven Design of Polymer-Based Biomaterials: High-Throughput Simulation, Experimentation, and Machine Learning. *ACS Appl. Bio Mater.* **2024**, *7* (2), 510–527. https://doi.org/10.1021/acsabm.2c00962.

(92) Li, B.; Raji, I. O.; Gordon, A. G. R.; Sun, L.; Raimondo, T. M.; Oladimeji, F. A.; Jiang, A. Y.; Varley, A.; Langer, R. S.; Anderson, D. G. Accelerating Ionizable Lipid Discovery for mRNA Delivery Using Machine Learning and Combinatorial Chemistry. *Nat. Mater.* **2024**, 1–7. https://doi.org/10.1038/s41563-024-01867-3.

(93) Aparna Loecher; Michael Bruyns-Haylett; Pedro J. Ballester; Salvador Borrós; Nuria Oliva. A Machine Learning Approach to Predict Cellular Uptake of pBAE Polyplexes. *Biomater. Sci.* **2023**, *11* (17), 5797–5808. https://doi.org/10.1039/d3bm00741c.

(94) Blakney, A. K.; McKay, P. F.; Ibarzo Yus, B.; Hunter, J. E.; Dex, E. A.; Shattock, R. J. The Skin You Are In: Design-of-Experiments Optimization of Lipid Nanoparticle Self-Amplifying RNA Formulations in Human Skin Explants. *ACS Nano* **2019**, *13* (5), 5920–5930. https://doi.org/10.1021/acsnano.9b01774.

(95) Kumar, R.; Lahann, J. Predictive Model for the Design of Zwitterionic Polymer Brushes: A Statistical Design of Experiments Approach. *ACS Appl. Mater. Interfaces* **2016**, *8* (26), 16595–16603. https://doi.org/10.1021/acsami.6b04370.

(96) Li, Z.; Guo, R.; Zhang, Z.; Yong, H.; Guo, L.; Chen, Z.; Huang, D.; Zhou, D. Enhancing Gene Transfection of Poly(β-Amino Ester)s through Modulation of Amphiphilicity and Chain Sequence. *J. Control. Release Off. J. Control. Release Soc.* **2024**, *368*, 131–139. https://doi.org/10.1016/j.jconrel.2024.02.002.

(97) Nguyen-Vu, V. L.; Pham, M. A.; Huynh, D. P. The Effects of Temperature, Feed Ratio, and Reaction Time on the Properties of Copolymer PLA-PEG-PLA. *Vietnam J. Sci. Technol. Eng.* **2019**, *61* (1), 9–13. https://doi.org/10.31276/VJSTE.61(1).09-13.

(98) Zhang, Q.; Jiang, X.; He, A. Synthesis and Characterization of Trans-1,4-Butadiene/Isoprene Copolymers: Determination of Monomer Reactivity Ratios and Temperature Dependence. *Chin. J. Polym. Sci.* **2014**, *32* (8), 1068–1076. https://doi.org/10.1007/s10118-014-1467-0.

(99) Jin, Y.; Wang, X.; Kromer, A. P. E.; Müller, J. T.; Zimmermann, C.; Xu, Z.; Hartschuh, A.; Adams, F.; Merkel, O. M. Role of Hydrophobic Modification in Spermine-Based Poly(β-Amino Ester)s for siRNA Delivery and Their Spray-Dried Powders for Inhalation and Improved Storage. *Biomacromolecules* **2024**, *25* (7), 4177–4191. https://doi.org/10.1021/acs.biomac.4c00283.

(100) Jin, Y.; Adams, F.; Isert, L.; Baldassi, D.; Merkel, O. M. Spermine-Based Poly(β-Amino Ester)s for siRNA Delivery against Mutated KRAS in Lung Cancer. *Mol. Pharm.* **2023**, *20* (9), 4505–4516. https://doi.org/10.1021/acs.molpharmaceut.3c00206.

(101) Jin, Y.; Adams, F.; Nguyen, A.; Sturm, S.; Carnerio, S.; Müller-Caspary, K.; Merkel, O. M. Synthesis and Application of Spermine-Based Amphiphilic Poly(β-Amino Ester)s for siRNA Delivery. *Nanoscale Adv.* **2023**, *5* (19), 5256–5262. https://doi.org/10.1039/D3NA00272A.

(102) Rampado, R.; Peer, D. Design of Experiments in the Optimization of Nanoparticle-Based Drug Delivery Systems. *J. Controlled Release* **2023**, *358*, 398–419. https://doi.org/10.1016/j.jconrel.2023.05.001.

(103) Bowden, G. D.; Pichler, B. J.; Maurer, A. A Design of Experiments (DoE) Approach Accelerates the Optimization of Copper-Mediated 18F-Fluorination Reactions of Arylstannanes. *Sci. Rep.* **2019**, *9* (1), 11370. https://doi.org/10.1038/s41598-019-47846-6.

(104) Hartl, N.; Adams, F.; Costabile, G.; Isert, L.; Döblinger, M.; Xiao, X.; Liu, R.; Merkel, O. M. The Impact of Nylon-3 Copolymer Composition on the Efficiency of siRNA Delivery to Glioblastoma Cells. *Nanomaterials* **2019**, *9* (7), 986. https://doi.org/10.3390/nano9070986.

(105) Craparo, E. F.; Drago, S. E.; Mauro, N.; Giammona, G.; Cavallaro, G. Design of New Polyaspartamide Copolymers for siRNA Delivery in Antiasthmatic Therapy. *Pharmaceutics* **2020**, *12* (2), 89. https://doi.org/10.3390/pharmaceutics12020089.

(106) Zhang, J.; Cai, X.; Dou, R.; Guo, C.; Tang, J.; Hu, Y.; Chen, H.; Chen, J. Poly(β-Amino Ester)s-Based Nanovehicles: Structural Regulation and Gene Delivery. *Mol. Ther. - Nucleic Acids* **2023**, *32*, 568–581. https://doi.org/10.1016/j.omtn.2023.04.019.

(107) Gogoi, R.; Alam, M. S.; Khandal, R. K.; Gogoi, R. Effect of Reaction Time on the Synthesis and Properties of Isocyanate Terminated Polyurethane Prepolymer. *Int. J. Eng. Res.* **2014**, *3* (5).

(108) Stille, J. K. Step-Growth Polymerization. *J. Chem. Educ.* **1981**, *58* (11), 862. https://doi.org/10.1021/ed058p862.

(109) Zamora, R.; Delgado, R. M.; Hidalgo, F. J. Model Reactions of Acrylamide with Selected Amino Compounds. *J. Agric. Food Chem.* **2010**, *58* (3), 1708–1713. https://doi.org/10.1021/jf903378x.

(110) Carothers, W. H. Polymers and Polyfunctionality. *Trans. Faraday Soc.* **1936**, *32* (0), 39–49. https://doi.org/10.1039/TF9363200039.

(111) Weiss, A. M.; Lopez, M. A.; Rawe, B. W.; Manna, S.; Chen, Q.; Mulder, E. J.; Rowan, S. J.; Esser-Kahn, A. P. Understanding How Cationic Polymers' Properties Inform Toxic or Immunogenic Responses via Parametric Analysis. *Macromolecules* **2023**, *56* (18), 7286–7299. https://doi.org/10.1021/acs.macromol.3c01223.

(112) Tanne, J. H. Covid-19: FDA Approves Pfizer-BioNTech Vaccine in Record Time. *BMJ* **2021**, *374*, n2096. https://doi.org/10.1136/bmj.n2096.

(113) Bowden-Reid, E.; Moles, E.; Kelleher, A.; Ahlenstiel, C. Harnessing Antiviral RNAi Therapeutics for Pandemic Viruses: SARS-CoV-2 and HIV. *Drug Deliv. Transl. Res.* **2025**. https://doi.org/10.1007/s13346-025-01788-x.

(114) Jin, Y.; Adams, F.; Möller, J.; Isert, L.; Zimmermann, C. M.; Keul, D.; Merkel, O. M. Synthesis and Application of Low Molecular Weight PEI-Based Copolymers for siRNA Delivery with Smart Polymer Blends. *Macromol. Biosci.* **2023**, *23* (2), 2200409. https://doi.org/10.1002/mabi.202200409.

(115) Toruntay, C.; Poyraz, F. S.; Susgun, S.; Yucesan, E.; Mansuroglu, B. Anticancer Effects of MAPK6 siRNA-Loaded PLGA Nanoparticles in the Treatment of Breast Cancer. *J. Cell. Mol. Med.* **2025**, *29* (2), e70309. https://doi.org/10.1111/jcmm.70309.

(116) Lahan, M.; Saikia, T.; Dutta, K.; Baishya, R.; Bharali, A.; Baruah, S.; Bharadwaj, R.; Medhi, S.; Sahu, B. P. Multifunctional Approach with LHRH-Mediated PLGA Nanoconjugate for Site-Specific Codelivery of Curcumin and BCL2 siRNA in Mice Lung Cancer. *Future J. Pharm. Sci.* **2024**, *10* (1), 163. https://doi.org/10.1186/s43094-024-00743-w.

(117) Kozielski, K. L.; Tzeng, S. Y.; Green, J. J. A Bioreducible Linear Poly(β-Amino Ester) for siRNA Delivery. *Chem. Commun.* **2013**, *49* (46), 5319. https://doi.org/10.1039/c3cc40718g.

(118) Jogdeo, C. M.; Siddhanta, K.; Das, A.; Ding, L.; Panja, S.; Kumari, N.; Oupický, D. Beyond Lipids: Exploring Advances in Polymeric Gene Delivery in the Lipid Nanoparticles Era. *Adv. Mater.* **2024**, *36* (31), 2404608. https://doi.org/10.1002/adma.202404608.

(119) Shan, X.; Cai, Y.; Zhu, B.; Zhou, L.; Sun, X.; Xu, X.; Yin, Q.; Wang, D.; Li, Y. Rational Strategies for Improving the Efficiency of Design and Discovery of Nanomedicines. *Nat. Commun.* **2024**, *15* (1), 9990. https://doi.org/10.1038/s41467-024-54265-3.

(120) Wan, Q.; Sun, Y.; Sun, X.; Zhou, Z. Rational Design of Polymer-Based mRNA Delivery Systems for Cancer Treatment. *Polym. Chem.* **2024**, *15* (24), 2437–2456. https://doi.org/10.1039/D4PY00206G.

(121) Wu, C.; Li, J.; Wang, W.; Hammond, P. T. Rationally Designed Polycationic Carriers for Potent Polymeric siRNA-Mediated Gene Silencing. *ACS Nano* **2018**, *12* (7), 6504–6514. https://doi.org/10.1021/acsnano.7b08777.

(122) Yan, Y.; Xiong, H.; Zhang, X.; Cheng, Q.; Siegwart, D. J. Systemic mRNA Delivery to the Lungs by Functional Polyester-Based Carriers. *Biomacromolecules* **2017**, *18* (12), 4307–4315. https://doi.org/10.1021/acs.biomac.7b01356.

(123) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57* (8), 1757–1772. https://doi.org/10.1021/acs.jcim.6b00601.

(124) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(125) Lin, Z.; Chou, W.-C.; Cheng, Y.-H.; He, C.; Monteiro-Riviere, N. A.; Riviere, J. E. Predicting Nanoparticle Delivery to Tumors Using Machine Learning and Artificial Intelligence Approaches. *Int. J. Nanomedicine* **2022**, *17*, 1365–1379. https://doi.org/10.2147/IJN.S344208.

(126) McDonald, S. M.; Augustine, E. K.; Lanners, Q.; Rudin, C.; Catherine Brinson, L.; Becker, M. L. Applied Machine Learning as a Driver for Polymeric Biomaterials Design. *Nat. Commun.* **2023**, *14* (1), 4838. https://doi.org/10.1038/s41467-023-40459-8.

(127) Kumar, R.; Le, N.; Tan, Z.; Brown, M. E.; Jiang, S.; Reineke, T. M. Efficient Polymer-Mediated Delivery of Gene-Editing Ribonucleoprotein Payloads through Combinatorial Design, Parallelized Experimentation, and Machine Learning. *ACS Nano* **2020**, *14* (12), 17626–17639. https://doi.org/10.1021/acsnano.0c08549.

(128) Gong, D.; Ben-Akiva, E.; Singh, A.; Yamagata, H.; Est-Witte, S.; Shade, J. K.; Trayanova, N. A.; Green, J. J. Machine Learning Guided Structure Function Predictions Enable *in Silico* Nanoparticle Screening for Polymeric Gene Delivery. *Acta Biomater.* **2022**, *154*, 349–358. https://doi.org/10.1016/j.actbio.2022.09.072.

(129) Kromer, A. P. E.; Sieber-Schäfer, F.; Farfan Benito, J.; Merkel, O. M. Design of Experiments Grants Mechanistic Insights into the Synthesis of Spermine-Containing PBAE Copolymers. *ACS Appl. Mater. Interfaces* **2024**, *16* (29), 37545–37554. https://doi.org/10.1021/acsami.4c06079.

(130) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. arXiv November 25, 2017. https://doi.org/10.48550/arXiv.1705.07874.

(131) *Stability, Intracellular Delivery, and Release of siRNA from Chitosan Nanoparticles Using Different Cross-Linkers | PLOS One*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128963 (accessed 2025-02-28).

(132) Müller, J. T.; Kromer, A. P. E.; Ezaddoustdar, A.; Alexopoulos, I.; Steinegger, K. M.; Porras-Gonzalez, D. L.; Berninghausen, O.; Beckmann, R.; Braubach, P.; Burgstaller, G.; Wygrecka, M.; Merkel, O. M. Nebulization of RNA-Loaded Micelle-Embedded Polyplexes as a Potential Treatment of Idiopathic Pulmonary Fibrosis. *ACS Appl. Mater. Interfaces* **2025**, *17* (8), 11861–11872. https://doi.org/10.1021/acsami.4c21657.

(133) Cirman, T.; Orešić, K.; Mazovec, G. D.; Turk, V.; Reed, J. C.; Myers, R. M.; Salvesen, G. S.; Turk, B. Selective Disruption of Lysosomes in HeLa Cells Triggers Apoptosis Mediated by Cleavage of Bid by Multiple Papain-like Lysosomal Cathepsins *. *J. Biol. Chem.* **2004**, *279* (5), 3578–3587. https://doi.org/10.1074/jbc.M308347200.

(134) Melo, F. R.; Lundequist, A.; Calounova, G.; Wernersson, S.; Pejler, G. Lysosomal Membrane Permeabilization Induces Cell Death in Human Mast Cells. *Scand. J. Immunol.* **2011**, *74* (4), 354–362. https://doi.org/10.1111/j.1365-3083.2011.02589.x.

(135) Sun, F.; Dong, B.; Zhang, H.; Tian, M. Permeability-Controlled Probe for Ratiometric Detection of Plasma Membrane Integrity and Late Apoptosis. *ACS Sens.* **2024**, *9* (11), 6092–6102. https://doi.org/10.1021/acssensors.4c01963.

(136) Bannigan, P.; Bao, Z.; Hickman, R. J.; Aldeghi, M.; Häse, F.; Aspuru-Guzik, A.; Allen, C. Machine Learning Models to Accelerate the Design of Polymeric Long-Acting Injectables. *Nat. Commun.* **2023**, *14* (1), 35. https://doi.org/10.1038/s41467-022-35343-w.

(137) Lu, W.-C.; Ji, X.-B.; Li, M.-J.; Liu, L.; Yue, B.-H.; Zhang, L.-M. Using Support Vector Machine for Materials Design. *Adv. Manuf.* **2013**, *1* (2), 151–159. https://doi.org/10.1007/s40436-013-0025-2.

(138) Kosmas, C. E.; Muñoz Estrella, A.; Sourlas, A.; Silverio, D.; Hilario, E.; Montan, P. D.; Guzman, E. Inclisiran: A New Promising Agent in the Management of Hypercholesterolemia. *Diseases* **2018**, *6* (3), 63. https://doi.org/10.3390/diseases6030063.

(139) Zhang, L.; Wu, T.; Shan, Y.; Li, G.; Ni, X.; Chen, X.; Hu, X.; Lin, L.; Li, Y.; Guan, Y.; Gao, J.; Chen, D.; Zhang, Y.; Pei, Z.; Chen, X. Therapeutic Reversal of Huntington's Disease by in Vivo Self-Assembled siRNAs. *Brain* **2021**, *144* (11), 3421–3435. https://doi.org/10.1093/brain/awab354.

(140) Dahm, R. Friedrich Miescher and the Discovery of DNA. *Dev. Biol.* **2005**, *278* (2), 274–288. https://doi.org/10.1016/j.ydbio.2004.11.028.

(141) Lou, W.; Zhang, L.; Wang, J. Current Status of Nucleic Acid Therapy and Its New Progress in Cancer Treatment. *Int. Immunopharmacol.* **2024**, *142*, 113157. https://doi.org/10.1016/j.intimp.2024.113157.

(142) Omo-Lamai, S.; Wang, Y.; Patel, M. N.; Essien, E.-O.; Shen, M.; Majumdar, A.; Espy, C.; Wu, J.; Channer, B.; Tobin, M.; Murali, S.; Papp, T. E.; Maheshwari, R.; Wang, L.; Chase, L. S.; Zamora, M. E.; Arral, M. L.; Marcos-Contreras, O. A.; Myerson, J. W.; Hunter, C. A.; Tsourkas, A.; Muzykantov, V.; Brodsky, I.; Shin, S.; Whitehead, K. A.; Gaskill, P.; Discher, D.; Parhiz, H.; Brenner, J. S. Lipid Nanoparticle-Associated Inflammation Is Triggered by Sensing of Endosomal Damage: Engineering Endosomal Escape Without Side Effects. bioRxiv April 18, 2024, p 2024.04.16.589801. https://doi.org/10.1101/2024.04.16.589801.

(143) Sharma, P.; Hoorn, D.; Aitha, A.; Breier, D.; Peer, D. The Immunostimulatory Nature of mRNA Lipid Nanoparticles. *Adv. Drug Deliv. Rev.* **2024**, *205*, 115175. https://doi.org/10.1016/j.addr.2023.115175.

(144) Cooper, B. M.; Putnam, D. Polymers for siRNA Delivery: A Critical Assessment of Current Technology Prospects for Clinical Application. *ACS Biomater. Sci. Eng.* **2016**, *2* (11), 1837–1850. https://doi.org/10.1021/acsbiomaterials.6b00363.

(145) Xiao, F.; Chen, Z.; Wei, Z.; Tian, L. Hydrophobic Interaction: A Promising Driving Force for the Biomedical Applications of Nucleic Acids. *Adv. Sci.* **2020**, *7* (16), 2001048. https://doi.org/10.1002/advs.202001048.

(146) Cheng Yan; Guoqiang Li. The Rise of Machine Learning in Polymer Discovery. *Adv. Intell. Syst.* **2023**, *5* (4). https://doi.org/10.1002/aisy.202200243.

(147) Chew, A. K.; Afzal, M. A. F.; Chandrasekaran, A.; Kamps, J. H.; Ramakrishnan, V. Designing the next Generation of Polymers with Machine Learning and Physics-Based Models. *Mach. Learn. Sci. Technol.* **2024**, *5* (4), 045031. https://doi.org/10.1088/2632-2153/ad88d7.

(148) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(149) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30* (8), 595–608. https://doi.org/10.1007/s10822-016-9938-8.

(150) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*; John Wiley & Sons, 2009.

(151) Hastie, T.; Friedman, J.; Tibshirani, R. Basis Expansions and Regularization. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Hastie, T., Friedman, J., Tibshirani, R., Eds.; Springer: New York, NY, 2001; pp 115–163. https://doi.org/10.1007/978-0-387-21606-5_5.

(152) Batista, G. E. A. P. A.; Prati, R. C.; Monard, M. C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor Newsl* **2004**, *6* (1), 20–29. https://doi.org/10.1145/1007730.1007735.

(153) Wu, K.; Yang, X.; Wang, Z.; Li, N.; Zhang, J.; Liu, L. Data-Balanced Transformer for Accelerated Ionizable Lipid Nanoparticles Screening in mRNA Delivery. *Brief. Bioinform.* **2024**, *25* (3), bbae186. https://doi.org/10.1093/bib/bbae186.

(154) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62* (23), 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.

(155) Steinegger, K. M.; Allmendinger, L.; Sturm, S.; Sieber-Schäfer, F.; Kromer, A. P. E.; Müller-Caspary, K.; Winkeljann, B.; Merkel, O. M. Molecular Dynamics Simulations Elucidate the Molecular Organization of Poly(Beta-Amino Ester) Based Polyplexes for siRNA Delivery. *Nano Lett.* **2024**, *24* (49), 15683–15692. https://doi.org/10.1021/acs.nanolett.4c04291.

(156) Bishop, C. J.; Kozielski, K. L.; Green, J. J. Exploring the Role of Polymer Structure on Intracellular Nucleic Acid Delivery via Polymeric Nanoparticles. *J. Controlled Release* **2015**, *219*, 488–499. https://doi.org/10.1016/j.jconrel.2015.09.046.

(157) Itaka, K.; Harada, A.; Yamasaki, Y.; Nakamura, K.; Kawaguchi, H.; Kataoka, K. *In Situ* Single Cell Observation by Fluorescence Resonance Energy Transfer Reveals Fast Intra-cytoplasmic Delivery and Easy Release of Plasmid DNA Complexed with Linear Polyethylenimine. *J. Gene Med.* **2004**, *6* (1), 76–84. https://doi.org/10.1002/jgm.470.

(158) Breunig, M.; Hozsa, C.; Lungwitz, U.; Watanabe, K.; Umeda, I.; Kato, H.; Goepferich, A. Mechanistic Investigation of Poly(Ethylene Imine)-Based siRNA Delivery: Disulfide Bonds Boost Intracellular Release of the Cargo. *J. Controlled Release* **2008**, *130* (1), 57–63. https://doi.org/10.1016/j.jconrel.2008.05.016.

(159) Teo, P. Y.; Yang, C.; Hedrick, J. L.; Engler, A. C.; Coady, D. J.; Ghaem-Maghami, S.; George, A. J. T.; Yang, Y. Y. Hydrophobic Modification of Low Molecular Weight Polyethylenimine for Improved Gene Transfection. *Biomaterials* **2013**, *34* (32), 7971–7979. https://doi.org/10.1016/j.biomaterials.2013.07.005.

(160) Liu, Z.; Zhang, Z.; Zhou, C.; Jiao, Y. Hydrophobic Modifications of Cationic Polymers for Gene Delivery. *Prog. Polym. Sci.* **2010**, *35* (9), 1144–1162. https://doi.org/10.1016/j.progpolymsci.2010.04.007.

(161) Thurston, T. L. M.; Wandel, M. P.; von Muhlinen, N.; Foeglein, Á.; Randow, F. Galectin 8 Targets Damaged Vesicles for Autophagy to Defend Cells against Bacterial Invasion. *Nature* **2012**, *482* (7385), 414–418. https://doi.org/10.1038/nature10744.

(162) Beach, M. A.; Nayanathara, U.; Gao, Y.; Zhang, C.; Xiong, Y.; Wang, Y.; Such, G. K. Polymeric Nanoparticles for Drug Delivery. *Chem. Rev.* **2024**, *124* (9), 5505–5616. https://doi.org/10.1021/acs.chemrev.3c00705.

(163) Wang, Q.; Bu, C.; Dai, Q.; Chen, J.; Zhang, R.; Zheng, X.; Ren, H.; Xin, X.; Li, X. Recent Progress in Nucleic Acid Pulmonary Delivery toward Overcoming Physiological Barriers and Improving Transfection Efficiency. *Adv Sci* **2024**, *11* (18), 2309748. https://doi.org/10.1002/advs.202309748.

(164) Wang, W.; Huang, Z.; Huang, Y.; Zhang, X.; Huang, J.; Cui, Y.; Yue, X.; Ma, C.; Fu, F.; Wang, W.; Wu, C.; Pan, X. Pulmonary Delivery Nanomedicines towards Circumventing Physiological Barriers: Strategies and Characterization Approaches. *Adv. Drug Deliv. Rev.* **2022**, *185*, 114309. https://doi.org/10.1016/j.addr.2022.114309.

(165) Whitsett, J. A.; Alenghat, T. Respiratory Epithelial Cells Orchestrate Pulmonary Innate Immunity. *Nat. Immunol.* **2015**, *16* (1), 27–35. https://doi.org/10.1038/ni.3045.

(166) Ding, L.; Tang, S.; Wyatt, T. A.; Knoell, D. L.; Oupický, D. Pulmonary siRNA Delivery for Lung Disease: Review of Recent Progress and Challenges. *J. Controlled Release* **2021**, *330*, 977–991. https://doi.org/10.1016/j.jconrel.2020.11.005.

(167) Merkel, O. M. Can Pulmonary RNA Delivery Improve Our Pandemic Preparedness? *J. Controlled Release* **2022**, *345*, 549–556. https://doi.org/10.1016/j.jconrel.2022.03.039.

(168) De Souza Carvalho, C.; Daum, N.; Lehr, C.-M. Carrier Interactions with the Biological Barriers of the Lung: Advanced in Vitro Models and Challenges for Pulmonary Drug Delivery. *Adv. Drug Deliv. Rev.* **2014**, *75*, 129–140. https://doi.org/10.1016/j.addr.2014.05.014.

(169) Prasher, P.; Sharma, M.; Singh, S. K.; Gulati, M.; Jha, N. K.; Gupta, P. K.; Gupta, G.; Chellappan, D. K.; Zacconi, F.; De Jesus Andreoli Pinto, T.; Chan, Y.; Liu, G.; Paudel, K. R.; Hansbro, P. M.; George Oliver, B. G.; Dua, K. Targeting Mucus Barrier in Respiratory Diseases by Chemically Modified Advanced Delivery Systems. *Chem. Biol. Interact.* **2022**, *365*, 110048. https://doi.org/10.1016/j.cbi.2022.110048.

(170) Liu, G.; Betts, C.; Cunoosamy, D. M.; Åberg, P. M.; Hornberg, J. J.; Sivars, K. B.; Cohen, T. S. Use of Precision Cut Lung Slices as a Translational Model for the Study of Lung Biology. *Respir. Res.* **2019**, *20* (1), 162. https://doi.org/10.1186/s12931-019-1131-x.

(171) Zimmermann, C. M.; Baldassi, D.; Chan, K.; Adams, N. B. P.; Neumann, A.; Porras-Gonzalez, D. L.; Wei, X.; Kneidinger, N.; Stoleriu, M. G.; Burgstaller, G.; Witzigmann, D.; Luciani, P.; Merkel, O. M. Spray Drying siRNA-Lipid Nanoparticles for Dry Powder Pulmonary Delivery. *J. Controlled Release* **2022**, *351*, 137–150. https://doi.org/10.1016/j.jconrel.2022.09.021.

(172) Merkel, O. M.; Beyerle, A.; Librizzi, D.; Pfestroff, A.; Behr, T. M.; Sproat, B.; Barth, P. J.; Kissel, T. Nonviral siRNA Delivery to the Lung: Investigation of PEG−PEI Polyplexes and Their In Vivo Performance. *Mol. Pharm.* **2009**, *6* (4), 1246–1260. https://doi.org/10.1021/mp900107v.

(173) Targeted Delivery of siRNA to Activated T Cells via Transferrin-Polyethylenimine (Tf-PEI) as a Potential Therapy of Asthma. *J. Controlled Release* **2016**, *229*, 120–129. https://doi.org/10.1016/j.jconrel.2016.03.029.

(174) Chung, E. J.; Kwon, S.; Reedy, J. L.; White, A. O.; Song, J. S.; Hwang, I.; Chung, J. Y.; Ylaya, K.; Hewitt, S. M.; Citrin, D. E. IGF-1 Receptor Signaling Regulates Type II Pneumocyte Senescence and Resulting Macrophage Polarization in Lung Fibrosis. *Int. J. Radiat. Oncol.* **2021**, *110* (2), 526–538. https://doi.org/10.1016/j.ijrobp.2020.12.035.

(175) Zhao, C.; Fang, X.; Wang, D.; Tang, F.; Wang, X. Involvement of Type II Pneumocytes in the Pathogenesis of Chronic Obstructive Pulmonary Disease.

Respir. Med. **2010**, *104* (10), 1391–1395. https://doi.org/10.1016/j.rmed.2010.06.018.

(176) Kim, J.; Minna, J. D. AP-1 Leads the Way in Lung Cancer Transformation. *Dev. Cell* **2022**, *57* (3), 292–294. https://doi.org/10.1016/j.devcel.2022.01.007.

(177) Baldassi, D.; Ambike, S.; Feuerherd, M.; Cheng, C.-C.; Peeler, D. J.; Feldmann, D. P.; Porras-Gonzalez, D. L.; Wei, X.; Keller, L.-A.; Kneidinger, N.; Stoleriu, M. G.; Popp, A.; Burgstaller, G.; Pun, S. H.; Michler, T.; Merkel, O. M. Inhibition of SARS-CoV-2 Replication in the Lung with siRNA/VIPER Polyplexes. *J. Controlled Release* **2022**, *345*, 661–674. https://doi.org/10.1016/j.jconrel.2022.03.051.

(178) Xie, L.; Tan, Y.; Wang, Z.; Liu, H.; Zhang, N.; Zou, C.; Liu, X.; Liu, G.; Lu, J.; Zheng, H. ε-Caprolactone-Modified Polyethylenimine as Efficient Nanocarriers for siRNA Delivery in Vivo. *ACS Appl. Mater. Interfaces* **2016**, *8* (43), 29261–29269. https://doi.org/10.1021/acsami.6b08542.

(179) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gomez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints.

(180) Tzeng, S. Y.; Green, J. J. Subtle Changes to Polymer Structure and Degradation Mechanism Enable Highly Effective Nanoparticles for siRNA and DNA Delivery to Human Brain Cancer. *Adv. Healthc. Mater.* **2013**, *2* (3), 468–480. https://doi.org/10.1002/adhm.201200257.

(181) Yan, Y.; Zhou, K.; Xiong, H.; Miller, J. B.; Motea, E. A.; Boothman, D. A.; Liu, L.; Siegwart, D. J. Aerosol Delivery of Stabilized Polyester-siRNA Nanoparticles to Silence Gene Expression in Orthotopic Lung Tumors. *Biomaterials* **2017**, *118*, 84–93. https://doi.org/10.1016/j.biomaterials.2016.12.001.

(182) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122* (31), 17575–17585. https://doi.org/10.1021/acs.jpcc.8b02913.

(183) Rotolo, L.; Vanover, D.; Bruno, N. C.; Peck, H. E.; Zurla, C.; Murray, J.; Noel, R. K.; O'Farrell, L.; Araínga, M.; Orr-Burks, N.; Joo, J. Y.; Chaves, L. C. S.; Jung, Y.; Beyersdorf, J.; Gumber, S.; Guerrero-Ferreira, R.; Cornejo, S.; Thoresen, M.; Olivier, A. K.; Kuo, K. M.; Gumbart, J. C.; Woolums, A. R.; Villinger, F.; Lafontaine, E. R.; Hogan, R. J.; Finn, M. G.; Santangelo, P. J. Species-Agnostic Polymeric Formulations for Inhalable Messenger RNA Delivery to the Lung. *Nat. Mater.* **2023**, *22* (3), 369–379. https://doi.org/10.1038/s41563-022-01404-0.

(184) Zhu, H.; Zhang, Y.; Li, W.; Huang, N. A Comprehensive Survey of Prospective Structure-Based Virtual Screening for Early Drug Discovery in the Past Fifteen Years. *Int. J. Mol. Sci.* **2022**, *23* (24), 15961. https://doi.org/10.3390/ijms232415961.

(185) Singh, N.; Chaput, L.; Villoutreix, B. O. Virtual Screening Web Servers: Designing Chemical Probes and Drug Candidates in the Cyberspace. *Brief. Bioinform.* **2021**, *22* (2), 1790–1818. https://doi.org/10.1093/bib/bbaa034.

(186) Seifert, M. H. J.; Wolf, K.; Vitt, D. Virtual High-Throughput *in Silico* Screening. *BIOSILICO* **2003**, *1* (4), 143–149. https://doi.org/10.1016/S1478-5382(03)02359-X.

(187) Binder, J.; Winkeljann, J.; Steinegger, K.; Trnovec, L.; Orekhova, D.; Zähringer, J.; Hörner, A.; Fell, V.; Tinnefeld, P.; Winkeljann, B.; Frieß, W.; Merkel, O. M. Closing the Gap between Experiment and Simulation—A Holistic Study on the Complexation of Small Interfering RNAs with Polyethylenimine. *Mol. Pharm.* **2024**, *21* (5), 2163–2175. https://doi.org/10.1021/acs.molpharmaceut.3c00747.

(188) Gao, X. J.; Ciura, K.; Ma, Y.; Mikolajczyk, A.; Jagiello, K.; Wan, Y.; Gao, Y.; Zheng, J.; Zhong, S.; Puzyn, T.; Gao, X. Toward the Integration of Machine Learning and

Molecular Modeling for Designing Drug Delivery Nanocarriers. *Adv. Mater.* **2024**, *36* (45), 2407793. https://doi.org/10.1002/adma.202407793.

(189) Esposito, C.; Wang, S.; Lange, U. E. W.; Oellien, F.; Riniker, S. Combining Machine Learning and Molecular Dynamics to Predict P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2020**, *60* (10), 4730–4749. https://doi.org/10.1021/acs.jcim.0c00525.

(190) Grogan, F.; Holst, M.; Lindblom, L.; Amaro, R. Reliability Assessment for Large-Scale Molecular Dynamics Approximations. *J. Chem. Phys.* **2017**, *147* (23), 234106. https://doi.org/10.1063/1.5009431.

(191) Sieber-Schäfer, F.; Jiang, M.; Kromer, A.; Nguyen, A.; Molbay, M.; Pinto Carneiro, S.; Jürgens, D.; Burgstaller, G.; Popper, B.; Winkeljann, B.; Merkel, O. M. Machine Learning-Enabled Polymer Discovery for Enhanced Pulmonary siRNA Delivery. https://doi.org/10.1002/adfm.202502805.

(192) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z. H.; Ji, C. MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network. *J. Chem. Inf. Model.* **2021**, *61* (7), 3159–3165. https://doi.org/10.1021/acs.jcim.1c00075.

(193) Nigam, A.; Pollice, R.; Krenn, M.; Passos Gomes, G. dos; Aspuru-Guzik, A. Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules Using SELFIES. *Chem. Sci.* **2021**, *12* (20), 7079–7090. https://doi.org/10.1039/D1SC00231G.

(194) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pKaprediction and Protonation State Generation for Drug-like Molecules. *J. Comput. Aided Mol. Des.* **2007**, *21* (12), 681–691. https://doi.org/10.1007/s10822-007-9133-z.

(195) Such, F. P.; Madhavan, V.; Conti, E.; Lehman, J.; Stanley, K. O.; Clune, J. Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning. arXiv April 20, 2018. http://arxiv.org/abs/1712.06567 (accessed 2024-03-11).

(196) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8. https://doi.org/10.1186/1758-2946-1-8.

(197) Coll, J. G.; Trousselier, P.; Pawar, S. D.; Bessin, Y.; Lichon, L.; Chain, J. L.; Sachon, E.; Bettache, N.; Ulrich, S. Amphiphilic Dynamic Covalent Polymer Vectors of siRNA. *Chem. Sci.* **2025**, *16* (5), 2413–2419. https://doi.org/10.1039/D4SC07668K.

(198) Hammond, S. M.; Bernstein, E.; Beach, D.; Hannon, G. J. An RNA-Directed Nuclease Mediates Post-Transcriptional Gene Silencing in Drosophila Cells. *Nature* **2000**, *404* (6775), 293–296. https://doi.org/10.1038/35005107.

(199) Ameres, S. L.; Martinez, J.; Schroeder, R. Molecular Basis for Target RNA Recognition and Cleavage by Human RISC. *Cell* **2007**, *130* (1), 101–112. https://doi.org/10.1016/j.cell.2007.04.037.

(200) Winkeljann, B.; Keul, D. C.; Merkel, O. M. Engineering Poly- and Micelleplexes for Nucleic Acid Delivery – A Reflection on Their Endosomal Escape. *J. Controlled Release* **2023**, *353*, 518–534. https://doi.org/10.1016/j.jconrel.2022.12.008.

(201) Grasso, G.; Deriu, M. A.; Patrulea, V.; Borchard, G.; Möller, M.; Danani, A. Free Energy Landscape of siRNA-Polycation Complexation: Elucidating the Effect of Molecular Geometry, Polymer Flexibility, and Charge Neutralization. *PLOS ONE* **2017**, *12* (10), e0186816. https://doi.org/10.1371/journal.pone.0186816.

(202) Skoraczyński, G.; Kitlas, M.; Miasojedow, B.; Gambin, A. Critical Assessment of Synthetic Accessibility Scores in Computer-Assisted Synthesis Planning. *J. Cheminformatics* **2023**, *15* (1), 6. https://doi.org/10.1186/s13321-023-00678-z.

(203) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. *J. Cheminformatics* **2020**, *12* (1), 35. https://doi.org/10.1186/s13321-020-00439-2.

(204) Wilson, E.; Goswami, J.; Baqui, A. H.; Doreski, P. A.; Perez-Marc, G.; Zaman, K.; Monroy, J.; Duncan, C. J. A.; Ujiie, M.; Rämet, M.; Pérez-Breva, L.; Falsey, A. R.; Walsh, E. E.; Dhar, R.; Wilson, L.; Du, J.; Ghaswalla, P.; Kapoor, A.; Lan, L.; Mehta, S.; Mithani, R.; Panozzo, C. A.; Simorellis, A. K.; Kuter, B. J.; Schödel, F.; Huang, W.; Reuter, C.; Slobod, K.; Stoszek, S. K.; Shaw, C. A.; Miller, J. M.; Das, R.; Chen, G. L. Efficacy and Safety of an mRNA-Based RSV PreF Vaccine in Older Adults. *N. Engl. J. Med.* **2023**, *389* (24), 2233–2244. https://doi.org/10.1056/NEJMoa2307079.

(205) Hashiba, K.; Taguchi, M.; Sakamoto, S.; Otsu, A.; Maeda, Y.; Ebe, H.; Okazaki, A.; Harashima, H.; Sato, Y. Overcoming Thermostability Challenges in mRNA–Lipid Nanoparticle Systems with Piperidine-Based Ionizable Lipids. *Commun. Biol.* **2024**, *7* (1), 556. https://doi.org/10.1038/s42003-024-06235-0.

(206) Omata, D.; Kawahara, E.; Munakata, L.; Tanaka, H.; Akita, H.; Yoshioka, Y.; Suzuki, R. Effect of Anti-PEG Antibody on Immune Response of mRNA-Loaded Lipid Nanoparticles. *Mol. Pharm.* **2024**, *21* (11), 5672–5680. https://doi.org/10.1021/acs.molpharmaceut.4c00628.

(207) Schaffert, D.; Troiber, C.; Salcher, E. E.; Fröhlich, T.; Martin, I.; Badgujar, N.; Dohmen, C.; Edinger, D.; Kläger, R.; Maiwald, G.; Farkasova, K.; Seeber, S.; Jahn-Hofmann, K.; Hadwiger, P.; Wagner, E. Solid-Phase Synthesis of Sequence-Defined T-, i-, and U-Shape Polymers for pDNA and siRNA Delivery. *Angew. Chem. Int. Ed.* **2011**, *50* (38), 8986–8989. https://doi.org/10.1002/anie.201102165.

(208) Lin, Y.; Li, M.; Luo, Z.; Meng, Y.; Zong, Y.; Ren, H.; Yu, X.; Tan, X.; Liu, F.; Wei, T.; Cheng, Q. Tissue-Specific mRNA Delivery and Prime Editing with Peptide–Ionizable Lipid Nanoparticles. *Nat. Mater.* **2025**. https://doi.org/10.1038/s41563-025-02320-9.

(209) Freitag, F.; Wagner, E. Optimizing Synthetic Nucleic Acid and Protein Nanocarriers: The Chemical Evolution Approach. *Adv. Drug Deliv. Rev.* **2021**, *168*, 30–54. https://doi.org/10.1016/j.addr.2020.03.005.

(210) Wagner, E. Have We Finally Found the Ideal Nucleic Acid Carrier with Lipo-Xenopeptides? *Expert Opin. Drug Deliv.* **2025**.

(211) Soares, T. A.; Nunes-Alves, A.; Mazzolari, A.; Ruggiu, F.; Wei, G.-W.; Merz, K. The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. *J. Chem. Inf. Model.* **2022**, *62* (22), 5317–5320. https://doi.org/10.1021/acs.jcim.2c01422.

(212) Schapin, N.; Majewski, M.; Varela-Rial, A.; Arroniz, C.; Fabritiis, G. D. Machine Learning Small Molecule Properties in Drug Discovery. *Artif. Intell. Chem.* **2023**, *1* (2), 100020. https://doi.org/10.1016/j.aichem.2023.100020.

(213) Maharjan, R.; Kim, K. H.; Lee, K.; Han, H.-K.; Jeong, S. H. Machine Learning-Driven Optimization of mRNA-Lipid Nanoparticle Vaccine Quality with XGBoost/Bayesian Method and Ensemble Model Approaches. *J. Pharm. Anal.* **2024**, *14* (11), 100996. https://doi.org/10.1016/j.jpha.2024.100996.

(214) Danishuddin; Khan, A. U. Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design. *Drug Discov. Today* **2016**, *21* (8), 1291–1302. https://doi.org/10.1016/j.drudis.2016.06.013.

(215) Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. In *Handbook of Computational Chemistry*; Leszczynski, J., Kaczmarek-Kedziera, A., Puzyn, T., G. Papadopoulos, M., Reis, H., K. Shukla, M., Eds.; Springer International Publishing: Cham, 2017; pp 2065–2093. https://doi.org/10.1007/978-3-319-27282-5_51.

(216) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2.

(217) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.

(218) Paloncýová, M.; Šrejber, M.; Čechová, P.; Kührová, P.; Zaoral, F.; Otyepka, M. Atomistic Insights into Organization of RNA-Loaded Lipid Nanoparticles. *J. Phys. Chem. B* **2023**, *127* (5), 1158–1166. https://doi.org/10.1021/acs.jpcb.2c07671.

(219) Cornebise, M.; Narayanan, E.; Xia, Y.; Acosta, E.; Ci, L.; Koch, H.; Milton, J.; Sabnis, S.; Salerno, T.; Benenato, K. E. Discovery of a Novel Amino Lipid That Improves Lipid Nanoparticle Performance through Specific Interactions with mRNA. *Adv. Funct. Mater.* **2022**, *32* (8), 2106727. https://doi.org/10.1002/adfm.202106727.

(220) Di Marco, S.; Aupič, J.; Bussi, G.; Magistrato, A. All-Atom Simulations Elucidate the Molecular Mechanism Underlying RNA–Membrane Interactions. *Nano Lett.* **2025**, *25* (11), 4628–4635. https://doi.org/10.1021/acs.nanolett.5c01254.

(221) Singh, A. P.; Prabhu, J.; Vanni, S. RNA Order Regulates Its Interactions with Zwitterionic Lipid Bilayers. *Nano Lett.* **2025**, *25* (1), 77–83. https://doi.org/10.1021/acs.nanolett.4c04153.

(222) Rodríguez-Clemente, I.; Karpus, A.; Buendía, A.; Sztandera, K.; Regulska, E.; Bignon, J.; Caminade, A.-M.; Romero-Nieto, C.; Steinmetz, A.; Mignani, S.; Majoral, J.-P.; Ceña, V. siRNA Interaction and Transfection Properties of Polycationic Phosphorus Dendrimers. *Biomacromolecules* **2025**, *26* (7), 4158–4173. https://doi.org/10.1021/acs.biomac.5c00171.

(223) Čechová, P.; Paloncýová, M.; Šrejber, M.; Otyepka, M. Mechanistic Insights into Interactions between Ionizable Lipid Nanodroplets and Biomembranes. *J. Biomol. Struct. Dyn.* **2024**, *0* (0), 1–11. https://doi.org/10.1080/07391102.2024.2329307.

(224) Carucci, C.; Philipp, J.; Müller, J. A.; Sudarsan, A.; Kostyurina, E.; Blanchet, C. E.; Schwierz, N.; Parsons, D. F.; Salis, A.; Rädler, J. O. Buffer Specificity of Ionizable Lipid Nanoparticle Transfection Efficiency and Bulk Phase Transition. Biophysics January 21, 2025. https://doi.org/10.1101/2025.01.17.633509.

(225) Zhao, Z.; Zhang, H.; Zhuang, X.; Yan, L.; Li, G.; Li, J.; Yan, H. In Silico Insights into the Membrane Disruption Induced by the Protonation of Ionizable Lipids. *J. Mol. Model.* **2025**, *31* (3), 81. https://doi.org/10.1007/s00894-025-06308-9.

(226) Zhao, Y.; Qu, Z.; Paloncýová, M.; Wang, Z.; Weng, B.; Zhao, Y.; Liu, L.; Song, D.; Wich, D.; Otyepka, M.; Xu, Q. Spatial Conformation of Ionizable Lipids Regulates Endosomal Membrane Disruption. *J. Am. Chem. Soc.* **2025**, jacs.5c10908. https://doi.org/10.1021/jacs.5c10908.

(227) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57* (4), 726–741. https://doi.org/10.1021/acs.jcim.6b00778.

(228) Szulc, N. A.; Mackiewicz, Z.; Bujnicki, J. M.; Stefaniak, F. Structural Interaction Fingerprints and Machine Learning for Predicting and Explaining Binding of Small Molecule Ligands to RNA. *Brief. Bioinform.* **2023**, *24* (4), bbad187. https://doi.org/10.1093/bib/bbad187.

(229) Chew, A. K.; Afzal, M. A. F.; Kaplan, Z.; Collins, E. M.; Gattani, S.; Misra, M.; Chandrasekaran, A.; Leswing, K.; Halls, M. D. Leveraging High-Throughput

Molecular Simulations and Machine Learning for the Design of Chemical Mixtures. *Npj Comput. Mater.* **2025**, *11* (1), 72. https://doi.org/10.1038/s41524-025-01552-2.

(230) Chew, A. K.; Sender, M.; Kaplan, Z.; Chandrasekaran, A.; Chief Elk, J.; Browning, A. R.; Kwak, H. S.; Halls, M. D.; Afzal, M. A. F. Advancing Material Property Prediction: Using Physics-Informed Machine Learning Models for Viscosity. *J. Cheminformatics* **2024**, *16* (1), 31. https://doi.org/10.1186/s13321-024-00820-5.

(231) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119* (43), 10509–10524. https://doi.org/10.1021/ja9718937.

(232) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Self-Organizing Neural Networks for Modeling Robust 3D and 4D QSAR: Application to Dihydrofolate Reductase Inhibitors. *Molecules* **2004**, *9* (12), 1148–1159. https://doi.org/10.3390/91201148.

(233) Caldas, G. B.; Ramalho, T. C.; da Cunha, E. F. F. Application of 4D-QSAR Studies to a Series of Benzothiophene Analogs. *J. Mol. Model.* **2014**, *20* (10), 2420. https://doi.org/10.1007/s00894-014-2420-4.

(234) Yavuz, S. C.; Sabanci, N.; Saripinar, E. Pharmacophore Modelling and 4D-QSAR Study of Ruthenium(II) Arene Complexes as Anticancer Agents (Inhibitors) by Electron Conformational- Genetic Algorithm Method. *Curr. Comput. - Aided Drug Des.* **2018**, *14* (1), 79–94. https://doi.org/10.2174/1573409913666170529103206.

(235) Ash, J.; Fourches, D. Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. *J. Chem. Inf. Model.* **2017**, *57* (6), 1286–1299. https://doi.org/10.1021/acs.jcim.7b00048.

(236) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Chief Elk, J.; Jerome, S. V.; Repasky, M. P.; Shelley, J. C. Epik: P K_a and Protonation State Prediction through Machine Learning. *J. Chem. Theory Comput.* **2023**, *19* (8), 2380–2388. https://doi.org/10.1021/acs.jctc.3c00044.

(237) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, Density Functional Theory-Based p K_a Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *J. Chem. Theory Comput.* **2016**, *12* (12), 6001–6019. https://doi.org/10.1021/acs.jctc.6b00805.

(238) Fourches, D.; Ash, J. 4D- Quantitative Structure–Activity Relationship Modeling: Making a Comeback. *Expert Opin. Drug Discov.* **2019**, *14* (12), 1227–1235. https://doi.org/10.1080/17460441.2019.1664467.

(239) Wu, Z.; Ramsundar, B.; N. Feinberg, E.; Gomes, J.; Geniesse, C.; S. Pappu, A.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513–530. https://doi.org/10.1039/C7SC02664A.

(240) Deng, J.; Yang, Z.; Wang, H.; Ojima, I.; Samaras, D.; Wang, F. A Systematic Study of Key Elements Underlying Molecular Property Prediction. *Nat. Commun.* **2023**, *14* (1), 6395. https://doi.org/10.1038/s41467-023-41948-6.

(241) Carrasco, M. J.; Alishetty, S.; Alameh, M.-G.; Said, H.; Wright, L.; Paige, M.; Soliman, O.; Weissman, D.; Cleveland, T. E.; Grishaev, A.; Buschmann, M. D. Ionization and Structural Properties of mRNA Lipid Nanoparticles Influence Expression in Intramuscular and Intravascular Administration. *Commun. Biol.* **2021**, *4* (1), 956. https://doi.org/10.1038/s42003-021-02441-2.

(242) Kovačič, T.; Haas, H.; Stotsky-Oterin, L.; Štrancar, A.; Bren, U.; Peer, D. The Impact of Chemical Reactivity on the Quality and Stability of RNA–LNP Pharmaceuticals. *Nat. Rev. Chem.* **2025**. https://doi.org/10.1038/s41570-025-00763-x.

(243) *RDKit*. https://www.rdkit.org/ (accessed 2025-11-11).

(244) Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Cisneros, G. A.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham, T. E. I.; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz, K. M. Jr. AmberTools. *J. Chem. Inf. Model.* **2023**, *63* (20), 6183–6191. https://doi.org/10.1021/acs.jcim.3c01153.

(245) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718. https://doi.org/10.1002/jcc.20291.

(246) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinforma. Oxf. Engl.* **2013**, *29* (7), 845–854. https://doi.org/10.1093/bioinformatics/btt055.

(247) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(248) *Building Biphasic Systems*. http://www.mdtutorials.com/gmx/biphasic/index.html (accessed 2025-11-11).

(249) Feng, S.; Park, S.; Choi, Y. K.; Im, W. CHARMM-GUI *Membrane Builder*: Past, Current, and Future Developments and Applications. *J. Chem. Theory Comput.* **2023**, *19* (8), 2161–2185. https://doi.org/10.1021/acs.jctc.2c01246.

(250) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(251) Liu, T.; Tian, Y.; Zheng, A.; Cui, C. Design Strategies for and Stability of mRNA–Lipid Nanoparticle COVID-19 Vaccines. *Polymers* **2022**, *14* (19), 4195. https://doi.org/10.3390/polym14194195.

(252) Akinc, A.; Maier, M. A.; Manoharan, M.; Fitzgerald, K.; Jayaraman, M.; Barros, S.; Ansell, S.; Du, X.; Hope, M. J.; Madden, T. D.; Mui, B. L.; Semple, S. C.; Tam, Y. K.; Ciufolini, M.; Witzigmann, D.; Kulkarni, J. A.; van der Meel, R.; Cullis, P. R. The Onpattro Story and the Clinical Translation of Nanomedicines Containing Nucleic Acid-Based Drugs. *Nat. Nanotechnol.* **2019**, *14* (12), 1084–1087. https://doi.org/10.1038/s41565-019-0591-y.

(253) Zhao, Q.; Peng, H.; Ma, Y.; Yuan, H.; Jiang, H. In Vivo Applications and Toxicities of AAV-Based Gene Therapies in Rare Diseases. *Orphanet J. Rare Dis.* **2025**, *20* (1), 368. https://doi.org/10.1186/s13023-025-03893-z.

(254) Xu, S.; Hu, Z.; Song, F.; Xu, Y.; Han, X. Lipid Nanoparticles: Composition, Formulation, and Application. *Mol. Ther. Methods Clin. Dev.* **2025**, *33* (2), 101463. https://doi.org/10.1016/j.omtm.2025.101463.

(255) Shepherd, S. J.; Han, X.; Mukalel, A. J.; El-Mayta, R.; Thatte, A. S.; Wu, J.; Padilla, M. S.; Alameh, M.-G.; Srikumar, N.; Lee, D.; Weissman, D.; Issadore, D.; Mitchell, M. J. Throughput-Scalable Manufacturing of SARS-CoV-2 mRNA Lipid Nanoparticle Vaccines. *Proc. Natl. Acad. Sci.* **2023**, *120* (33), e2303567120. https://doi.org/10.1073/pnas.2303567120.

(256) Xu, Y.; Ma, S.; Cui, H.; Chen, J.; Xu, S.; Wang, K.; Varley, A.; Lu, R. X. Z.; Wang, B.; Li, B. AGILE Platform: A Deep Learning-Powered Approach to Accelerate LNP

Development for mRNA Delivery. bioRxiv June 2, 2023, p 2023.06.01.543345. https://doi.org/10.1101/2023.06.01.543345.

(257) Cui, H.; Xu, Y.; Pang, K.; Li, G.; Gong, F.; Wang, B.; Li, B. LUMI-Lab: A Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. bioRxiv February 16, 2025, p 2025.02.14.638383. https://doi.org/10.1101/2025.02.14.638383.

(258) Han, X.; Zhang, H.; Butowska, K.; Swingle, K. L.; Alameh, M.-G.; Weissman, D.; Mitchell, M. J. An Ionizable Lipid Toolbox for RNA Delivery. *Nat. Commun.* **2021**, *12* (1), 7233. https://doi.org/10.1038/s41467-021-27493-0.

(259) Rademacker, S.; Pinto Carneiro, S.; Molbay, M.; Catapano, F.; Forné, I.; Imhof, A.; Wibel, R.; Heidecke, C.; Hölig, P.; Merkel, O. M. The Impact of Lipid Compositions on siRNA and mRNA Lipid Nanoparticle Performance for Pulmonary Delivery. *Eur. J. Pharm. Sci.* **2025**, *212*, 107182. https://doi.org/10.1016/j.ejps.2025.107182.

(260) Hanari, N.; Mihandoost, S.; Rezvantalab, S. Intelligence Prediction of Microfluidically Prepared Nanoparticles. *Sci. Rep.* **2025**, *15* (1), 37512. https://doi.org/10.1038/s41598-025-21471-y.

(261) Hoseini, B.; Jaafari, M. R.; Golabpour, A.; Momtazi-Borojeni, A. A.; Karimi, M.; Eslami, S. Application of Ensemble Machine Learning Approach to Assess the Factors Affecting Size and Polydispersity Index of Liposomal Nanoparticles. *Sci. Rep.* **2023**, *13* (1), 18012. https://doi.org/10.1038/s41598-023-43689-4.

(262) Hanafy, B. I.; Munson, M. J.; Soundararajan, R.; Pereira, S.; Gallud, A.; Sanaullah, S. M.; Carlesso, G.; Mazza, M. Advancing Cellular-Specific Delivery: Machine Learning Insights into Lipid Nanoparticles Design and Cellular Tropism. *Adv. Healthc. Mater.* **2025**, *14* (18), 2500383. https://doi.org/10.1002/adhm.202500383.

(263) Chan, A.; Kirtane, A. R.; Qu, Q. R.; Huang, X.; Woo, J.; Subramanian, D. A.; Dey, R.; Semalty, R.; Bernstock, J. D.; Ahmed, T.; Honeywell, R.; Hanhurst, C.; Diaz Becdach, I.; Prizant, L. S.; Brown, A. K.; Song, H.; Law Cobb, J.; DeRidder, L. B.; Santos, B.; Jimenez, M.; Sun, M.; Huang, Y.; Byrne, C.; Traverso, G. Designing Lipid Nanoparticles Using a Transformer-Based Neural Network. *Nat. Nanotechnol.* **2025**, *20* (10), 1491–1501. https://doi.org/10.1038/s41565-025-01975-4.

(264) Witten, J.; Raji, I.; Manan, R. S.; Beyer, E.; Bartlett, S.; Tang, Y.; Ebadi, M.; Lei, J.; Nguyen, D.; Oladimeji, F.; Jiang, A. Y.; MacDonald, E.; Hu, Y.; Mughal, H.; Self, A.; Collins, E.; Yan, Z.; Engelhardt, J. F.; Langer, R.; Anderson, D. G. Artificial Intelligence-Guided Design of Lipid Nanoparticles for Pulmonary Gene Therapy. *Nat. Biotechnol.* **2024**. https://doi.org/10.1038/s41587-024-02490-y.

(265) Leong, H. S.; Butler, K. S.; Brinker, C. J.; Azzawi, M.; Conlan, S.; Dufés, C.; Owen, A.; Rannard, S.; Scott, C.; Chen, C.; Dobrovolskaia, M. A.; Kozlov, S. V.; Prina-Mello, A.; Schmid, R.; Wick, P.; Caputo, F.; Boisseau, P.; Crist, R. M.; McNeil, S. E.; Fadeel, B.; Tran, L.; Hansen, S. F.; Hartmann, N. B.; Clausen, L. P. W.; Skjolding, L. M.; Baun, A.; Ågerstrand, M.; Gu, Z.; Lamprou, D. A.; Hoskins, C.; Huang, L.; Song, W.; Cao, H.; Liu, X.; Jandt, K. D.; Jiang, W.; Kim, B. Y. S.; Wheeler, K. E.; Chetwynd, A. J.; Lynch, I.; Moghimi, S. M.; Nel, A.; Xia, T.; Weiss, P. S.; Sarmento, B.; das Neves, J.; Santos, H. A.; Santos, L.; Mitragotri, S.; Little, S.; Peer, D.; Amiji, M. M.; Alonso, M. J.; Petri-Fink, A.; Balog, S.; Lee, A.; Drasler, B.; Rothen-Rutishauser, B.; Wilhelm, S.; Acar, H.; Harrison, R. G.; Mao, C.; Mukherjee, P.; Ramesh, R.; McNally, L. R.; Busatto, S.; Wolfram, J.; Bergese, P.; Ferrari, M.; Fang, R. H.; Zhang, L.; Zheng, J.; Peng, C.; Du, B.; Yu, M.; Charron, D. M.; Zheng, G.; Pastore, C. On the Issue of Transparency and Reproducibility in Nanomedicine. *Nat. Nanotechnol.* **2019**, *14* (7), 629–635. https://doi.org/10.1038/s41565-019-0496-9.

(266) Soneson, C.; Gerster, S.; Delorenzi, M. Batch Effect Confounding Leads to Strong Bias in Performance Estimates Obtained by Cross-Validation. *PLOS ONE* **2014**, *9* (6), e100335. https://doi.org/10.1371/journal.pone.0100335.

(267) Sheshanarayana, R.; You, F. Molecular Representation Learning: Cross-Domain Foundations and Future Frontiers. *Digit. Discov.* **2025**, *4* (9), 2298–2335. https://doi.org/10.1039/D5DD00170F.

(268) Zhang, W.; Deng, L.; Zhang, L.; Wu, D. A Survey on Negative Transfer. *IEEECAA J. Autom. Sin.* **2023**, *10* (2), 305–329. https://doi.org/10.1109/JAS.2022.106004.

(269) Gharoun, H.; Momenifar, F.; Chen, F.; Gandomi, A. H. Meta-Learning Approaches for Few-Shot Learning: A Survey of Recent Advances. *ACM Comput Surv* **2024**, *56* (12), 294:1-294:41. https://doi.org/10.1145/3659943.

(270) Stanley, M.; Bronskill, J. F.; Maziarz, K.; Misztela, H.; Lanini, J.; Segler, M.; Schneider, N.; Brockschmidt, M. FS-Mol: A Few-Shot Learning Dataset of Molecules; 2021.

(271) Ru, X.; Xu, L.; Han, W.; Zou, Q. *In Silico* Methods for Drug-Target Interaction Prediction. *Cell Rep. Methods* **2025**, *5* (10), 101184. https://doi.org/10.1016/j.crmeth.2025.101184.

(272) Qian, X.; Ju, B.; Shen, P.; Yang, K.; Li, L.; Liu, Q. Meta Learning with Attention Based FP-GNNs for Few-Shot Molecular Property Prediction. *ACS Omega* **2024**, *9* (22), 23940–23948. https://doi.org/10.1021/acsomega.4c02147.

(273) Vella, D.; Ebejer, J.-P. Few-Shot Learning for Low-Data Drug Discovery. *J. Chem. Inf. Model.* **2023**, *63* (1), 27–42. https://doi.org/10.1021/acs.jcim.2c00779.

(274) Bellamy, H.; Rehim, A. A.; Orhobor, O. I.; King, R. Batched Bayesian Optimization for Drug Design in Noisy Environments. *J. Chem. Inf. Model.* **2022**, *62* (17), 3970–3981. https://doi.org/10.1021/acs.jcim.2c00602.

(275) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. *Drug Discov. Today* **2015**, *20* (4), 458–465. https://doi.org/10.1016/j.drudis.2014.12.004.

(276) Van Tilborg, D.; Grisoni, F. Traversing Chemical Space with Active Deep Learning: A Computational Framework for Low-Data Drug Discovery. February 23, 2024. https://doi.org/10.26434/chemrxiv-2023-wgl32-v3.

(277) Reker, D.; Schneider, P.; Schneider, G. Multi-Objective Active Machine Learning Rapidly Improves Structure–Activity Models and Reveals New Protein–Protein Interaction Inhibitors. *Chem. Sci.* **2016**, *7* (6), 3919–3927. https://doi.org/10.1039/C5SC04272K.

(278) Loeffler, H. H.; Wan, S.; Klähn, M.; Bhati, A. P.; Coveney, P. V. Optimal Molecular Design: Generative Active Learning Combining REINVENT with Precise Binding Free Energy Ranking Simulations. *J. Chem. Theory Comput.* **2024**, *20* (18), 8308–8328. https://doi.org/10.1021/acs.jctc.4c00576.

Acknowledgements

First, I would like to express my deepest gratitude to my supervisor, Olivia Merkel. Without your continuous support and encouragement throughout the years, none of this would have been possible. I sincerely hope that our future work together will be just as inspiring, motivating, and fruitful as it has been so far. My heartfelt thanks also go to Ben for mentoring me and supporting me across my seemingly endless list of projects. I truly appreciate that you always found the time to help, no matter how many different things I brought to your desk.

My deep gratitude extends to all my co-authors. Adrian, your clarity of thought and execution is admirable. Min, your diligence and constant positivity make working with you an absolute pleasure. Jonas, the project we built together with B2B was one of the most productive and creative experiences of my entire PhD, thank you for being as enthusiastic about bold ideas and new technologies as I am. Nora, your support, organization, and creative input made our collaboration wonderful, and I am genuinely grateful for our work-related and non-work-related conversations that helped me switch off when needed.

Lasse and Leon, thank you for sharing countless moments of fun, laughter, and friendship-both in the synthesis lab and far beyond.

A big thank you goes to my ML bro Fabi, for the daily visits and endless discussions about machine learning and everything else. Special thanks also to everyone in our ML Journal Club.

Thank you, Kathi, for providing such a friendly yet productive working environment, and Katrin, for our long and insightful conversations during AFL1 teaching. You helped me see many things from a different perspective. Thank you, Moritz, for the endless snacks I consumed over the past year. And thank you, Prüßi, for being such a humorous prostdoc and for all your help, including the many car and bike rides. Thanks, Stina(t), for the lovely work chats/rants and the off-work conversations.

To the entire AK Merkel: thank you for creating such a wonderful work (and off-work) environment. I genuinely enjoy coming to the lab every day, and you are a major reason for that.

A special thanks goes to Joschka for the extremely productive time we spent working on the EXIST grant in recent months, especially for taking over tasks when I was stuck in revisions or thesis writing. I hope our future collaboration will be equally successful. I would also like to acknowledge Flo and Min once more, your dedication to the team from day one is truly appreciated.

I thank Prof. Friess for his support and for our many conversations, sometimes highly productive, sometimes less so, but always appreciated. I would like to thank the whole AK Friess. Your team spirit is unique and admirable.

My gratitude also goes to AK Wagner, especially Sophie, for the friendly collaboration. I wish you all the best for your upcoming defenses.

A special thank you to my Master student Johan and my Bachelor student Tim. Your hard work and support made my life significantly easier.

I would like to acknowledge the support of AI-based tools (ChatGPT, Gemini, DeepL) for language refinement and structural assistance during the preparation of this thesis. All responsibility for the scientific content lies solely with me. I also thank BioRender.com for enabling the creation of several figures.

Thank you to my friends outside of university life. Your support, your interest in me as a person, and your understanding of the limited time I had over the past years have meant a great deal to me. I truly value and enjoy the moments we get to spend together.

Finally, I want to thank my family. To my parents: your never-ending love and support made everything possible. Without your belief in me, I would not have achieved any of this. To my brother Julian, thank you for the deep connection and countless conversations we share. To my sister Caro, one of the strongest and funniest women I know. To Luci, who inspires me every day. May your dedication and enthusiasm for the things you love never fade. To my grandparents, cousins, aunts, and uncles: thank you for being the reason our family is such a wonderful place.

And finally, my wife, Janine. You are the most amazing person I have ever met. Your love, kindness, and unwavering support in both good and bad times are the foundation that made all of this possible.

List of Publications

Kromer APE, **Sieber-Schäfer F**, Farfan Benito J, Merkel OM. Design of Experiments Grants Mechanistic Insights into the Synthesis of Spermine-Containing PBAE Copolymers. ACS Appl Mater Interfaces. 2024;16(29):37545–37554. doi: 10.1021/acsami.4c06079.

Steinegger KM, Allmendinger L, Sturm S, **Sieber-Schäfer F**, Kromer APE, Müller-Caspary K, Winkeljann B, Merkel OM. Molecular Dynamics Simulations Elucidate the Molecular Organization of Poly(beta-amino ester) Based Polyplexes for siRNA Delivery. Nano Lett. 2024;24(49):15683–15692. doi: 10.1021/acs.nanolett.4c04291

Sieber-Schäfer F, Jiang M, Kromer APE, Nguyen A, Molbay M, Pinto Carneiro S, Jürgens D, Burgstaller G, Popper B, Winkeljann B, Merkel OM. Machine Learning-Enabled Polymer Discovery for Enhanced Pulmonary siRNA Delivery. Adv Funct Mater. 2025;. doi: 10.1002/adfm.202502805.

Sieber-Schäfer F, Kromer APE, Molbay M, Carneiro S, Jiang M, Nguyen A, Müller J, Farfan Benito J, Merkel OM. Machine Learning on a Small Orthogonal Polymer Library Reveals Functional Insights and Optimizes PBAE Copolymer Synthesis and Performance. SSRN. 2025; doi: 10.2139/ssrn.5320789.

Jiang M, **Sieber-Schäfer F**, Carneiro SP, Matzek D, Nguyen A, Porras-Gonzalez DL, Verma AK, Kolog-Gulko M, Jürgens DC, Burgstaller G, Popper B, Sun X, Merkel OM. A hybrid polymeric system for pulmonary mRNA delivery: Advancing mucosal vaccine development. Cell Biomaterials. 2026;. doi: 10.1016/j.celbio.2025.100311.

Sieber-Schäfer F, Binder J, Münchrath T, Steinegger KM, Jiang M, Winkeljann B, Friess W, Merkel OM. From Bits to Bonds – High-throughput Virtual Screening of RNA Nanocarriers Using a Combinatorial Approach of Machine Learning and Molecular Dynamics. J Am Chem Soc. 2025; doi: 10.1021/jacs.5c12694.