Sampling-Based Approaches for the Computation of Reaction Paths, Stacking Interactions, and IR-Spectra Applied to the Fields of Origins of Life and Enzymatic Catalysis

Beatriz von der Esch

München 2023

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

Sampling-Based Approaches for the Computation of Reaction Paths, Stacking Interactions, and IR-Spectra Applied to the Fields of Origins of Life and Enzymatic Catalysis

Beatriz von der Esch

aus

Asunción, Paraguay

2023

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Prof. Dr. Christian Ochsenfeld betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 28.02.2023

Beatriz von der Esch

Dissertation eingereicht am:	10.03.2023
1. Gutachterin/Gutachter:	Prof. Dr. Christian Ochsenfeld
2. Gutachterin/Gutachter:	Prof. Dr. Regina de Vivie-Riedle
Mündliche Prüfung am:	26.05.2023

Danksagung

"Geteilte Freude ist doppelte Freude!" Ich möchte mich bei allen bedanken, die mich während meiner Promotion begleitet haben, all denjenigen, die mich unterstützt haben, mit denen ich diskutieren konnte, die tausend Mal Korrektur gelesen haben und mich motiviert haben.
An erster Stelle möchte ich mich bei **Prof. Dr. Christian Ochsenfeld** bedanken für die Möglichkeit, meine Doktorarbeit in seinem Arbeitskreis anzufertigen. Ich möchte mich herzlich dafür bedanken, das er mich bestärkt hat, meinen Ideen nachzugehen, mich dazu ermutigt hat, selbständig zu forschen und gleichzeitig immer ein offenes Ohr für mich hatte. Des Weiteren möchte ich mich bei **Prof. Dr. Regina de Vivie-Riedle** für die Anfertigung des Zweitgutachtens bedanken. Ein besonderer Dank gilt meinen Kooperationspartnern **Johannes Dietschreit**, **Laurens Peters**, **Lena Sauerland**, **Alexandra Stan**, **Florian Kruse** und **Prof. Dr. Oliver Trapp** sowie **Avinash Dass** und **Prof. Dr. Dieter Braun**. Zudem möchte ich mich beim gesamten **AK Ochsenfeld** für zahlreiche Anregungen, die lustige Zeit und das kollegiale Arbeitsklima bedanken. Besonders möchte ich mich hier auch bei meinen Kollegen **Katalin Szántó**, **Gökçen Savaşçı**, **Henryk Laqua** und **Andreas Hulm** danken.
Ich möchte mich bei meinen **Freunden** und meiner **Familie** vor allem bei **Brigitta Bachmair**, **Katja Csizi**, **Alexandra** und **Elisabeth von der Esch** für die Unterstützung während des Studiums bedanken. Zu guter Letzt möchte ich mich bei **Daniel Stenzel** für das Verständnis, den Rückhalt und die vielen motivierenden Gespräche während der Promotion bedanken.

Abstract

The characterization of chemical phenomena is at the core of computational chemistry. The goal is to examine, elucidate, and predict different properties and behaviors based on physically accurate descriptions of chemical systems which consequently mimic experimental behavior and provide a deeper understanding of various processes. In quantum-chemical studies, static calculations are often used to characterize molecular properties, reactions, and other chemical transformations. Continuous improvement of computational infrastructure and the development of highly efficient quantum-chemistry programs now allow for studying these processes dynamically rather than statically. Because of the vast configurational space of most chemical systems and the statistical nature of chemistry, dynamic sampling-based approaches can reflect chemical processes more accurately. Therefore, they provide access to new and more exact properties, thereby strengthening the links between theory and experiment.
Here, sampling-based quantum-chemical methods are presented to compute infrared (IR) spectra, investigate nucleotide assemblies, characterize reaction mechanisms and explore chemical reaction space for the discovery of new pathways towards probable precursors for the building blocks of life. Enhanced sampling techniques are applied to the post-translational enzymatic desuccinylation reaction of protein lysine side chains by sirtuin 5 and a prebiotically plausible synthesis of the canonical deoxyribonucleosides. Furthermore, different 3',5'-cyclic ribonucleotide assemblies were investigated using molecular dynamics and examined with regards to their stability and suitability for polymerization.

A key component of all studies was the efficient use and analysis of the vast amount of data generated by sampling. A protocol for preprocessing and quantitative comparison of simulated and measured spectra was introduced, highlighting the superiority of IR-spectra obtained from molecular dynamics simulations. In addition, data-driven techniques have been developed (1) to identify reactive configurations using a machine learning model trained to relate reactant geometries to activation barriers and (2) to build complex reaction networks based solely on the evolution of bond orders during molecular dynamics calculations with induced reactivity. Enhancing the reactivity while ensuring the stability of molecular dynamics runs was achieved by using a newly designed periodic smooth step function in combination with optimized simulation parameters. This optimization, as well as the construction of reaction networks encompassing hundreds of compounds and chemical transformations was enabled by fully automated post-processing.

List of Publications

In the scope of this cumulative dissertation, the following five articles published in peer-reviewed journals will be presented. Author contributions for all publications are provided in the list below.

I **B. von der Esch**, J. C. B. Dietschreit, L. D. M. Peters, C. Ochsenfeld,
"Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5",
J. Chem. Theory Comput., 15, 6660–6667 (2019).
Contribution by the Author: *Project design, implementation, calculations, and shared writing of the manuscript with J. C. B. Dietschreit who actively assisted in all stages of the research project*

II J. C. B. Dietschreit, **B. von der Esch**, C. Ochsenfeld,
"Exponential Averaging Versus Umbrella Sampling for Computing the QM/MM Free Energy Barrier of the Initial Step of the Desuccinylation Reaction Catalyzed by Sirtuin 5"
Phys. Chem. Chem. Phys., 24, 7723-7731 (2022).
Contribution by the Author: *Conjoint concept with J. C. B. Dietschreit, parts of the implementation, half of the calculations and analyses, as well as writing of the manuscript*

III A. Stan, **B. von der Esch**, C. Ochsenfeld,
"Fully Automated Generation of Prebiotically Relevant Reaction Networks from Optimized Nanoreactor Simulations"
J. Chem. Theory Comput., 18, 6700-6712 (2022).
Contribution by the Author: *Conceptual idea, assistance in the implementation, computations, and analysis as well as shared writing of the manuscript*

IV A. V. Dass, S. Wunnava, J. Langlais, **B. von der Esch**, M. Krushe, L.d Ufer, N. Chrisam, R. C. A. Dubini, F. Gartner, S. Angerpointer, C. F. Dirscherl, P. Rovó, C. B. Mast, C. Ochsenfeld, E. Frey, D. Braun.
"RNA Oligomerisation without Added Catalyst from 2',3'-Cyclic Nucleotides by Drying at Air-Water Interfaces"
ChemSystemsChem, 5, e202200026 (2022).
Contribution by the Author: *Participation in designing the research project and writing of the manuscript. All computations of molecular systems and their evaluations*

V **B. von der Esch**, L. Sauerland, L. D. M. Peters, C. Ochsenfeld,
"Quantitative Comparison of Experimental and Computed IR-Spectra Extracted from Ab Initio Molecular Dynamics",
J. Chem. Theory Comput., 17, 985–995 (2021).
Contribution by the Author: *Project design, most of the calculations as well as the analyses and writing of the manuscript*

Publications not related to this thesis:

VI P. Jerabek, **B. von der Esch**, H. Schmidbaur, P. Schwerdtfeger,
"Influence of Relativistic Effects on Bonding Modes in M(II) Dinuclear Complexes (M = Au, Ag, and Cu)"
Inorg. Chem., 56, 14624–14631 (2017).

Contents

Chapter 1

Introduction

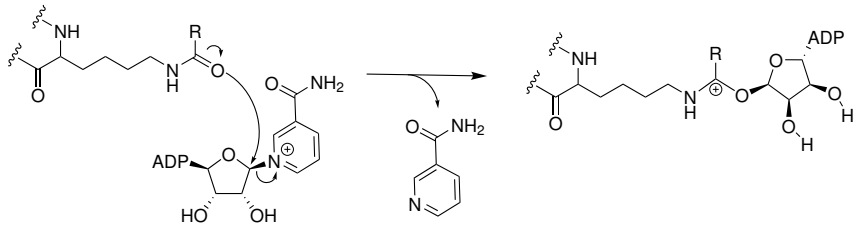
When conceptualizing quantum-chemical studies, a trade-off between description of the electronic structure and the amount of configurations taken into account has to be made. In the past, this meant that studies which both reflect the statistical nature of chemistry, with properties ensuing from thermodynamic ensemble averages, and describing the system accurately were restricted to very small systems comprised of only a few atoms. However, today, computational chemists can grow their repertoire of approaches by harnessing the power of recent advances in the development of efficient low-scaling methods, enhanced sampling protocols, and the rising performance of computational resources. Thereby, allowing for the application of extensive sampling at reasonable level of theory to increasingly large systems.
A given chemical system can, for instance, be sampled using adiabatic ground state *ab initio* molecular dynamics (AIMD). In these simulations the system is propagated on a potential energy surface (PES) determined by the electronic structure. To do so, the motion of the nuclei and electrons can be separated. In Born–Oppenheimer molecular dynamics (BOMD) the electrons are treated quantum mechanically and generate the PES, while the movement of the nuclei is described classically obeying Newton's equations of motion. The resulting trajectories, time-series of molecular geometries, and ensemble averages provide insight into the dynamic behaviour of the system and can be used to extract various properties which can be related to experimental findings.

In the scope of this cumulative dissertation, five peer-reviewed publications and one ongoing project are presented. The compiled studies aim to show the aptitude of sampling based routines for various chemical systems and objectives, bridging gaps between computational and experimental studies. Because ample sampling at *ab initio* level still requires excessive computational resources, enhanced sampling techniques are used and combined with data analysis methods to optimally exploit the generated data.

Publications **I** and **II** highlight the benefits and importance of thorough sampling for the characterization of enzymatic reactions. The large system extent of proteins, their flexibility and implied high dimensional underlying potential energy surfaces pose many

challenges. Among these is the choice of the enzyme-substrate complex starting configuration from which the reaction is simulated.

In the first study (Publication **I**), the dependence of the computed reaction barrier on the starting structure was shown using the initial step of the post-translational desuccinylation reaction of lysine residues catalyzed by SIRT5 (sirtuin 5) as model reaction. SIRT5 belongs to the protein family of sirtuins, categorized as class III lysine deacylases (KDACs).[1] There are seven sirtuin isoforms in mammals. Other than their classification suggests, these also catalyze, e.g., desuccinylations and demyristoylations.[1,2] Sirtuins are located in various cell compartments where they take part in different biological processes. Similar to SIRT3 and SIRT4, SIRT5 is found in mitochondria, it has only weak deacetylase activity and mainly catalyzes demalonylation and desuccinylation reaction of proteins.[3–6] In the catalyzed reaction the removed acyl-group is transferred to the NAD^+ co-substrate leading to the release of nicotinamide and 2'-O-acyl-ADP-ribose.[5] The reaction begins with the cleavage of the glycosidic bond forming an α-1'-O-alkylamidate intermediate (scheme 1.1). The study highlighted that the selection of the starting structure is determinant for



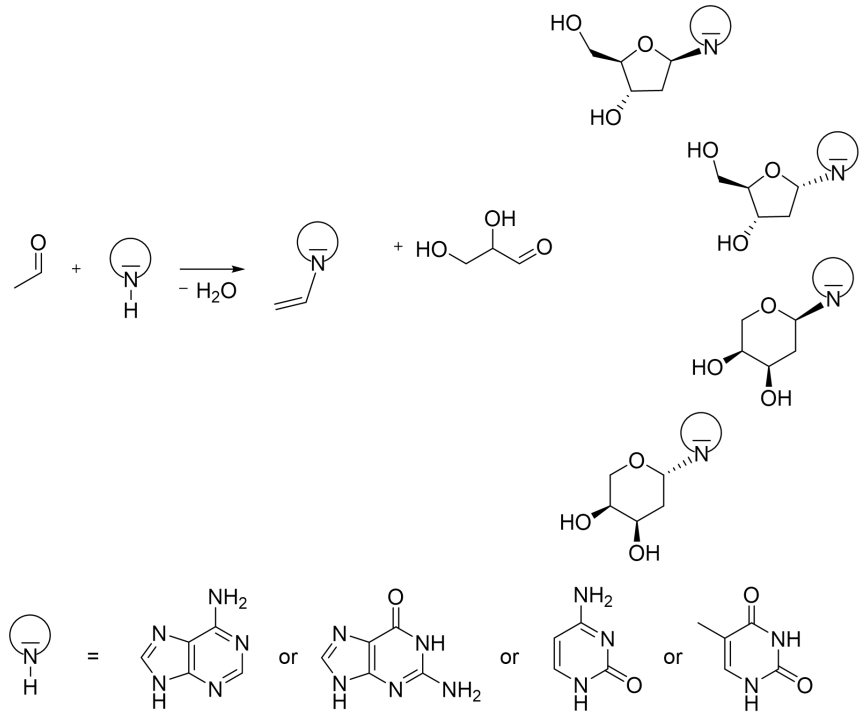
Scheme 1.1: Initial step of the deacylation reaction catalyzed by sirtuins.

the success of the computational characterization of enzymatic reactions. As a solution, we proposed a novel machine learning (ML) based approach to identify a reactive starting conformation and discover distinct structural features that govern the reactivity. The ML model was trained to link geometrical features to activation barriers. The barrier heights were obtained from 150 minimum energy paths started from different snapshots extracted from a molecular mechanics molecular dynamics trajectory of the SIRT5-NAD^+-substrate complex solvated in water. In the reaction path calculations the active center was treated at HF-3c/minix[7] level of theory and the environment described by a classical force field. Using the ML model we were able to find multiple potentially reactive configurations, which were validated by subsequent computation of the respective paths.

In the follow-up study (Publication **II**), we focused on determining the reaction free energy barrier, which can be estimated *via* the exponential average over minimum energy paths or using enhanced sampling methods.[8] We computed the free energy reaction profile and the according transition barrier by QM/MM umbrella sampling[9] simulations evaluated using the Multistate Bennett's Acceptance Ratio.[10] Thereby, we were able to directly compare the 'true' free energy barrier at the chosen level of theory to the transition barriers

calculated from the exponential average of the 150 minimal energy paths computed in the preceding study to build a machine learning model as well as to predict 7501 barriers with the respective model. This comparison, again, underlined the need for extensive sampling and showed that the previously computed barrier from 150 minimum energy paths underestimated the effective free energy barrier. Furthermore, the computational investigation showed that the initial reaction step, the nicotinamide cleavage, is of an S_N2 reaction type and is highly conserved among the sirtuins.

The lessons learned from investigating the initial step of the desuccinylation reaction catalyzed by SIRT5 were applied to the characterization of the stereo- and regioselective synthesis of DNA nucleosides under prebiotic conditions, as proposed by Teichert *et al.*[11] This ongoing project is outlined in chapter 3. In the presented synthesis route, starting from a nucleobase and acetaldehyde, first, the vinylated nucleobase is formed which subsequently reacts with glyceraldehyde to the β-deoxyribonucleoside (scheme 1.2). In the



Scheme 1.2: Chemical pathway towards β-deoxyribonucleosides and its possible side products, α-deoxyribonucleosides and pentopyranosyl-isomers.[11]

preceding study it was shown that exclusively the β-furanose form is obtained.[11] Here, this selectivity was computationally studied using well-tempered metadynamics extended-system adaptive biasing force (WTM-eABF)[12,13] simulations of all reaction steps yielding the reaction free energy profiles.

In publication **III**, the computational sampling of chemical space using ground state *ab-initio* molecular dynamics was used to explore entangled reaction networks. In this study different external potentials are applied to a collection of encircled starting compounds. By choosing periodic potentials the available space is repeatedly reduced and expanded, thereby the probability of collisions between the molecules is increased. In turn, numerous reaction events are observed at reduced time scales. This approach, termed molecular nanoreactor, was pioneered by Wang *et al.*[14] In addition to presenting an alternate implementation of the computational nanoreactor a fully automated evaluation was developed which allows to systematically discuss the influence of various simulation parameters. The post-processing provides a detailed qualitative and quantitative overview of all observed reaction events. Furthermore, in publication **III** the molecular nanoreactor approach is applied to systems of prebiotic interest, a collection of HCN molecules, as well as a mixture of formaldehyde and glycolaldehyde, which are staring compounds for the formose reaction network.[15–18] The goal was to observe key reaction steps towards the formation of pentoses, hexoses, nitrogen-rich heterocycles, and other prebiotically relevant molecules that allow us to verify existing hypotheses as well as potentially discover novel reaction pathways.

Following the formation of the organic molecules constituting the sub-units of the biopolymers found in living organisms, they must self-assemble and polymerize. In publication **IV** the non-enzymatic polymerization of 2',3'-cyclic nucleotides is presented. As a possible arrangement of the monomers, intercalated stacks were computationally investigated as starting point for polymerization. Stacking interaction energies calculated from the energy difference between complex and monomers were complemented by molecular dynamics simulations of the complexes in a water sphere where the compactness of the nucleotide stacks and the time until dissociation were monitored to assess the relative stability.

While publications **I**, **II**, and **III** showed the benefits and importance of thorough sampling for the characterization of reaction paths in publication **V**, we employed sampling to compute infrared (IR) spectra. When extracting IR-spectra from *ab initio* molecular dynamics, a continuous spectrum is obtained which accounts for anharmonic effects as well as possible flexibility of the given compounds.[19] In the presented study, the necessary preprocessing steps to compare spectra are discussed and various quantitative measures are examined to assess the accordance between computed spectra and experimentally recorded spectra. Furthermore, publication **V** shows that, on average, there is a greater similarity between spectra obtained using sampling and measured IR-spectra in comparison to computed spectra relying entirely on the harmonic approximation and a single structure.

Chapter 2

Theoretical Basis

2.1 Calculation of Minimum Energy Paths

Modeling chemical reactions allows to connect theory and experiment, and learn about aspects of these processes that are difficult, or impossible to observe experimentally. We can verify or propose new mechanisms, gain insights into kinetics, compare reactivities, and learn about structural motifs that influence reactivity. There exist several methods to model chemical reactions, some of which are presented in the following subsections. When characterizing a minimum energy path (MEP) we aim to find a continuous pathway on the potential energy surface (PES) connecting the reactant and product state. The direction of the path is given by an intrinsic reaction coordinate. The routines available to find MEPs can be divided into two main categories: (1) propagation methods, that drive the system along a chosen transition coordinate, and (2) interpolation-based approaches, that require at least a reactant and product structure. For simple systems, where a good estimate of the transition state (TS) is possible, the MEP can also be obtained by tracing the reaction coordinate from the TS to the reactant and product states.[20]

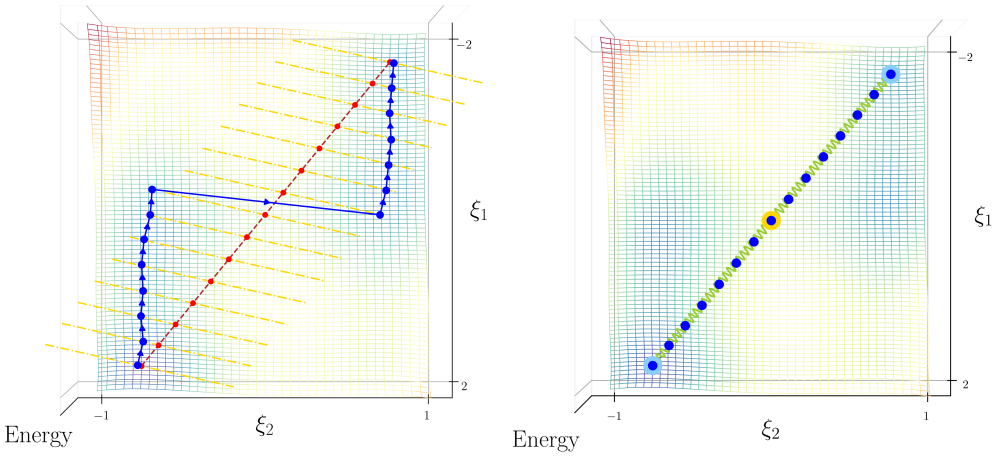
2.1.1 Coordinate-Driven and Chain-of-States Routines

A straightforward approach for computing MEPs is to define an intrinsic coordinate, and perform restrained optimizations in a sequential fashion along the path given by the coordinate. The transition coordinate ξ, which is a function of the system configuration $\mathbf{x}$, effectively maps the high dimensional PES on to lower dimensional representation, $\mathbf{z}$. 2.2.

$$\xi(\mathbf{x}) = \mathbf{z}. \tag{2.1}$$

In principle, this can be any function of phase-space. When computing reaction paths, a lower dimensional representation, e.g., a geometric feature, is chosen aiming to best describe the transition between reactant and product state. Lower dimensional representations, also termed collective variables (CV), to describe chemical transitions are equivalently used in coordinate-driven enhanced sampling methods to compute free energy profiles and surfaces

which will be described in section 2.2.
Starting from either the reactant or product state, the path of slowest ascent is followed by step-wise changing the restraint, given by different values of the transition coordinate, while minimizing all remaining degrees of freedom, resulting in the minimum energy profile by adiabatic mapping when performed on the ground state potential energy surface.[21]
The transition coordinate must represent the change between reactant and product state. However, selecting, as well as defining an optimal reaction coordinate is quite difficult and introduces a strong user-bias. Using adiabatic mapping a poor choice of the reaction coordinate can lead to "hysteresis", resulting in discord of the MEP depending on the simulation direction (increasing and decreasing of the reaction coordinate) and sudden changes of the geometry and energy while only marginally changing the transition coordinate.[20] In Figure 2.1 (a) it is shown for a simple two dimensional example that the diagonal which clearly discerns the two end-states leads to hysteresis as described above. Close to the TS, a large structural rearrangement results from a small increase of the driving variable. For the chosen example this problem can not be solved by a smaller step size. This example highlights, that selecting an appropriate driving coordinate is often difficult, especially for concerted reactions, where several bonds are broken and formed simultaneously. To date, transition coordinates are mainly selected based on chemical intuition and trial and error. However, several methods have been developed to aid the selection of reaction coordinates, in particular for the use in coordinate-driven enhanced sampling methods.[22–25]



(a) Adiabatic Mapping. (b) Nudged Elastic Band Method.

Figure 2.1: Comparison of the coordinate driven adiabatic mapping method (a) and a starting scenario for the nudged elastic band approach (b), belonging to the category of chain-of-states methods. In subfigure (a) the transition coordinate, the linear combination of ξ_1 and ξ_2, is indicated by the red line. As described before the system is optimized while restraining for a series of values of the CV yielding the blue points. In subfigure (b) images are shown in blue, the reactant and product configuration are indicated by a light blue border. The climbing image is marked by a yellow border. The "springs" are drawn in green.

As an alternative to coordinate-driven approaches, the so-called chain-of-states methods can be used to compute MEPs. A widely applied representative of this category is the Nudged Elastic Band (NEB) method.[26,27] To remedy the difficult selection of the reaction coordinate a preliminary path is generated via interpolation between the reactant and the product state (fig. 2.1 (b)). The initial set of images, the collection of intermediate configurations, can also be generated by a previous coordinate-driven calculation. Then a target function, ζ_{NEB}, is defined and minimized, which sums over the energies of all intermediate states M and incorporates harmonic potentials, ensuring that the images remain spaced along the reaction path (eq. 2.2).[26,27]

$$\zeta_{\mathrm{NEB}}(\mathbf{x}_1, ..., \mathbf{x}_M) = \sum_{i=1}^{M} E(\mathbf{x}_i) + \sum_{i=1}^{M-1} \frac{1}{2} k_i (\mathbf{x}_{i+1} - \mathbf{x}_i)^2 \tag{2.2}$$

These penalty terms can be envisioned as elastic bands. The spring constants k may be constant or varied depending on the energy, resulting in an uniform or TS-concentrated spacing of the images. The spring constant must be selected with care, as a too large k may lead to "corner cutting", and a too weak k results in the images "sliding down", thereby not adequately resolving the higher energy region around the TS.[28] In principle, this problem can be solved by incorporating more images, which however leads to increasingly computationally demanding optimization of ζ_{NEB} ($3M_{\mathrm{images}}N_{\mathrm{atoms}}$ variables). To lessen this problem, efficient "nudging" is allowed in the NEB method. This is done by only using the component of the NEB force (F_i^{NEB}) which is parallel to the tangent (τ_i) of the path (F_i^{S}), defined by the difference vector of two neighboring images and only the perpendicular component of the gradient ($F_i^{\perp}$) when minimizing the target function (fig. 2.2).[29–31]

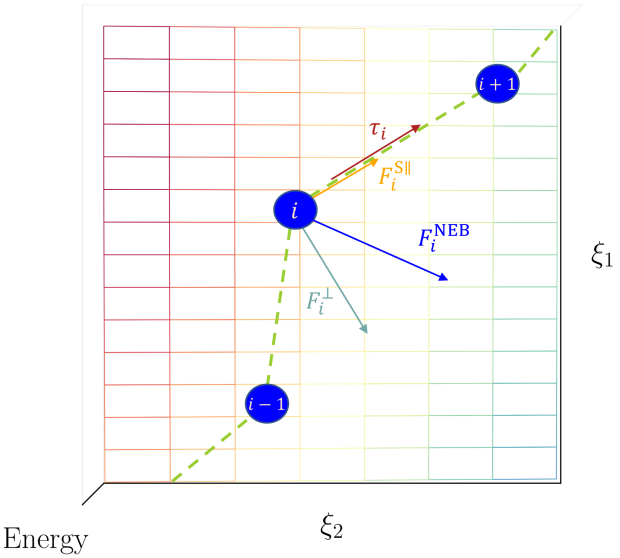


Figure 2.2: Visualisation of components used to allow for nudging. The NEB force F_i^{NEB}, is composed of the component of the spring force $F_i^{\mathrm{S}\parallel}$, parallel to the tangent τ_i, and the perpendicular component of the gradient $F_i^{\perp}$. Adapted from Sheppard *et al.*[30]

In addition, a climbing image may be defined, which is allowed to move more freely along the path, aiming to find the exact TS.[27] In fig. 2.1 (b) the blue markers represent the images determined by linear interpolation, the green springs indicate the penalty potentials, and the climbing image is marked by a yellow border.

2.1.2 Identifying Reactive Configurations by Machine Learning

Both in static and dynamic procedures for modeling chemical reactions, there is a ubiquitous dependency of the result on the chosen starting configuration, which is a stationary point on the PES and defines the orientation of the reacting molecules. Selecting a suitable starting structure as well as overcoming this dependency becomes increasingly demanding with the system extent.
In recent years, machine learning (ML) has gained much popularity for solving a wide variety of problems in computational chemistry. In a review by Keith *et al.* an extensive overview of applications of ML in computational chemistry is given.[32]
In publication **I**, machine learning was used to aid the selection of reactive configurations by investigating the structure-reactivity relationship by linking inter-atomic distances within the active center of sirtuin 5 to the reaction barrier. In machine learning terms the inter-atomic distances are the features and the activation energy is the target. Since the target is a numerical entity, predicting the transition barrier height is a typical regression problem. There are several regression models that are suitable to fit different features to target relationships. The spectrum of algorithms differ in the function estimate, as well as the included error function, the simplest being Linear Regression.[33] Equation 2.3 shows the estimated function used in Linear Regression. In eq. 2.3 and 2.4, x denotes the input feature data, β regression coefficients, and y the output target values. The regression coefficients are optimized so that the error function, here, the residual sum of squares, is minimized for the training data. Therefore, Linear Regression is also termed Least-Squares.

$$\begin{aligned} &\text{Estimated function :} && f_\beta^{\text{linear}}(x_i) = \beta_0 + \beta_1 x_i. \\ &\text{Error function :} && \text{argmin}_\beta \left\{ \sum_i ||y_i - f_\beta^{\text{linear}}(x_i)||^2 \right\}. \end{aligned} \tag{2.3}$$

Polynomial regression fits a polynomial function of k^{th}-order by minimizing the same error function as Least-Squares regression. Alternatively, non-linear dependencies between features and targets can be handled by applying the so-called 'kernel-tick'.[33]
Lasso, Ridge, Elastic Net, Logistic, and Bayesian regression rely on a different error function being minimized. In eq. 2.4, the error functions for Lasso, Ridge and Elastic Net regression are given. In Lasso regression, the L1 norm ($||\beta||_1 = \sum_{j=0}^{k} |\beta_j|$) is used, in Ridge Regression, the L2 norm ($||\beta||_2^2 = \sum_{j=0}^{k} \beta_j^2$). Elastic Net regression includes both the $||\beta||_1$

and $||\beta||_2^2$ terms.

$$\begin{aligned}
&\text{Lasso :} && \sum_{i=0}^{m} ||y_i - f_\beta(x_i)||^2 + \lambda \sum_{j=0}^{k} |\beta_j|, \\
&\text{Ridge :} && \sum_{i=0}^{m} ||y_i - f_\beta(x_i)||^2 + \lambda \sum_{j=0}^{k} \beta_j^2, \\
&\text{Elastic Net :} && \sum_{i=0}^{m} ||y_i - f_\beta(x_i)||^2 + \lambda_1 \sum_{j=0}^{k} |\beta_j| + \lambda_2 \sum_{j=0}^{k} \beta_j^2.
\end{aligned} \tag{2.4}$$

These norms are introduced for regularization, which addresses the problem of over-fitting. This issue arises when small amount of data is provided or the trained model has low bias and high variance.[34] The strength of regularization can be controlled by the hyper-parameter λ or in the case of Elastic Net regression the interplay of λ_1 and λ_2.
In machine learning terms, bias is the deviation between the average model prediction and the underlying ground truth and variance is the variability in the model predictions. When a model has low bias and high variance, it performs well on the training set and produces large errors on the test data, which means that it fails to generalize predictions. When increasing the bias, the variance is reduced automatically. L2 Regularization reduces the feature weights. The L1 regularization term, in addition, introduces sparsity to the weights. When more weights equal zero, the number of significant features is reduced, thereby simplifying the system and suppressing over-fitting.
Assuming expressive features, increasing the sample to feature ratio in general enhances the predictive power of ML models. Therefore, the use of dimensionality reduction routines such as the principal component analysis[35] are often applied when preparing the fitted data aiming to reduce the number of input features while maintaining or optimizing their quality and expressiveness.
For chemical systems, several representations, that can be used for machine learning, are well established such as molecular graphs,[36] Coulomb matrices,[33,37] Bag of Bonds,[38] SMILES,[39,40] and many more.[41–45] Depending on the application in mind, a different representation might be needed. Furthermore, different outputs of quantum chemical calculations may be used as features for machine learning procedures in computational chemistry. In publication **I**, an Elastic Net machine learning model was trained to predict energy barriers from configurations taken from a classical MD simulation based on only 15 inter-atomic distances. These were determined by correlation-based feature reduction. Inter-atomic distances were selected to represent structural changes invariant to translation and rotation.

2.1.3 Importance and Limitations of Minimum Energy Paths

The computation of minimum energy paths (MEP) belongs to the standard repertoire of computational chemistry. MEPs allow to check the plausibility of proposed reaction mechanisms. According to Transition State Theory (TST), the reaction rate, k_{rct}, can be derived from the free energy difference between the transition and reactant state, the

highest point along the MEP dividing the surface into reactant and product state. The geometrical configuration at this point is the transition state structure. The probability of a system taking on a specific configuration is related to a Boltzmann distribution. Therefore, the macroscopic rate constant can be expressed using the Eyring equation.[46]

$$k_{\mathrm{rct}} = \frac{\kappa k_{\mathrm{B}} T}{h} \mathrm{e}^{-\Delta G^{\ddagger}/(RT)}, \tag{2.5}$$

where κ is the transition coefficient, which is usually set to one,[47] k_{B} is the Boltzmann constant, T is the Temperature, R the gas constant, and $\Delta G^{\ddagger}$ the Gibbs free energy difference between transition and reactant state. In terms of enthalpy H and entropy S the Gibbs free energy is given as

$$G = H - TS. \tag{2.6}$$

Therefore, vibrational, rotational, and translational contributions must be determined. Usually, rigid-rotor and harmonic approximations are used to include thermodynamic corrections. Within the TST, re-crossings of the transition-state dividing plane are not allowed when κ is set to one. Thereby, eq. 2.5 provides an upper bound to the true reaction rate. In order to account for re-crossings and tunneling, dynamic effects have to be taken into account to obtain transmission coefficients. However, this is only useful for very simple systems, where the activation energy can be determined with extremely high accuracy.[46]

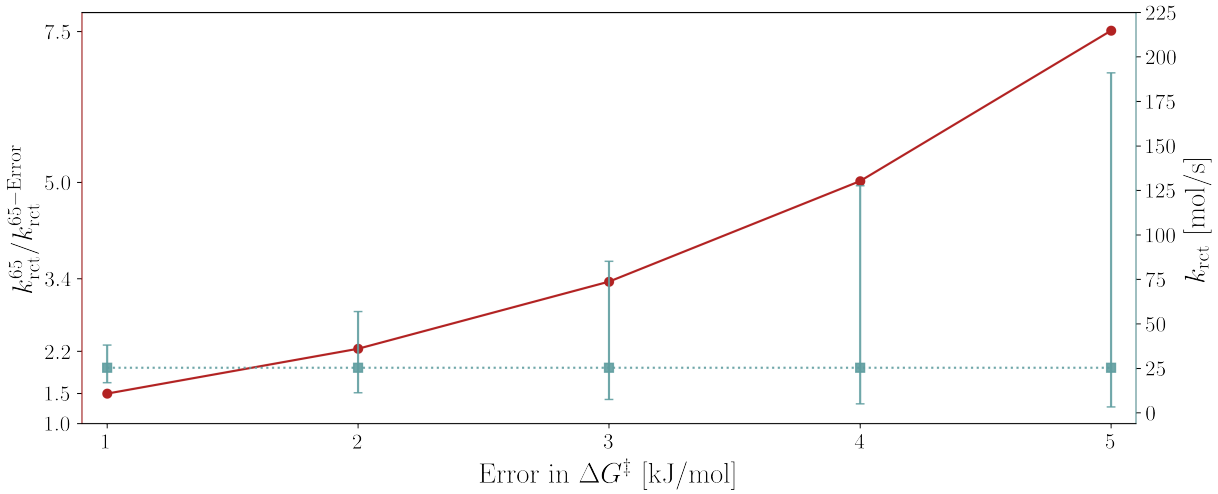


Figure 2.3: Assuming $\Delta G^{\ddagger} = 65$ kJ/mol and T=298.15 K the resulting k_{rct}, calculated according to eq. 2.5, of 25.41 mol/s is indicated by dotted line. The effect of over or underestimating $\Delta G^{\ddagger}$ is indicated by errors-bars for deviations between 1 and 5 kJ/mol. In red the factor of the error is given ($k_{\mathrm{rct}}^{65}/k_{\mathrm{rct}}^{65-\mathrm{Error}}$).

Fig. 2.3 shows how the error in $\Delta G^{\ddagger}$ translates to the reaction rate. For example, an error in $\Delta G^{\ddagger}$ of only 1 kJ/mol results in an error of a factor of 1.5 at T=298.15 K in the reaction rate. An error of 5 kJ/mol leads to an error of factor 7.5, and an error of 10 kJ/mol in an error of factor >50.[46]

Because small errors in $\Delta G^{\ddagger}$ result in large deviations of k_{rct}, interpretation of the obtained reaction rates, which rely on a highly accurate energy difference between reactant and transition state, is often problematic, especially for larger systems where geometry optimizations, as well as the computation of force constants needed for the thermodynamic corrections, become increasingly demanding. This is particularly problematic for biomolecular reactions, where reactant state configurations are often extracted from a previous MD simulation. Although lower energy states have an exponentially higher chance of being visited, the much larger number of high energy configurations diminishes the chances of selecting a structure close to a minimum energy geometry. Furthermore, kinetic trapping of the system in unrepresentative/unreactive regions of phase space, and the possibility of having selected a structure in a local minimum along the reaction path makes the calculation of reliable activation barriers extremely difficult. Another structure-based problem is that non-dynamic methods are unable to reflect large structural or electronic re-configurations which might play a central role in reactions for some complex systems.
In addition, there are limitations with regards to the electronic structure method, QM-region size in QM/MM models,[48–50] and the challenge of finding an optimal reaction coordinate.
The large error in single MEPs for extended systems such as enzymes was addressed by Ryde.[8] Under the assumption that the activation energies have a Gaussian shaped distribution, Ryde discusses how many conformations need to be included in order to diminish structure-based errors in the activation energy depending on the standard deviation of the distribution. Using the exponential average of the collection of energy barriers the free energy barrier can be estimated, which is however ill-conditioned meaning that the number of necessary samples increases more than exponentially for increasing standard deviations. Alternative averaging techniques such as the arithmetic mean converge much faster. However, it was shown that the exponential average provides a better estimate.[51]
In publication **I**, 150 energy barriers, for the initial step of the desuccinalytion reaction of lysine residues catalyzed by sirtuin 5, were computed. For the obtained distribution a standard deviation of 22.9 kJ/mol was obtained, meaning that 10^6 MEPs would be needed in order for the exponential average to provide an estimate of the transition barrier height with an accuracy of 4 kJ/mol.[52] In publication **II**, the estimate resulting from the previously discussed 150 MEPs was compared to the free energy barrier obtained by umbrella sampling[9] (see section 2.2.1) and the exponential average of 7501 barriers predicted by the ML used in publication **I** to find reactive configurations. The comparison showed that the approximation based on 150 MEPs underestimated the free energy barrier.[52,53]
In summary, dynamic methods which take into account a larger number of configurations and inherently give rise to thermodynamic contributions might be necessary to make accurate predictions of reaction rates, where energy differences can no longer be determined with high certainty.

2.2 Calculation of Free Energy Paths

The previously described methods seek to find the MEP on the potential energy surface, then thermodynamic corrections are used to compute free energies of the stationary points. Alternatively, free energy methods may be used to compute the Helmholtz free energy A with respect to a chosen internal coordinate. $A(z)$ is termed free energy surface (FES) or potential mean force (PMF).[54,55]

$$A = U - TS = -\frac{1}{\beta}\ln Q. \tag{2.7}$$

$$Q = \int e^{\beta H(\mathbf{r},\mathbf{p})} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{p}. \tag{2.8}$$

The free energy may be given in terms of the internal energy U, minus the temperature T, multiplied by the entropy S, or using the canonical equipartion function Q (eq. 2.8), where β denotes the inverse of the Boltzmann's constant multiplied by the temperature, $(k_B T)^{-1}$ (eq. 2.7). Because the entropy and the equipartion function are measures of the available phase space of a system, in theory, complete sampling of phase space is needed, which makes the calculation of free energies extremely daunting. The sampling is usually done using molecular dynamics where the probability of a point in phase space being visited is dependent on the energy of the given configuration weighted by the Boltzmann factor,

$$P(E) \propto e^{-\beta E}. \tag{2.9}$$

Because of this relation, the probability distributions $\rho(z)$ may be used to compute free energy differences, which are more accessible than absolute free energies, as well as the FES:

$$\Delta A_{1\to 2} = A_1 - A_0 = -\frac{1}{\beta}\ln\frac{Q_1}{Q_0} = -\frac{1}{\beta}\ln\frac{\rho_1}{\rho_0}, \tag{2.10}$$

$$A(z) = -\frac{1}{\beta}\ln\rho(z) = -\frac{1}{\beta}\ln\int \delta(\xi(\mathbf{x}) - z) e^{-\beta U(\mathbf{x})} \mathrm{d}\mathbf{x}. \tag{2.11}$$

Ample sampling along the entire coordinate ξ is a precondition for computing free energy profiles according to eq. 2.11. However, because of the exponentially decreasing probability with rising energy of visiting a certain state according to the relation eq. 2.9, continuous sampling is non-trivial. To do so, importance-sampling techniques, based on the definition of a collective variable, have been developed to enable efficient sampling of the relevant regions of phase space by encouraging the system to overcome meta-stable states.[55]

2.2.1 Umbrella sampling

The main issue of computing reaction paths *via* dynamics is the large amount of sampling needed especially to ensure adequate sampling of higher energy regions such as transition barriers. As the number of degrees of freedom rises with system size the problem of insufficient sampling becomes increasingly difficult to overcome by sheer computational power. Several enhanced sampling techniques have been developed to overcome this issue. Torrie and Valleau introduced the umbrella sampling method.[9] Here, the reaction path is divided into several windows which are sampled by independent molecular dynamics simulations which can be performed in parallel.
In each umbrella window i a biasing potential ω is applied which alters the true PES U^{u} and thereby constrains the system to successive regions in phase space along the reaction coordinate $\xi(\mathbf{x})$.

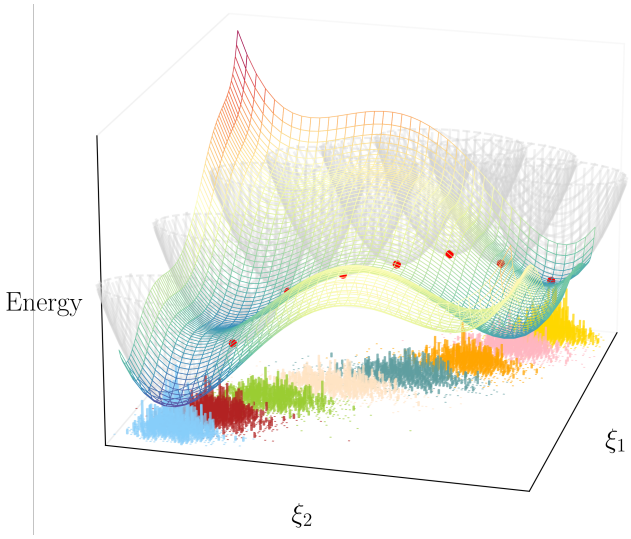


Figure 2.4: In umbrella sampling exploration of a desired area is ensured by a sequence of biasing potentials employed in independent simulations each centered in one of the windows, i, connecting the two minimum energy regions. From the biased sampling, biased distributions $\rho_i^{\mathrm{b}}(z)$ are obtained. By post-processing, the unbiased FES, $A^{\mathrm{u}}(z)$ can be recovered.

In principle any functional form can be chosen for the biasing potential, the simplest and most commonly chosen function is a harmonic potential (eq. 2.13), where k_i is the force constant which determines the strength of the constraining potential and therefore has to be chosen carefully to ensure that the system remains close to the desired region while allowing for meaningful sampling.

$$U^{\mathrm{b}} = U^{\mathrm{u}} + \omega_i. \tag{2.12}$$

$$\omega_i(z) = \frac{1}{2} k_i (\xi(\mathbf{x}) - z_i)^2. \tag{2.13}$$

Furthermore, overlap between the windows has to be achieved. The quality of the sampling is dependent on the window placement and the strength of the force constants and therefore benefits from prior knowledge of the free-energy landscape at hand. Fig. 2.4 illustrates the described approach. On a model potential, the space between the two minimum energy states is sampled continuously by introducing harmonic biasing potentials.

The free energy A along the reaction coordinate then has to be recovered from the biased distributions $\rho_i^{\mathrm{b}}(z)$ resulting from the series of biased molecular dynamics simulations (eq. 2.14).

$$\rho_i^{\mathrm{b}}(z) = \frac{\int \delta(\xi(\mathbf{x}) - z) e^{-\beta(U_0(\mathbf{x}) + \omega_i(\xi(\mathbf{x})))} \mathrm{d}\mathbf{x}}{\int e^{-\beta(U_0(\mathbf{x}) + \omega_i(\xi(\mathbf{x})))} \mathrm{d}\mathbf{x}}. \tag{2.14}$$

To do so, the biases have to be removed and the factors F_i have to be determined in order to align and recombine the separate simulations (eq. 2.15).

$$A_i^{\mathrm{u}}(z) = -\frac{1}{\beta} \ln \rho_i^{\mathrm{b}}(z) - \omega_i(z) + F_i. \tag{2.15}$$

2.2.2 Adaptive Biasing Force

There are several alternative methods to umbrella sampling,[9] such as Targeted Molecular Dynamics,[56] Metadynamics (MtD),[57] and the Adaptive Biasing Force (ABF) Method.[58] In the latter, a biasing force is constructed throughout a dynamics simulation. The biasing force is calculated using a local running average so that it cancels the free energy force along a chosen CV ($z = \xi(\mathbf{x})$). Thereby, the system is able to escape kinetic traps and eventually free to move along the predefined CV. Unhindered diffusion indicates convergence. Subsequently, the free energy profile can be determined from the biasing force. Other than for umbrella sampling, no prior knowledge of the free energy landscape is needed, which is one of its many advantages.[55,59] The estimate is calculated according to eq. 2.16, where |J| is the determinant of the Jacobian needed for the transformation from generalized to cartesian coordinates.

$$\frac{\partial A(z)}{\partial z} = -\langle F \rangle_z = \left\langle \frac{\partial U(z)}{\partial z} \right\rangle_z - \left\langle \beta^{-1} \frac{\partial \ln |\mathbf{J}|}{\partial z} \right\rangle_z. \tag{2.16}$$

$\overline{F}(z)$ is used to approximate the z-conditioned ensemble average $\langle F \rangle_z$, where N_{s} denotes the number of simulations steps.

$$\langle F \rangle_z \approx \overline{F}(z) = \frac{1}{N_{\mathrm{s}}} \sum_{\mu=1}^{N_{\mathrm{s}}} F_\mu \tag{2.17}$$

ABF was introduced by Darve *et al.*[58] and since then many further developments have been made. To make the ABF more easily applicable, the extended-system ABF (eABF) was proposed by Lesage *et al.*, where a fictitious particle is introduced which is coupled to the CV by a harmonic restraint, alleviating some of the initial limitations for the CV resulting

from the Jacobian term (eq. 2.16).[60] In eABF the biasing force acts on the fictitious particle (λ) instead of the physical system. The extended potential, incorporating the fictitious particle is defined as

$$U_{\text{ext}}(\mathbf{x}, \lambda, t) = U(\mathbf{x}) + \frac{1}{2\beta\sigma^2}(\xi(\mathbf{x}) - \lambda)^2 + U_b(\lambda, t), \tag{2.18}$$

where σ is the thermal coupling width. Tight coupling ensures that efficient sampling of λ translates to efficient sampling along the CV. Therefore, a biasing potential U_b can be chosen that only affects the dynamics of the non-physical particle directly. The eABF method was further enhanced by combining it with metadynamics (eq. 2.19) and its well-tempered variant, yielding the WTM-eABF method.[12,13].

$$F_{\text{meta-eABF}}(\lambda) = F_{\text{eABF}}(\lambda) + F_{\text{MtD}}(\lambda). \tag{2.19}$$

In well-tempered metadynamics periodically repulsive gaussian kernels are placed at the current location along the CV, resulting in an adaptive external biasing potential equalizing the underlying free energy surface.[61] The WTM-eABF method, thereby significantly accelerates the convergence by "Shaving Barriers, and Flooding Valleys".[12,13]
In this work, WTM-eABF was applied to compute the formation of deoxyribonucleosides from acetaldehyde, glyceraldehyde and the respective canonical nucleobases and explore the regio- and stereo-selectivity of the proposed synthesis route. The project is summarized in chapter 3.

2.2.3 Multistate Bennett's Acceptance Ratio

Umbrella sampling and the variations of ABF have in common that they alter the underlying free energy surface and thereby enable sampling of higher energy regions that would otherwise only be visited infrequently. As these procedures produce biased distributions, evaluation routines are required to recover the original weights of each sample in order to determine unbiased free energy surfaces and ensemble averages. This can be done using the Multistate Bennett's Acceptance Ratio (MBAR),[10] which originates from the Bennett's Acceptance Ratio (BAR)[62] introduced in 1976.
The free energy in each window, A_i, can be self-consistently calculated using the MBAR equation:

$$e^{-\beta A_i} = \sum_{j=1}^{S}\sum_{n=1}^{N_j} \frac{e^{-\beta\omega_i(\mathbf{x}_{jn})}}{\sum_{k=1}^{K} N_k e^{\beta A_k - \omega_k(\mathbf{x}_{jn})}}. \tag{2.20}$$

S denotes the total number of windows, N_j is the number of samples in window j and ω_i is the value of the biasing potential for the i^{th} frame in window j. The unbiased FES $A^u(z)$ can then be obtained using

$$\begin{aligned} A^u(z) &= -\beta^{-1}\ln\, \rho_0(z) \\ &= -\beta^{-1}\ln \sum_{j=1}^{S}\sum_{n=1}^{N_j} \frac{\delta\left(\xi(j,n) - z\right)}{\sum_l^S N_l\ e^{\beta A_l - \beta\omega_l(j,n)}}. \end{aligned} \tag{2.21}$$

2.3 The Computational Nanoreactor

While the previous sections discussed aspects of the characterization of specific reaction mechanisms, the molecular nanoreactor approach pioneered by Wang *et al.*[14] in 2014 seeks to automatically discover novel reactions computationally from a feed-stock of starting molecules. In the proposed routine, high-temperature *ab initio* MD simulations are conducted, while the molecules are constrained to a periodically contracting sphere. Both the elevated temperature and the compression of the available space are employed to increase the probability of reaction events and thereby decrease the required timescale to observe several chemical transformations.[14,63–66]
The automated exploration of reaction space has a rich history, first attempts were made as early as 1994 by Broadbelt *et al.*[36] Because the performance of extensive *ab initio* sampling was still prohibitively expensive as computational hardware and quantum chemistry code were less developed, these early routines employed many heuristic rules in order to construct chemical reaction networks.[36,67–71] Today, rule-based routines still play a major role, especially in the discovery of novel pharmaceutical candidates. However, heuristic models are not as general as the proposed nanoreactor approach, and less exotic reactions are expected as they are based on rules derived from already established chemistry.[72] The nanoreactor approach was also taken up by other groups such as Grimme and co-workers, who have introduced reactivity to their simulations by employing meta-dynamics using the RMSD as collective variable driving the system into new regions of chemical space, while ensuring that reactants remain within a capsule by applying a constant wall potential.[73,74]
In the initial nanoreactor concept, a modified Heaviside step function is applied to switch between two predefined radii. On atoms that exceed the set radius r_0 a mass-weighted harmonic potential U, is applied, pushing these towards the sphere center (eq. 2.22).[14]

$$\begin{aligned} V^{\mathrm{RW}}(r,t) &= f(t)U(r,r_{\max},k_{\max}) + (1-f(t))U(r,r_{\min},k_{\min}), \\ U(r,r_0,k) &= \frac{mk}{2}(r-r_0)^2\theta(r-r_0),\ \ f(t) = \theta\left(\left\lfloor\frac{t}{t_{\mathrm{total}}}\right\rfloor - \frac{t}{t_{\mathrm{total}}} + \frac{t_{\exp}}{t_{\mathrm{total}}}\right). \end{aligned} \tag{2.22}$$

This approach leads to abrupt accelerations. A less aggressive alternative is the application of a cosine function in order to smoothly switch between an extended and contracted sphere (V^{CW}, eq. 2.23).

$$\begin{aligned} V^{\mathrm{CW}}(r,r_0(t),k) &= \frac{mk}{2}\left[\max\left(0,r-r_0(t)\right)\right]^2, \\ r_0(t) &= r_{\min} + \frac{r_{\max}-r_{\min}}{2}\left[1+\cos\left(\frac{t}{t_{\mathrm{total}}}2\pi\right)\right]. \end{aligned} \tag{2.23}$$

However, this leads to a reduced time in the compressed state, in which the reactions are initiated, as well as in the expanded phase, necessary for the system to relax and stable species to form.
An alternative function was proposed in publication **III**, which enables smoother transitions to and from the minimal radius, where harsh accelerations affect the simulation the

most while allowing for the system to remain in the contracted and extent phase for more time-steps.

$$\begin{aligned} V^{\mathrm{SC}}(r, r_0(t), k) &= \frac{mk}{2} \left[\max\left(0, r - r_0(t)\right)\right]^2, \\ r_0(t) &= \min\left[r_{\max} + (r_{\max} - r_{\min}) \sin\left(\frac{\pi}{2} \cos\left(\frac{t}{t_{\mathrm{total}}} 2\pi\right)\right), r_{\max}\right]. \end{aligned} \tag{2.24}$$

The previously proposed rectangular wave function (V^{RW}, eq. 2.22), the smooth cosine wave (V^{CW}, eq. 2.23), and limited triangular function (V^{SC}, eq. 2.24) are compared in fig. 2.5.

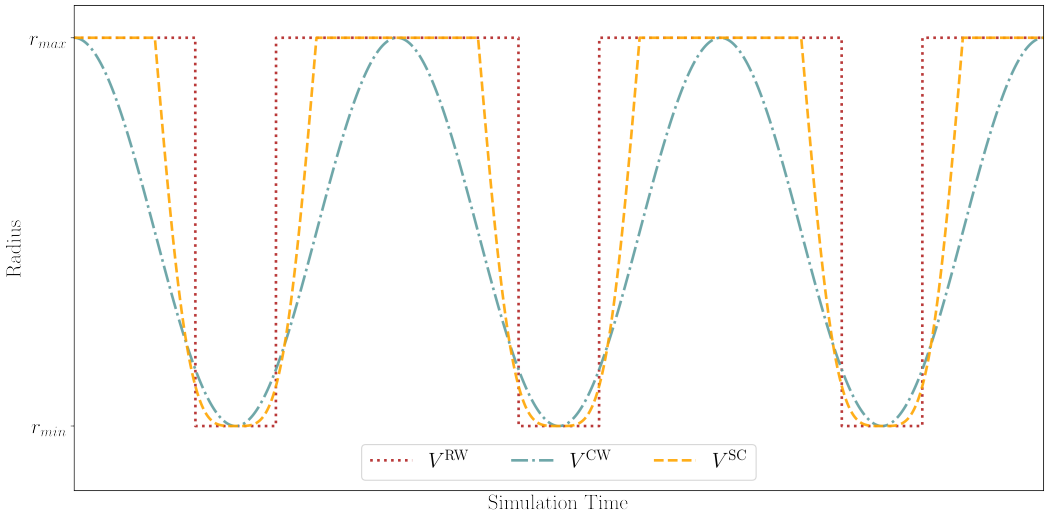


Figure 2.5: Collection of boundary potentials used to periodically decrease the space available to the molecular system in a nanoreactor simulation. Compared are V^{RW}, V^{CW}, and V^{SC} as defined in eq. 2.22, eq. 2.23, and eq. 2.24, respectively.

Manual evaluation of the events in nanoreactor simulations is very tedious. To automate the identification of reaction events and novel species, molecules have to be effectively recognized. This was previously done based on inter-atomic distances.[14,65,66,75] A more general approach could be the use of Wiberg bond orders W_{AB},[76] which can be obtained with minimal additional computational cost from the density matrix P,

$$W_{AB} = \sum_{\mu \in A} \sum_{\nu \in B} P_{\mu\nu}^2. \tag{2.25}$$

From the Wiberg bond orders connectivity matrices are obtained which can be used to derive the molecular species. Using the Python library RDKit,[77] the molecules can be transformed to MOL-objects, which can be used to visualize these, generate respective SMILES[39,40] and perform further qualitative and quantitative analyses as showcased in publication **III**.

2.4 Intermolecular Interactions of Nucleotides

For nucleotides to polymerize to RNA or DNA, these first have to form suitable aggregates. Intermolecular interactions such as π-stacking and hydrogen bridges dictate the pre-organization, and are therefore central in the self-polymerization step necessary for life to emerge.
Intrinsic interaction energies characterize the direct forces between sub-units and are governed by short-range exchange repulsion, dispersion attraction, and electrostatic forces. Besides the direct forces between the monomers, thermodynamic effects play a major role in intermolecular interactions.[78]
The computed energies heavily depend on the molecular structure. For complex systems with many degrees of freedom finding a representative minimum energy structure is non-trivial, therefore a major bias is introduced by the selected configuration, which cannot be alleviated reliably by geometry optimization.

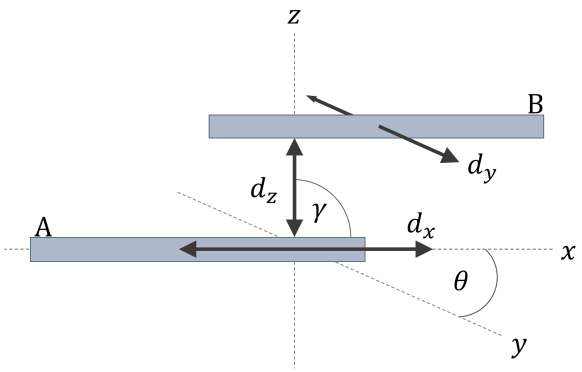


Figure 2.6: Possible displacements of two subunits A and B. Several variables may change simultaneously which results in an extensive collection of possible aggregates as well as a large number of local and relevant minima.

In previous studies, scans were conducted for several possible displacements.[79] However, for systems consisting of several subunits the number of scans required to fully characterize all possible assemblies quickly becomes unfeasible (fig. 2.6). Alternatively, the relative stability of nucleotide assemblies was studied using molecular dynamics simulation. Based on the simple idea that stronger interactions lead to more stable aggregates, the time a system remains in an assembly can be assessed to compare inter-molecular interactions. In addition, this approach allows for straight-forward incorporation of explicit solvent, which is challenging in static studies as the placement of solvent-molecules may influence results drastically. In publication **IV** this dynamic approach was chosen to compare the stability of various homogeneous and heterogeneous intercalated stacked tetramers.

2.5 Computation and Quantitative Comparison of IR-Spectra

Computation of spectroscopic data is an excellent link to experiments and to reassure the veracity of a proposed hypothesis as well as assess the quality of quantum chemical calculations. The interaction of matter and radiation can be described theoretically using different methods, e.g., by a perturbative approach. Many (spectroscopic) properties can be derived from a Taylor series of the perturbed energy $E(\xi)$ with respect to the unperturbed ground state energy $E(0)$ describing the electronic response.
The perturbation ξ can be manifold, time-independent or dependent, e.g., a static external electric field, a magnetic field, changes in nuclear spins, or an electromagnetic wave. In addition, perturbations may be combined, giving access to further properties. If ξ is a change in molecular geometry $\mathbf{R}$ the following Taylor expansion is obtained:

$$E(\mathbf{R}) = E(\mathbf{R_0}) + \frac{\partial E}{\partial \mathbf{R}}(\mathbf{R}-\mathbf{R_0}) + \frac{1}{2}\frac{\partial^2 E}{\partial \mathbf{R}^2}(\mathbf{R}-\mathbf{R_0})^2 + \frac{1}{6}\frac{\partial^3 E}{\partial \mathbf{R}^3}(\mathbf{R}-\mathbf{R_0})^3 + ... \tag{2.26}$$

The first derivative of the energy with respect to the nuclear coordinates is the gradient. The second derivative is the Hessian $\mathbf{H}$ and contains the force constants. The higher order terms give access to anharmonic effects such as e.g., Fermi resonance. At a minimum energy geometry and under the assumption that the potential energy surface at such a stationary point can be approximated by a harmonic potential, harmonic vibrational frequencies can be determined based on the Hessian. To do so, the force constant matrix is mass weighted and diagonalized. The resulting eigenvalues ϵ_k are related to the vibrational normal modes $\tilde{\nu}_k$ as given in eq. 2.27.

$$\tilde{\nu}_k = \frac{1}{2\pi c}\sqrt{\epsilon_k}, \tag{2.27}$$

The respective IR intensities are proportional to the change of the dipole moment along the according eigenvector. Similarly, Raman intensities may be obtained from the derivative of the polarizability with respect to the normal mode vector.
Alternatively, to this established computation of harmonic frequencies, the IR-spectrum can be extracted from *ab initio* molecular dynamics simulation by computing the Fourier transform of the auto-correlation function of the time derivative of the dipole moment $\boldsymbol{\mu}$,[19,80]

$$I_{\mathrm{IR}}(\omega) \propto \int \langle \dot{\boldsymbol{\mu}}(\tau)\dot{\boldsymbol{\mu}}(t+\tau)\rangle_\tau e^{-i\omega t}\mathrm{d}t. \tag{2.28}$$

Based on the Wiener-Khintchine theorem[81,82] the auto-correlation of a time-dependent entity χ is given by[19]

$$\langle \chi(\tau)\chi(t+\tau)\rangle_\tau = \frac{1}{2\pi}\int \left| \int \chi(t) e^{-i\omega t}\mathrm{d}t \right|^2 e^{-i\omega t}\mathrm{d}\omega. \tag{2.29}$$

Extracting e.g., IR- and Raman-spectra from dynamics has several advantages, such as the inclusion of anharmonic effects, as the harmonicity of the potential energy is not a pre-assumption, and the straightforward possibility to incorporate experimental conditions

such as temperature and solvent, as well as the consideration of different conformers by design. Furthermore, extracting IR-spectra from dynamics avoids the need to find a true minimum energy structure and having to compute the Hessian, which comes at an immense computational cost for extended systems. Other than from the normal mode analysis a continuous spectrum is obtained, which gives access to peak areas, for additional analysis and makes peak broadening steps superfluous.
Prior to the quantitative comparison of measured and computed spectra, the theoretical IR-spectra are scaled to compensate for systematical deficiencies of the applied electronic structure method.[83–88] Experimental spectra are usually prepared by baseline correction and peak smoothing.[89–92]
The quantitative agreement of spectra is given by a 'Hit Quality Index' (HQI), which usually has a value between zero and one. A higher HQI indicates a higher degree of agreement.[93–95] There are several measures available that can be used as HQIs such as the Euclidean distance the Root Mean Square Deviation, the Absolute Difference Value Search,[93] the Kullback-Leibler Divergence,[96] the Jeffrey Divergence,[97] or the Earth Mover Distance.[98] The Pearson correlation coefficient (M_{PCC}) (eq. 2.30) is another possible measure that is neither distance-based, nor relies on peak picking and matching but on the correlation between two spectra (s and r).

$$M_{\mathrm{PCC}} = \frac{\sum_{i=1}^{n}(s_i - \overline{s})(r_i - \overline{r})}{\sqrt{\sum_{i=1}^{n}(s_i - \overline{s})^2 \sum_{i=1}^{n}(r_i - \overline{r})^2}} \tag{2.30}$$

The similarity indicators penalize dissimilarities differently and therefore all lead to non-identical scores, as well as disparate responses to possible differences. In turn, the appropriate measure must be selected for each application (e.g., library search or comparison of computational approach).
In publication **V**, the processing of IR-spectra is presented in detail and several similarity indicators are tested. The study found that the slowly decreasing M_{PCC} is best for comparing calculated and measured spectra and for evaluating the quality of different calculated IR-spectra.[99]

Chapter 3

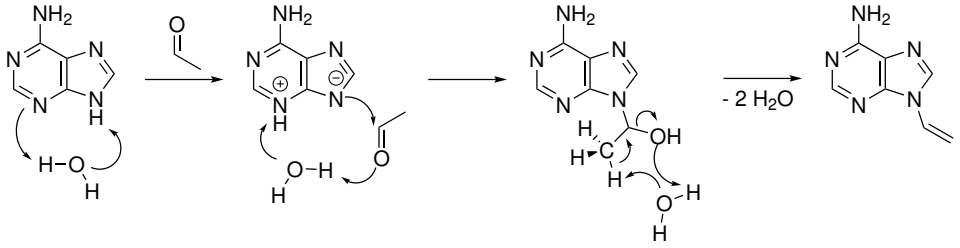
Characterization of a Prebiotic Pathway to Deoxyribonucleosides

In a previous work by Teichert *et al.*[11] it was shown experimentally that deoxyribonucleosides can be synthesized under ambient conditions in an aqueous solution from the corresponding canonical nucleobases, acetaldehyde, and glyceraldehyde. To date, it is widely argued that RNA probably arose prior to DNA as it has the ability to self-catalyze its replication, which is referred to as the "RNA-World" hypothesis.[15] Furthermore, it is discussed if alternative nucleotides could have constituted a variety of early nucleic acid polymers, which were ancestors to, or co-existed with RNA and DNA. This variety of information-carrying alternatives to RNA and DNA could have differed in the carbohydrate backbone, the nucleobases, as well as the phosphodiester junctions.[100]
Most proposed prebiotic chemical pathways to ribonucleosides, other than the proposed synthetic route towards deoxyribonucleosides, require, to our knowledge, scarce D-ribose. In addition, many suggested pathways lack chemical selectivity and fast degradation of intermediates.[101]
In the initiating work, it was already shown for deoxyadenosine that the proposed synthesis exclusively forms the β-furanose form.[11] In this consecutive computational study, the high regio- and stereoselectivity observed experimentally was studied using WTM-eABF simulations (see section 2.2.2) at ω-B97M-V/def2-TZVP[102–105] level of theory.

3.1 Simulation Details

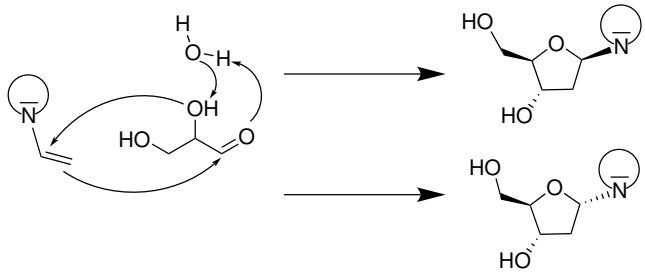
System Setup

The formation of the vinylated nucleobases was characterized in three steps using adenine as an example. Three systems were created as starting points for each of the subreactions (scheme 3.1) including (1) adenine, (2) 3H-adenine and acetaldehyde, and (3) 9-acetyladenine. Each system included an additional water molecule to facilitate the transfer of protons.

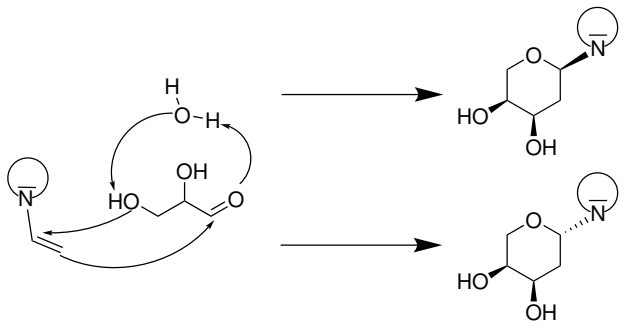


Scheme 3.1: Three step formation of N-9-vinyladenine.

The final condensation step was, in addition, also sampled with five, seven, and nine water molecules to investigate the effect of surrounding water on the reaction.
For all four canonical bases, adenine (A), cytosine (C), guanine (G), and thymine (T), the α- and β-deoxyfuranoses and pentopyranosyl-isomers were created. All systems included one additional water molecule to aid the proton transfer (scheme 3.2, 3.3) and were used as starting points for molecular dynamics simulations after geometry optimization.



Scheme 3.2: Formation of α- and β-deoxyfuranoses.



Scheme 3.3: Side reaction to pentopyranosyl-isomers.

Computational Details

All systems were optimized at PBEh-3c/def2-mSVP[106] level of theory. Starting from the optimized structures, ω-B97M-V/def2-TZVP[102–105] molecular dynamics simulations including VV10 dispersion correction[107] were conducted, comprised of heating and production. All simulations were conducted using the FermiONs++[108,109] program package. The computation of the exchange integrals was accelerated using the sn-LinK[110] procedure. Furthermore, the extended-Lagrangian[111,112] approach with the k-order 9 was employed. Implicit solvation with water using the COSMO[113] continuum solvation model was used in all molecular dynamics simulations.
The system, propagated using the Velocity Verlet integrator,[114] was heated to 323.15 K over the course of 3230 0.1 fs MD-steps by increasing the temperature by 1 K every 10 steps. The initial momenta were drawn randomly from a Maxwell-Boltzmann distribution. During the heating procedure, the system was constrained within the reaction path by a harmonic potential with k=500 kJ $mol^{-1}Å^{-2}$.
For the characterization of the enamine formation (scheme 3.1), six 20 ps WTM-eABF[12,13] simulations were run for each reaction step. Here, the time-step was set to 0.5 fs and the Langevin thermostat was used to maintain a target temperature of 323.15 K. The ABF force was scaled by a linear ramp and fully applied after 200 samples. The initial height of the Gaussian hills, with variance 0.1 Å, depositioned every 20 steps, was set to 0.5 kJ/mol and scaled down during the course of the well-tempered simulation. The effective temperature of the WTM-MtD was set to 2000 K. The force of confinement was set to 5000 kJ/mol per bin width. Similar settings were chosen for the formation of the furanose (scheme 3.2) and pyranose (scheme 3.3) rings.
For each system, four WTM-eABF simulations were performed where the dissolution of the sugar ring is observed resulting in the vinylated nucleobase and glyceraldehyde. These were started from the product configuration and aimed to show a single transition in order to ensure that only the transition to a selected isomer is observed. The four walkers were run independently and varied in the initial momenta. For the furanose systems, these simulations were 20 ps long, and for pyranose 30 ps. To guarantee sufficient sampling along the entire reaction path additional 20 ps WTM-eABF simulations were run confined to the transition state region until at least 200 samples per bin were collected. In total seven to ten independent simulations were run for each system to investigate the sugar ring formation.

Analysis

Reaction profiles were obtained by analysis using MBAR (see sec. 2.2.3).[10,62,115] Reaction free energies and activation free energies are calculated according to eq. 3.1 and eq. 3.2, respectively,[53,116]

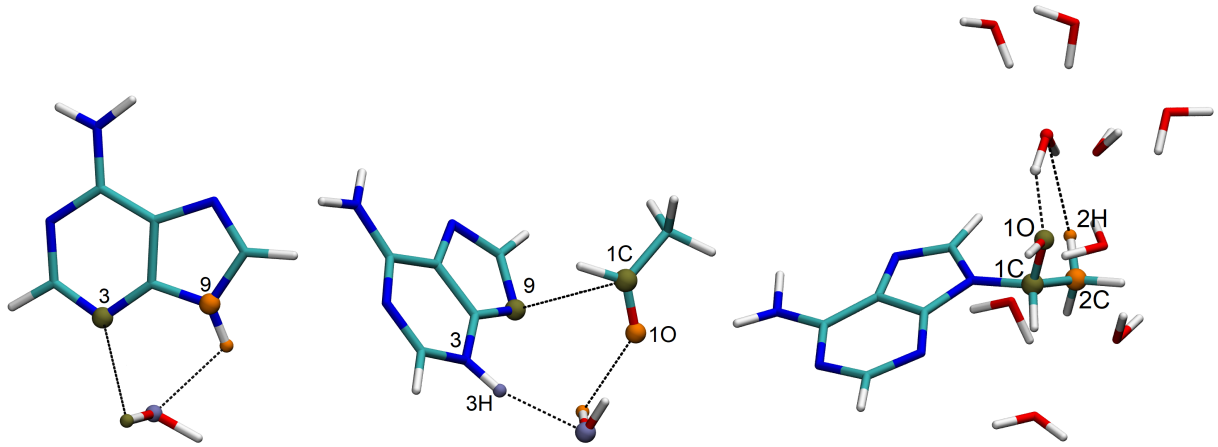
$$\Delta A^{\mathrm{rct}} = -\beta^{-1} \ln \frac{\int_{\mathrm{Product}} \mathrm{d}z \; e^{-\beta A(z)}}{\int_{\mathrm{Reactant}} \mathrm{d}z \; e^{-\beta A(z)}}, \tag{3.1}$$

$$\Delta A^{\ddagger} = \beta^{-1} \ln \frac{\rho(z_{TS}) \langle \lambda_\xi \rangle_{z_{TS}}}{\int_{\mathrm{Reactant}} \mathrm{d}z \; e^{-\beta A(z)}}, \tag{3.2}$$

where the reaction energy is computed from the non-overlapping integrals over product and reactant state.[53] In eq. 3.2, $\rho(z_{TS})$ is the normalized probability density at the transition state, and $\langle \lambda_\xi \rangle_{z_{TS}}$ is the z-conditioned average of $\lambda_\xi = \sqrt{h^2/2\pi k_b T m_\xi}$. Here, m_ξ is the effective mass of the pseudo-particle associated with the transition coordinate.[116]

Reaction coordinates

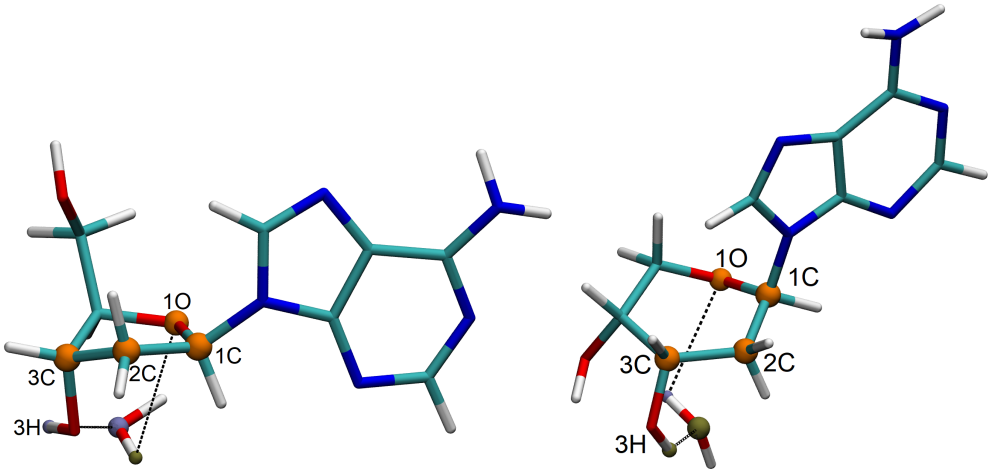
For the initial step of the formation of the vinylated nucleobase (scheme 3.1), the proton transfer from 3N to 9N is facilitated by a water molecule. To model this reaction, the $(d_{\mathrm{3NH}} + d_{\mathrm{OH}}) - d_{\mathrm{9NH}}$ linear combination of distances was chosen as the reaction coordinate and examined between values of 0.0 and 3.5 Å.
For the following step, the formation of the hemiaminal, comprising of the C-N bond formation and proton transfer from N3 to the carbonyl group, the $d_{\mathrm{1C9N}} + d_{\mathrm{1OH}} + d_{\mathrm{O3H}}$ coordinate was selected as collective variable in a range between -3.2 and 1.5 Å during the simulation. However, as multiple hydrogen atoms are available and this coordinate is not generalized, the resulting reaction free energy profile is evaluated along the 1C-9N interatomic distance in order to avoid samples being misplaced with respect to the reaction progress.
In the concluding condensation step yielding the vinylated nucleobase, the 1C-1O and one 2C-2H bond were cleaved and a proton was transferred from the surrounding water to the OH-leaving group. The resulting OH^- can then accept the proton from the methyl group. The $(d_{\mathrm{1OH}} + d_{\mathrm{O2H}}) - (d_{\mathrm{2C2H}} + d_{\mathrm{1C1O}})$ reaction coordinate was sampled between -3.2 and 3.2 Å. Similar to before the reaction coordinate was changed for evaluation. The coordinate was generalized to the 1C-1O distance and the third shortest 2C-H distance. Therefore, samples are correctly sorted even if a different proton is abstracted or back-transferred to the methyl group. The reaction coordinates used for sampling are visualized in fig. 3.1.



(a) $(d_{3\mathrm{NH}} + d_{\mathrm{OH}}) - d_{9\mathrm{NH}}$ (b) $d_{1\mathrm{C9N}} + d_{1\mathrm{OH}} + d_{\mathrm{O3H}}$ (c) $(d_{1\mathrm{OH}} + d_{\mathrm{O2H}}) - (d_{2\mathrm{C2H}} + d_{1\mathrm{C1O}})$

Figure 3.1: Summary of reaction coordinates used to simulate the three step formation of vinylated adenine.

The reactions to deoxyfuranoses (scheme 3.2) and pyranoses (scheme 3.3) were started from the product configurations in order to compare reactivities. As reaction coordinate, the distance between the 1C2C and 3C1O center of mass and the 1OH, O3H distances was selected ($d_{3\mathrm{C1O}-1\mathrm{C2C}} - (d_{1\mathrm{OH}} + d_{\mathrm{O3H}})$, see fig 3.2). The FES was then constructed along the center of mass distance between 1C-2C and 3C-1O.



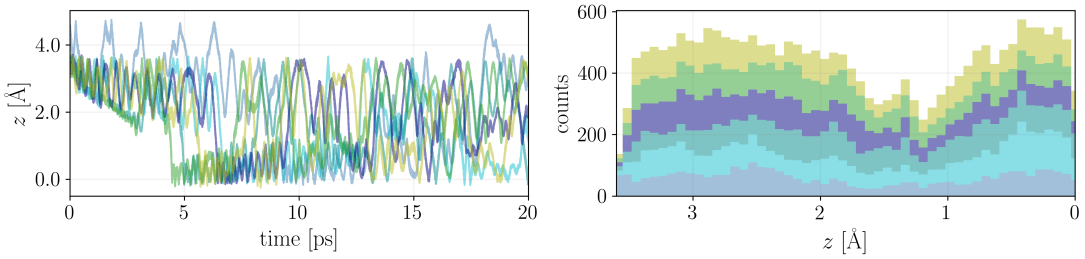
(a) $d_{3\mathrm{C1O}-1\mathrm{C2C}} - (d_{1\mathrm{OH}} + d_{\mathrm{O3H}})$ (b) $d_{3\mathrm{C1O}-1\mathrm{C2C}} - (d_{1\mathrm{OH}} + d_{\mathrm{O3H}})$

Figure 3.2: Visualization of the reaction coordinates used for the sugar ring formations yielding (a) furanose and (b) pyranose isomers.

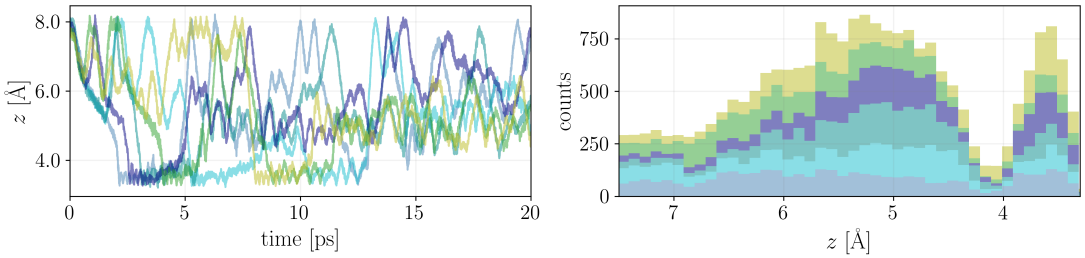
3.2 Current Results

Formation of the Vinyl Nucleobase

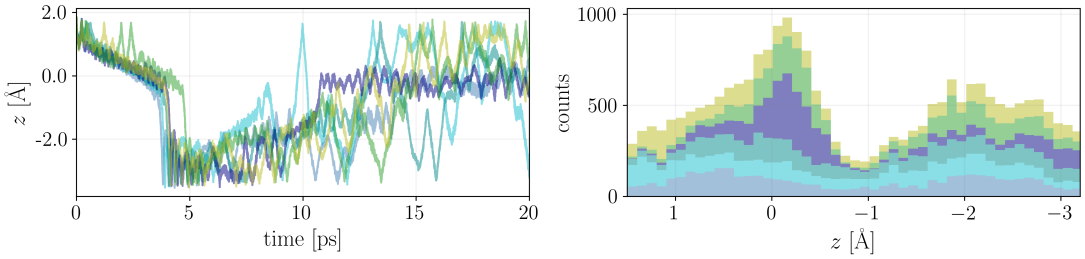
WTM-eABF calculations were run in order to collect samples along a chosen transition coordinate. Fig. 3.3 shows the sampling along the collective variable throughout the enhanced sampling molecular dynamics simulations and the distribution of samples.



(a) First step



(b) Second step

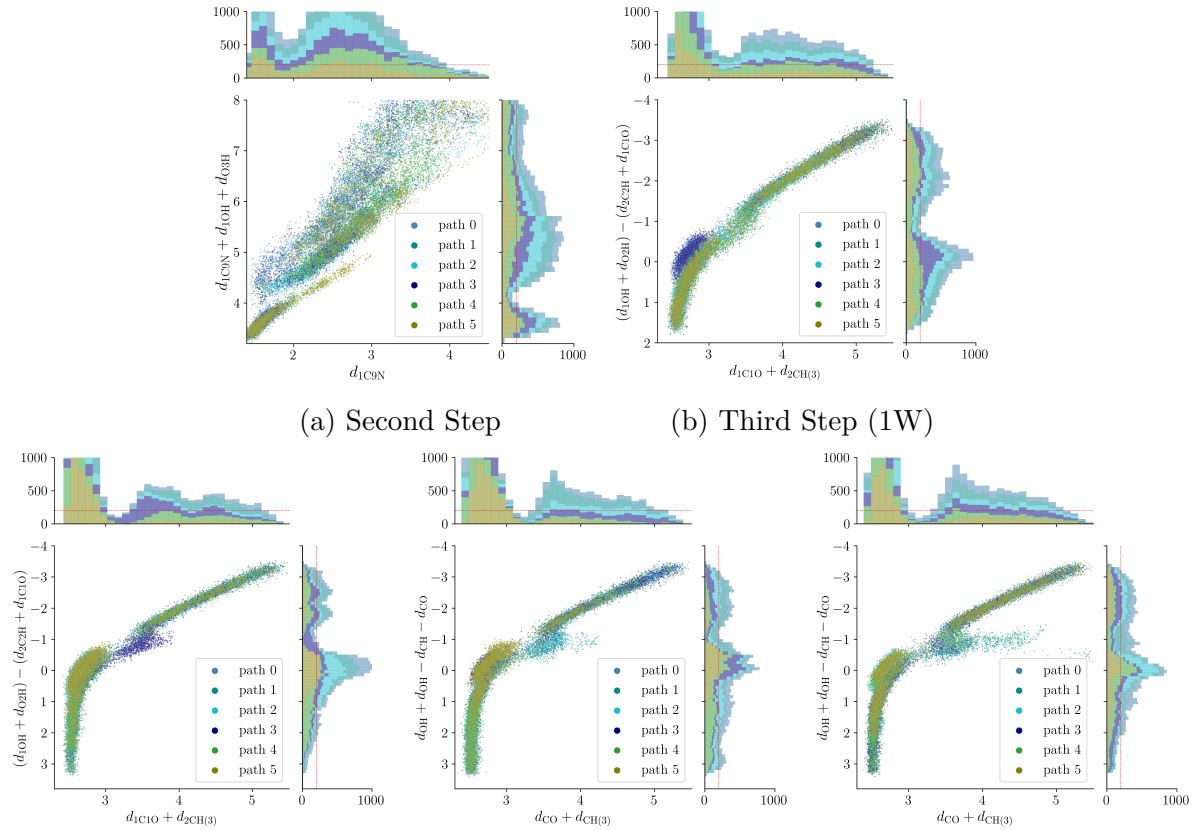


(c) Third step

Figure 3.3: On the left side the value of the reaction coordinate z is shown throughout the simulation. On the right side summarizing stacked bar plots are displayed, giving an overview of the sample distribution throughout the CV. Results are given for all three reaction steps towards N-9-vinyladenine for systems including a single water molecule. The first row (a) shows the results for the proton transfer from 9N to 3N, the second row (b) for the subsequent formation of the hemiaminal and (c) the third step for the formation of the enamine species. The different colors differentiate the results of independent walkers.

Best sampling was achieved for the first reaction step of the enamine formation. Nearly

uniform sampling was attained, showing the desired effect of the WTM-eABF method. For the following steps, greater differences were observed in the distribution of samples throughout the range of the CV. These steps are more complex, including several concerted bond rearrangements, making the definition of an appropriate CV more challenging.
Furthermore, the non-generalized CVs, meaning that specific protons were included in the transition coordinate, result in problems when other protons participate in the reaction. Therefore, other reaction coordinates were selected for the subsequent construction of reaction free energy profiles. The relationship between the coordinates used during the WTM-eABF simulations and the alternative coordinates tested for post-processing is visualized in fig. 3.4.



(a) Second Step (b) Third Step (1W)

(c) Third Step (5W) (d) Third Step (7W) (e) Third Step (9W)

Figure 3.4: Comparison of the CV used for sampling and the coordinate selected for analysis for the second and the third reaction step. Results are shown for systems containing one (1W), five (5W), seven (7W), and nine (9W) water molecules. As visual guide a line at 200 samples is added in the histograms.

It was found, that the formation of the vinylated nucleobase could be initiated by the proton transfer from the N-9 position to the N-3 position. This was discovered while simulating the second reaction step, where the 3-H-adenine species was consistently obtained

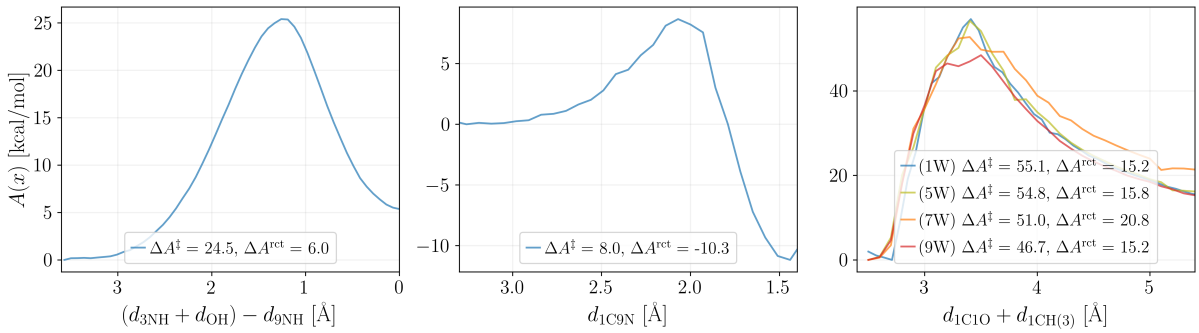


Figure 3.5: Obtained reaction free energy profiles, activation energies and reaction energies for the reaction steps yielding the vinylated adenine intermediate. For the third step results are shown for systems containing one (1W), five (5W), seven (7W) and nine (9W) water molecules.

after the first observed back-reaction. For this initiating proton transfer, a reaction barrier of 25 kcal/mol and a reaction energy of 6 kcal/mol were determined. For the subsequent formation of the hemiaminal species, a transition barrier height of 8 kcal/mol and a reaction energy of -10 kcal/mol were obtained. The reaction free energy profiles are shown in fig. 3.5.
For the third step, a high barrier of 55 kcal/mol was determined. However, by conducting further simulations incorporating more explicit water molecules it was found that this barrier lowers due to stabilization of the transition state via a network of hydrogen bonds (see fig. 3.6).

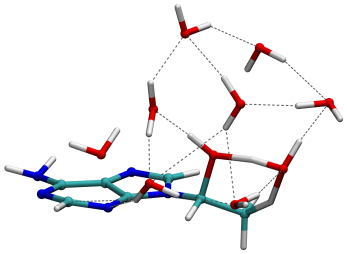


Figure 3.6: Snapshot from WTM-eABF molecular dynamics simulation showing hydrogen bond network stabilizing the transition state during the condensation reaction from the hemiaminal to the enamine species.

When adding eight water molecules (nine in total) to the system setup the barrier was lowered by 8 kcal/mol to 47 kcal/mol relative to the original activation energy obtained for the minimal system including only a single water molecule. We assume that this effect

is even higher in bulk water. However, due to the computational expense of *ab initio* MD simulations and the increasing demand of sampling needed if high amounts of water are introduced, this effect was so far only studied using minimal amounts of water molecules (five, seven, nine). In addition, more generalized CVs are needed for sampling when the system includes more explicit solvent to allow the reaction to proceed incorporating any of the available, e.g., water molecules. It should be noted that the stabilization of the water environment through hydrogen bonds can not be fully approximated using the given implicit solvent model. We further assume that the bulk water could have a catalyzing effect on all proton transfers, quantifying this thoroughly would be an interesting starting point for further investigation.

Assessment of the Stereoselectivity

To assess the stereoselectivity of the furanose ring formation from D-glyceraldehyde and the vinylated nucleobase WTM-eABF simulations were run capturing the reaction process yielding the α- and β-deoxyribonucleosides. Assuming a pericyclic cycloaddition, the linear combination of the distance between the center of mass of the 1O, 3C and 1C, 2C atoms and two O-H interatomic distances to transfer the proton from the 3O to the O1 atom *via* a solvent water molecule was chosen ($d_{3C1O-1C2C} - (d_{1OH} + d_{O3H})$).

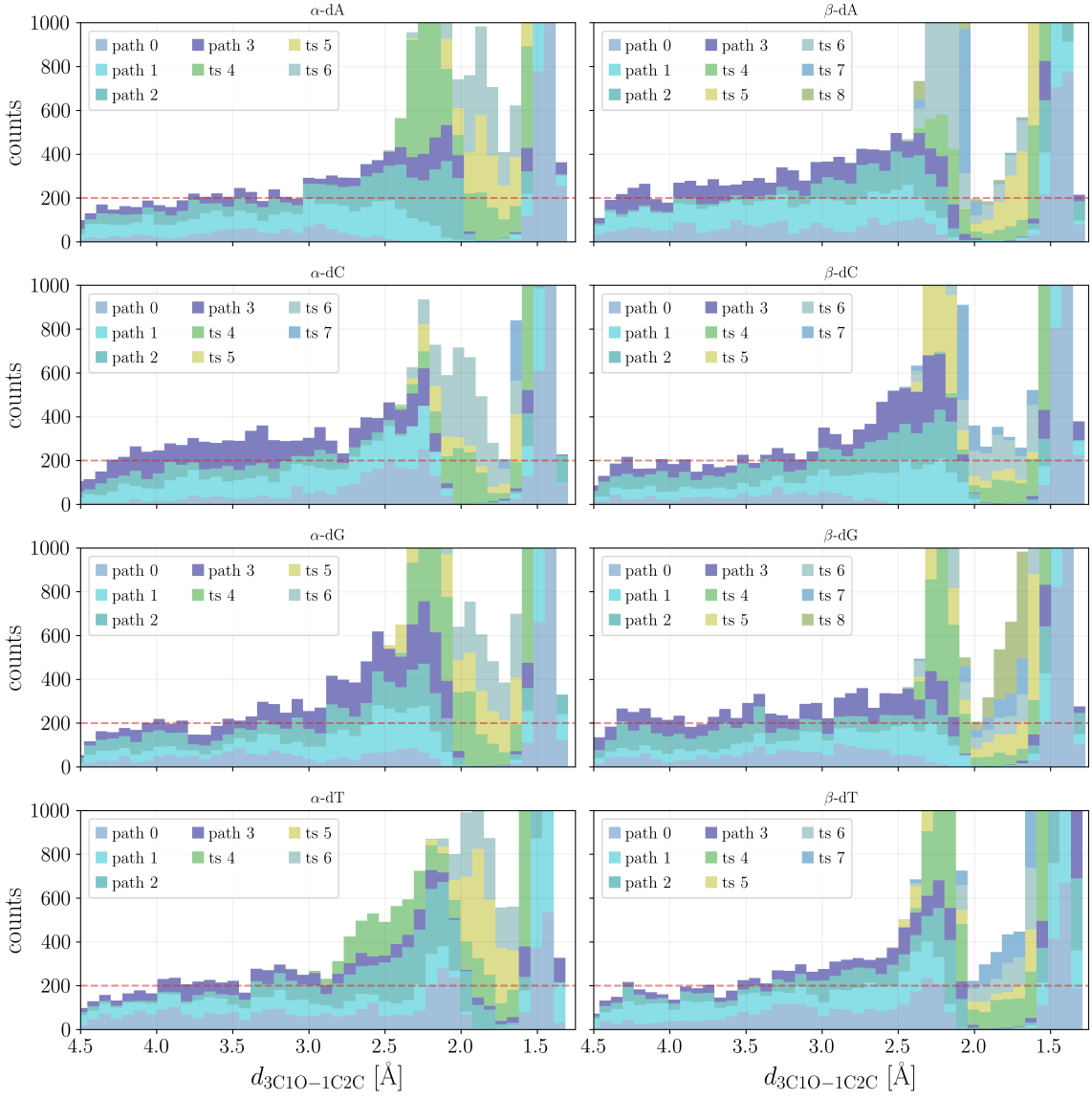


Figure 3.7: Stacked barplot showing the distribution of samples along the $d_{3C1O-1C2C}$ center of mass distance. The samples from seven independent simulations per system are shown marked by different colors.

Because the catalytic solute molecule can move freely and thereby disproportionately influence the value of the reaction coordinate, it was excluded during the evaluation procedure. However, when the O-H distances were not included in the reaction coordinate the reaction process was not observed within the same simulation time. For evaluation, the center of mass distance $d_{3C1O-1C2C}$ was deemed more reliable as transition coordinate. Fig. 3.7 shows the distribution of samples along the coordinate. Each color indicates the collected samples from an independent simulation. Because we aimed to evaluate the stereo-selectivity, the dissolution of the sugar ring was simulated making additional restrains, that ensure that a certain anomer is formed, superfluous. For the same reason, a single transition process was desired when the entire range of the reaction coordinate was sampled. As a consequence, between 2.0 and 1.5 Å additional simulations were needed in order to assure sufficient sampling in the transition state region. Because in this restricted region of the transition the reactants do not separate completely, several ring formations and openings can be sampled.
After the MBAR analysis was carried out along the original reaction coordinate, the free energy profile was constructed using the determined unbiased weights. Fig. 3.8 shows the free energy reaction profiles for the formation of α-dA, α-dC, α-dG, α-dT, β-dA, β-dC, β-dG, and β-dT. In addition, the activation and reaction energies are given.

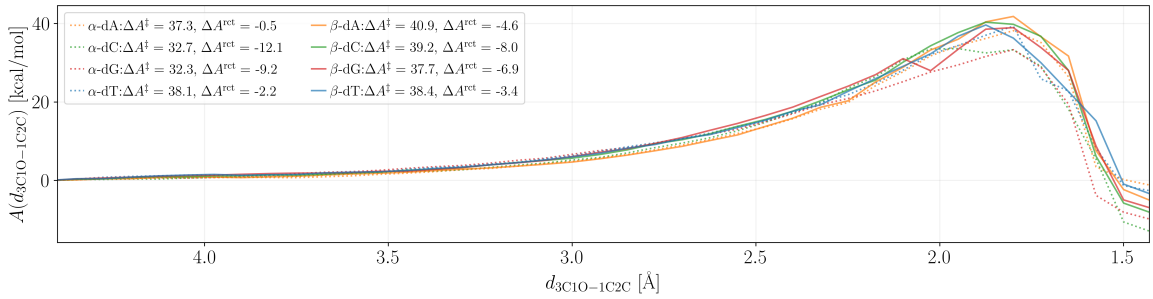


Figure 3.8: Results are given for the formation of the α- (dashed) and β-anomers (solid). Activation and reaction energies are given in kcal/mol.

For all four canonical nucleobases the formation of the α-isomer is favored, which is in disagreement with the experimentally observed stereo-selectivity. Both, the transition barriers and reaction free energy do not provide a clear indication that the β-isomer is expected as the predominant product. While this could be an artifact of insufficient sampling or an inadequate CV for this reaction mechanism, the findings could also suggest that important aspects of the reaction are not yet taken into account.
To investigate the disagreement between the experimental findings and the computational results (1) an estimate of the error of the free energy profile has to be added to the evaluation procedure, (2) the reaction should be studied involving additional solvent molecules as it was already shown that the surrounding water influences the obtained results, and (3) a sequential mechanism should be reconsidered, where first an open chained deoxyribose system is formed, which then undergoes cyclization. In principle, this is also allowed in

the current setup, however, the charge-separation in this mechanism could be disfavored by the lack of bulk solvent.

Assessment of the Regioselectivity

The alternate formation of the six-memberd rings yielding pentopyranosyl-isomers was modeled and evaluated analogously to the formation of the five-membered sugar rings.

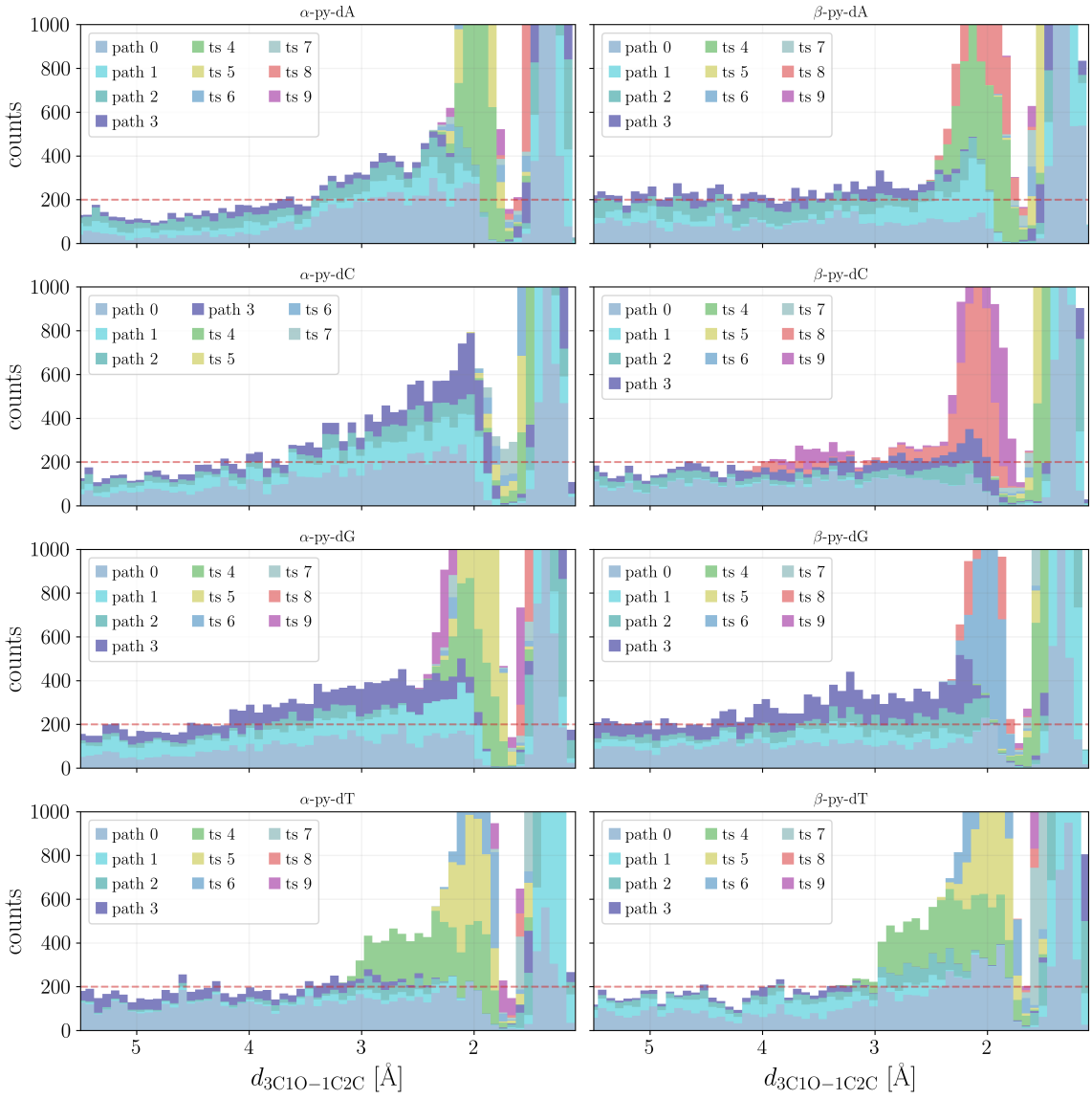


Figure 3.9: Stacked barplots showing the distributions of samples along the $d_{3C1O-1C2C}$ center of mass distance.

Samples were collected from eight to ten independent simulations. More simulations

were carried out if significantly less then 200 samples (indicated by red line, fig. 3.9) were collected in the transition state region. However, for α-py-A, α-py-G, α-py-T, β-py-A, β-py-C, and β-py-G still less then 200 samples were collected in a bin in the transition-state region. The insufficient sampling leads to inconsistencies in the resulting free energy profiles. The distribution of samples is shown in fig. 3.9. The recovered free energy profiles are given in fig. 3.10.

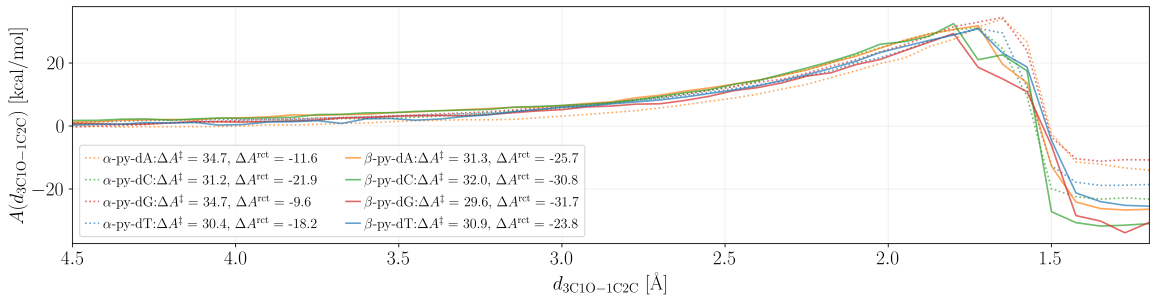


Figure 3.10: Results are given for the formation of the α- (dashed) and β-py-anomers (solid). Activation and reaction energies are given in kcal/mol.

Other than for the furanose isomers, here, the formation of the β-pyranose species seems to be favored, supported by slightly lower transition barriers and greater reaction energies than for the α-form. Furthermore, for all four canonical nucleobases the formation of the β-ribopyranoside constitution isomer seems to be both kinetically and thermodynamically favored over the formation of the expected β-deoxyribonucleoside.
However, as written before, at this point we have not fully exploited all computational possibilities. Further investigations are needed to clarify if the chosen setup is adequate to characterize the reactions sufficiently accurately to draw conclusions about the selectivity of the synthesis route and why the experimental findings and computational results, at this stage, are in disagreement.

3.3 Concluding Remarks

To investigate the reaction mechanism yielding deoxyribonucleosides from prebiotically available nucleobases, acetaldehyde, and D-glyceraldehyde, reaction free energy profiles were calculated using WTM-eABF simulations and subsequent MBAR analyses. We were able to observe all reaction steps computationally, supporting the suggested mechanism based on experimental observations and thereby the possibility that the deoxyribonucleosides developed earlier than assumed so far.
Furthermore, it was found that surrounding water plays a catalytic role and stabilizes the observed transitions. Therefore, to refine the results, further simulations with extended water spheres should be conducted. With the inclusion of a higher number of water molecules the need for more generalized reaction coordinates arises to efficiently sample the reactions and deliver reliable results. Adding more explicit water molecules might also change the conformational stabilities (e.g., chair conformations) of the reactants and products, influencing the results.
To asses the quality of the free energy profiles, in future work, an error estimate has to be implemented to measure the statistical error. As the effect of surrounding water has to be investigated further and the error in the free energy profiles is to be determined, so far, no clear conclusions can be made concerning the regio- and stereoselectivity of the postulated prebiotic pathway to DNA nucleosides.

Chapter 4

Publications

4.1 Publication I: Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5

Beatriz von der Esch, Johannes C. B. Dietschreit, Laurens D. M. Peters, Christian Ochsenfeld

"Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5"
J. Chem. Theory Comput. **2019**, *15*, 6660-6667.

Abstract: Sirtuin 5 is a class III histone deacetylase that, unlike its classification, mainly catalyzes desuccinylation and demanoylation reactions. It is an interesting drug target that we use here to test new ideas for calculating reaction pathways of large molecular systems such as enzymes. A major issue with most schemes (e.g., adiabatic mapping) is that the resulting activation barrier height heavily depends on the chosen educt conformation. This makes the selection of the initial structure decisive for the success of the characterization. Here, we apply machine learning to a large number of molecular dynamics frames and potential energy barriers obtained by QM/MM calculations in order to identify (1) suitable start-conformations for reaction path calculations and (2) structural features relevant for the first step of the desuccinylation reaction catalyzed by Sirtuin 5. The latter generally aids the understanding of reaction mechanisms and important interactions in active centers. Using our novel approach, we found eleven key features that govern the reactivity. We were able to estimate reaction barriers with a mean absolute error of 3.6 kcal/mol and identified reactive configurations.

JCTC Journal of Chemical Theory and Computation
Article
Cite This: J. Chem. Theory Comput. 2019, 15, 6660−6667
pubs.acs.org/JCTC

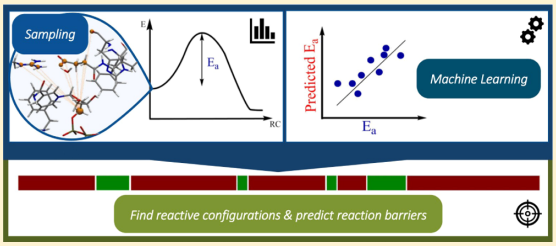
Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5

Beatriz von der Esch,[†] Johannes C. B. Dietschreit,[†] Laurens D. M. Peters, and Christian Ochsenfeld*

Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany

Supporting Information

ABSTRACT: Sirtuin 5 is a class III histone deacetylase that, unlike its classification, mainly catalyzes desuccinylation and demanoylation reactions. It is an interesting drug target that we use here to test new ideas for calculating reaction pathways of large molecular systems such as enzymes. A major issue with most schemes (e.g., adiabatic mapping) is that the resulting activation barrier height heavily depends on the chosen educt conformation. This makes the selection of the initial structure decisive for the success of the characterization. Here, we apply machine learning to a large number of molecular dynamics frames and potential energy barriers obtained by quantum mechanics/molecular mechanics calculations in order to identify (1) suitable start-conformations for reaction path calculations and (2) structural features relevant for the first step of the desuccinylation reaction catalyzed by Sirtuin 5. The latter generally aids the understanding of reaction mechanisms and important interactions in active centers. Using our novel approach, we found eleven key features that govern the reactivity. We were able to estimate reaction barriers with a mean absolute error of 3.6 kcal/mol and identified reactive configurations.



INTRODUCTION

Computationally obtained reaction barriers are an excellent link to experiment. They allow us to verify or propose new reaction mechanisms, gain insights into kinetics, or compare reactivities. However, the calculation of reliable activation energies is a demanding task, especially for large molecular systems, for example, enzymes.

There exists a large collection of static and dynamic methods to model chemical reactions (e.g., adiabatic mapping,[1] nudged elastic band,[2,3] string methods,[4,5] transition path sampling,[6] umbrella sampling,[7] metadynamics,[8] and many more). Regardless of the method, there are two major challenges: (1) the choice of the theoretical description, and (2) the sampling bottleneck that leads to a ubiquitous dependency on the chosen start conformation. One of the main tools of choice for studying enzymatic systems is the combination of quantum mechanics and molecular mechanics (QM/MM) (see e.g., refs 9−11). The description of the QM part varies between semi-empirical (e.g., AM1 or SCC-DFTB)[12−15] and ab initio methods (e.g., HF or DFT).[16−18] Besides the level of theory chosen for the QM region, the extent of the QM sphere[19−21] and the treatment of the boundary region play an important role.[9−11] With increasing computational power and novel efficient methods, we are able to increase our attention to detail (e.g., solvent effects), and apply higher level theoretical methods. However, the second issue of selecting an initial configuration is nearly as important as the accuracy of the description of the electronic structure. In order to circumvent the need to search for suitable starting structures from a vast number of frames,[22,23] extensive sampling would be needed. Unfortunately, it becomes more and more demanding to sample the phase space with increasing system size and accuracy of the Hamiltonian. Therefore, for extended system such as enzymes, exploring the entire phase space remains prohibitively expensive at the QM/MM level. Start configurations can be taken from an MM-MD simulation. Alternatively, it has been suggested to start from the crystal structure, avoiding the selection problem entirely (see e.g., ref 24). However, the X-ray structures, which often differ from structures in solution, are not guaranteed to be reactive.[11] Even if they are suited for the initial step of a reaction, problems might arise for subsequent reaction steps.

Thus, it is paramount to develop a straight forward approach for pinpointing reactive configurations visited during the MM-MD, which are located at the beginning of reaction paths. The work of Lodola et al.[25] supports the importance of exploring the influence of conformational changes. They show the power of statistical tools, for example, principal component analysis, to identify conformational changes dominating enzymatic reactivity.[25] In a recent study, Bonk et al.[26] tried to link geometry and reactivity using machine learning during extensive transition interface sampling which enabled them to find reactive trajectories more often.

Received: August 30, 2019
Published: November 25, 2019

Here, we apply QM/MM adiabatic mapping to a large selection of MM-MD frames to obtain an estimate of the reaction barrier starting from these snapshots. Adiabatic mapping is a straight forward approach to calculate the potential energy profile of a reaction, where a predefined reaction coordinate is gradually changed while the remaining system is relaxed. It should be noted that adiabatic mapping is not suitable for modelling reactions involving large structural rearrangements or changes in solvation.[11] We relate the initial structures taken from the MD trajectory and the calculated transition barriers using simple machine learning in order to understand which conformational changes influence the reactivity, and build a predictive model for activation energies. The model is subsequently applied to all MD frames in order to identify reactive regions within the trajectory. This set up is intended to help identify suitable start frames and therefore alleviate the need of extensive sampling, which is a true limitation at the QM/MM level.

As a model reaction, we investigate the first step of the desuccinylation reaction catalyzed by Sirt5, which belongs to the class of histone deacetylases.[27,28] Despite what its enzyme class name suggests, Sirt5 mainly catalyzes the desuccinylation or demanoylation of lysines and not a deacetylation.[29] This desuccinylation is thought to be a three step reaction which is initiated by a nucleophilic attack of the substrate on the NAD^+ cofactor that leads to the dissociation of nicotinamide.[27,30−32]

METHODS

Data Acquisition. *Structure Preparation.* The crystal structure (PDB: 3RIY[33]) consists of a dimer of Sirt5 in the complex with a histone tail peptide containing a succinylated lysine (SLL) as well as NAD^+. We selected the first monomer in the file (chain A for Sirt5 and chain D for the peptide) as well as the respective NAD^+. Hydrogen atoms were added using the program tleap from the program suite Amber16.[34] The protonation state of titratable residues was set according to PropKa.[35,36] The zinc finger in Sirt5 was parametrized using the ZAFF (Zinc Amber Force Field) parameters.[37] For the residue SLL, GAFF (Generalized Amber Force Field) parameters[38] were assigned using the Antechamber code, which determined the atomic partial charges from an AM1[12] calculation with bond-charge corrections (AM1−BCC).[39] The parameters for NAD^+ were taken from the AMBER parameter database.[40,41] All other parameters were taken from AmberFF14.[42] Finally, the system was solvated by placing it in a TIP3P[43] water box with a distance of 17 Å in all three dimensions at a density of 0.832 g/cm^3. The system was neutralized with one chloride ion.

MM-MD Simulation. Two minimizations (10 000 steps) were carried out to prepare the solvated system. During the first minimization, the protein was constrained and only the water molecules were optimized. In the second step, the entire system was subjected to the minimization. The system was heated to 300 K by increasing the temperature by 1 K every 100 fs. Afterward, the system was equilibrated for 100 000 time steps. During heating and equilibration, the temperature was controlled with simple velocity rescaling. The following production run was performed in the *NPT* ensemble for 200 ns. The pressure was kept at one atmosphere and the temperature at 300 K with the Langevin Piston barostat and Langevin thermostat implemented in NAMD.[44] The time step for equilibration and production was set to 2 fs. Nonbonded interactions were evaluated explicitly within 10 Å and smoothly switched off at 12 Å. A Verlet nearest neighbor list[45] with a radius of 13.5 Å was used to speed up the computations. Periodic boundary conditions were used in all three directions. Electrostatic interactions were evaluated with the particle mesh Ewald method[46] and an interpolation of the sixth order. The MD simulations were carried out with the NAMD[44] program package.

QM/MM Calculations. We selected frames (every 0.5 ns) from the production run as starting points for QM/MM calculations. All the frames were minimized twice at the MM level for 10 000 steps, again minimizing first only the solvent and then the full system. Subsequently, the frames were subjected to a QM/MM optimization. The QM region always included the residues Arg105, Phe70, Phe223, His158, part of NAD^+, and the succinyl-lysine residue, as well as all water molecules within 4 Å of the C1′ atom of the ribose in NAD^+, which are in total 139−151 atoms, depending on the number of water molecules in the active site (Supporting Information, SFigure 1 shows the QM region). The QM region was described at the HF-3c[47] level of theory and the MM region as specified in the section "Structure Preparation". The two subsystems were coupled via electrostatic embedding. The QM/MM calculations were performed within the ChemShell[48] code, with the QM part treated by the program package FermiONs++.[49,50]

We performed a small benchmark comparing HF-3c with higher level DFT methods to show that it is well suited for our endeavor. HF-3c consistently overestimates the reaction barrier. Trends in higher and lower barriers are reflected properly compared to DFT (see the Supporting Information for more details).

The optimized structures were used as starting points for adiabatic mapping pathways. The reaction coordinate was defined as the difference between the C1′−O bond and the C1′−N bond. While the C1′−O distance was reduced, the C1′−N bond was elongated. In each step, the bond difference was changed by 0.2 Å and fixed, while the remaining system was minimized.

Machine Learning. *Data Preprocessing.* We are interested in the relation between the educt configuration and the reaction barrier. Therefore, a representation of the geometry is needed that is suited to describe structural changes. There are several representations which are well established for chemical investigations such as Bag of Bonds,[51] *XYZ*-coordinates, Coulomb-matrices,[52,53] or SMILES.[54] Each of these representations is appropriate for different problems. Even though there is a number of established representations, we decided to simply select the distances between all nonhydrogen atoms within the QM region to describe the geometry in the active site. This representation allows for a preliminary correlation analysis which reduces the number of features in our system (see next section). Additionally, no further calculation of, for example, atomic charges is needed (which are heavily influenced by the employed QM method). The collection of interatomic distances is invariant to translation and rotation, and therefore, avoids any problems that might otherwise occur. Additionally, the number of water molecules within the QM region was considered as an additional feature. All in all, this added up to 2629 features.

Dimensionality Reduction. Because the outcome of a machine learning fit is dependent on expressive features and can be impaired by redundant or even insignificant variables, the features were purged. The dimensions were reduced by a

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

Journal of Chemical Theory and Computation Article

simple correlation analysis. All features with absolute correlations <0.375 to the reaction barrier were omitted. This value was chosen quite low to ensure that most of the variations were captured. Further, the remaining features were checked for strong absolute correlations >0.9 among each other. If a pair of features were strongly correlated, one of them was omitted. This resulted in a subset containing only 15 features out of the original 2629.

Model Selection, Refinement, and Application. There exists a vast number of machine learning algorithms to choose from. Because we want to predict activation energies, we are trying to solve a typical regression problem from the mathematical point of view. There are different types of regression models, simple linear regression (least squares), polynomial regression, support vector regression, decision tree regression, to name a few.[55,56] The predictive power of the different machine learning models depends strongly on the structure and size of the data, and the relation between the target and feature variables. All machine learning scripts were performed in python with a combination of pandas[57] and scikit-learn.[58] We tested different supervised learning regression techniques, the results can be found in section "Machine Learning Model Comparison" of the Supporting Information. After testing we chose a sparse regression model, the elastic net regressor.[58,59]

Elastic net regression includes variable selection and regularization, which leads to a greater predictive power and enhances the interpretability of the results. Methods including regularization are especially suited for problems where little data is available. They suppress overfitting by introducing a cost function.[59] Based on all 150 samples, an elastic net model was built. The hyperparameter α, which controls the strength of the bias, and the l1_ratio (the ratio between the l1- and l2-type cost functions) were determined using fivefold cross validation. To evaluate the performance of the model the mean-absolute-error (MAE), RMSE, and R^2 value were calculated using threefold cross validation. To additionally visualize the skill of the machine learning model on new samples, the data set was randomly divided into a training and testing set (2:1), fitted to the training set and applied to the test set.

Lastly, the model was fitted to all the available data (all 150 frames), with the previously determined hyperparameters. The final model was then used to predict the reaction barrier for every MD frame (every 10 ps of the trajectory). For ten frames with low predicted reaction barriers, adiabatic mapping as described in the section "QM/MM Calculations" was carried out to show that the model helps to find reactive frames. The model generated here is not transferable, but the presented protocol can be employed for other extended systems.

RESULTS AND DISCUSSION

Reaction Barriers Obtained by Adiabatic Mapping. The combined QM/MM adiabatic mapping calculation of 250 reaction pathways starting from snapshots taken from an MM-MD simulation gave reaction barrier heights between 22 and 80 kcal/mol. Figure 1 shows how the calculated reaction barriers increase with an increasing number of water molecules.

As the MD simulation advances, the peptide and NAD^+ slightly unbind and more water molecules coordinate the carbonyl-oxygen involved in the first reaction step, and thus, its nucleophilicity decreases. After 75 ns, the adiabatic mapping approach was mostly incapable of describing the nicotinamide

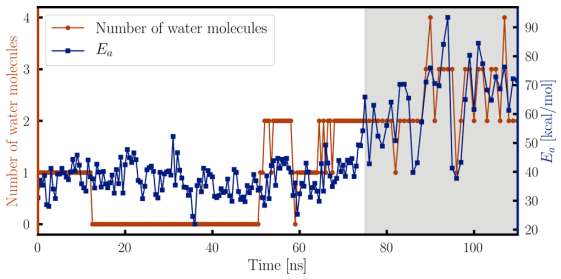


Figure 1. Number of water molecules within 4 Å of the C1′ atom of the ribose in NAD^+ is shown in red. The computed HF-3c activation energies are plotted in blue. The shaded area highlights the region that was not included in the subsequent machine learning steps.

cleavage, the desired products were no longer obtained. The incapability to model the reaction expresses itself in very high reaction barriers. Only the first 150 reaction pathways, starting from snapshots taken within the first 75 ns of the MD-trajectory, were included in the data-set for machine learning.

Figure 1 also shows that extended periods of the MD trajectory are especially nonreactive. This underlines that if only very few frames are picked or a very short MD simulation is used as basis for further calculations, one can miss reactive periods completely. The first 150 samples, each 0.5 ns apart along the MD-trajectory, yield energy barriers between 21 and 60 kcal/mol. The distribution of the barrier heights is shown in Figure 2. It highlights that educt configurations that lead to a

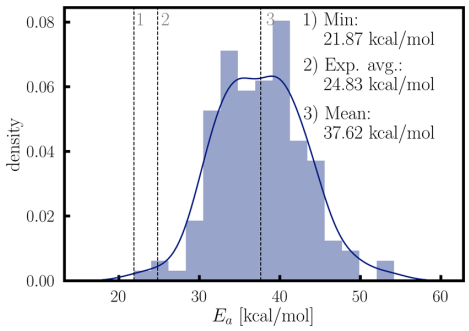


Figure 2. Distribution of the calculated energy barriers (adiabatic mapping with HF-3c). The blue line indicates a smooth distribution function fitted to the histogram. (1) Lowest barrier found, (2) exponentially averaged barrier, (3) mean barrier.

low energy transition are extremely rare. This emphasizes how difficult it is to find an appropriate start frame, that closely resembles the reactive enzyme complex and provides a reasonable energy barrier. The large variation of reaction profiles obtained with different initial configurations, and therefore the importance of suited starting points has been recognized early on (see e.g., refs 9, 25, 60−62).

As presumed by Ryde,[63] the energy barriers roughly form a Gaussian distribution. The arithmetic average is 37.62 kcal/mol, and the minimum activation barrier is 21.87 kcal/mol. The exponential average gives a good estimate for the barrier and is suitable for comparison with experiments, under the condition that the picked snapshots are well chosen;[64] here we obtain a value of 24.83 kcal/mol. The

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

standard deviation, σ, within these 150 samples is 5.40 kcal/mol. Based on the conclusions by Ryde,[63] more than 10^6 samples would be needed to obtain an estimate of the activation barrier within chemical accuracy (within 1 kcal/mol) of the exponential average. In contrast, most studies have only used a few snapshots (about 3−10).[65−68] To avoid having to calculate millions of pathways, we propose a strategic, scalable approach. We suggest using machine learning based on selected distances to pinpoint reactive regions within the MM trajectory. This allows for strategic sampling of reaction pathways that contribute significantly to the exponential average, giving a more accurate estimate of the energy barrier using less samples. Alternatively, productive snapshots from the MD trajectory found by the machine learning model, can be used for further (more accurate) QM/MM studies.

Machine Learning Performance. Using Elastic Net regression with 15 input features and 150 samples, a model was built to predict reaction energies from geometrical features. With α set to 0.09 and l1_ratio equal to 0.5, a MAE of 3.58 kcal/mol and a root mean squared error (RMSD) 4.46 kcal/mol is obtained. The R^2 score, which describes the percentage of the response variable variation that is explained, is 0.28. In general an R^2 score of 0.28 may be regarded as very poor. However, with respect to the complexity of the system and the very limited number of training points (100) a score of 0.28 is surprisingly high. Additionally, only 15 features were needed to describe the problem to this level of accuracy, which is possibly influenced by a much greater collection of residues.

To visualize the performance of the machine learning model, the data set was randomly spilt into a test (50 samples) and a training set (100 samples). This 1:2 division is similar to the one made during one cycle of the threefold cross validation used to assess performance. Subsequently the model was fitted on the training set and applied to the test set. The predictions for the test set versus the activation barriers calculated using adiabatic mapping are shown in Figure 3. The predictions of

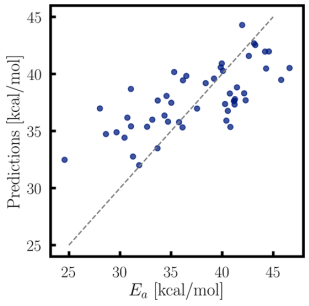


Figure 3. Scatter plot showing the performance of the Elastic Net Regressor on a test set (50 random points which were not used for learning).

the regression model are in quite good agreement with the actual activation barriers. The model is least accurate for the extreme barriers, it overestimates low barriers and underestimates high barriers. These are the regions in the training distribution of reaction barriers with the least number of samples. In order to increase the predictive power of the model, without prior knowledge of the system, more training points are needed. We suggest to test if semi-empirical methods might be a solution to the sampling bottleneck. Another way to increase the predictive power is to iteratively apply the model: calculate frames with predicted low barrier heights, add the results to the training data, and enhance its performance with every cycle. However, enhancing the performance of the model is only necessary if the goal is to predict a final energy barrier using this approach. That being said, the model is able to differentiate between less and more reactive frames. Therefore, this straight forward approach is sufficient to identify regions of interest within the MD trajectory, which is the intent of this work. Appropriate starting geometries identified by the built model can then be used for involved QM/MM studies, for example, aiming at calculating a free energy reaction profile.

Analysis of the Resulting Feature Subset. The group of features that remained after the two selection steps (explained in section "Dimensionality Reduction") is shown in the following three tables (Tables 1−3). For each of the features, the indices of the involved atoms (pdb file of the entire system is attached to the Supporting Information), the Pearson correlation coefficient to the activation energy, the elastic net coefficient, and an explanatory figure are given. The distances are grouped into 3 categories. The first category contains distances between the binding pocket and either the substrate or the cofactor (Table 1). The second group consists of intramolecular distances of SLL and NAD^+ (Table 2). The last group contains the intermolecular interactions between SLL and NAD^+ (Table 3).

The features given in Table 1 are distances from the substrate and the cofactor to the surrounding amino acids. Two distances are between atoms of PHE 70 to SLL (1) and NAD^+ (2), which are anticorrelated to the activation energy and contribute to the predictive model. The two features indicate that PHE 70 and the attached backbone must allow enough space for the nicotinamide leaving group to move out of the binding pocket. Hence, if the SLL−PHE 70 and the NAD^+−PHE 70 distances increase, the activation barriers become lower. Feature 3 and 4 show that the binding pocket has to be compact and the NAD^+ cofactor has to be located deep in the active center for the reaction to take place.

The second category includes intramolecular distances. It shows that small conformational changes within the reactants clearly influence the reactivity. Features 5 and 6 express the relative position of the nicotinamide to the ribose ring. As they are very similar, feature 6 was eliminated by the elastic net model due to its redundancy.

The alignment of the succinyl group plays a major role. Feature 7 has the highest absolute coefficient of all the features and therefore has the greatest impact on the predicted transition barrier. Feature 7 expresses the distance between the C4 atom and the terminal carboxyl group. This distance is anticorrelated to the activation barrier, and thus the barrier is lowest when the negatively charged carboxyl group is furthest away from the reactive centers.

The last group is the largest, it contains eight features which describe the relative positions of NAD^+ and SLL. Features 8, 9, and 10 are related to the previously explained feature 7. These distances are also a measure of the relative position of the carboxyl group, and therefore redundant, their coefficients are small or zero. The other five distances between NAD^+ and SLL show all positive correlation to the energy barrier. They indicate that the substrate and the cofactor have to be sufficiently aligned in order for the reaction to take place. Additionally, based on the large number of features containing

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

Journal of Chemical Theory and Computation Article

Table 1. Features 1−4 Used in the Elastic Net Model[a]

Number	1	2	3	4
Atom indices[b]	4192, 584	4281, 594	4279, 1165	4280, 2022
Corr. to E_a	-0.38	-0.39	0.39	0.38
Coefficient	-1.12	-0.78	0.82	1.19

[a]These four features describe the overall configuration of the active site. The atoms in between which the distance is measured is colored in green. [b]Atom indices as in the pdb-file (see Supporting Information).

Table 2. Features 5−7 Used in the Elastic Net Model for Describing Interactions Within SLL and NAD$^+$[a]

Number	5	6	7
Atom indices[b]	4294, 4284	4293, 4284	4194, 4191
Corr. to E_a	0.41	0.40	-0.45
Coefficient	0.66	0.0	-2.99

[a]The atoms in between which the distance is measured is colored in green. [b]Atom indices as in the pdb-file (see Supporting Information).

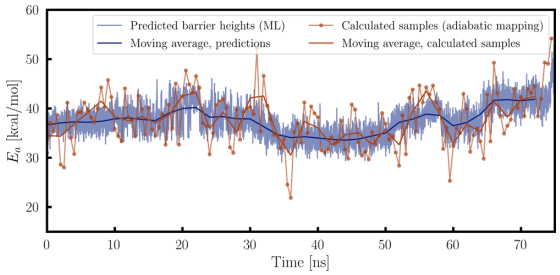


Figure 4. Section from the entire MD for which the activation energies were predicted with the previously built model. The red dots indicate the energy barriers calculated with adiabatic mapping.

the ribose ring, we suspect that the pucker of the ring plays an important role.

Application of the Trained Model to the Entire MD Simulation. The final model, which was trained on all 150 samples, was applied to the entire MD-trajectory. The predicted barrier heights for the initial step of the dessuccinylation are shown in Figure 4. One can see the general agreement between predicted (blue) and calculated MD frames (orange). The changes in the reactivity are captured and reflected by the estimated barriers. It is interesting to note that there are periods in the MD trajectory which are either reactive or nonreactive, and others in which the reactivity oscillates very strongly.

The distribution of the predicted activation energies is shown in Figure 5 on the left. It is compared to the initial collection of barrier heights used for learning. The comparison shows that the distribution of the predictions is much narrower. This already suggests that the model will overestimate low energy transitions and underestimate high barriers.

In order to check the reliability of the model for predicting reactive regions within the MD trajectory, we selected 10 frames for which a low barrier was forecast and three additional snapshots to represent the frames with higher predicted activation energies. These three additional samples are the frames at 25, 50, and 75% of the distribution of predicted transition barriers. Starting from these snapshots, adiabatic mapping calculations were carried out. The results for the picked frames that were modeled are shown in Figure 5 on the right. The predicted (ML) values and the calculated results (adiabatic mapping with HF-3c) are compared. They are put

Table 3. Features 8−15 Used in the Elastic Net Model for Describing the Interactions between SLL and NAD$^+$[a]

Number	8	9	10	11
Atom indices[b]	4279, 4188	4279, 4190	4279, 4195	4279, 4189
Corr. to E_a	0.43	0.41	0.42	0.43
Coefficient	0.0	-0.16	0.0	1.81
Number	12	13	14	15
Atom indices[b]	4282, 4188	4284, 4188	4286, 4188	4286, 4191
Corr. to E_a	0.45	0.43	0.54	0.48
Coefficient	0.92	0.0	0.80	0.10

[a]The atoms in between which the distance is measured is colored in green. [b]Atom indices as in the pdb-file (see Supporting Information).

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

Journal of Chemical Theory and Computation Article

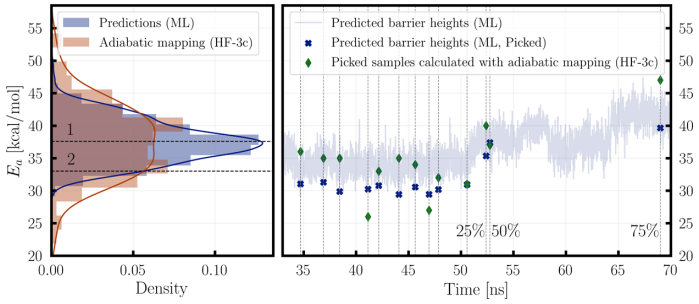


Figure 5. Left: Distribution of the predicted barrier heights (blue). For comparison the distribution of the initially calculated barriers (adiabatic mapping with HF-3c), used for learning, is given (orange). The arithmetic mean (1) and exponential average (2) of the predicted barriers are 37.60 and 33.02 kcal/mol, respectively. Right: Comparison of predicted (ML) and calculated (adiabatic mapping with HF-3c) reaction barriers for 10 frames with low energy transitions and three additional representative snapshots. The values shown here are listed in Table 4.

into context with the predicted barriers for all frames of the MM trajectory and the samples originally given to the model for training.

The predicted barrier heights and the calculated reaction barriers for the thirteen frames are listed in Table 4.

Table 4. Comparison of Predicted and Calculated Barrier Heights for Ten Frames with Low Estimated Reaction Barriers by the ML Model[a]

time [ns]	$E_a^{Pred.}$	$E_a^{Calc.}$ [kcal/mol]	ΔE
34.70	31	36	5
36.89	31	35	4
38.44	30	35	5
41.14	30	**26**	−4
42.16	31	33	2
44.08	30	35	5
45.64	31	34	3
46.96	29	**27**	−2
47.85	30	32	2
50.58	31	31	0
52.39[b]	35	40	5
52.76[c]	37	34	−3
69.00[d]	40	47	7

[a]Three additional values are given for frames from 25, 50, and 75% of the distribution of predicted transition barriers. Bold numbers indicate calculated barriers that are below 30 kcal/mol. In general, all calculated activation energies are close to the predicted values. The MAE for these 13 samples is 3.6 kcal/mol. [b]25%. [c]50%. [d]75%.

The comparison of the calculated and predicted activation energies shows that the designed model overestimates low energy transitions. The start geometries that lead to low transitions are few compared to the number of snapshots that are unsuitable starting points for QM/MM reaction path studies. From the original 150 samples only 9 had energy barriers below 30 kcal/mol. Using the machine learning model 2 out of the 10 frames, thought to be suited, lead to barriers lower than 30 kcal/mol. Therefore, the model allows us to identify relevant frames that will contribute significantly to the exponential average of the reaction barrier. For an accurate estimate of the exponential average, more data points used for training would be required. Improving the predictive model and subsequently calculating the exponential average from all predicted barriers could be an interesting approach to approximate the true activation barrier, which then can be compared to experiments. Overall, we are able to meet our goal to strategically find reactive regions within the MD-trajectory. Using the model, we are able to exclude the majority of frames without needing to calculate them specifically.

CONCLUSIONS

Using simple machine learning techniques, we are able to find reactive periods within the MD trajectory without prior knowledge of the structural factors that govern the reactivity of Sirtuin 5. The applied protocol enables us to identify the structural features that stabilize the transition state, and thus enhance the reactivity.

We found that the cofactor NAD^+ and the substrate SLL have to be located close together and be well aligned; therefore, the compactness of the binding pocket is a prerequisite. At the same time, there has to be sufficient room for nicotinamide, the leaving group, to exit the active site. Configurational changes within NAD^+ and SLL are also connected to the reactivity. The relative position of the nicotinamide to the ribose ring in NAD^+, the orientation of the terminal carboxyl group of SLL and its salt bridge to the neighboring ARG 105 are important structural features. Using measurements of these changes we were able to estimate activation energies with a MAE of 3.6 kcal/mol. For the initial step of the desuccinylation, we found transitions with barriers as low as 26 kcal/mol. We expect that the inclusion of dynamic effects through free energy simulations and even more accurate methods will yield a more reliable transition barrier than found in the scope of this work. These results also support the assumption that the dessucinylation investigated here has a reaction mechanism which is analogous to the deacetylation by Sirtuin 2, which has already been studied in greater detail.[30−32] The straightforward approach we applied here to estimate transition barriers is transferable to any extended system. It greatly simplifies the search for appropriate educt conformations, which significantly influences the outcome of most QM/MM-schemes to model enzymatic reaction mechanisms. The approach is scalable and can be easily customized to meet individual needs, by employing other descriptions for the MM or the QM part, adjusting the number of samples or adding further features.

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.9b00876.

Visualization of the QM region, benchmark: HF-3c versus other functionals, and machine learning model comparison (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: christian.ochsenfeld@uni-muenchen.de.

ORCID

Johannes C. B. Dietschreit: 0000-0002-5840-0002

Christian Ochsenfeld: 0000-0002-4189-6558

Author Contributions

†B.v.d.E. and J.C.B.D. contributed equally to this work

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support by the SFB 1309 "Chemical Biology of Epigenetic Modifications" (DFG) and the DFG cluster of excellence (EXC 114) "Center for Integrated Protein Science Munich" (CIPSM). C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

■ REFERENCES

(1) Ranaghan, K. E.; Mulholland, A. J. Investigations of enzyme-catalysed reactions with combined quantum mechanics/molecular mechanics (QM/MM) methods. *Int. Rev. Phys. Chem.* **2010**, *29*, 65−133.

(2) Mills, G.; Jónsson, H. Quantum and thermal effects in H_2 dissociative adsorption: Evaluation of free energy barriers in multidimensional quantum systems. *Phys. Rev. Lett.* **1994**, *72*, 1124.

(3) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978.

(4) Weinan, E.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 052301.

(5) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Efficient exploration of reaction paths via a freezing string method. *J. Chem. Phys.* **2011**, *135*, 224108.

(6) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **1998**, *108*, 1964−1977.

(7) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(8) Laio, A.; Parrinello, M.; Li, Y.; Zhang, R.; Du, L.; Zhang, Q.; Wang, W. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562.

(9) Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. Enzyme Mechanisms with Hybrid Quantum and Molecular Mechanical Potentials. I. Theoretical Considerations. *Int. J. Quantum Chem.* **1996**, *60*, 89−1200.

(10) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.

(11) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. A practical guide to modelling enzyme-catalysed reactions. *Chem. Soc. Rev.* **2012**, *41*, 3025−3038.

(12) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(13) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58*, 7260−7268.

(14) Hur, S.; Bruice, T. C. The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12015−12020.

(15) Guo, H.; Cui, Q.; Lipscomb, W. N.; Karplus, M. Understanding the Role of Active-Site Residues in Chorismate Mutase Catalysis from Molecular-Dynamics Simulations. *Angew. Chem., Int. Ed.* **2003**, *42*, 1508−1511.

(16) Blank, I. D.; Sadeghian, K.; Ochsenfeld, C. A Base-Independent Repair Mechanism for DNA Glycosylase−No Discrimination Within the Active Site. *Sci. Rep.* **2015**, *5*, 10369.

(17) Roßbach, S.; Ochsenfeld, C. Quantum-Chemical Study of the Discrimination against dNTP in the Nucleotide Addition Reaction in the Active Site of RNA Polymerase II. *J. Chem. Theory Comput.* **2017**, *13*, 1699−1705.

(18) Kreppel, A.; Blank, I. D.; Ochsenfeld, C. Base-Independent DNA Base-Excision Repair of 8-Oxoguanine. *J. Am. Chem. Soc.* **2018**, *140*, 4522−4526.

(19) Roßbach, S.; Ochsenfeld, C. Influence of Coupling and Embedding Schemes on QM Size Convergence in QM/MM Approaches for the Example of a Proton Transfer in DNA. *J. Chem. Theory Comput.* **2017**, *13*, 1102−1107.

(20) Sumowski, C. V.; Ochsenfeld, C. A Convergence Study of QM/MM Isomerization Energies with the Selected Size of the QM Region for Peptidic Systems. *J. Phys. Chem. A* **2009**, *113*, 11734−11741.

(21) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381−11394.

(22) Sadiq, S. K.; Coveney, P. V. Computing the Role of Near Attack Conformations in an Enzyme-Catalyzed Nucleophilic Bimolecular Reaction. *J. Chem. Theory Comput.* **2015**, *11*, 316−324.

(23) Santos-Martins, D.; Calixto, A. R.; Fernandes, P. A.; Ramos, M. J. A Buried Water Molecule Influences Reactivity in α-Amylase on a Subnanosecond Time Scale. *ACS Catal.* **2018**, *8*, 4055−4063.

(24) Neves, R. P. P.; Fernandes, P. A.; Ramos, M. J. Mechanistic insights on the reduction of glutathione disulfide by protein disulfide isomerase. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E4724−E4733.

(25) Lodola, A.; Sirirak, J.; Fey, N.; Rivara, S.; Mor, M.; Mulholland, A. J. Structural Fluctuations in Enzyme-Catalyzed Reactions: Determinants of Reactivity in Fatty Acid Amide Hydrolase from Multivariate Statistical Analysis of Quantum Mechanics/Molecular Mechanics Paths. *J. Chem. Theory Comput.* **2010**, *6*, 2948−2960.

(26) Bonk, B. M.; Weis, J. W.; Tidor, B. Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* **2019**, *141*, 4108−4118.

(27) Schemies, J.; Uciechowska, U.; Sippl, W.; Jung, M. NAD^+-dependent histone deacetylases (sirtuins) as novel therapeutic targets. *Med. Res. Rev.* **2009**, *30*, 861−889.

(28) Parihar, P.; Solanki, I.; Mansuri, M. L.; Parihar, M. S. Mitochondrial sirtuins: Emerging roles in metabolic regulations, energy homeostasis and diseases. *Exp. Gerontol.* **2015**, *61*, 130−141.

(29) Nakagawa, T.; Guarente, L. Sirtuins at a glance. *J. Cell Sci.* **2011**, *124*, 833−838.

(30) Liang, Z.; Shi, T.; Ouyang, S.; Li, H.; Yu, K.; Zhu, W.; Luo, C.; Jiang, H. Investigation of the Catalytic Mechanism of Sir2 Enzyme with QM/MM Approach: SN1 vs SN2? *J. Phys. Chem. B* **2010**, *114*, 11927−11933.

(31) Hawse, W. F.; Hoff, K. G.; Fatkins, D. G.; Daines, A.; Zubkova, O. V.; Schramm, V. L.; Zheng, W.; Wolberger, C. Structural Insights into Intermediate Steps in the Sir2 Deacetylation Reaction. *Structure* **2008**, *16*, 1368−1377.

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

(32) Wang, Y.; Fung, Y. M. E.; Zhang, W.; He, B.; Chung, M. W. H.; Jin, J.; Hu, J.; Lin, H.; Hao, Q. Deacylation Mechanism by SIRT2 Revealed in the 1-SH-2-O-Myristoyl Intermediate Structure. *Cell Chem. Biol.* **2017**, *24*, 339−345.
(33) Du, J.; Zhou, Y.; Su, X.; Yu, J. J.; Khan, S.; Jiang, H.; Kim, J.; Woo, J.; Kim, J. H.; Choi, B. H.; et al. Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. *Science* **2011**, *334*, 806−809.
(34) Case, D.; Betz, R.; Cerutti, D.; Cheatham, T. E.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Homeyer, N.; et al. *AMBER*, 2016.
(35) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.
(36) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284−2295.
(37) Peters, M. B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M. N.; Merz, K. M., Jr. Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF). *J. Chem. Theory Comput.* **2010**, *6*, 2935−2947.
(38) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(39) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(40) Pavelites, J. J.; Gao, J.; Bash, P. A.; MacKerell, A. D. A molecular mechanics force field for NAD^+ NADH, and the pyrophosphate Groups of nucleotides. *J. Comput. Chem.* **1997**, *18*, 221−239.
(41) Walker, R. C.; De Souza, M. M.; Mercer, I. P.; Gould, I. R.; Klug, D. R. Large and fast relaxations inside a protein: Calculation and measurement of reorganization energies in alcohol dehydrogenase. *J. Phys. Chem. B* **2002**, *106*, 11658−11665.
(42) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(44) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(45) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98−103.
(46) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(47) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672−1685.
(48) Sherwood, P.; De Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; et al. QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis. *J. Mol. Struct.: THEOCHEM* **2003**, *632*, 1−28.
(49) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, 134114.
(50) Kussmann, J.; Ochsenfeld, C. Preselective Screening for Linear-Scaling Exact Exchange-Gradient Calculations for Graphics Processing Units and General Strong-Scaling Massively Parallel Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 918−922.
(51) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(52) Rupp, M.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(53) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(54) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(55) Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.
(56) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(57) McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 2010; pp 51−56.
(58) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(59) Hastie, T.; Zou, H. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301−320.
(60) Scheiner, S. Comparison of proton transfers in heterodimers and homodimers of NH_3 and OH_2. *J. Chem. Phys.* **1982**, *77*, 4039−4050.
(61) Scheiner, S. Proton Transfers in Hydrogen-Bonded Systems. Cationic Oligomers of Water. *J. Am. Chem. Soc.* **1981**, *103*, 315−320.
(62) Scheiner, S.; Harding, L. B. Proton Transfers in Hydrogen-Bonded Systems. 2. Electron Correlation Effects in $(N_2H_7)^+$. *J. Am. Chem. Soc.* **1981**, *103*, 2169−2173.
(63) Ryde, U. How Many Conformations Need to Be Sampled to Obtain Converged QM/MM Energies? the Curse of Exponential Averaging. *J. Chem. Theory Comput.* **2017**, *13*, 5745−5752.
(64) Cooper, A. M.; Kästner, J. Averaging Techniques for Reaction Barriers in QM/MM Simulations. *ChemPhysChem* **2014**, *15*, 3264−3269.
(65) Lonsdale, R.; Houghton, K. T.; Żurek, J.; Bathelt, C. M.; Foloppe, N.; de Groot, M. J.; Harvey, J. N.; Mulholland, A. J. Quantum Mechanics/Molecular Mechanics Modeling of Regioselectivity of Drug Metabolism in Cytochrome P450 2C9. *J. Am. Chem. Soc.* **2013**, *135*, 8001−8015.
(66) Sokkar, P.; Boulanger, E.; Thiel, W.; Sanchez-Garcia, E. Hybrid Quantum Mechanics/Molecular Mechanics/Coarse Grained Modeling: A Triple-Resolution Approach for Biomolecular Systems. *J. Chem. Theory Comput.* **2015**, *11*, 1809−1818.
(67) Lonsdale, R.; Reetz, M. T. Reduction of α,β-Unsaturated Ketones by Old Yellow Enzymes: Mechanistic Insights from Quantum Mechanics/Molecular Mechanics Calculations. *J. Am. Chem. Soc.* **2015**, *137*, 14733−14742.
(68) Li, Y.; Zhang, R.; Du, L.; Zhang, Q.; Wang, W. Insight into the Catalytic Mechanism of Meta-Cleavage Product Hydrolase BphD: A Quantum Mechanics/Molecular Mechanics Study. *RSC Adv.* **2015**, *5*, 66591−66597.

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

Supporting Information: "Finding reactive configurations: A machine learning approach for estimating energy barriers applied to Sirutin 5"

Beatriz von der Esch,[†,‡] Johannes C. B. Dietschreit,[†,‡] Laurens D. M. Peters,[†] and Christian Ochsenfeld[*,†]

†*Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany*

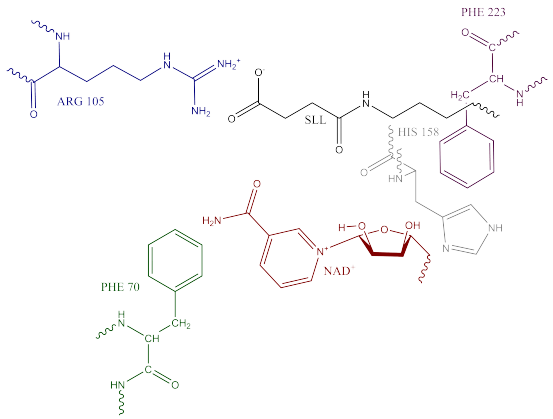
‡*These authors contributed equally to this work*

E-mail: christian.ochsenfeld@uni-muenchen.de

Visualisation

All images of molecular geometries were generated using *VMD*.[1] All plots were produced using the python-packages *matplotlib*[2] and *seaborn*. The chemical structures were drawn with *ChemDraw*.

QM-region

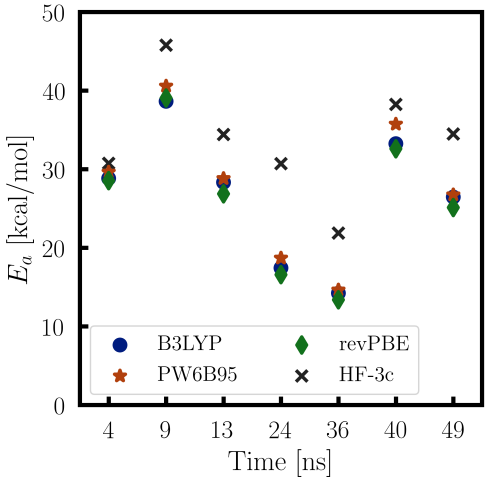


SFigure 1: Visualisation of the QM-region. The substrate SLL is shown in black and the co-factor NAD^+ in red. Additionally, four amino-acids and zero to four water molecules were included.

Besides the two reactants, SLL and the co-factor NAD+, four additional amino-acids and zero to four water molecule were included. Therefore, the number of atoms included in the QM-region varied from 139 to 152. The residues contained in the QM-region are shown in Figure 1. The substrate SLL is shown in black, NAD^+ in red, HIS 158 in grey, ARG 105 in blue, PHE 223 in magenta and PHE 70 in green. The residues were chosen based on proximity to the reactive centers.

Benchmark: HF-3c vs other functionals

To assess the accuracy of the HF-3c/minix for the description of the QM region, seven frames that covered a 25 kcal/mol range were compared to results obtained by higher theory methods. For those frames single point calculations were carried out for the educt and the transition state at the B3LYP-D3/def2-tzvp, revPBE-D3/def2-tzvp, and PW6B95-D3/def2-tzvp level of theory.[3–11] The functionals were selected because of their general use (B3LYP or revPBE) or because they were especially created for kinetic barriers (PW6B95). The activation barriers were calculated from the single point energies. The QM/MM partitioning and all interactions were treated as described in section "QM/MM Simulations".



SFigure 2: Comparison of predicted barrier heights based on the QM/MM adiabatic mapping paths generated with HF-3c. In all cases HF-3c is an upper limit to the barrier height, and thus it consistently overestimates the activation energy. The values on the x-axis show when the frames were picked from the MD-trajectory.

Figure 2 clearly shows that HF-3c is always proportional to the energy barriers estimated with the other methods and consistently overestimates the barrier height. This consistency allows us to use HF-3c/minix to distinguish frames higher and lower barriers, as we do not aim to use it in order to estimate a value comparable to experiment.

Machine Learning Model Comparison

Listed in the Table 1 are the results for the tested regression models. The numerical hyperparameters were determined using 5-fold cross validation. The MAE, RMSE, and R2 value were calculated using 3-fold cross validation.

STable 1: Summary of the tested machine learning models. The mean absolute error (MAE), the root-mean-squared-error (RMSE) and the R2 value for each model are listed. Besides these measures of performance the chosen hyperparameters are given.

Model	Hyperparameters	MAE [kcal/mol]	RMSE [kcal/mol]	R2
Linear Regression		4.28	5.41	-0.06
Descision Tree Regression	max depth=9	5.08	6.91	-0.54
Ridge Regression	$\alpha = 20$	3.57	4.46	0.28
Lasso Regression	$\alpha = 0.1$	3.71	4.59	0.23
Kernel Ridge Regression	$\alpha = 20$, kernel='linear'	3.55	4.44	0.28
Elastic Net Regression	$\alpha = 0.06$, l1 ratio=0.5	3.59	4.46	0.28

References

(1) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(2) Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* **2007**, *9*, 90–95.

(3) Vosko, S.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(4) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(5) Becke, A. D. Densityfunctional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(6) Stephens, P.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(7) Zhang, Y.; Yang, W. Comment on Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1998**, *80*, 890.

(8) Zhao, Y.; Truhlar, D. G. The Design of Density Functionals that are Broadly Accurate for Thermochemistry, Thermochemical Kinetics, and Nonbonded Interactions. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.

(9) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–305.

(10) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*.

(11) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

4.2 Publication II: Exponential Averaging Versus Umbrella Sampling for Computing the QM/MM Free Energy Barrier of the Desuccinylation Reaction Catalyzed by Sirtuin 5

Johannes C. B. Dietschreit, **Beatriz von der Esch**, Christian Ochsenfeld
"Exponential Averaging Versus Umbrella Sampling for Computing the QM/MM Free Energy Barrier of the Desuccinylation Reaction Catalyzed by Sirtuin 5"
Phys. Chem. Chem. Phys. **2022**, *24*, 7723-7731.

Abstract: The computational characterization of enzymatic reactions poses a great challenge which arises from the high dimensional and often rough potential energy surfaces commonly explored by static QM/MM methods such as adiabatic mapping (AM). The present study highlights the difficulties in estimating free energy barriers via exponential averaging over AM pathways. Based on our previous study [v. d. Esch *et al.*, *JCTC*, 2019, **15**, 6660-6667], where we analyzed the first reaction step of the desuccinylation reaction catalyzed by human Sirtuin 5 by means of QM/MM adiabatic mapping and machine learning, we use, here, Umbrella Sampling to compute the free energy profile of the initial reaction step. The computational investigation leads to the conclusion that the NAD^+ transfer, the first step of the deacylation reaction, is highly conserved among all sirtuins and proceeds via an S_N2-type reaction mechanism in SIRT5. In addition, the direct comparison of the extrapolated free energy barrier from minimal energy paths and the computed free energy path from umbrella sampling further underlines the importance of extensive sampling.

Johannes C. B. Dietschreit, Beatriz von der Esch, Christian Ochsenfeld
"Exponential Averaging Versus Umbrella Sampling for Computing the QM/MM Free Energy Barrier of the Desuccinylation Reaction Catalyzed by Sirtuin 5"
Phys. Chem. Chem. Phys. **2022**, *24*, 7723-7731.

PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2022, **24**, 7723

Exponential averaging *versus* umbrella sampling for computing the QM/MM free energy barrier of the initial step of the desuccinylation reaction catalyzed by sirtuin 5†

Johannes C. B. Dietschreit, ‡[a] Beatriz von der Esch ‡[a] and Christian Ochsenfeld *[ab]

Received 2nd November 2021,
Accepted 11th February 2022

DOI: 10.1039/d1cp05007a

rsc.li/pccp

The computational characterization of enzymatic reactions poses a great challenge which arises from the high dimensional and often rough potential energy surfaces commonly explored by static QM/MM methods such as adiabatic mapping (AM). The present study highlights the difficulties in estimating free energy barriers *via* exponential averaging over AM pathways. Based on our previous study [von der Esch *et al.*, *J. Chem. Theory Comput.*, 2019, **15**, 6660–6667], where we analyzed the first reaction step of the desuccinylation reaction catalyzed by human sirtuin 5 (SIRT5) by means of QM/MM adiabatic mapping and machine learning, we use, here, umbrella sampling to compute the free energy profile of the initial reaction step. The computational investigations show that the initial step of the desuccinylation reaction proceeds *via* an S_N2-type reaction mechanism in SIRT5, suggesting that the first step of the deacylation reactions catalyzed by sirtuins is highly conserved. In addition, the direct comparison of the extrapolated free energy barrier from minimal energy paths and the computed free energy path from umbrella sampling further underlines the importance of extensive sampling.

1 Introduction

Post-translational modifications (PTMs) describe the chemical alteration of proteins after their expression. They greatly increase the variety of a cell's proteome by expanding the chemical space of the 20 canonical amino acids and play an important role in, for example, protein activity, cell signaling, or transcription.[1] A frequently modified residue is lysine. Best known is the interplay of lysine acetylation[2,3] and methylation[4,5] fixing its charge state to either neutral or positively charged, especially in histone tails.

Acetylation is one of the possible modifications subsumed under the group of ε-*N*-acylation of lysine. In humans, there are 18 lysine deacylases (KDACs). They can be divided into four classes. Classes I, II, and IV are Zn^{2+}-dependent enzymes; their active site contains a catalytically active zinc ion. Class III KDACs, known as sirtuins, also contain Zn^{2+}, but they are NAD^+-dependent. The catalytic center is located next to an NAD^+-binding Rossmann-fold subdomain, whereas the zinc binding motif is spatially separated and ensures the structural integrity of the enzymes.[6] Sirtuins are the mammalian homologs of the silent information regulator 2 (Sir2), a highly conserved family of proteins found in archaea and eukaryotes.[7,8] There are seven different sirtuin isoforms in mammals (SIRT1-7) that cover a wide range of lysine deacylations. They not only catalyze lysine deacetylation, but also, for example, desuccinylation and demyristoylation.[9,10] In line with their wide range of catalytic activity, sirtuins can be found in several different cell compartments such as the nucleus or the mitochondria,[11] where they are involved in various biological processes.[12,13]

This paper focuses on the catalytic activity of SIRT5, which shows no detectable deacetylation but rather demalonylation and desuccinylation activity.[14] It is located in the mitochondria and its main target is the carbamoyl phosphate synthetase 1 (CPS1).[15] Its active site consists of a hydrophobic pocket with a positively charged arginine (Arg105) at the end. Together with Tyr102, those two residues position the negatively charged end of the dicarboxylic acid modification for removal and have been identified to govern the selectivity of SIRT5.[16] SIRT5 transfers succinyl (and malonyl) to its cosubstrate by cleaving the ribosyl bond in NAD^+ and thereby generating nicotinamide, a natural sirtuin inhibitor,[17,18] and a mixture of 2′- and 3′-*O*-succinyl-ADP-ribose.[14]

[a] *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany*
[b] *Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany. E-mail: christian.ochsenfeld@uni-muenchen.de*

† Electronic supplementary information (ESI) available: Containing details on the HB-R1 interaction, umbrella window placement, choice of force constants, and the effect of the selected bin-size on the FES constructed using MBAR. See DOI: 10.1039/d1cp05007a

‡ These authors contributed equally to this work.

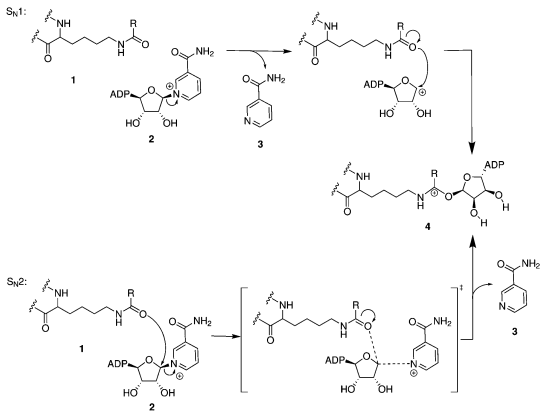


Fig. 1 Reaction scheme for the first of several reaction steps catalyzed by sirtuins. In this initial step the acylated lysine substrate **1** reacts with NAD^+ **2** which results in the α-1′-*O*-alkylamidate intermediate **4** and the release of nicotinamide **3**. Depicted are both theoretically possible reaction types, the stepwise S_N1 and the concerted S_N2. Similar schemes can be found, *e.g.*, in ref. 9, 21, 22, 25, and 26. SIRT1-3,6 act as deacetylases (R = CH_3), SIRT2 can also remove fatty acids (*e.g.*, R = $C_{14}H_{29}$), SIRT5 removes succinyl and malonyl modifications (R = $CH_2CH_2COO^-$, CH_2COO^-), and SIRT4 is an ADP-ribosyltransferase (similar to the product of the reaction step depicted here).[11,27]

A mutagenesis study of the His116 in the active site, modification of NAD^+, and sirtuin crystal structures strongly suggest that no residue in the catalytic pocket takes actively part in the first step of the reaction,[19,20] namely the cleavage of the glycosidic bond between ribose and nicotinamide, and the addition of the substrates amide carbonyl oxygen to ribose, forming an iminium adduct (henceforth called intermediate). Said intermediate was captured by using thioamide substrate analogs.[21,22] The NAD^+ exchange reaction can either proceed *via* an S_N1-like step-wise or an S_N2-like concerted mechanism (see Fig. 1). So far computational studies have only focused on the initial step of the deacetylation reaction in the bacterial sirtuin analogue Sir2Tm[23] and the yeast homolog yHst2.[24] Both concluded that the first step is very likely concerted. Since all following reaction steps are intramolecular rearrangements and proton transfers, it is assumed that the initial step is the rate limiting step.

In our previous publication we have analyzed the first reaction step catalyzed by SIRT5 by means of quantum mechanics/molecular mechanics (QM/MM).[28] We calculated minimal potential energy paths for the first reaction step by means of adiabatic mapping (AM).[29] AM calculations minimise the energy of the system while constraining a collective variable to a specific set of values (in our case the difference between the breaking glycosidic bond and the forming bond between the amide carbonyl oxygen and C1′ of ribose). For these paths we used 150 different reactant configurations which were extracted from a MM-molecular dynamics (MD) simulation of the SIRT5–substrate complex solvated in water. The study connected the configuration of the active site with the calculated activation energy by means of machine learning (see ref. 28). We were able to identify interactions of the substrate (a succinylated peptide) and residues within the active site that could increase or decrease the activation barrier. Due to the complexity of the high-dimensional potential energy surface (PES), the procedure drags the system from reactant to intermediate by visiting many local minima. This leads to a large scattering of the activation barriers as the minimised reactant geometries also correspond to many different local minima.

The effective free energy activation barrier can best be estimated by the exponential average from many of these minimal energy barriers.[30] However, Ryde[31] has cautioned that one needs quite a large number of minimal energy activation barriers as the exponential average is ill-conditioned and converges very slowly. Ryde pointed out that many computational studies based on minimal energies have very large error bars, because of their very low number of calculated paths, so that their conclusions are questionable. Therefore, we study the actual free-energy profile (FEP) for this system as a function of the reaction coordinate, in order to be able to compare the FEP to the results of the previous study, which was one of the biggest in scale to date. This comparison will clearly show if the increase in number of paths by one to two orders of magnitude (compared to those listed in ref. 31) has improved its predictive power. We use umbrella sampling[32] and the same QM/MM setup as in our previous study[28] to explore important regions of configuration space and evaluate the free energy as a function of the reaction coordinate by means of Multistate Bennett's Acceptance Ratio (MBAR).[33–35]

The manuscript starts with a brief introduction into the difficulty of predicting effective energy barriers using exponential averaging and then outlines the equations employed to compute the FEP based on QM/MM umbrella sampling calculations. After reviewing the computational details in Section 3, the obtained FEP of the initial NAD^+ exchange reaction and the resulting free energy activation barrier is compared to the previously determined minimum energy path and exponentially averaged effective barrier.

2 Theory and methods

2.1 The problem of the Ill-conditioned exponential average

If the minimal energy activation barrier of the single adiabatic mapping path i is denoted with $\Delta E_i^\ddagger$, then the average activation barrier for n samples is

$$\langle \Delta E^\ddagger \rangle = \frac{1}{n}\sum_i^n \Delta E_i^\ddagger \tag{1}$$

and its variance

$$\sigma^2 = \langle (\Delta E^\ddagger)^2 \rangle - \langle \Delta E^\ddagger \rangle^2. \tag{2}$$

The exponential average (EA) for this set of energies is then computed as

$$\Delta E_{\mathrm{EA,num}}^\ddagger = -\beta^{-1} \ln\left(\frac{1}{n}\sum_i^n \mathrm{e}^{-\beta \Delta E_i^\ddagger}\right), \tag{3}$$

where $\beta \equiv 1/k_B T$, with k_B being the Boltzmann constant and T the absolute temperature, which is fixed to 300 K within the scope of

this work. As there are several local minima along each degree of freedom (DoF) orthogonal to the reaction coordinate into which the system is minimized, and the number of these DoF is very large in extended biomolecular systems, one can assume that the minimal energy reaction barriers are normal distributed based on the central limit theorem.[36] The exponential average of normal distributed reaction barriers can be calculated analytically using the arithmetic mean $\langle \Delta E^{\ddagger} \rangle$ and the variance σ^2.

$$\Delta E^{\ddagger}_{\mathrm{EA,ana}} = \langle \Delta E^{\ddagger} \rangle - \frac{1}{2}\beta\sigma^2 \tag{4}$$

Ryde[31] performed numerical experiments, drawing random numbers from a normal distribution and computed the EA using eqn (3) and (4). Ryde found that he needed more than an exponentially increasing large number of samples for increasing σ to converge the exponential average within 95% confidence of the known result. This slow convergence of the exponential average is the same which also impedes the computation of absolute free energies. Mean and variance converge much faster than the exponential average, and thus the analytical expression (eqn (4)) using the first and second moment of the underlying distribution is more robust, but can only be employed if the distribution of activation barriers is indeed Gaussian.

2.2 Multistate Bennett's acceptance ratio

In the advanced sampling scheme employed in this manuscript, each single umbrella simulation i (called umbrella window) is associated with a biasing potential B_i, which modifies the original Born–Oppenheimer QM/MM potential energy surface (PES) U_0 to

$$U_i = U_0 + B_i. \tag{5}$$

In order to recover the unbiased data, we use binless WHAM/MBAR[33–35] to estimate the (relative) free energies A_i of each window. The free energy A_i of one window is implicitly defined as a function of all simulation frames and all free energies

$$\mathrm{e}^{-\beta A_i} = \sum_{j}^{S}\sum_{k}^{n_j} \frac{\mathrm{e}^{-\beta B_i(j,k)}}{\sum_{l}^{S} n_l \mathrm{e}^{\beta A_l - \beta B_l(j,k)}}, \tag{6}$$

where S is the number of windows, n_i the number of frames in window i, and $B_i(j, k)$ the value of the biasing function of window i for frame k from simulation window j. Eqn (6) has to be solved self-consistently, but can alternatively be recast into a minimization problem

$$g_i = n_i - \sum_{j}^{S}\sum_{k}^{n_j} \frac{n_i \mathrm{e}^{\beta A_i - \beta B_i(j,k)}}{\sum_{l}^{S} n_l \mathrm{e}^{\beta A_l - \beta B_l(j,k)}} = 0, \tag{7}$$

where all g_i's are zero at the exact solution. The unbiased free energy as a function of the collective variable ξ is recovered using

$$\beta A_0(\xi) = -\ln \sum_{j}^{S}\sum_{k}^{n_j} \frac{\delta(\xi(j,k) - \xi)}{\sum_{l}^{S} n_l \mathrm{e}^{\beta A_l - \beta B_l(j,k)}}. \tag{8}$$

The Dirac delta function is evaluated with finite resolution using an indicator function $\mathbf{1}_{\xi \in [\xi_{\min}, \xi_{\max}]}$, which is equal to one if $\xi \in [\xi_{\min}, \xi_{\max}]$ and otherwise zero. We refer to $\delta\xi = \xi_{\max} - \xi_{\min}$ as the bin width at which we compute the free energy surface.

3 Computational details

3.1 QM/MM setup

As reference geometries for the umbrella simulations, we used the adiabatic mapping path with the lowest activation barrier from our previous study.[28] We chose this particular path in order to show that the barrier is significantly underestimated due to the minimizer identifying a local minimum with a high energy as reactant rather than the lower basin containing most reactant configurations. The same protein residues within the active site, namely Arg105, His158, Phe170, and Phe223, as well as succinyl-lysine (SLL) and the ribose-nicotinamide part of NAD^+ were included in the QM region (113 atoms in total).

The QM/MM separation is shown in Fig. 2. We only modified the location of the QM/MM border compared to our previous study, avoiding a cut through the peptide bonds along the protein backbone and placed it between C_α and C_β. The QM region is described with HF-3c/minix,[37] which has been shown to yield accurate chemistry but elevated energies for transition states.[28] The activation free energy is therefore expected to be higher than one computed with a more accurate quantum mechanical method as, *e.g.*, the one obtained in ref. 23. However, the free energy surface will be qualitatively correct, and we expect that S_N1 and S_N2 can be correctly discerned. The MM parameters for all standard protein residues were taken from AmberFF14,[38] those for NAD^+ from the AMBER parameter database.[39,40] SLL is described with GAFF[41] parameters and AM1-BCC[42] charges. For the zinc finger we use ZAFF[43] parameters. The employed water model is TIP3P.[44] For the full

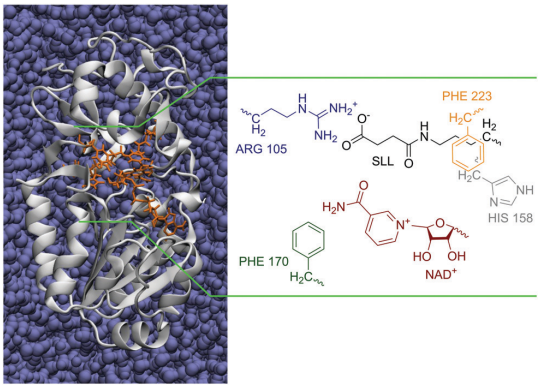


Fig. 2 Visualisation of the QM/MM subsystem division. On the left, the protein (white) is shown embedded in water (iceblue). Arg105, His158, Phe170, and Phe223, as well as the succinyl-lysine substrate and the NAD^+ co-factor are marked in orange. These residues were partially included in the QM-region as defined on the right, with the QM-subsystem shown in detail.

original MM setup see von der Esch *et al.*[28] All QM/MM calculations were performed with our in-house program suite FermiONs++[45–47] which uses the OpenMM 7.3 library[48–50] to evaluate the MM subsystem (Fig. 2).

3.2 Details of the QM/MM umbrella simulations

For the restrained QM/MM-MD simulations we used a Python interface for FermiONs++,[45–47] which allows low-level access to the QM engine. The propagation of atomic coordinates, application of a thermostat, and evaluation of the umbrella potential were done within Python, only for the QM energy and gradient evaluations the PyFermiONsInterface was used.

Each umbrella window simulation consists of three parts: (i) heating, (ii) equilibration, and (iii) production. The system was propagated using the velocity Verlet algorithm[51] and the temperature was controlled using the Langevin thermostat.[52] The initial forces assigned to the active atoms were randomly chosen from the Maxwell–Boltzmann distribution at 1 K. During heating the time step was set to 0.1 fs and no thermostat was used. Every 10 time steps the velocities were rescaled in 1 K increments, reaching 300 K after 3000 time steps.

For equilibration and production, the time step was set to 0.5 fs and the Langevin friction constant to 1 ps^{-1}. For increased speed and stability, we used the fully converged extended Lagrangian method[53] implemented in FermiONs++.[54] The equilibration period was 1 ps long. The production runs were at least 10 ps and a maximum of 20 ps long. Simulations were terminated before the 20 ps limit, if the Mann–Kendall[55–57] test indicated that the mean of the two biased bond lengths had converged, and therefore equilibrium within the window had been reached. Outputs were written every 2 fs.

The umbrella simulations were started from structures taken from the previously obtained minimal energy path with the lowest barrier height.[28] All residues within 10 Å of the QM subsystem were chosen to be active, thereby ensuring that there is always a layer of frozen atoms enclosing the active atoms. This ensures that no molecule can escape into the vacuum surrounding the simulation box, as no periodic boundary conditions were employed.

In order to distinguish between S_N1 and S_N2 reaction type (Fig. 1), the sampling was conducted along two dimensions, the breaking C1′–N bond between ribose and nicotinamide and the forming bond O–C1′ between the carbonyl oxygen and ribose. Hence, each umbrella window i was biased with two harmonic functions

$$B_i(\mathbf{x}) = \frac{1}{2}k_{1,i}(d_1 - d_{1,i})^2 + \frac{1}{2}k_{2,i}(d_2 - d_{2,i})^2, \qquad (9)$$

with $\mathbf{x}$ being a point in configuration space, $d_1 = d$(O–Cl′), $d_2 = d$(Cl′–N), as well as $k_{j,i}$ and $d_{j,i}$ being the force constant and equilibrium bond length of bond j in the biasing potential i. The force constants range from 200 to 700 kJ mol^{-1} $Å^{-2}$, adapting to the slope of the local PES. In total, 106 umbrella simulations were carried out. The force constants and locations of the minima are shown pictorially in Fig. S2 and summarized in Table S1 (ESI†).

3.3 MBAR analysis

As only the relative values of the A_i's calculated with MBAR of each umbrella window are meaningful, the free energy of the first window is set to zero. The starting guess is zero for all windows. Eqn (6) was solved self-consistently; suggested minimization algorithms such as Newton–Raphson[34] or DIIS[58] were not needed. As convergence criterion of self-consistent iterations we used the largest absolute change in the βA_i's per step and $\mathbf{g}^T = (g_1, g_2, \ldots, g_S)$ (g_i's as defined in eqn (7)). Convergence was reached when $\max|\Delta A_i|$ dropped under 10^{-7} and the norm of $\mathbf{g}$ was below 10^{-4}, ensuring that a stable minimum had been found.

The numerical errors of each bin were computed *via* bootstrapping[59] analysis. Ten bootstrapping runs were performed, drawing random frames from each simulation with replacing and then performing ten additional MBAR analyses. The standard deviation between the bootstrap samples of the free energy within each bin was used as statistical error estimate.

4 Results and discussion

4.1 Interaction of SLL with protein residues

The interactions between Arg105, Tyr102, and the succinylated substrate were identified by experiments focussing on the cause of SIRT5 selectivity.[16] Based on our extensive QM/MM sampling of the desuccinylation reaction, we are able to study these interactions as the reaction progresses. It should be noted that while SLL and Arg105 are part of the QM subsystem, the Tyr102 residue was treated at the MM level.

For each frame the hydrogen–acceptor (O–H), hydrogen–donor (H–X), and the donor–acceptor (O–X) distances were measured, as well as the hydrogen bond angle O–H–X (*cf.* Fig. 3). Subsequently, the samples were binned along the bond length difference d(Cl′–O) − d(N–Cl′) (bin width = 0.05 Å). The results are shown in Fig. 4.

The interaction between SLL and ARG105 contains, due to the two-prong nature of these residues, two hydrogen bonds, which are labelled with HB-R1 and HB-R2, respectively. The hydrogen bond between SLL and TYR102 is denoted with HB-Y1. HB-R1 and HB-Y1 involve the same carboxyl oxygen

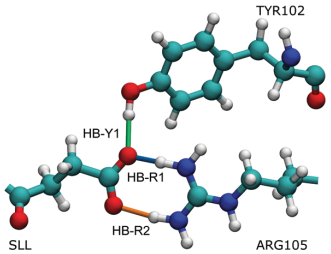


Fig. 3 Hydrogen bond and salt bridge-like interactions between SLL and the protein residues ARG105 (HB-R1 and HB-R2) and TYR102 (HB-Y1). The three bonds correspond to those analyzed in Fig. 4.

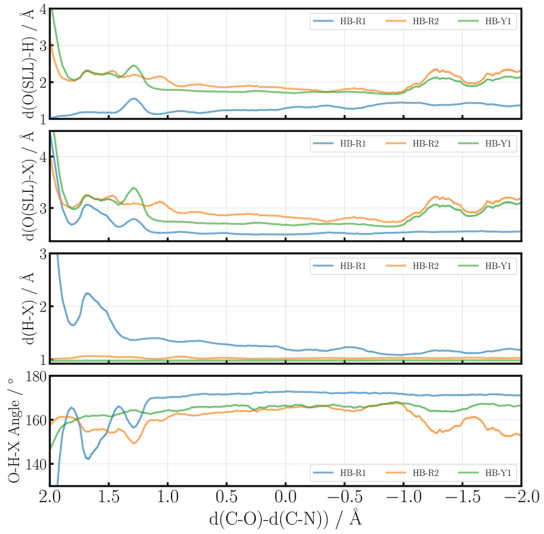


Fig. 4 Evolution of the interactions between SLL and the protein residues ARG105 (HB-R1 and HB-R2) and TYR102 (HB-Y1) over the course of the reaction. The hydrogen bonds are described by the hydrogen–acceptor (O–H), hydrogen–donor (H–X), donor–acceptor (O–X) distances, as well as the hydrogen bond angle O–H–X along the bond length difference d(Cl′–O) − d(N–Cl′) (reaction coordinate).

of SLL. Since SLL rotates in some simulations, the labels were assigned based on the shorter distances measured for each SLL oxygen to all interaction partners. At the beginning of the reaction (left-hand side of Fig. 4 or large values on the abscissa), SLL moves closer to the two protein residues. This is best shown by the decrease of all three distances in the second panel from the top of Fig. 4. Additionally one can see, when comparing HB-R1 and HB-R2, ARG105 and SLL clearly bind one proton each (see opposite behaviour of O–H and H–X distances in the first and third panels). This means that the (in the apo-form) charged residues have changed to a neutral state and are held together by two hydrogen bonds, which is expected in a hydrophobic environment such as this active site.[60,61] The instability of the charged ARG105 in the largely hydrophobic pocket was shown in ref. 16.

As the reaction takes place (d(Cl′–O) − d(N–Cl′) < 1.25 Å), judging by the O–X distance and the O–H–X angle, the hydrogen bond HB-R1 is stronger than HB-R2. It is important to note that especially directly after the onset of the reaction (bond difference between 1 and 2 Å), the interaction is not as stable as before and after the reaction. The carboxyl group rotates over the course of a single simulation window, thereby switching interaction partners. This causes the noticeable, but artificial bumps in this region.

As the nucleophilic attack progresses, the hydrogen in the HB-R1 bond is slowly shifted towards arginine (see Fig. S1, ESI†), finally leading to a salt bridge like state for the product. This change in the nature of the interaction can be explained by the development of a positive charge on the attacking carbonyl oxygen of SLL. Additionally, SLL becomes slightly twisted after this reaction step to accommodate the free nicotine amide. The change at the site of the nucleophilic attack might change the character of the binding site, from a neutral to a generally more charged one, and therefore stabilise the usually stronger salt bridge over the neutral double hydrogen bond.

4.2 Free energy surface of the initial reaction step in SIRT5

The umbrella sampling method allows for easy parallelization during the exploration of the free energy surface. However, we still performed these simulations in consecutive batches, filling in gaps between sampled areas that had been left unexplored by the previous set of simulations. In total 106 umbrella windows were included.

This large number was needed to properly map out not only the very low but also higher energy regions of the FES. The surface obtained here is much steeper and therefore harder to sample than the one of Sir2Tm.[23] This can have several reasons: (i) the QM-method employed here overestimates the energy of stretched bonds and makes all free energy valleys more narrow and (ii) the QM region includes significantly more atoms and therefore describes the interaction between the reactive center and the surroundings differently. In contrast to ref. 23, we included 113 instead of 65 QM atoms, around 65 000 instead of 9000 MM atoms, and sampled cumulatively for 2 ns instead of 720 ps.

The algorithms WHAM[33] or MBAR[34] assume that the input data describe the simulated system in equilibrium and that they are uncorrelated. We calculated the decorrelation times of the biasing potential of each umbrella window, the mean was 23 fs. Hence, the statistical inefficiency[34,62] was 47 fs. Based on these findings, we used data 40 fs apart to construct the free energy surface. For completeness, results based on the full data set can be found in the ESI† (Fig. S4). After determining the relative free energies A_i, we used a bin width of 0.075 Å for both bond lengths to evaluate eqn (8) (see Fig. 5A).

To obtain the minimal free energy path (MFEP) for the nucleophilic substitution, we used Dijkstra's algorithm[63] to find the lowest energy path (shown in Fig. 5B, corresponding to the grey line in Fig. 5A) connecting the lowest point of the reactant basin (d(Cl′–N) < 2 Å and d(O–Cl′) > 2.5 Å) with the lowest point of the intermediate basin (d(Cl′–N) > 2.75 Å and d(O–Cl′) < 1.75 Å).

The position on the free energy surface of the line connecting the educt and product of the investigated reaction step very clearly indicates a concerted mechanism. The energy changes first very little as the carbonyl oxygen approaches, but then the shortening of the (C1′–O)-bond length is directly proportional to the elongation of the (C1′–N)-bond in NAD^+. After the new bond has been formed, the energy decreases slightly further by nicotinamide moving away from the ribose. We can therefore conclude that the reaction mechanism is always of S_N2 type disregarding of whether sirtuins catalyse a deacetylation or desuccinylation. The changes within the active site that lead to the different substrate specificity of the seven sirtuins do not change the overall conserved reaction mechanism.

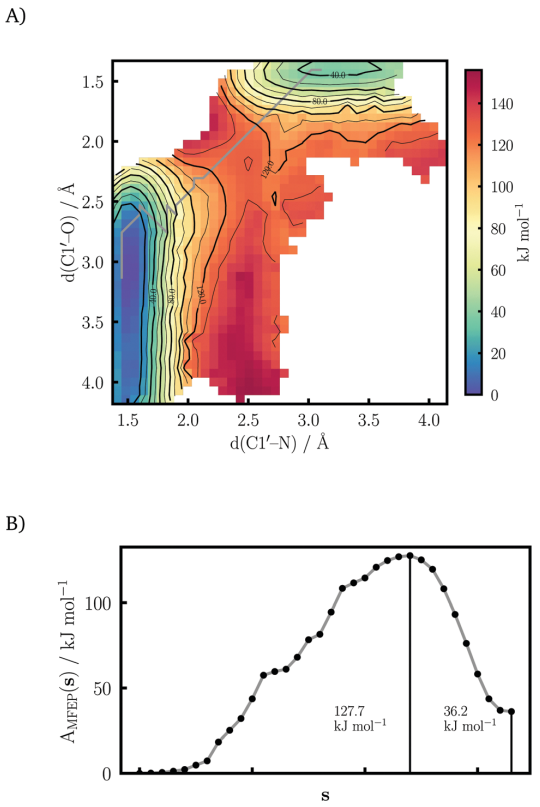


Fig. 5 (A) Free energy surface of the first reaction step catalyzed by SIRT5 calculated with HF-3c/MM. The minimal free energy path (MFEP) connecting the reactant and intermediate state is shown in grey. White areas were not visited during the simulations. Bins, which were not adequately explored, but have at least three fully sampled neighbors, were filled with the mean free energy of the adjacent bins. The original surface is given in the ESI† (Fig. S3). (B) The free energy profile along the MFEP (most likely reaction path *s*), corresponding to the gray line in (A). The difference of well depths and barrier height along the MFEP are given explicitly.

We additionally performed binning along the bond difference (d(Cl′–N) − d(O–Cl′)) to obtain a one-dimensional free-energy profile, which is shown in Fig. 6. The values of the 1D FEP are quite similar to the MFEP obtained from the 2D surface, but it extends beyond the lowest points in the reactant and product basins showing two smooth minima.

As reaction free energy we obtain from our simulations

$$\Delta A = -\beta^{-1} \ln \frac{\int_{\text{Product}} \mathrm{d}\xi \mathrm{e}^{-\beta A(\xi)}}{\int_{\text{Reactant}} \mathrm{d}\xi \mathrm{e}^{-\beta A(\xi)}} = 37.0 \text{ kJ mol}^{-1}. \quad (10)$$

4.3 Free energy paths *vs.* minimal energy paths

The AM path that provided the starting configurations for the umbrella windows on the d(O–Cl′) − d(Cl′–N)-surface, had predicted an activation energy of 91.6 kJ mol^{-1},[28] which is

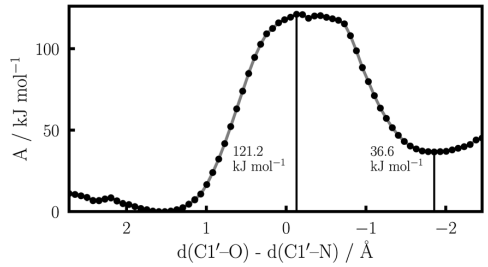


Fig. 6 One-dimensional free energy profile for the initial step of the desuccinylation catalyzed by SIRT5 calculated with HF-3c/MM. The simulation data was binned along the bond length difference of the breaking and forming bond. The educt minimum has been aligned with 0 kJ mol^{-1} and the product minimum lies 36.6 kJ mol^{-1} higher.

around 30–40 kJ mol^{-1} smaller than the approximate energy barrier taken from the MFEP (127.7 kJ mol^{-1}, Fig. 5B) or the 1D FEP (121.2 kJ mol^{-1}, Fig. 6). Both, the AM, as well as, the umbrella sampling simulations were performed at HF-3c/MM level of theory. The AM path is obtained from a sequence of minimizations along a the d(O–Cl′) − d(Cl′–N) distance while the MFEP is extracted from the PES based on MD simulations at 300 K. The much lower AM energy barrier is caused by the path starting off in a local minimum that is already much higher in energy than the majority of configurations forming the reactant basin. We want to underline here that the umbrella windows were started from the nearest points along this path. This means that the non-zero temperature in the umbrella sampling simulations has caused the system to escape the local minima, in which the AM path was stuck, and find the broad reactant basin. In conclusion, the reaction path obtained from umbrella sampling offers a more realistic characterization of the reaction.

This result strongly suggests that predictions of reaction barriers or even reaction mechanisms based on minimal energy paths can be misleading, as has already been hinted at by the strong scattering of minimal reaction barrier values in our previous paper.[28] The exponentially averaged barrier, $\Delta E^{\ddagger}_{\text{EA,num}}$, which combines all 150 paths from the previous study, is also lower than the free energy barrier obtained here (see Table 1). Employing Ryde's considerations,[31] the numerical exponential average has, because of the large variance and comparably low number of frames, a 95% confidence interval of roughly 2000 kJ mol^{-1}. A broad distribution, like the one we obtained, would require billions of paths to achieve chemical accuracy. In light of this, free energy methods seem to be an attractive alternative even though they are usually perceived to be costly for QM/MM studies. The analytical EA, $\Delta E^{\ddagger}_{\text{EA,ana}}$, (based on the Gaussian approximation) is much lower than $\Delta E^{\ddagger}_{\text{EA,num}}$ due to the extremely large scattering of the computed barriers (large variance). The fact that the distribution of the 150 frames is bi-modal calls the applicability of the analytical formula into question, which assumes a normal distribution. Therefore, the value of $\Delta E^{\ddagger}_{\text{EA,ana}}$ for the 150 adiabatic mapping values is regarded as nonsensical.

Table 1 The numerical results of our previous machine learning supported study on the reaction barriers of the first reaction step are summarized by their mean ($\langle \Delta E^{\ddagger} \rangle$, eqn (1)), standard deviation (σ, eqn (2)), numerical exponential average ($\Delta E^{\ddagger}_{EA,num}$, eqn (3)), exponential average assuming a Gaussian distribution ($\Delta E^{\ddagger}_{EA,ana}$, eqn (4)), and the width of the 95% confidence interval ($\Delta\Delta E^{\ddagger}_{EA,num}$, $\Delta\Delta E^{\ddagger}_{EA,ana}$). All numbers are given in kJ mol^{-1}. The values of $\Delta\Delta E^{\ddagger}_{EA}$ are estimated based on the results given in ref. 31

	AM	ML
Samples	150	7501
$\langle \Delta E^{\ddagger} \rangle$	157.4	157.3
σ	22.9	13.4
$\Delta E^{\ddagger}_{EA,num}$	104.0	138.4
$\Delta\Delta E^{\ddagger}_{EA,num}$	$\sim$2000	$\sim$18
$\Delta E^{\ddagger}_{EA,ana}$	52.0	120.8
$\Delta\Delta E^{\ddagger}_{EA,ana}$	$\sim$20	$\sim$3

We also want to stress that non-MD-based methods like AM do not have to be abandoned altogether, as they are well suited for initial exploration. The distribution of energy barriers predicted by our ML model for the entire classical MD simulation is uni-modal and more narrow than the ground-truth distribution, as the fit underestimates high and overestimates low barriers. Its EA result, as given in Table 1, is much closer to the free energy barrier based on umbrella sampling derived in our present study. The low-dimensional ML model yields, to some degree surprisingly, a more realistic barrier estimate. With its few features it cannot incorporate the many DoF orthogonal to the reaction coordinate, and thus effectively averages over them creating a Gaussian distribution one would expect in the high-sampling regime. By its inability to fit the complexity of the biological system, it reduces the noise from the many DoFs and helps to get a more realistic barrier. Finally, it is important to stress again that the free energy barrier reported here is expected to be an upper limit to the true reaction barrier, due to the minimal basis-set employed in HF-3c.

In our previous study, we computed 150 AM paths with around 25 images each. On average 35 optimization steps were needed per image along a path, which accumulates to roughly 131 250 QM/MM energy and force calculations. For the construction of the 2D free energy surface, several million QM/MM-MD-time steps were required. The ML builds on top of the AM results and can therefore not be done without it, but after having performed many AM scans the ML comes at negligible additional cost. While the umbrella sampling is most reliable, it comes at a significantly higher computational cost, therefore working on the improvement of ML techniques based on reaction path scans may provide a cost-effective alternative to determine free energy barriers of extended systems.

5 Conclusions

Through computation of the FEP by means of QM/MM-MD simulations and subsequent evaluation using MBAR we have characterized the initial step of the desuccinylation reaction catalyzed by SIRT5. The results indicate that analogously to the first step of the deacetylation reaction, the NAD^+ transfer step of the desuccinylation reaction is of S_N2 type. This suggests that the differences in the active site, which give rise to varying substrate specificities within the sirtuin enzyme family, do not change the reaction mechanism. Therefore, the first of several desuccinylation reaction steps has now been shown to be independent of sirtuin specificity. A future study has to identify the exact mechanism of the remaining reaction steps.

The computation of the FEP (and of the MFEP connecting reactant and intermediate) allowed us to evaluate the quality of free energy activation barriers estimated by means of exponential averaging. It was shown that the previously computed barrier based on 150 adiabatic mapping pathways underestimated the effective free energy barrier. This calls generally the reliability of reaction barriers and mechanisms based on minimal energy paths into question.

Because of the high computing effort of free energy methods, we are currently still limited to cost-effective methods such as HF-3c or smaller QM regions. The development of ever faster QM codes enables the exploration of increasingly complex system. A complementary approach will be free energy surface reweighing techniques suitable for extended systems, allowing us to extrapolate more accurate results from low-level sampling.

Author contributions

J. C. B. D. and B. v. d. E. conceived the project, performed all calculations, analysed the data, and prepared the original draft. All authors reviewed the manuscript and participated actively in the discussion of the results.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Financial support was provided by ''Deutsche Forschungsgemeinschaft'' (DFG, German Research Foundation) – SFB 1309-325871075 Chemical Biology of Epigenetic Modifications. C. O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

References

1 V. Uversky, *Brenner's Encyclopedia of Genetics*, Academic Press, San Diego, 2nd edn, 2013, pp. 425–430.
2 C. Choudhary, B. T. Weinert, Y. Nishida, E. Verdin and M. Mann, *Nat. Rev. Mol. Cell Biol.*, 2014, **15**, 536–550.
3 M. Schiedel and S. J. Conway, *Curr. Opin. Chem. Biol.*, 2018, **45**, 166–178.
4 K. Zhang and S. Y. Dent, *J. Cell. Biochem.*, 2005, **96**, 1137–1148.
5 E. L. Greer and Y. Shi, *Nat. Rev. Genet.*, 2012, **13**, 343–357.

6 B. D. Sanders, B. Jackson and R. Marmorstein, *Biochim. Biophys. Acta, Proteins Proteomics*, 2010, **1804**, 1604–1616.
7 R. A. Frye, *Biochem. Biophys. Res. Commun.*, 1999, **260**, 273–279.
8 R. A. Frye, *Biochem. Biophys. Res. Commun.*, 2000, **273**, 793–798.
9 J. Schemies, U. Uciechowska, W. Sippl and M. Jung, *Med. Res. Rev.*, 2010, **30**, 861–889.
10 M. Schiedel, D. Robaa, T. Rumpf, W. Sippl and M. Jung, *Med. Res. Rev.*, 2018, **38**, 147–200.
11 W. Dang, *Drug Discovery Today: Technol.*, 2014, **12**, e9–e17.
12 S. Michan and D. Sinclair, *Biochem. J.*, 2007, **404**, 1–13.
13 A. Chalkiadaki and L. Guarente, *Nat. Rev. Cancer*, 2015, **15**, 608–624.
14 J. Du, Y. Zhou, X. Su, J. J. Yu, S. Khan, H. Jiang, J. Kim, J. Woo, J. H. Kim, B. H. Choi, B. He, W. Chen, S. Zhang, R. A. Cerione, J. Auwerx, Q. Hao and H. Lin, *Science*, 2011, **334**, 806–809.
15 R. H. Houtkooper, E. Pirinen and J. Auwerx, *Nat. Rev. Mol. Cell Biol.*, 2012, **13**, 225–238.
16 J. Yu, M. Haldar, S. Mallik and D. K. Srivastava, *PLoS One*, 2016, **11**, 1–26.
17 K. J. Bitterman, R. M. Anderson, H. Y. Cohen, M. Latorre-Esteves and D. A. Sinclair, *J. Biol. Chem.*, 2002, **277**, 45099–45107.
18 M. T. Schmidt, B. C. Smith, M. D. Jackson and J. M. Denu, *J. Biol. Chem.*, 2004, **279**, 40122–40129.
19 J. Min, J. Landry, R. Sternglanz and R.-M. Xu, *Cell*, 2001, **105**, 269–279.
20 M. D. Jackson, M. T. Schmidt, N. J. Oppenheimer and J. M. Denu, *J. Biol. Chem.*, 2003, **278**, 50985–50998.
21 Y. Zhou, H. Zhang, B. He, J. Du, H. Lin, R. A. Cerione and Q. Hao, *J. Biol. Chem.*, 2012, **287**, 28307–28314.
22 Y. Wang, Y. M. E. Fung, W. Zhang, B. He, M. W. H. Chung, J. Jin, J. Hu, H. Lin and Q. Hao, *Cell Chem. Biol.*, 2017, **24**, 339–345.
23 P. Hu, S. Wang and Y. Zhang, *J. Am. Chem. Soc.*, 2008, **130**, 16721–16728.
24 Z. Liang, T. Shi, S. Ouyang, H. Li, K. Yu, W. Zhu, C. Luo and H. Jiang, *J. Phys. Chem. B*, 2010, **114**, 11927–11933.
25 S. Lee, Z. Chen and G. Zhang, *Cell Chem. Biol.*, 2017, **24**, 248–249.
26 P. Hu, S. Wang and Y. Zhang, *J. Am. Chem. Soc.*, 2008, **130**, 16721–16728.
27 E. Fiorino, M. Giudici, A. Ferrari, N. Mitro, D. Caruso, E. De Fabiani and M. Crestani, *IUBMB Life*, 2014, **66**, 89–99.
28 B. von der Esch, J. C. B. Dietschreit, L. D. M. Peters and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2019, **15**, 6660–6667.
29 K. E. Ranaghan and A. J. Mulholland, *Int. Rev. Phys. Chem.*, 2010, **29**, 65–133.
30 A. M. Cooper and J. Kästner, *ChemPhysChem*, 2014, **15**, 3264–3269.
31 U. Ryde, *J. Chem. Theory Comput.*, 2017, **13**, 5745–5752.
32 G. M. Torrie and J. P. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.
33 S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *J. Comput. Chem.*, 1992, **13**, 1011–1021.
34 M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 1–10.
35 B. Roux, *Comput. Phys. Commun.*, 1995, **91**, 275–282.
36 *Central Limit Theorem*, ed. Y. Dodge, Springer New York, New York, NY, 2008, pp. 66–68.
37 R. Sure and S. Grimme, *J. Comput. Chem.*, 2013, **34**, 1672–1685.
38 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.
39 J. J. Pavelites, J. Gao, P. A. Bash and A. D. MacKerell, *J. Comput. Chem.*, 1997, **18**, 221–239.
40 R. C. Walker, M. M. De Souza, I. P. Mercer, I. R. Gould and D. R. Klug, *J. Phys. Chem. B*, 2002, **106**, 11658–11665.
41 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.
42 A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2002, **23**, 1623–1641.
43 M. B. Peters, Y. Yang, B. Wang, L. Füsti-Molnár, M. N. Weaver and K. M. Merz Jr., *J. Chem. Theory Comput.*, 2010, **6**, 2935–2947.
44 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.
45 J. Kussmann and C. Ochsenfeld, *J. Chem. Phys.*, 2013, **138**, 134114.
46 J. Kussmann and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2015, **11**, 918–922.
47 J. Kussmann and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2017, **13**, 3153–3159.
48 M. S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. LeGrand, A. L. Beberg, D. L. Ensign, C. M. Bruns and V. S. Pande, *J. Comput. Chem.*, 2009, **30**, 864–872.
49 V. Pande and P. Eastman, *Comput. Sci. Eng.*, 2010, **12**, 34–39.
50 P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLoS Comput. Biol.*, 2017, **13**, 1–17.
51 W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *J. Chem. Phys.*, 1982, **76**, 637–649.
52 M. Kröger, *Models for Polymeric and Anisotropic Liquids*, Springer-Verlag, Berlin Heidelberg, 1st edn, 2005.
53 A. M. N. Niklasson, P. Steneteg, A. Odell, N. Bock, M. Challacombe, C. H. Tymczak, E. Holmström, G. Zheng and V. Weber, *J. Chem. Phys.*, 2009, **130**, 214109.
54 L. D. M. Peters, J. Kussmann and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5479–5485.
55 H. Mann, *Econometrica*, 1945, **13**, 163–171.
56 M. Kendall, *Rank Correlation Methods*, Charles Griffin, London, 4th edn, 1975.
57 R. Gilbert, *Statistical Methods for Environmental Pollution Monitoring*, Wiley, New York, 1987.
58 C. Zhang, C.-L. Lai and B. M. Pettitt, *Mol. Simul.*, 2016, **42**, 1079–1089.
59 B. Efron, *Ann. Stat.*, 1979, **7**, 1–26.

60 A. Melo and M. Ramos, *Chem. Phys. Lett.*, 1995, **245**, 498–502.
61 A. Melo, M. Ramos, W. B. Floriano, J. Gomes, J. Leão, A. Magalhães, B. Maigret, M. C. Nascimento and N. Reuter, *THEOCHEM*, 1999, **463**, 81–90.
62 W. Janke, *Quantum Simulations Complex Many-Body Syst. From Theory to Algorithms*, John von Neumann Institute for Computing, Jülich, 2002, vol. 10, pp. 423–445.
63 E. W. Dijkstra, *Numer. Math.*, 1959, **1**, 269–271.

Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information: Exponential Averaging Versus Umbrella Sampling for Computing the QM/MM Free Energy Barrier of the Initial Step of the Desuccinylation Reaction Catalyzed by Sirtuin 5

Johannes C. B. Dietschreit,[a,‡] Beatriz von der Esch,[a,‡] Christian Ochsenfeld[*,a,b]

[a]Chair of Theoretical Chemistry, Department of Chemistry,
University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany
[b]Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany
[‡]Contributed equally to this work
[*]E-Mail: christian.ochsenfeld@uni-muenchen.de

Contents

1 Hydrogenbond between Succinyl Group and Arg105 (HB-R1)

The hydrogen bond between the succinyl group and Arg105 located in the active center of SIRT5 changes its character as the reaction progresses. This is discussed in section 4.1. In Figure S1 the distance between the oxygen of the succinyl group and the hydrogen as well as the distance between the nitrogen belonging to Arg105 and the hydrogen are shown. In the reactant state, both are neutrally charged. During the reaction, the hydrogen becomes more and more associated with the Arg105 residues, which results in higher charge separation. The cross-over takes place close to the transition state region, $\mathrm{d(C1' - O) - d(N - C1')} \approx 0.25$ Å.

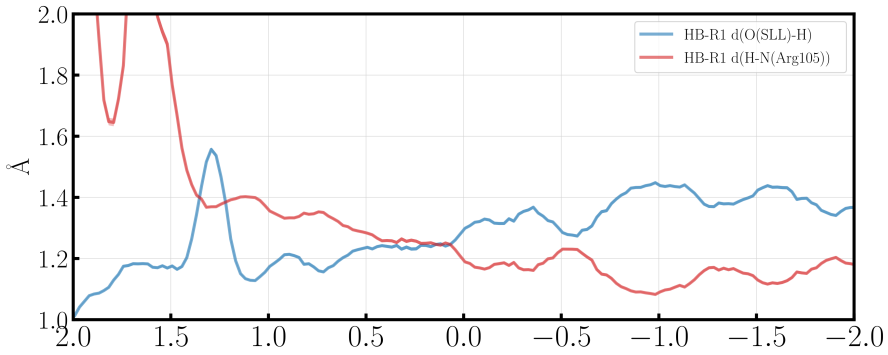


Figure S1: Change of d(O(SLL)-H) and d(H-N(Arg105)),associated with the HB-R1 interaction, during the progression of first step of the desuccinylation reaction.

2 Window Placement and Deviation of Umbrella Window Mean from Bias Potential Minimum Position

The umbrella windows have to be placed so that the space between reactant and product state is seamlessly sampled. The simulations were submitted in batches. Therefore, we were able to set our simulation widows along the becoming more and more apparent MFEP. In addition, several windows were placed at d(C-N) = 2.5, to unequivocally exclude the alternative SN1-reaction type.

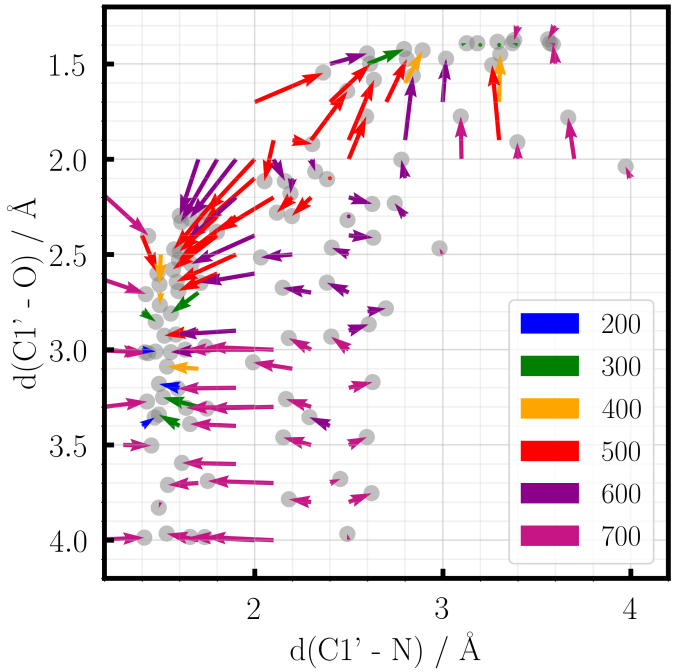


Figure S2: The origin of each arrow indicates the original window placement, and therefore the center of each biasing potential $d_{j,i}$. The arrow's color corresponds to the force constant in kJ mol^{-1} Å^{-2}. The arrow head points to the mean d(C1'-O)/d(C1'-N) sampled in each umbrella simulation.

Figure S2 visualizes the deviation of the mean along $d(\mathrm{C1'-N})$ and $d(\mathrm{O-C1'})$ within each umbrella window and the minimum of the biasing potential. Windows placed near the high energy transition state region or in one of the basins (either reactant or intermediate) show very little deviations between intended window mean (arrow base) and the computed mean (arrow tip). Windows placed in regions, where the free-energy profile changes rapidly, deviate more strongly even if large force constants have been used. This is due to the overestimation of the transition barrier energy by HF-3c and corresponding large forces. In contrast, much lower force constants (160 kJ mol^{-1} Å^{-2}) were used by Hu et al. [**Hu2008**] for the one-dimensional umbrella simulations (24 umbrella windows) of Sir2Tm, where they employed B3LYP/6-31G* and calculated a free energy barrier of only 65.8 kJ/mol for the deacetylation.

Table S 1: List of the exact biasing potential parameters used in the different Umbrella windows. The equilibrium distances d_0 are given in Å and the force constants in kJ/mol Å^2

d_0(C1'-O)	k_{CO}	d_0(C1'-N)	k_{CN}
1.3	700.0	3.4	700.0
1.3	700.0	3.6	700.0
1.4	300.0	3.1	300.0
1.4	300.0	3.2	300.0
1.4	300.0	3.3	300.0
1.4	300.0	3.4	300.0
1.4	700.0	3.6	700.0
1.5	600.0	2.4	600.0
1.5	300.0	2.6	300.0
1.5	700.0	3.6	700.0
1.6	400.0	2.8	400.0
1.7	500.0	2.0	500.0
1.7	500.0	2.4	500.0
1.7	500.0	2.7	500.0
1.7	600.0	3.0	600.0
1.7	400.0	3.3	400.0
1.9	500.0	2.1	500.0
1.9	500.0	2.2	500.0
1.9	500.0	2.3	500.0
1.9	500.0	2.5	500.0
1.9	600.0	2.8	600.0
1.9	500.0	3.3	500.0
2.0	600.0	1.7	600.0
2.0	600.0	1.8	600.0
2.0	600.0	1.9	600.0
2.0	500.0	2.0	500.0
2.0	600.0	2.1	600.0
2.0	600.0	2.3	600.0
2.0	500.0	2.5	500.0
2.0	700.0	3.1	700.0
2.0	700.0	3.4	700.0
2.0	700.0	3.7	700.0
2.1	700.0	1.1	700.0
2.1	500.0	2.0	500.0
2.1	600.0	2.2	600.0
2.1	500.0	2.4	500.0
2.1	600.0	2.8	600.0
2.1	700.0	4.0	700.0
2.2	500.0	1.8	500.0
2.2	600.0	1.9	600.0
2.2	500.0	2.1	500.0
2.2	500.0	2.2	500.0
2.2	500.0	2.3	500.0
2.2	600.0	2.5	600.0
2.3	500.0	1.8	500.0
2.3	500.0	1.9	500.0
2.3	600.0	2.5	600.0

Continued on next page

Table S1 – *Continued from previous page*

d_0(C1'-O)	k_{CO}	d_0(C1'-N)	k_{CN}
2.3	600.0	2.8	600.0
2.4	500.0	1.4	500.0
2.4	500.0	1.8	500.0
2.4	600.0	2.0	600.0
2.4	600.0	2.5	600.0
2.5	400.0	1.5	400.0
2.5	500.0	1.9	500.0
2.5	600.0	2.2	600.0
2.5	600.0	2.5	600.0
2.5	700.0	3.0	700.0
2.6	700.0	1.1	700.0
2.6	500.0	1.8	500.0
2.6	600.0	2.0	600.0
2.7	400.0	1.5	400.0
2.7	300.0	1.7	300.0
2.7	600.0	2.3	600.0
2.7	600.0	2.5	600.0
2.8	300.0	1.4	300.0
2.8	600.0	2.6	600.0
2.9	500.0	1.7	500.0
2.9	600.0	1.9	600.0
2.9	600.0	2.5	600.0
3.0	700.0	1.1	700.0
3.0	700.0	1.2	700.0
3.0	200.0	1.4	200.0
3.0	600.0	1.8	600.0
3.0	700.0	2.0	700.0
3.0	700.0	2.1	700.0
3.0	700.0	2.3	700.0
3.0	700.0	2.5	700.0
3.1	400.0	1.7	700.0
3.1	700.0	2.2	700.0
3.2	200.0	1.6	200.0
3.2	700.0	1.9	700.0
3.2	700.0	2.5	700.0
3.3	700.0	1.2	700.0
3.3	300.0	1.7	500.0
3.3	700.0	2.0	700.0
3.3	700.0	2.1	700.0
3.3	700.0	2.3	700.0
3.4	200.0	1.4	200.0
3.4	300.0	1.6	400.0
3.4	700.0	1.9	700.0
3.4	600.0	2.4	600.0
3.5	700.0	1.3	700.0
3.5	700.0	2.3	700.0
3.5	700.0	2.5	700.0
3.6	700.0	1.9	700.0
3.7	700.0	1.7	700.0
3.7	700.0	2.1	700.0

Continued on next page

Table S1 – *Continued from previous page*

d_0(C1'-O)	k_{CO}	d_0(C1'-N)	k_{CN}
3.7	700.0	2.4	700.0
3.8	700.0	1.5	700.0
3.8	700.0	2.3	700.0
3.8	700.0	2.5	700.0
4.0	700.0	1.1	700.0
4.0	700.0	1.7	700.0
4.0	700.0	2.0	700.0
4.0	700.0	2.1	700.0
4.0	700.0	2.5	700.0

3 Free energy surface of the initial reaction step catalyzed by SIRT5

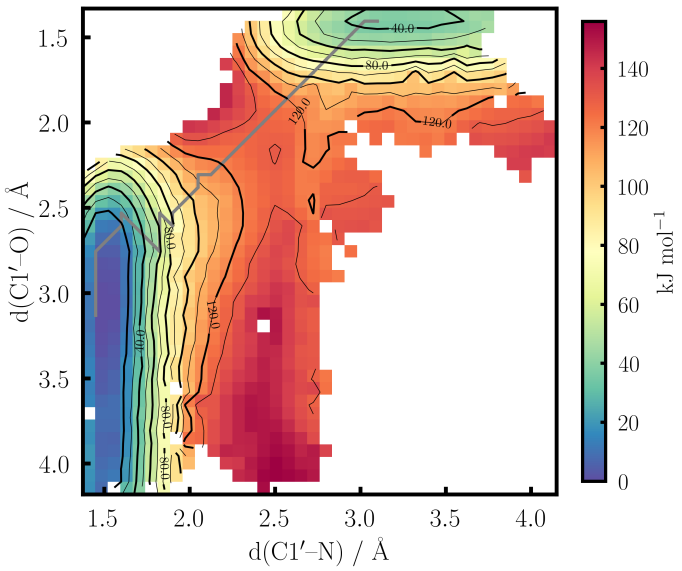


Figure S 3: Original free energy surface of the first reaction step catalyzed by SIRT5 calculated with HF-3c/MM. The minimal free energy path (MFEP) connecting the reactant and intermediate state is shown in grey. White areas were not visited during the simulations.

4 Influence of Bin Width and Sample Number

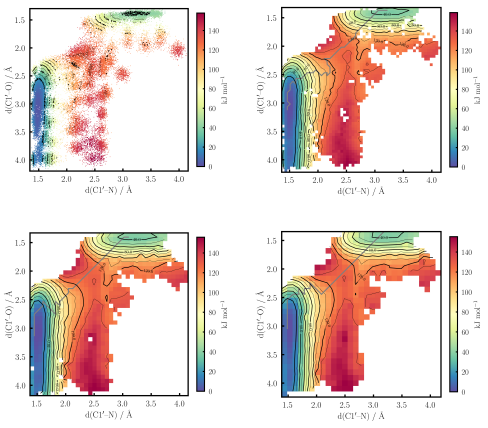


Figure S4: All plots are based on the full data set (data points are 2 fs apart). The bin sizes used for the surfaces are the same along $d(\mathrm{O}-\mathrm{C1}')$ and $d(\mathrm{C1}'-\mathrm{N})$. The sizes are from left to right and top to bottom 0.01 Å, 0.05 Å, 0.075 Å, and 0.1 Å, respectively.

The influence of bin size on the free activation energy as well as the location of the minimal free energy path connecting the two minima on the surface was tested. Figure S4 indicates that there is no significant influence.

4.3 Publication III: Fully Automated Generation of Prebiotically Relevant Reaction Networks from Optimized Nanoreactor Simulations

Alexandra Stan, **Beatriz von der Esch**, Christian Ochsenfeld
"Fully Automated Generation of Prebiotically Relevant Reaction Networks from Optimized Nanoreactor Simulations"
J. Chem. Theory Comput. **2022**, *18*, 6700-6712.

Abstract:
The nanoreactor approach first introduced by the group of T. J. Martínez [Wang *et al.*, *Nat. Chem.*, **2014**, *6*, 1044–1048] has recently attracted much attention because of its ability to accelerate the discovery of reaction pathways. Here, we provide a comprehensive study of various simulation parameters and present an alternative implementation for the reactivity-enhancing spherical constraint function, as well as for the detection of reaction events. In this context, a fully automated post-simulation evaluation procedure based on RDKit and NetworkX analysis is introduced. The chemical and physical robustness of the procedure is examined by investigating the reactivity of selected homogeneous systems. The optimized procedure is applied at the GFN2-xTB level of theory to a system composed of HCN molecules and argon atoms, acting as a buffer, yielding prebiotically plausible primary and secondary precursors for the synthesis of RNA. Furthermore, the formose reaction network is explored leading to numerous sugar precursors. The discovered compounds reflect experimental findings, however, new synthetic routes and a large collection of exotic, highly reactive molecules are observed, highlighting the predictive power of the nanoreactor approach for unraveling the reactive manifold.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

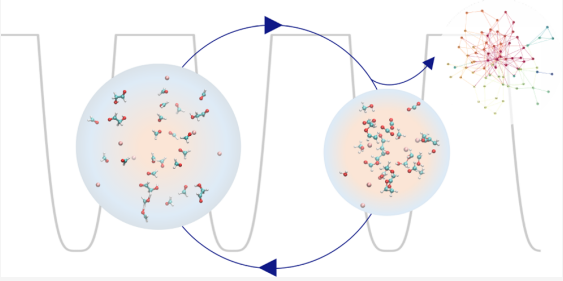
Fully Automated Generation of Prebiotically Relevant Reaction Networks from Optimized Nanoreactor Simulations

Alexandra Stan, Beatriz von der Esch, and Christian Ochsenfeld*

Cite This: J. Chem. Theory Comput. 2022, 18, 6700−6712 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The nanoreactor approach first introduced by the group of Martínez [Wang et al. *Nat. Chem.* **2014,** *6,* 1044−1048] has recently attracted much attention because of its ability to accelerate the discovery of reaction pathways. Here, we provide a comprehensive study of various simulation parameters and present an alternative implementation for the reactivity-enhancing spherical constraint function, as well as for the detection of reaction events. In this context, a fully automated postsimulation evaluation procedure based on RDKit and NetworkX analysis is introduced. The chemical and physical robustness of the procedure is examined by investigating the reactivity of selected homogeneous systems. The optimized procedure is applied at the GFN2-xTB level of theory to a system composed of HCN molecules and argon atoms, acting as a buffer, yielding prebiotically plausible primary and secondary precursors for the synthesis of RNA. Furthermore, the formose reaction network is explored leading to numerous sugar precursors. The discovered compounds reflect experimental findings; however, new synthetic routes and a large collection of exotic, highly reactive molecules are observed, highlighting the predictive power of the nanoreactor approach for unraveling the reactive manifold.



INTRODUCTION

At the core of chemical research is the deepening of the understanding of chemical reactions and exploring the chemical space.[1] Quantum chemistry has so far mainly played a role in characterizing reactions that were previously discovered by experiments, taking more of an explaining and validating rather than an exploratory and discovering role. In recent years, the computational molecular nanoreactor approach was introduced by Martínez et al.[2,3] This method aims to observe novel reactions within reactive ab initio molecular dynamics (MD) simulations. Therefore, a periodic external potential is applied to a collection of encapsulated starting compounds which leads to the contraction and expansion of the available space. In turn, the probability of collisions between the atoms is increased, which results in numerous reaction events.[2,4,5]

The original nanoreactor approach, as pioneered by Wang et al.[2] to enhance reactivity and explore chemical space, was applied to several systems: (1) a mixture of HCN and water,[6,7] (2) a homogeneous collection of acetylene molecules,[2] (3) a Miller−Urey[8,9] type system (mixture of H_2, CH_4, H_2O, NH_3, and CO),[2] and (4) for graphene synthesis via detonation at different oxygen/acetylene ratios.[10]

Recently, Grimme proposed a nanoreactor approach that employs metadynamics[11] as driving force for reactivity on an encapsulated system. He applied this to the thermal decomposition of benzene and ferrocene, ethyne polymerization, oxidation of cyclohexane, a Miller−Urey type system, as well as a model system to study the enzymatic oxidation of testosterone mediated by P450.[12]

Pieri et al.[13] have combined the metadynamics driven nanoreactor approach with nonadiabatic MD, allowing the exploration of photochemical processes. Using the nonadiabatic molecular nanoreactor, they explored the rich excited state chemistry of benzene and were able to confirm the existence of previously described conical intersections.

Alternatively, a large collection of heuristically based methods for the prediction of mono- and bimolecular reactions, focusing on the discovery of possible transition states, has been developed throughout the years.[14,15] First attempts to explore new mechanisms computationally were made by reducing the multidimensional problem of finding transition states to a two-dimensional matrix representation, which allows to generate intermediates at minimal computational cost. This approach was implemented by Broadbelt et al. in NetGen[16] and served as a starting point for numerous other

Received: July 21, 2022
Published: October 21, 2022

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

heuristically based methods, leading to the discovery of new reactions, for example, the Reaction Mechanism Generator by Green et al.[17,18] Later methods have proceeded to incorporate more general chemical principles, for example, electron flow in polar reactions,[19] rather than solely encoded elementary steps for the generation of possible intermediates.

Newer methods aiming to identify transition states and thereby discover reaction mechanisms rely on automated exploration of the potential energy surface by performing high-energy dynamics, as implemented in the TSSCDS routine by Martínez-Núñez,[20−23] which has been recently improved,[24] or by applying external forces as in the well-known adaptive force-induced reaction (AFIR) method[25] for bimolecular reactions. Further developments have been achieved in the Reiher group[26] by discriminating reactive sites based on predefined reactivity measures derived from the electronic wave function, which generate high-energy "reaction structures" for further optimization and IRC calculations.

Furthermore, efforts were made to predict reactions from databases of published reactions using fingerprint methods combined with statistical tools, such as machine learning techniques.[27] One of the earliest algorithms, SYNCHEM, was published in 1990 by Gelernter et al.,[28] and it is based on a vast database, created by inductive and deductive generalization algorithms. Neural networks have also gained attention for the prediction of possible products, as they can be easily trained with literature-known data.[29−32] While these approaches require less computational effort than the dynamic methods presented before, they rely on vast amount of carefully curated input data and specialized training.

In the scope of this work, we revise several aspects of the molecular nanoreactor approach in detail and introduce alternative implementations for the spherical constraint function and the postprocessing. The novel postprocessing provides the user with an automatically generated overview of all obtained molecular species and their abundance, a reaction library and network, as well as an illustrative graphical video description. Furthermore, the introduction of helium and argon atoms as buffer atoms is considered, and the role of the used electronic structure method is examined. Here, we compare the results from RHF/3-21G,[33] GFN2-xTB,[34−36] and PBEh-3c/def2-mSVP[37] simulations. Aiming to provide a comprehensive overview of the approach and the parameter selection, the procedure was tested using various homogeneous systems and is discussed in detail. The optimized procedure is applied to a simple system containing HCN and argon. Here, the formation of relevant primary and secondary precursors for the prebiotic synthesis of RNA[38] such as cyanogen and formamidine is observed. In addition, the formose reaction network is explored,[39,40] yielding several postulated compounds, for example, aldoses, as well as other small reactive species.

THEORY AND METHODS

Electronic Structure Method. Thousands to millions of energy and force evaluations are executed during an MD simulation. Therefore, the chosen ab initio method must be cost-efficient to enable meaningful, yet achievable time scales.

So far, the nanoreactor simulations, as presented by the group of Martínez, employed Hartree−Fock (HF) in combination with small basis sets and GPU-acceleration, as well as high temperatures to increase the kinetic energy and to allow for faster sampling.[2,7,10] Alternatively, Grimme used his highly efficient semiempirical tight-binding method, GFN2-xTB, aiming to optimize the cost-accuracy ratio in metadynamics-based nanoreactor computations.[12]

In this work, we performed high-temperature MD simulations as presented by Wang et al.[2] at the DFT level using the PBEh-3c/def2-mSVP method and compared the results with calculations at the RHF/3-21G and GFN2-xTB levels of theory. The computation of the exact exchange energy for HF and PBEh-3c was accelerated using the sn-LinK method, recently introduced by Laqua et al.[41] To compare computation time and assess the quality of results, we have chosen a series of compounds, namely C_2H_2, HCN, CO, H_2O, and NH_3, and generated homogeneous systems with 156 atoms each.

Initialization Procedure. The initialization of the molecules within a given spherical radius is important to ensure optimal spacing and low forces acting on the atoms in the nanoreactor. Otherwise, convergence problems may be encountered. Furthermore, the initial configuration influences the obtained results. We further elaborate on this matter when discussing our results.

Here, we introduce a novel algorithm for placing a given amount of specified molecules in a sphere based on mapping the Fibonacci lattice on the surface of corresponding subspheres as given by the golden angle.[42] The latter is defined as the angle between two arcs of a circle whose lengths behave to each other according to the golden ratio. This means, that the ratio between the length of the smaller arc and the length of the bigger arc is the same as the ratio between the length of the bigger arc and the length of the full circle. The golden angle φ is defined as ≈137.508° and can be calculated from the golden ratio ϕ as follows

$$\varphi = 2\pi\left(1 - \frac{1}{\phi}\right) = \pi(3 - \sqrt{5}) \tag{1}$$

To avoid crowding, the maximal radius is not given as a variable but calculated based on the amount of molecules to be placed and an interspherical distance given by the user for the subspheres. The interspherical distance corresponds to the distance between two molecules placed on neighboring subspheres at the same angular coordinates. The total amount of molecules to be placed is further divided according to the Fibonacci series to avoid crowding in the most inner shell with two molecules being placed on the surface of the smallest subsphere at the center of the nanoreactor. The molecules are shuffled prior to placing so that many different configurations can be generated, and the obtained setup is independent of the user-specified order. The corresponding algorithm can be found in the Supporting Information.

Spherical Confinement. A spherical constraint function in form of an external potential is used to induce the contraction of the nanoreactor sphere and defines the forces of confinement in the simulation.

This virtual piston can be represented by a step function as previously suggested by Wang et al.,[2] who uses a mass-weighted harmonic potential to generate the forces. The switch between the large and the small radius of the sphere is given by a modified Heaviside step function $f(t)$. However, this results in a harsh transition.

$$V(r, t) = f(t)U(r, r_{max}, k_{max}) + (1 - f(t))U(r, r_{min}, k_{min}) \tag{2}$$

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

where

$$U(r, r_0, k) = \frac{mk}{2}(r - r_0)^2\theta(r - r_0) \tag{3}$$

$$f(t) = \theta\left(\left\lfloor \frac{t}{t_{\text{total}}} \right\rfloor - \frac{t}{t_{\text{total}}} + \frac{t_{\text{exp}}}{t_{\text{total}}}\right) \tag{4}$$

Equation 2 summarizes the rectangular wave potential,[2] where r is the radial coordinate of the atom of interest, r_{max} and r_{min} are the selected maximal and minimal radius of the nanoreactor sphere, respectively, and k_{max} and k_{min} represent the chosen force constants for the mass-weighted harmonic potential to confine the atoms to the corresponding radius. In the effective harmonic potential $U(r,r_0,k)$, r_0 is either r_{max} or r_{min}. The mass-weighting is necessary to ensure equal acceleration for all atoms at a given radial coordinate r. Even though this is not a prerequisite for expansion and contraction, exclusion of the mass-weighting would lead to lighter species accumulating in the center of the nanoreactor sphere. The custom Heaviside step function $f(t)$ in eq 4 takes as an argument a time-dependent expression, which evaluates if the current time step t belongs to the contraction phase t_{con} or to the expansion phase t_{exp}. The total period of a contraction−expansion cycle is given by t_{total}.

Here, we introduce an alternative by using a mass-weighted harmonic potential, which can be combined with a continuous function that smoothly transitions between the two states. We introduce this smooth transition as a cosine wave, where the amplitude controls the radii and t_{total} defines the period. However, due to the much smoother transition, the effective time spent at the two target radii is low compared to the transition process. Hence, there is less time for reactions to occur and for subsequent relaxation, which is a disadvantage with regards to reactivity-enhancement.

$$V(r, r_0(t), k) = \frac{mk}{2}[\max(0, r - r_0(t))]^2 \tag{5}$$

where

$$r_0(t) = r_{\text{min}} + \frac{r_{\text{max}} - r_{\text{min}}}{2}\left[1 + \cos\left(\frac{t}{t_{\text{total}}}2\pi\right)\right] \tag{6}$$

Therefore, we propose a further spherical constraint function which combines smooth transitions, as given by the periodic cosine function, with the literature-known rectangular wave potential (eq 2) and therefore exploits the advantages of both methods. For this purpose, we decided to use a combined sine and cosine function to provide a smooth transition to the minimal radius while also allowing the system to stay at this radius for a longer time than the simple cosine function presented in eqs 5 and 6. To ensure that the time spent in the expanded state is longer than in the contracted state, the symmetry of the function is broken by introducing a cutoff at r_{max}, as given below and shown in Figure 1.

$$V(r, r_0(t), k) = \frac{mk}{2}[\max(0, r - r_0(t))]^2 \tag{7}$$

where

$$r_0(t) = \min\left[r_{\text{max}} + (r_{\text{max}} - r_{\text{min}})\sin\left(\frac{\pi}{2}\cos\left(\frac{t}{t_{\text{total}}}2\pi\right)\right), r_{\text{max}}\right] \tag{8}$$

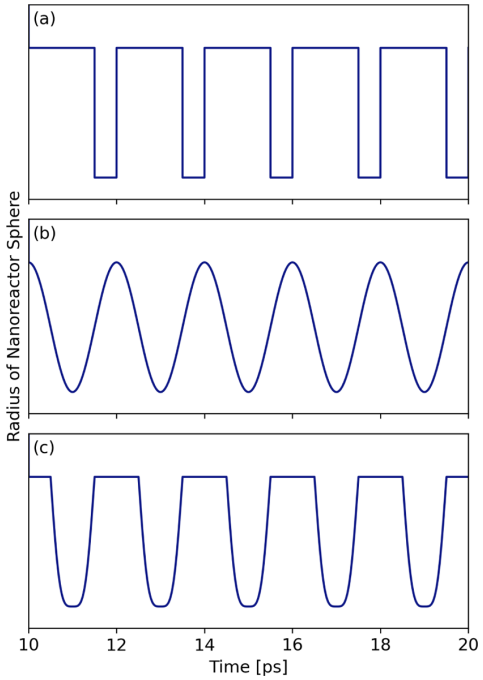


Figure 1. Time evolution of the three implemented types of spherical constraint functions: (a) rectangular wave function,[2] (b) cosine wave, and (c) smooth-step function. In all cases, an external mass-weighted harmonic potential is imposed upon contraction of the nanoreactor sphere.

Fragment Recognition. To process the nanoreactor simulations, the molecular species in each step have to be defined and isolated. Previously, interatomic distances were used to distinguish between molecules.[2,6,7,10] However, standard interatomic distances vary greatly between atom types and hybridization states, as shown in Table 1.

Table 1. Experimentally Determined Standard Bond Distances Given in Å for the Most Frequent Elements in Organic Systems[44]

	H	C	N	O	S
H	0.741	1.099	1.012	0.967	1.345
C		1.530	1.484	1.432	1.809
N			1.425	1.463	1.710
O				1.208	1.432
S					2.048

This problem has been addressed by Hutchings et al.,[43] suggesting the use of a bond-order time series. While imposing a fixed upper threshold on a Mayer bond order time-series, defined as the mean value of the oscillating time-series, has proven to not be reliable enough due to high dependence on the bond type, taking the first derivative of the bond order time-series provides sharp peaks, based on which reaction events could be defined.

The measure used to determine which atoms belong to a fragment should add as little computational cost as possible to the simulations, while at the same time being general. We suggest the use of Wiberg bond orders (WBOs)[45] instead of

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

the initially used covalent interatomic distances as basis for the fragment identification.

WBOs[45] are calculated as the sum over the squared orthonormalized density matrix elements $P_{\mu\nu}$ and describe covalent bonds between atoms

$$W_{AB} = \sum_{\mu \in A} \sum_{\nu \in B} P_{\mu\nu}^2 \tag{9}$$

They represent the most simple type of bond orders described so far and are a special case of the Mayer bond orders[46] (eq 10), which can be computed using nonorthogonalized matrices, as they directly employ the overlap matrix S.

$$M_{AB} = 2 \sum_{\mu \in A} \sum_{\nu \in B} (PS)_{\mu\nu} (PS)_{\nu\mu} \tag{10}$$

Based on the calculated WBO matrix, the obtained molecules can then be reconstructed by imposing a minimal threshold to define a bond. This absolute threshold was chosen as $\geq$0.5, based on the definition that a bond should have a WBO of 1,[45] and it does not depend on the bond type. However, cases were encountered where atoms could not be assigned based on this definition due to their WBOs not exceeding 0.5 to any other atom. For these cases, a second condition has been implemented based on interatomic distance. Herein, first, all interatomic distances to the unassigned atom are calculated; then, the standard deviations to the default bond lengths (given in Table 1) are computed. Based on the lowest standard deviation, the atom is assigned to a fragment (list of atom indices). By not employing fixed thresholds for bond lengths but relative deviations, the procedure remains generally applicable. The use of bond orders solves the problem of high-amplitude molecular vibrations, and spurious species[2] do not occur in first place, making the use of a Hidden Markov Model (HMM) in the evaluation procedure superfluous.

Using the Python3 library pandas,[47,48] the gathered information is stored in a data frame, which is the starting point for all further analysis. A function, based on the RDKit module,[49] has been developed to compute the SMILES[50,51] string starting from the stored WBO matrix and the on-the-fly computed fragments, which are represented by grouped atom indices. SMILES strings provide information about the connectivity of the elements and can be easily converted to chemical sum formulas and molecular structures.[50,51] During this second step of molecule parsing, charges are added based on predefined valence rules, if necessary. The stored bond order matrix of the fragment is used to construct a `mol` object, which then yields a correct SMILES. Therefore, the first step of grouping atoms into list of atoms representing the found fragments is making sure that all atoms have been assigned, while in the second part, the WBOs are used to determine the presence of covalent or ionic bonds. The RDKit package is further used to print the molecular structures of the encountered species on a grid, providing a comprehensive visual summary and allowing for quick interpretation of the results. Furthermore, several visual aids are automatically generated, that is, a continuous bar plot, providing an overview over the events during the simulation, an automatically generated video of the trajectory, where species are color-labeled based on their SMILES, as well as a network for a detailed analysis of the (new-)found reaction paths.

Reaction Events and Network Construction. To identify reaction events, only time steps at the end of the expansion are considered to allow the molecules to relax after the contraction has taken place. A reaction is detected if an atom is assigned to a different SMILES in a consecutive expansion time step, which is defined as the product time step.

When a reaction event is identified, an iterative procedure begins used to find all molecules participating in the transformation. While the cumulative collection of atom indices of the products is unequal to that of the reactants, we iterate through all fragments searching for the molecules containing unmatched atom indices. When a fragment is found containing a missing index (1) the fragment is added either to the reactants or products depending on the time step at which it has been found, (2) the SMILES of the fragment is stored, and (3) the corresponding set of atom indices is updated. This search is conducted bidirectionally. The loop stops when all atoms of both reactants and products have been assigned and a stoichiometrically correct reaction has been written. Each identified reaction is considered only at the time step it has first occurred.

The adjacency matrix for the resulting reaction network is generated by looking at the obtained SMILES and the corresponding atoms. Each node represents a unique SMILES. An edge is defined between two SMILES which have at least one mutual atom, meaning they are involved in a chemical transformation. To avoid creating false edges between molecules based only on mutual atoms, the network is constructed in a stepwise fashion.

The reaction network is generated with NetworkX,[52] and the underlying information, the full list of reactions, is stored as a JSON file to allow for further graph analysis. JSON[53] is a data interchange format which provides a facile way to store and share complex data types across different programing languages. The nodes of the network are color-coded based on the time step at which they first occur, allowing us to retrace the chronology of the events in the simulation.

Introduction of Buffer Atoms. One of the advantages of the molecular nanoreactor as compared to rule-based and coordinate-driven exploration methods is its unbiased sampling, which also allows for the discovery of multimolecular transition states.[14] In this way, novel reaction paths have been reported where numerous molecules concomitantly and actively participate in a chemical transformation.[2,7]

However, transition states involving more than three molecules are rather unlikely under physical conditions. To avoid these nonphysical reactions, we propose adding helium or argon atoms to the simulation system in small amounts. These serve as buffer atoms and are expected to be strongly inert. In addition, they are helpful for assessing the overall reasonableness of the reactor design: if the inert noble gas atoms start to considerably participate in reactions, the imposed forces and parameters are deemed as inadequate to provide meaningful results that reflect the true chemical reaction space.

Computational Details. The RHF and DFT calculations were conducted with the program package FermiONs++[54,55] and the LibXC library.[56] The acceleration procedure for calculating exact exchange sn-LinK[41] and the resolution-of-the-identity for the Coulomb integral (RI_J) were used.[57] For DFT, the gm3 grid was employed, and gm2 was used for sn-LinK. The SCF convergence criterion was defined as the root mean square (RMS) of the FPS commutator, and it was set to 10^{-6} a.u. However, it was lowered temporarily for a maximum of five consecutive steps to 10^{-5} a.u. if convergence

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

during the MD simulations could not be achieved otherwise. For the GFN2-xTB simulations, our in-house MD engine was interfaced to the `xtb` package.[36] The SCF convergence criterion and the electronic temperature were set to default values, 10^{-6} a.u. and 300 K, if not stated otherwise. The initialization in the nanoreactor sphere was done using preoptimized molecules at the PBEh-3c/def2-mSVP level of theory.

Screening and Application Setups. For all parameter tests, the system consisting of 39 acetylene molecules (156 atoms) using a rectangular wave spherical constraint (see Figure 1a) (T_{target} = 2000 K, γ = 7 ps^{-1}, k_{min} = 0.5 kcal mol^{-1} Å^{-2}, k_{max} = 1.0 kcal mol^{-1} Å^{-2}, r_{min} = 8 Å, r_{max} = 14 Å, t_{con} = 0.5 ps, and t_{exp} = 1.5 ps) was selected as reference, and the given parameters were varied. GFN2-xTB was used to compute the energies and forces during the MD simulation to keep the computational effort at a minimum.

To investigate the differences between the three spherical constraint functions, results from five simulations with different initial configurations for each constraint type were averaged. As a measure of stability, the mean and standard deviation of the temperature and pressure were assessed. The reactivity was evaluated both qualitatively in terms of the chemical nature of the observed species and quantitatively by the number of unique species obtained on the automatically generated molecular grid.

To study the effect of different electronic structure methods on the outcome, calculations were performed using GFN2-xTB, RHF/3-21G, and PBEh-3c/def-mSVP on simple homogeneous test systems. As a starting point, the system consisting of 39 acetylene molecules was chosen. The number of atoms in the simulations (156 atoms), as well as all settings, were then kept constant to avoid introducing biases besides the different electronic structure method and change of spherical constraints. In addition, further homogeneous systems were considered, consisting of HCN, CO, H_2O, and NH_3 molecules.

To test the effect of buffer atoms, simulations of HCN with helium and argon atoms were performed with the smooth-step spherical constraint and compared to the HCN-only equivalent (T_{target} = 2000 K, γ = 7 ps^{-1}, k = 1.0 kcal mol^{-1} Å^{-2}, r_{min} = 8 Å, r_{max} = 14 Å, t = 2.0 ps). Different percentages of added helium and argon atoms, ranging from 5 to 25%, were screened to determine the optimal amount of buffer atoms for nanoreactor simulations. The total number of atoms was kept constant to simulate the same compression degree. To reduce biases introduced by the initial configuration, all numerical results were averaged from a total of five simulations each.

Furthermore, simulations of HCN with argon buffer were performed and analyzed with regards to the presence of prebiotic primary and secondary RNA precursors. Lastly, the formose reaction was investigated starting from systems containing formaldehyde and glycolaldehyde in a ratio of 4:1 and argon buffer atoms.

All tests regarding the effect of buffer atoms, the HCN/Ar, and formose/Ar applications were performed with GFN2-xTB, along with the smooth-step spherical constraint function and its optimized parameters. A complete overview on the used parameters both for the screening and applications systems is provided in the Supporting Information (Tables S1−S12).

Postprocessing. After conducting the MD simulations, the trajectories and bond order files were processed automatically, and a set of visual representations was generated comprising a grid of the resulting species provided by the `MolsToGridImage` function of the RDKit Python3 package, a continuous bar plot for a first overview of the reaction events, a movie of the trajectory generated with PyMOL and OpenCV, as well as an interactive reaction network constructed using NetworkX, accompanied by a list of reactions using SMILES.

Here, it should be noted that consecutive intermediate monomolecular transformations were excluded from the network representation to decongest these and facilitate evaluation of the results.

RESULTS AND DISCUSSION

Simulation Parameter Tuning. Previous applications presented in the literature[2,7,10] revealed that the obtained results and the stability of the nanoreactor simulations heavily depend on the chosen spherical constraint and employed settings. Therefore, we provide a systematic study to investigate these effects.

Langevin Thermostat. To achieve expressive results within feasible simulation times for the nanoreactor simulations, the temperature must be kept high to increase reactivity and, therefore, speed up reaction events. Different target temperatures T_{target} were tested (500, 1000, 3000, and 4000 K) and compared to the most frequently used temperature of 2000 K, while all other parameters were kept as listed above. The simulations run at 3000 and 4000 K provided the greatest variety of molecular species, but the temperature and pressure throughout the simulation revealed that the thermostat was not able to handle the highly increased kinetic energy after 200 ps, which resulted in immense fluctuations of both quantities. This was not observed with lower target temperatures.

Figure 2a shows the obtained number of molecular species versus simulation temperatures. To avoid distortion of the results by single outliers, the interquartile range (IQR) method[58] was used before determining the mean and standard deviation of the temperature. With increasing thermostat temperature, the number of unique species highly increases until a saturation is reached, here at about T_{target} = 2000 K. A further increase in temperature leads to the aforementioned fluctuations and instability of the simulation. The obtained results with different friction constants and a target temperature of 2000 K are depicted in Figure 2b. As expected, the friction constant plays an important role in enhancing the reactivity and the variety of obtained products throughout the simulation due to increasing the Brownian motion in the context of the Langevin thermostat. While a higher friction constant has a positive effect on the reactivity, an increase in temperature over ∼2500 K is unfavorable due to higher instability of the MD simulation.

From a qualitative point of view, the abundance in the obtained species switches from macrocycles to long chains when increasing temperature and from polyunsaturated chain molecules to increasingly complex aromatic cycles for higher friction constants.

Spherical Constraint Function. The choice of the parameters for the spherical confinement plays an important role for the outcome. This includes selecting appropriate minimal and maximal radii for the nanoreactor sphere, as well as adjusting the periodic length for the contraction−expansion cycles and the strength of the applied force constant(s).

Variation of the radii for the rectangular wave (compare eq 2) revealed that small radii favor (poly-)cyclic molecules, while there is a tendency for the formation of chain polymers when the atoms are given more room to propagate. Experimenting

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

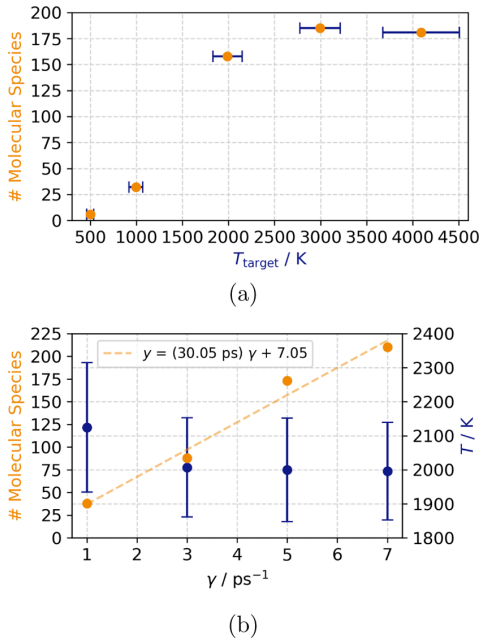


Figure 2. Effect of the Langevin thermostat parameters, (a) target temperature T_{target} (γ = 7 ps^{-1}) and (b) friction constant γ, on the outcome of computational nanoreactor simulations. For the temperature, shown in dark blue, standard deviations and the arithmetic mean are given. Target temperatures T_{target} of 500, 1000, 2000, 3000, and 4000 K were tested. For the comparison of the friction constants, T_{target} was set to 2000 K. As the friction constant is increased, the fluctuations of the temperature T throughout the simulations decrease, and the variety of observed species, shown in orange, increases. The linear fit between molecular species and the friction constant has an R^2 score of 0.98, highlighting this relation. Outliers have been excluded from the statistical treatment according to the IQR[58] method.

with different time periods for the contraction−expansion cycles led to the conclusion that longer contraction periods are favorable for the reactivity, as expected, but the variety of obtained molecular species decreases for the same total simulation time, which in turn leads to the discovery of fewer reactions at same computational cost. In addition, we found that the expansion should last longer than the contraction to allow for the molecules to relax.

Finally, the influence of the force constant of the external harmonic potential was investigated, and here, the behavior of both temperature and pressure was analyzed as an indicator of simulation stability. The results are depicted in Figure 3, where contraction periods are underlined in light blue. An increase in the standard deviation due to more fluctuations in both temperature and pressure when choosing higher force constants was found. The bottom subplot in both figures indicates no advantage to choosing different k_{min} and k_{max} for the rectangular wave constraint. Significant peaks in temperature and pressure were observed during contraction periods regardless of the employed setting.

Therefore, our goal was to introduce a milder function for the spherical confinement, which should provide similar results, while also reducing the number of necessary parameters. Besides the rectangular wave, here, a pure cosine function was tested, which led to rather poor results as the system only briefly visits the contracted state, and thereby, the reactivity is tremendously decreased. As a consequence, we sought to combine the smooth behavior of the cosine function with longer times spent in the contracted state. This goal has been achieved in the form of the sine of a cosine function presented in eqs 7 and 8.

As shown in Figure 4, the obtained number of distinct molecular species highly decreases when using the cosine wave constraint. The rectangular wave function and the newly introduced smooth-step constraint provide the same variety in terms of obtained species, but the latter exhibits less fluctuations in the measured pressure, which is because of the milder switch between the expanded and contracted state of the sphere. This in turn leads to more reproducible simulations and increased stability. Therefore, the smooth-step mass-weighted function represents a good alternative for performing nanoreactor simulations of reasonable length under milder conditions.

It should be further noted that there is a dependency of the simulation outcome on the starting geometry. This can be evaluated in terms of the obtained amount of unique species on the grid. For the simulations carried out to assess the role of the spherical constraint used to plot Figure 4, fluctuations ranging between −75 and +64% from the presented mean number of species were observed. Therefore, it is recommended to use several simulations with different initial configurations for applications.

Comparison of Electronic Structure Methods. In the following, we investigate the influence of different levels of theory for producing a meaningful nanoreactor simulation and to avoid accumulation of nonphysical molecular structures while maintaining the computational cost on a reasonable scale.

All simulations were performed with the three available options for the spherical confinement. An initial configuration was generated for each system and used in all simulations of this species to assess the effect of the chosen electronic structure alone, while avoiding any deviations that might result from varying initial arrangements. The relative number of obtained species to the simulation length for each setup is summarized in Table 2.

For the acetylene systems, most distinct species were obtained for the cosine and rectangular wave constraint using PBEh-3c/def2-mSVP, while in the case of the smooth-step function, GFN2-xTB performed best. However, in the case of RHF/3-21G and PBEh-3c/def2-mSVP, mostly small cycles with little experimental relevance could be identified, which was not the case for GFN2-xTB, where a great variety of polymers and complex structures were found. Overall, for acetylene, most species were observed using the smooth-step function for confinement and GFN2-xTB, while out of a qualitative point of view, all constraints delivered a variety of polymerization products and (condensed) cyclic molecules, among which ethylene, unsaturated cyclobutane derivatives, along with cyclohexene derivatives, and allenes could be identified. The H_2O and NH_3 simulations were completely inert lacking even proton transfers. However, the ability of GFN2-xTB for describing proton transfers in this context was tested and confirmed through further simulations containing protonated water and ammonia molecules besides their uncharged counterparts. Against chemical intuition, the CO systems exhibited the greatest reactivity out of all tested homogeneous collections when described using GFN2-xTB. A

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

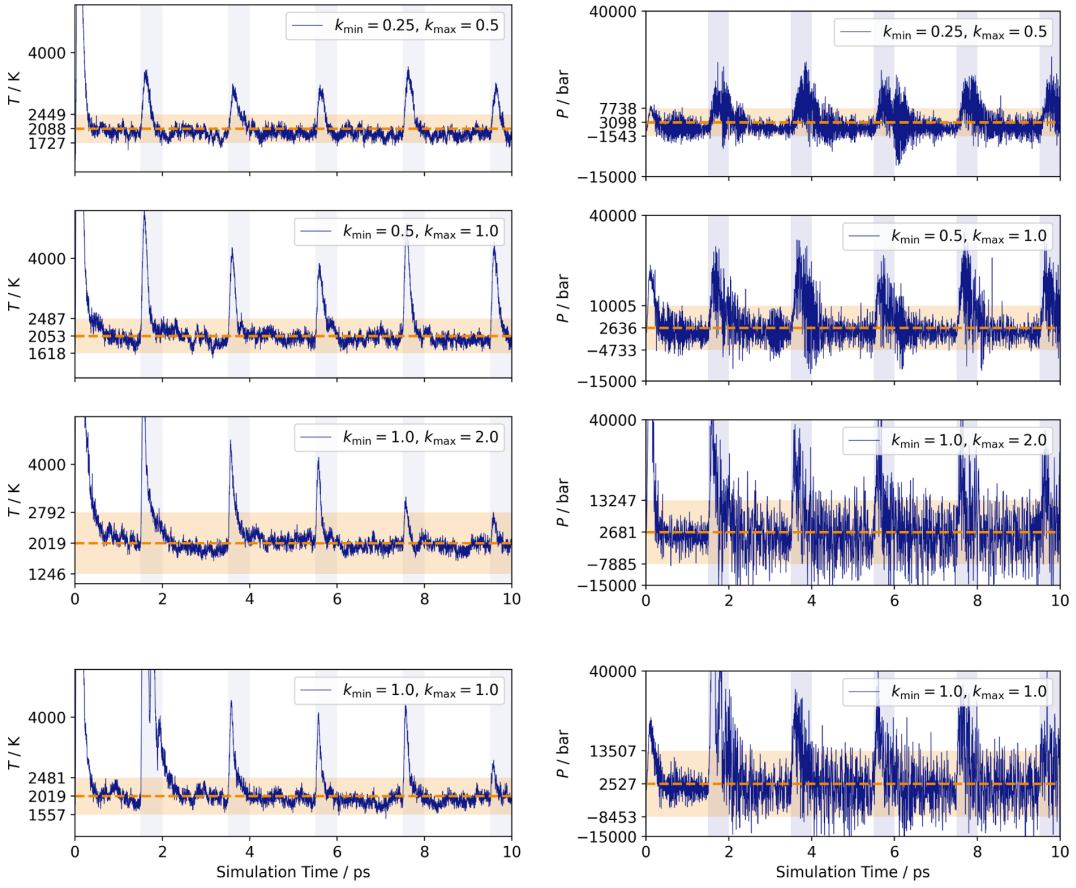


Figure 3. Fluctuations of temperature (a) and pressure (b) in the first 10 ps of a nanoreactor simulation using the rectangular wave function employing different force constants for the harmonic potentials, confining the system to the minimum and maximum radii (T_{target} = 2000 K and γ = 7 ps^{-1}). The means are shown by a dashed line, and the standard deviations are indicated by the areas shaded in orange. The areas shaded in light blue represent contraction phases. All force constants are given in kcal mol^{-1} $Å^{-2}$.

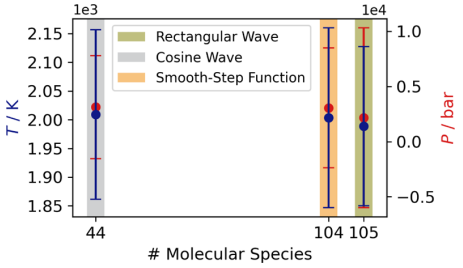


Figure 4. Behavior of temperature and pressure in acetylene simulations using the three different spherical constraints: rectangular wave in olive, cosine wave in grey, and the smooth-step function in orange. The mean and standard deviation are shown as a function of the used spherical constraint and obtained mean number of molecular species. All shown values have been obtained as a mean from five simulations with different initial configurations (T_{target} = 2000 K, γ = 7 ps^{-1}, k_{min} = k_{max} = 1.0 kcal mol^{-1} $Å^{-2}$, r_{min} = 8 Å, and r_{max} = 14 Å). Outliers have been excluded from the statistical treatment according to the IQR[58] method.

Table 2. Number of Molecular Species per 100 ps Simulation Time for Homogeneous Systems of Acetylene, Cyanhydric Acid, Carbon Monoxide, Water, and Ammonia Using Different Functions for Spherical Confinement and Varying the Electronic Structure Method[a]

# distinct species/100 ps	C_2H_2	HCN	CO	H_2O	NH_3
Cosine Wave					
GFN2-xTB	29	55	79	0	0
RHF/3-21G	30	2	3	2	4
PBEh-3c/def2-mSVP	39	29	3	3	2
Rectangular Wave					
GFN2-xTB	24	150	37	0	0
RHF/3-21G	15	2	23	3	1
PBEh-3c/def2-mSVP	34	21	13	6	2
Smooth-Step Function					
GFN2-xTB	59	96	85	0	0
RHF/3-21G	4	3	2	2	4
PBEh-3c/def2-mSVP	38	8	7	3	2

[a]A relative representation was chosen for better comparison due to different simulation lengths. Absolute numbers are given in Table S13.

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

great variety of unexpected polymerization products and cyclic species were isolated which points to a poor description of the electronic structure of the CO molecule at this level of theory. The cyanhydric acid simulations were evaluated with respect to possible primary and secondary RNA precursors. Here, GFN2-xTB performed best as complex reaction networks leading to heterocyclic species, as well as to prebiotic precursors, were formed.

In contrast, the simulations run with RHF/3-21G and PBEh-3c/def2-mSVP were overall less reactive, while the computational effort was 10−20 times higher. However, using these two methods, proton transfers were observed for the simulations containing H_2O and NH_3 along with a few dimerization reactions in contrast to the GFN2-xTB simulations where such processes had to be confirmed through the specially designed setups mentioned above. In the context of reaction path discovery for the C_2H_2 and HCN systems, both RHF/3-21G and PBEh-3c/def2-mSVP failed to provide the same compound and reaction variety in the given simulation time as the semi-empirical method GFN2-xTB especially out of a qualitative point of view. The results obtained with the aforementioned methods include mainly isomers of reactive small molecules (3-rings) rather than the formation of novel, larger, and stable compounds.

While GFN2-xTB resulted in promising results for the reactive systems and confirmed the low reactivity of the water and ammonia arrangements, it failed to describe the electronic structure of the CO molecules correctly resulting in very improbable species. Performing the simulation at higher levels of theory using RHF/3-21G and PBEh-3c/def2-mSVP proved to be much more sensitive to the used parameters and decreased the variety in the obtained molecular species and the reactivity while also increasing the total computational effort. To compensate for the low reactivity, longer time scales are required to obtain meaningful results using the given settings.

By design, Fermi smearing[59] is used in GFN2-xTB.[35] Because in the existing literature, Fermi smearing (see refs 2 and 7) was not used reportedly for RHF in nanoreactor simulations, the same setting was chosen here, both for RHF and PBEh-3c. We expect, that the use of thermal smearing could greatly impact the results obtained with RHF and PBEh-3c. Therefore, we plan to investigate Fermi smearing and further settings, as well as perform a more in depth analysis of the role of the electronic structure method in a future work.

Introducing Buffer Atoms. To circumvent nonphysical reactions with a large number of simultaneously reacting molecules, the addition of buffer atoms to the setup was considered. For this purpose, HCN systems were used as basis, keeping the number of atoms of interest (H, C, and N in this case) constant, in order to qualitatively compare the results.

The simulations containing more than 15% added buffer atoms were not successful due to increased effective compression on the system. As buffer atoms, helium and argon were compared in terms of obtained species and degree of inertness. Helium buffer simulations yielded smaller products than the corresponding systems containing argon atoms regardless of the used amount of buffer atoms. Helium also displayed higher reactivity yielding nonphysical species such as helium-substituted ammonia when the amount of buffer is increased. As we did not employ mass repartitioning, this could be an effect of the much lower mass of helium compared to argon.

Simulations containing argon resulted in promising molecular species, for example, cyanogen, methane, acetylene, and methyl amine, regardless of the buffer concentration. By increasing the amount of argon, the resulting species qualitatively shifted, from mainly acyclic polymers to relevant N-heterocycles while the stability of the simulations decreased. Therefore, 10% of added argon atoms were found to be a good compromise between reactivity and stability.

After having determined the suitable parameters for employing buffer atoms, the effect of argon on the resulting reactions was assessed. For this, the obtained chemical transformations were categorized in monomolecular, bimolecular, and termolecular reactions (Figure 5). Buffer atoms

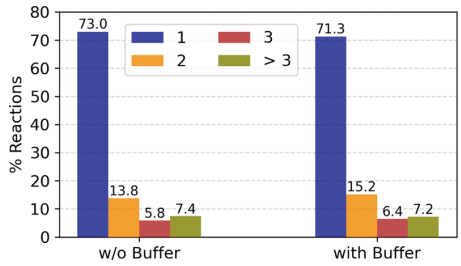


Figure 5. Obtained number of mono- and multimolecular reactions for a HCN system with and without argon buffer. The buffer-containing simulations were run with 50 HCN molecules and 15 argon atoms, this being equivalent to 10% of added buffer. To simulate the same compression degree, the simulations without buffer were run with 55 HCN molecules. All numbers were averaged from five simulations each. The different colors correspond to different amounts of reactants in the observed reactions.

slightly decrease the number of monomolecular transformations, such as isomerizations, while also increasing the occurrence of reactions of interest, such as bi- and termolecular reactions, which are relevant for reaction path discovery.

Reactions with more than three participating molecules were summarized in the green bar in Figure 5. The obtained slightly higher number of such reactions is an effect of summarizing over all subcategories, where each type occurred with lower probability than the ternary reactions. Furthermore, the property of the computational nanoreactor method to support the occurrence of multimolecular reactions has been observed in previous studies[2,7] and could be attributed to the extreme conditions employed, as well as to the formation of preassociated complexes.

On the basis of the HCN/Ar simulations presented above, we have quantified the amount of reactions per unit of computing power, where the computing power was defined as the total wall time needed for all simulations steps in seconds. Here, very similar results were obtained with and without argon buffer, with a mean of 1.00×10^{-4} reactions per second for the setups without buffer and 1.21×10^{-4} when employing 10% added argon buffer. By keeping the percentage of buffer atoms low, the minimal rise in required computing time per time step (using a minimal basis) is overcompensated by the advantages regarding the type and quantity of reactions observed within the same total simulation time.

These findings suggest that the usefulness of the results from the nanoreactor approach can be improved by the addition of a small amount of argon buffer.

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

Figure 6. Obtained reaction networks for the different initial configurations of the reactants. On average, 9 novel molecules were detected every 100 ps. The first network (far left) was obtained from a 243 ps simulation, while the other five were constructed from 750 ps simulations. Each node represents a molecular species colored based on the time step where it first occurred. Early time steps have red hues, while late time steps are represented by blue tones. Edges encode molecular transformations and are colored according to the starting node. Consecutive intermediate monomolecular transformations were excluded from the networks. An enlarged view along with the corresponding molecular structures is provided in the Supporting Information.

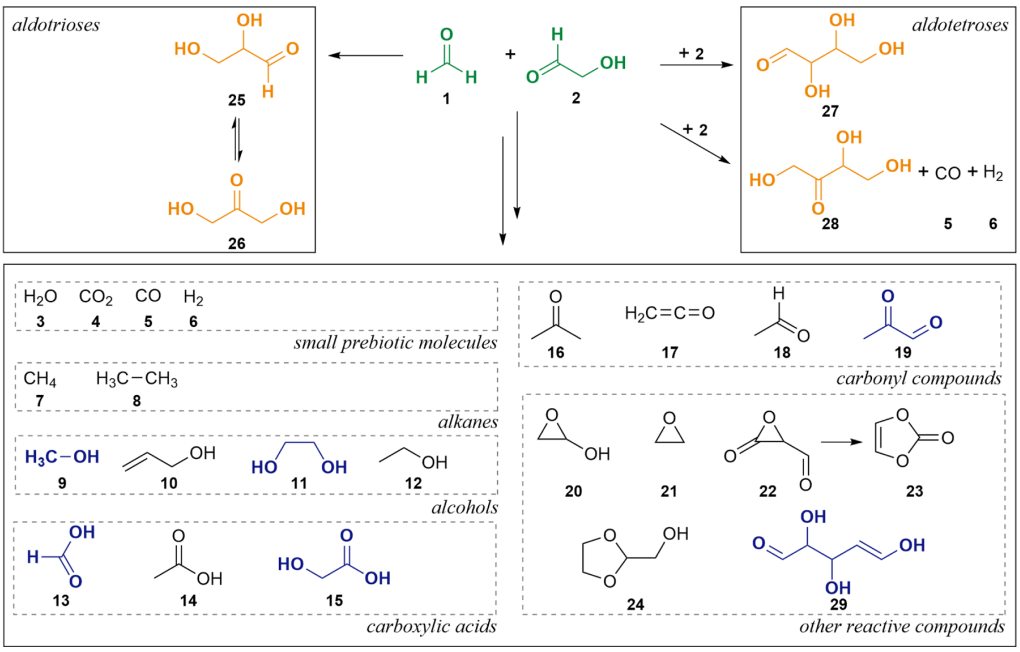


Figure 7. Overview of relevant molecular species obtained from six formose reaction network simulations starting from formaldehyde and glycolaldehyde at a ratio of 4:1 and 5% of added argon atoms. Feedstock compounds are highlighted in green, while obtained aldoses are colored in orange. The depicted reaction paths to the aldotrioses and aldotetroses were extracted as such from the generated networks. Side products of the formose reaction network are depicted in dark blue.

Formose Reaction Network. The self-catalyzing formose reaction network suggested as source for ribose and other sugars in prebiotic chemistry[39,40] was investigated using the computational nanoreactor. In the setup, argon buffer atoms were included, and the newly introduced smooth-step spherical constraint and postprocessing procedure were used. To account for the statistical nature and the reported dependency on the initialization of the nanoreactor, the results were acquired by six simulations with distinct starting configurations.

The simulations, AsForm1 to AsForm6, starting from a mixture of formaldehyde (**1**) and glycolaldehyde (**2**) at a ratio of 4:1 and 5% of added argon atoms provided a great variety of prebiotically relevant compounds, including aldoses of various chain lengths and several small organic molecules. While AsForm1 had an MD simulation length of 243 ps, 750 ps were chosen for AsForm2 to AsForm6. Even though the formose reaction is known to require basic catalysis, the presented results have been obtained under neutral conditions as addition of catalytic amounts of OH^- ions has highly decreased the stability of the simulations. Here, the extreme conditions employed in the simulations are expected to initialize the reaction network without the basic ions present. The lack of basic catalysis also has the advantage that the Cannizzaro reaction is not favored, which is an undesired side reaction in experimental setups of the formose reaction. The obtained number of species in the network varied from 22 to 81 with a mean of 9 novel molecules detected every 100 ps (see Figure 6). On average, 56 events, that is, unique reactions, were identified.

From the multitude of obtained organic species, important prebiotic compounds, which were observed to be highly reactive, such as water (**3**), carbon dioxide (**4**), carbon monoxide (**5**), and hydrogen (**6**) were selected. The structural variety of the found species was broad, ranging from alkanes, of which methane (**7**) and ethane (**8**) were most abundant, to alcohols **9** to **12** and carboxylic acids **13** to **15**. Several known side products of the formose reaction were found, among

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

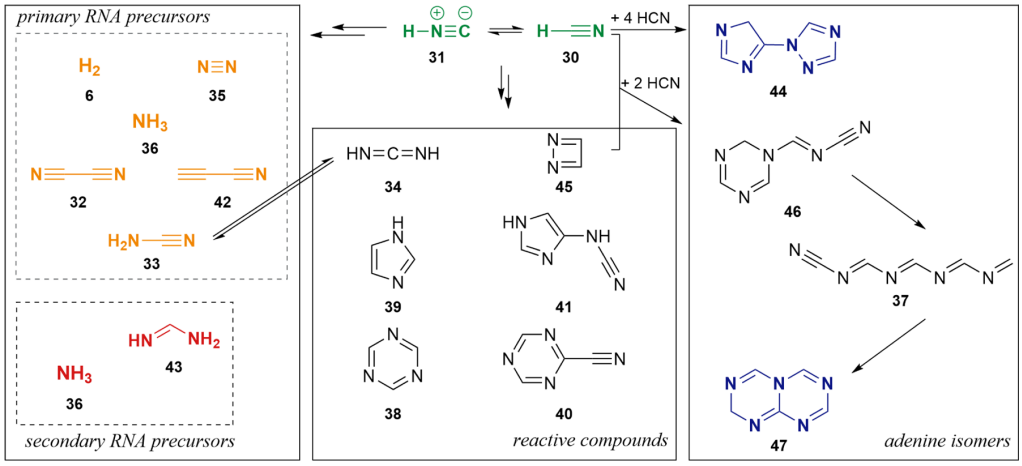


Figure 8. Overview of relevant molecular species from five HCN reaction network simulations with 10% of added argon atoms. The feedstock compound is highlighted in green, while obtained primary RNA precursors are colored in orange, and secondary RNA precursors are shown in red. The depicted reaction paths to adenine isomers (dark blue) were extracted as such from the obtained networks. Hereby, charged intermediates were omitted in this depiction. Ammonia is listed both as a primary and secondary precursor (hydrolysis product of formamide, urea, or other compounds in a prebiotic context) based on the classification by Benner et al.[38]

which the Cannizzaro reaction products of reactants **1** and **2**, that is, methanol (**9**) and glycol (**11**) along with the corresponding carboxylic acid components, formic acid (**13**) and glycolic acid (**15**), were present (Figure 7).

Furthermore, small carbonyl compounds **16** to **19** were encountered, of which acetone (**16**) and acetaldehyde (**18**) are important molecules bearing structural information needed for aldol reactions. Also, dicarbonyl compounds such as 2-oxopropanal (**19**), which can form through β-elimination from glyceraldehyde isomers, were isolated. Here, **19** was formed through addition of a CO molecule to a previously formed acetaldehyde.

(Un)substituted oxiranes **20** to **22** along with carbonates and their derivatives **23** and **24** played an important role in terms of reactivity. Vinylidene carbonate (**23**) is known to undergo polymerization. Several polymeric addition products were seen to support the synthesis of intermediates on the way to aldoses of different length. Glyceraldehyde (**25**) was obtained by aldol addition from formaldehyde to glycolaldehyde. It further isomerized to dihydroxyacetone (**26**). Aldotetrose **27** and erythrulose (**28**) formed directly from the initial compounds as postulated.[40] Only precursors of aldopentoses such as 2,3-hydroxypentanedial (**29**) could be found. The enol form of this compound bears a reactive double bond prone to addition of water under acidic conditions resulting in ribose. Aldohexoses were missing altogether, which was in accord with experimental findings, as aldohexoses are known to form only in very small amounts as part of the formose reaction network.[40] Nevertheless, we expect further aldopentoses and small amounts of aldohexoses to form at longer time scales and with varying ratios of the initial reactants.

HCN Reaction Networks. The ribonucleic acid (RNA) first hypothesis is supported by the vast presence of RNA cofactors and catalysts in the present biosphere, implying that on the early Earth, genetic evolution started with this molecule. Therefore, abiotic pathways to the components of RNA are needed.[38]

Starting from our parameter tests with homogeneous systems, we have performed further simulations of HCN (**30**) with argon buffer assuming a very simple model of a reductive atmosphere, and evaluated them focusing on the presence of primary and secondary RNA precursors, as well as nucleobase scaffolds. The simulation lengths of ASHCN1 to ASHCN5 ranged from 109 to 250 ps (see Table S11 for further details). The results were collected from a total of five simulations performed with different initial configurations. The obtained number of molecular species in the networks varied between 126 and 210 with a mean of 100 species found every 100 ps. The expected isomerization of cyanhydric acid to isocyanhydric acid (**31**) was observed, which opened up new reaction avenues.

Figure 8 contains selected compounds, which have been previously suggested to have played an important role as primary and secondary precursors in the synthesis of RNA components[38] and have here also been observed successfully in silico using the computational nanoreactor approach.

From the known primary RNA precursors,[38] all compounds lacking oxygen were retrieved from the simulations. Here, great amounts of cyanogen (**32**) and cyanamide (**33**) formed along with its isomer carbodiimide (**34**) and catalytically active molecules such as H_2 (**6**), N_2 (**35**), and NH_3 (**36**). All reaction paths first led from HCN to polymeric structures such as **37**, which underwent subsequent fragmentation to the presented precursors or cyclization to various highly reactive compounds, such as 1,3,5-triazine (**38**) or imidazole (**39**) and corresponding derivatives **40** and **41**. Small amounts of cyanoacetylene (**42**) were also retrieved. Secondary RNA precursors, which form through reduction from the primary precursors,[60] were detected, among which formamidine (**43**), usually a product of ammonia and **33**, is to be mentioned.

Purine and pyrimidine scaffolds could not be detected as such due to the unfavorable ratio between carbon, nitrogen, and hydrogen. Adenine, being the only nucleobase lacking oxygen and therefore an expected product in experimental setups, was not obtained in the performed simulations. However, isomers of adenine ($C_5H_5N_5$) were present and

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

formed both directly in a more concerted fashion and through multistep processes with stable intermediates starting from HCN. Compound **44** consisting of a 4H-imidazole and a 1,2,4-triazole group was obtained through charged intermediates from five molecules of HCN. The great reactivity of triazole could potentially lead to a ring opening with subsequent isomerization to adenine. Diazete (**45**) also played an important role in the formation of $C_5H_5N_5$ scaffolds as an intermediate which further reacts with HCN to yield the substituted 6-ring triazole **46**. This reactive compound undergoes two rearrangements, first to the linear conjugated structure **37** and second to compound **47**. The latter provides the right conformational setup for a further potential isomerization to adenine.

CONCLUSIONS

The objective of this work was to provide a thorough discussion of the nanoreactor approach, which allows the automated exploration of reaction space given a feedstock of starting materials. We investigated several different spherical constraint functions and tuned the respective parameters by monitoring the quantitative and qualitative effects on the resulting productivity of the nanoreactor simulations and their stability. Furthermore, the use of buffer atoms was introduced, which led to a slightly improved number of relevant bi- and termolecular reactions, while the monomolecular transformations were reduced. The inertness exhibited by the buffer atoms during the simulation was also assessed as an indicator for the suitability of the chosen parameters. The quantitative comparisons were enabled by our fully automated evaluation procedure, which provides us with a list of all occurring reaction events and an overview of the newly found compounds as well as their abundance. The postprocessing foots on the connectivity matrices built using Wiberg bond orders calculated throughout the simulations, which are then translated into molecules with the Python3 library RDKit. This initial reduction in dimensionality from 3D to 1D enables the construction of corresponding reaction networks and list of reactions, while the 3D information is preserved and stored in the trajectory.

Further, we applied the optimized approach at the GFN2-xTB level of theory to homogeneous HCN systems, where the formation of prebiotically relevant primary and secondary RNA precursors, such as cyanogen, cyanamide, formamidine, and isomers of adenine were observed. In addition, simulations were carried out starting from formaldehyde and glycolaldehyde in a ratio of 4:1 aiming to reproduce the formose reaction network. Here, reaction paths to aldotrioses and aldotetroses could be determined, and precursors of aldopentoses were identified. Furthermore, side products such as dicarboxyl species and Cannizzaro reaction products were present.

In the future, we aim to further optimize the found reaction paths in order to add kinetic information based on free-energy simulations to the constructed reaction networks and develop an automated pipeline connecting the two parts of the nanoreactor procedure.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00754.

All algorithms used, simulation details, and continuous bar plots (PDF)

Reaction networks for the HCN and formose simulations and full molecular grids (ZIP)

AUTHOR INFORMATION

Corresponding Author

Christian Ochsenfeld − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany; Max Planck Institute for Solid State Research, D-70569 Stuttgart, Germany;* orcid.org/0000-0002-4189-6558; Email: christian.ochsenfeld@uni-muenchen.de

Authors

Alexandra Stan − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0003-3542-9993

Beatriz von der Esch − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0002-8366-5272

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00754

Notes

The authors declare no competing financial interest.

A maintained and updated version of the postprocessing code is available at https://github.com/ochsenfeld-lab/nanoreactor_processing.

ACKNOWLEDGMENTS

The authors thank Laurens Peters for his useful suggestions and Jörg Kussmann (LMU Munich) for providing a development version of the FermiONs++ program package. The authors acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project-ID 364653263—TRR 235 "Emergence of Life" and SFB 1309-325871075 "Chemical Biology of Epigenetic Modifications". C.O. acknowledges further support as a Max-Planck fellow at the MPI-FKF Stuttgart.

REFERENCES

(1) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. *Annu. Rev. Phys. Chem.* **2020**, *71*, 121−142.

(2) Wang, L. P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **2014**, *6*, 1044−1048.

(3) Wang, L. P.; McGibbon, R. T.; Pande, V. S.; Martínez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **2016**, *12*, 638−649.

(4) Goldman, N. A virtual squeeze on chemistry. *Nat. Chem.* **2014**, *6*, 1033−1034.

(5) Martínez, T. J. Ab Initio Reactive Computer Aided Molecular Design. *Acc. Chem. Res.* **2017**, *50*, 652−656.

(6) Meisner, J.; Zhu, X.; Martínez, T. J. Computational Discovery of the Origins of Life. *ACS Cent. Sci.* **2019**, *5*, 1493−1495.

(7) Das, T.; Ghule, S.; Vanka, K. Insights into the Origin of Life: Did It Begin from HCN and H_2O? *ACS Cent. Sci.* **2019**, *5*, 1532−1540.

(8) Miller, S. L. A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science* **1953**, *117*, 528−529.

(9) Miller, S. L.; Urey, H. C. Organic Compound Synthesis on the Primitive Earth. *Science* **1959**, *130*, 245−251.

(10) Lei, T.; Guo, W.; Liu, Q.; Jiao, H.; Cao, D. B.; Teng, B.; Li, Y. W.; Liu, X.; Wen, X. D. Mechanism of Graphene Formation via Detonation Synthesis: A DFTB Nanoreactor Approach. *J. Chem. Theory Comput.* **2019**, *15*, 3654−3665.
(11) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 826−843.
(12) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 2847−2862.
(13) Pieri, E.; Lahana, D.; Chang, A. M.; Aldaz, C. R.; Thompson, K. C.; Martínez, T. J. The non-adiabatic nanoreactor: Towards the automated discovery of photochemistry. *Chem. Sci.* **2021**, *12*, 7294−7307.
(14) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, No. e134.
(15) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123*, 385−399.
(16) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates. *Ind. Eng. Chem. Res.* **1994**, *33*, 790−799.
(17) Matheu, D. M.; Dean, A. M.; Grenda, J. M.; Green, W. H. Mechanism Generation with Integrated Pressure Dependence: A New Model for Methane Pyrolysis. *J. Phys. Chem. A* **2003**, *107*, 8552−8565.
(18) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212−225.
(19) Rappoport, D.; Galvin, C. J.; Zubarev, D. Y.; Aspuru-Guzik, A. Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry. *J. Chem. Theory Comput.* **2014**, *10*, 897−907.
(20) Martínez-Núñez, E. An automated method to find transition states using chemical dynamics simulations. *J. Comput. Chem.* **2015**, *36*, 222−234.
(21) Martínez-Núñez, E. An automated transition state search using classical trajectories initialized at multiple minima. *Phys. Chem. Chem. Phys.* **2015**, *17*, 14912−14921.
(22) Vázquez, S. A.; Martínez-Núñez, E. HCN elimination from vinyl cyanide: Product energy partitioning, the role of hydrogen-deuterium exchange reactions and a new pathway. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6948−6955.
(23) Varela, J. A.; Vázquez, S. A.; Martínez-Núñez, E. An automated method to find reaction mechanisms and solve the kinetics in organometallic catalysis. *Chem. Sci.* **2017**, *8*, 3843−3851.
(24) Martínez-Núñez, E.; Barnes, G. L.; Glowacki, D. R.; Kopec, S.; Peláez, D.; Rodríguez, A.; Rodríguez-Fernández, R.; Shannon, R. J.; Stewart, J. J.; Tahoces, P. G.; Vazquez, S. A. AutoMeKin2021: An open-source program for automated reaction discovery. *J. Comput. Chem.* **2021**, *42*, 2036−2048.
(25) Maeda, S.; Morokuma, K. Finding Reaction Pathways of Type A + B → X: Toward Systematic Prediction of Reaction Mechanisms. *J. Chem. Theory Comput.* **2011**, *7*, 2335−2345.
(26) Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M. Heuristics-Guided Exploration of Reaction Mechanisms. *J. Chem. Theory Comput.* **2015**, *11*, 5712−5722.
(27) Feng, F.; Lai, L.; Pei, J. Computational Chemical Synthesis Analysis and Pathway Design. *Front. Chem.* **2018**, *6*, 199.
(28) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 492−504.
(29) Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. *J. Am. Chem. Soc.* **1997**, *119*, 4033−4042.
(30) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(31) Ji, W.; Deng, S. Autonomous Discovery of Unknown Reaction Pathways from Data by Chemical Reaction Neural Network. *J. Phys. Chem. A* **2021**, *125*, 1082−1092.
(32) Bort, W.; Baskin, I. I.; Gimadiev, T.; Mukanov, A.; Nugmanov, R.; Sidorov, P.; Marcou, G.; Horvath, D.; Klimchuk, O.; Madzhidov, T.; Varnek, A. Discovery of novel chemical reactions by deep generative recurrent neural network. *Sci. Rep.* **2021**, *11*, 3178.
(33) Binkley, J. S.; Pople, J. A.; Hehre, W. J. Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *J. Am. Chem. Soc.* **1979**, *102*, 939−947.
(34) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1−86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.
(35) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(36) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e143.
(37) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*, 054107.
(38) Benner, S. A.; Bell, E. A.; Biondi, E.; Brasser, R.; Carell, T.; Kim, H.; Mojzsis, S. J.; Omran, A.; Pasek, M. A.; Trail, D. When Did Life Likely Emerge on Earth in an RNA-First Process? *ChemSystemsChem* **2020**, *2*, No. e1900035.
(39) Breslow, R. On the mechanism of the formose reaction. *Tetrahedron Lett.* **1959**, *1*, 22−26.
(40) Haas, M.; Lamour, S.; Christ, S. B.; Trapp, O. Mineral-mediated carbohydrate synthesis by mechanical forces in a primordial geochemical setting. *Commun. Chem.* **2020**, *3*, 140.
(41) Laqua, H.; Thompson, T. H.; Kussmann, J.; Ochsenfeld, C. Highly Efficient, Linear-Scaling Seminumerical Exact-Exchange Method for Graphic Processing Units. *J. Chem. Theory Comput.* **2020**, *16*, 1456−1468.
(42) González, Á. Measurement of Areas on a Sphere Using Fibonacci and Latitude-Longitude Lattices. *Math. Geosci.* **2010**, *42*, 49−64.
(43) Hutchings, M.; Liu, J.; Qiu, Y.; Song, C.; Wang, L. P. Bond-Order Time Series Analysis for Detecting Reaction Events in Ab Initio Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2020**, *16*, 1606−1617.
(44) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1−S19.
(45) Wiberg, K. B. Application of the pople-santry-segal CNDO method to the cyclopropylcarbinyl and cyclobutyl cation and to bicyclobutane. *Tetrahedron* **1968**, *24*, 1083−1096.
(46) Mayer, I. Bond order and valence indices: A personal account. *J. Comput. Chem.* **2007**, *28*, 204−221.
(47) McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 2010; pp 56−61, DOI: 10.25080/Majora-92bf1922-00a.
(48) The pandas development team, pandas-dev/pandas: Pandas 1.4.3. 2022, https://zenodo.org/record/6702671 (accessed September 05, 2022).
(49) Landrum, G. rdkit/rdkit: 2021_03_4 (Q1 2021) Release. 2021, https://zenodo.org/record/5085999 (accessed September 05, 2022).
(50) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

(51) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(52) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA, 2008; pp 11−15.
(53) Pezoa, F.; Reutter, J. L.; Suarez, F.; Ugarte, M.; Vrgoč, D. Foundations of JSON Schema. *Proceedings of the 25th International Conference on World Wide Web*. Republic and Canton of Geneva, CHE, 2016; pp 263−273, DOI: 10.1145/2872427.2883029.
(54) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, 134114.
(55) Kussmann, J.; Ochsenfeld, C. Preselective Screening for Linear-Scaling Exact Exchange-Gradient Calculations for Graphics Processing Units and General Strong-Scaling Massively Parallel Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 918−922.
(56) Lehtola, S.; Steigemann, C.; Oliveira, M. J.; Marques, M. A. Recent developments in libxc - A comprehensive library of functionals for density functional theory. *SoftwareX* **2018**, *7*, 1−5.
(57) Kussmann, J.; Laqua, H.; Ochsenfeld, C. Highly Efficient Resolution-of-Identity Density Functional Theory Calculations on Central and Graphics Processing Units. *J. Chem. Theory Comput.* **2021**, *17*, 1512−1521.
(58) Haslwanter, T. *An Introduction to Statistics with Python*, 1st ed.; Springer Cham, 2016.
(59) Mermin, N. D. Thermal Properties of the Inhomogeneous Electron Gas. *Phys. Rev.* **1965**, *137*, A1441−A1443.
(60) Slebocka-Tilk, H.; Sauriol, F.; Monette, M.; Brown, R. S. Aspects of the hydrolysis of formamide: revisitation of the water reaction and determination of the solvent deuterium kinetic isotope effect in base. *Can. J. Chem.* **2011**, *80*, 1343−1350.

https://doi.org/10.1021/acs.jctc.2c00754
J. Chem. Theory Comput. 2022, 18, 6700−6712

Supporting Information: "Fully Automated Generation of Prebiotically Relevant Reaction Networks from Optimized Nanoreactor Simulations"

Alexandra Stan,[†] Beatriz von der Esch,[†] and Christian Ochsenfeld[*,†,‡]

†*Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany*

‡*Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany*

E-mail: christian.ochsenfeld@uni-muenchen.de

Simulation Details and Continuous Bar Plots

Parameter Simulations

In all tables below the use of the rectangular wave function, cosine function and smooth-step function is indicated by rect_wave, cos_wave, and smooth_step, respectively.

Table S1: Employed parameters for temperature test simulations.

Thermostat Temperature T					
Simulation	CST1	CST2	CST3	CST4	CST5
Molecules	C_2H_2	C_2H_2	C_2H_2	C_2H_2	C_2H_2
# Atoms	156	156	156	156	156
Constraint Fct.	rect_wave	rect_wave	rect_wave	rect_wave	rect_wave
k/kcal/(molÅ^2)	0.5/1.0	0.5/1.0	0.5/1.0	0.5/1.0	0.5/1.0
r_{min}/Å	8	8	8	8	8
r_{max}/Å	14	14	14	14	14
T_{target}/K	500	1000	2000	3000	4000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	0.5/1.5	0.5/1.5	0.5/1.5	0.5/1.5
Δt/fs	0.5	0.5	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG	STO-mG	STO-mG
Length/ps	250	250	250	250	250

Table S2: Employed parameters for friction constant test simulations.

Friction Constant γ				
Simulation	CSgamma1	CSgamma2	CSgamma3	CSgamma4
Molecules	C_2H_2	C_2H_2	C_2H_2	C_2H_2
# Atoms	156	156	156	156
Constraint Fct.	rect_wave	rect_wave	rect_wave	rect_wave
k/kcal/(molÅ^2)	0.5/1.0	0.5/1.0	0.5/1.0	0.5/1.0
r_{min}/Å	8	8	8	8
r_{max}/Å	14	14	14	14
T_{target}/K	2000	2000	2000	2000
γ/fs^{-1}	1×10^{-3}	3×10^{-3}	5×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	0.5/1.5	0.5/1.5	0.5/1.5
Δt/fs	0.5	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG	STO-mG
Length/ps	250	250	250	250

Table S3: Employed parameters for spherical radii test simulations.

Radii $r_{\mathrm{min}}/r_{\mathrm{max}}$					
Simulation	CSr1	CSr2	CSr3	CSr4	CSr5
Molecules	C_2H_2	C_2H_2	C_2H_2	C_2H_2	C_2H_2
# Atoms	156	156	156	156	156
Constraint Fct.	rect_wave	rect_wave	rect_wave	rect_wave	rect_wave
k/kcal/(molÅ^2)	0.5/1.0	0.5/1.0	0.5/1.0	0.5/1.0	0.5/1.0
r_{min}/Å	6	8	10	8	14
r_{max}/Å	14	14	14	20	20
T_{target}/K	2000	2000	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	0.5/1.5	0.5/1.5	0.5/1.5	0.5/1.5
Δt/fs	0.5	0.5	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG	STO-mG	STO-mG
Length/ps	250	250	250	250	250

Table S4: Employed parameters for contraction and expansion period test simulations.

Period $t_{\mathrm{con}}/t_{\mathrm{exp}}$				
Simulation	CSt1	CSt2	CSt3	CSt4
Molecules	C_2H_2	C_2H_2	C_2H_2	C_2H_2
# Atoms	156	156	156	156
Constraint Fct.	rect_wave	rect_wave	rect_wave	rect_wave
k/kcal/(molÅ^2)	0.5/1.0	0.5/1.0	0.5/1.0	0.5/1.0
r_{min}/Å	8	8	8	8
r_{max}/Å	14	14	14	14
T_{target}/K	2000	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	1.0/1.0	1.0/3.0	2.0/2.0
Δt/fs	0.5	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG	STO-mG
Length/ps	250	250	250	250

Table S5: Employed parameters for force constant test simulations.

Force Constant k				
Simulation	CSk1	CSk2	CSk3	CSk4
Molecules	C_2H_2	C_2H_2	C_2H_2	C_2H_2
# Atoms	156	156	156	156
Constraint Fct.	rect_wave	rect_wave	rect_wave	rect_wave
k/kcal/(molÅ^2)	0.25/0.5	0.5/1.0	1.0/1.0	1.0/2.0
r_{min}/Å	8	8	8	8
r_{max}/Å	14	14	14	14
T_{target}/K	2000	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	0.5/1.5	0.5/1.5	0.5/1.5
Δt/fs	0.5	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG	STO-mG
Length/ps	250	250	250	250

Table S6: Homogeneous systems for test simulations with GFN2-xTB. X = C_2H_2, HCN, CO, H_2O, NH_3

Homogeneous Simulations - GFN2-xTB/STO-mG			
Simulation	HSxTB1	HSxTB2	HSxTB3
Molecules	X	X	X
# Atoms	156	156	156
Constraint Fct.	rect_wave	cos_wave	smooth_step
k/kcal/(molÅ^2)	0.5/1.0	1.0	1.0
r_{min}/Å	8	8	8
r_{max}/Å	14	14	14
T_{target}/K	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	2.0	2.0
Δt/fs	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG

Table S7: Homogeneous systems for test simulations with RHF/3-21G. X = C_2H_2, HCN, CO, H_2O, NH_3

Homogeneous Simulations - RHF/3-21G			
Simulation	HSHF1	HSHF2	HSHF3
Molecules	X	X	X
# Atoms	156	156	156
Constraint Fct.	rect_wave	cos_wave	smooth_step
k/kcal/(molÅ^2)	0.5/1.0	1.0	1.0
r_{min}/Å	8	8	8
r_{max}/Å	14	14	14
T_{target}/K	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	2.0	2.0
Δt/fs	0.5	0.5	0.5
Method	RHF	RHF	RHF
Basis Set	3-21G	3-21G	3-21G

Table S8: Homogeneous systems for test simulations with PBEh-3c/def2-mSVP. X = C_2H_2, HCN, CO, H_2O, NH_3

Homogeneous Simulations - PBEh-3c/def2-mSVP			
Simulation	HSDFT1	HSDFT2	HSDFT3
Molecules	X	X	X
# Atoms	156	156	156
Constraint Fct.	rect_wave	cos_wave	smooth_step
k/kcal/(molÅ^2)	0.5/1.0	1.0	1.0
r_{min}/Å	8	8	8
r_{max}/Å	14	14	14
T_{target}/K	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	0.5/1.5	2.0	2.0
Δt/fs	0.5	0.5	0.5
Method	PBEh-3c	PBEh-3c	PBEh-3c
Basis Set	def2-mSVP	def2-mSVP	def2-mSVP

Table S9: Employed parameters for helium buffer test simulations.

Buffer Atoms - Helium			
Simulation	CSBHe5%	CSBHe10%	CSBHe15%
Molecules	HCN, He	HCN, He	HCN, He
# Atoms	150, 7	150, 15	150, 22
Constraint Fct.	smooth_step	smooth_step	smooth_step
k/kcal/(molÅ^2)	1.0	1.0	1.0
r_{min}/Å	8	8	8
r_{max}/Å	14	14	14
T_{target}/K	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	2.0	2.0	2.0
Δt/fs	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG
Length/ps	250	250	250

Table S10: Employed parameters for argon buffer test simulations.

Buffer Atoms - Argon				
Simulation	CSBAr5%	CSBAr10%	CSBAr15%	CSBw/oAr
Molecules	HCN, Ar	HCN, Ar	HCN, Ar	HCN
# Atoms	150, 7	150, 15	150, 22	165
Constraint Fct.	smooth_step	smooth_step	smooth_step	smooth_step
k/kcal/(molÅ^2)	1.0	1.0	1.0	1.0
r_{min}/Å	8	8	8	8
r_{max}/Å	14	14	14	14
T_{target}/K	2000	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\mathrm{con}}/t_{\mathrm{exp}}$)/ps	2.0	2.0	2.0	2.0
Δt/fs	0.5	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG	STO-mG
Length/ps	250	250	161	250

Application Simulations

Table S11: Parameters for the simulations run on the HCN systems with argon buffer.

HCN Systems					
Simulation	ASHCN1	ASHCN2	ASHCN3	ASHCN4	ASHCN5
Molecules			HCN, Ar		
# Atoms			150, 15		
Constraint Fct.	smooth_step	smooth_step	smooth_step	smooth_step	smooth_step
k/kcal/(molÅ^2)	1.0	1.0	1.0	1.0	1.0
r_{min}/Å	8	8	8	8	8
r_{max}/Å	14	14	14	14	14
T_{target}/K	2000	2000	2000	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\text{con}}/t_{\text{exp}}$)/ps	2.0	2.0	2.0	2.0	2.0
Δt/fs	0.5	0.5	0.5	0.5	0.5
Method	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG	STO-mG	STO-mG	STO-mG
Length/ps	250	187	150	250	109

Table S12: Parameters for the simulations run on the formose reaction network with argon buffer.

Formose Reaction Network		
Simulation	ASForm1	ASForm2–ASForm6
Molecules	formaldehyde, glycolaldehyde, Ar	
# Atoms	80, 20, 6	
Constraint Fct.	smooth_step	smooth_step
k/kcal/(molÅ^2)	2.0	2.0
r_{min}/Å	6	6
r_{max}/Å	14	14
T_{target}/K	2000	2000
γ/fs^{-1}	7×10^{-3}	7×10^{-3}
t_{total} ($t_{\text{con}}/t_{\text{exp}}$)/ps	2.0	2.0
Δt/fs	0.5	0.5
Method	GFN2-xTB	GFN2-xTB
Basis Set	STO-mG	STO-mG
Length/ps	243	750

Table S13: Number of molecular species obtained for acetylene, cyanhydric acid, carbon monoxide, water, and ammonia homogeneous simulations using different functions for spherical confinement and varying the electronic structure method. The total duration of each simulation is given in parentheses.

# Species (Duration / ps)	C_2H_2	HCN	CO	H_2O	NH_3
			Cosine Wave		
GFN2-xTB	72 (250)	137 (250)	197 (250)	0 (250)	0 (250)
RHF/3-21G	30 (101)	3 (177)	4 (123)	6 (250)	9 (250)
PBEh-3c/def2-mSVP	26 (67)	62 (217)	5 (154)	5 (192)	5 (250)
			Rectangular Wave		
GFN2-xTB	60 (250)	375 (250)	92 (250)	0 (250)	0 (250)
RHF/3-21G	13 (89)	4 (189)	29 (128)	6 (202)	3 (250)
PBEh-3c/def2-mSVP	31 (90)	22 (107)	20 (155)	8 (140)	6 (250)
			Smooth-Step Function		
GFN2-xTB	147 (250)	239 (250)	213 (250)	0 (250)	0 (250)
RHF/3-21G	6 (155)	5 (180)	4 (250)	6 (250)	9 (250)
PBEh-3c/def2-mSVP	43 (113)	9 (110)	10 (149)	5 (187)	5 (210)

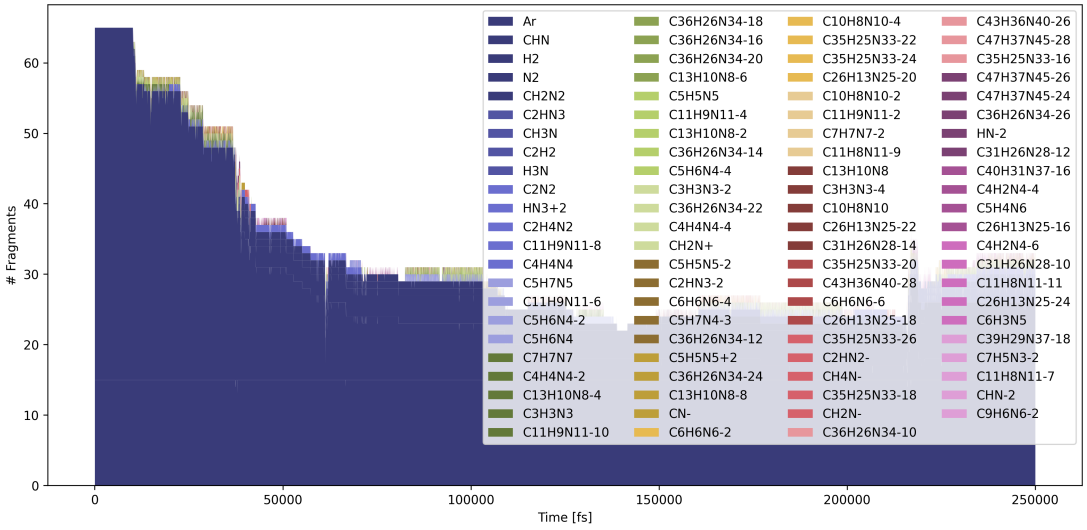


Figure S1: Continuous bar plot for ASHCN1.

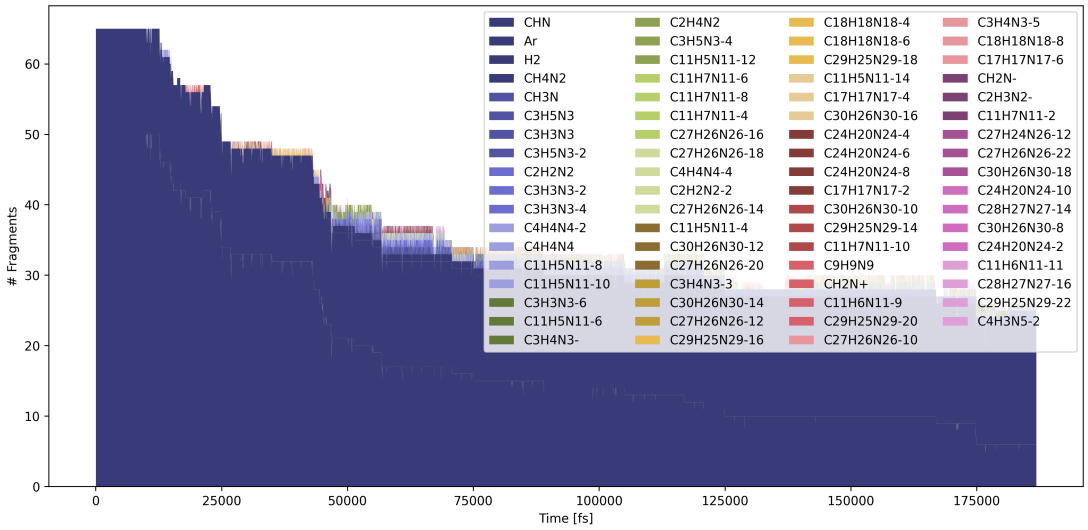


Figure S2: Continuous bar plot for ASHCN2.

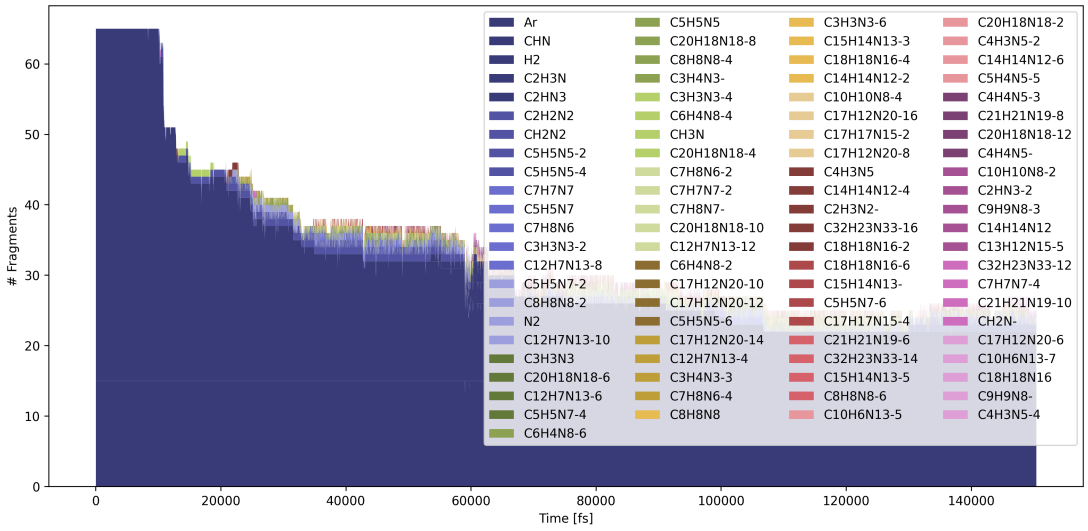


Figure S3: Continuous bar plot for ASHCN3.

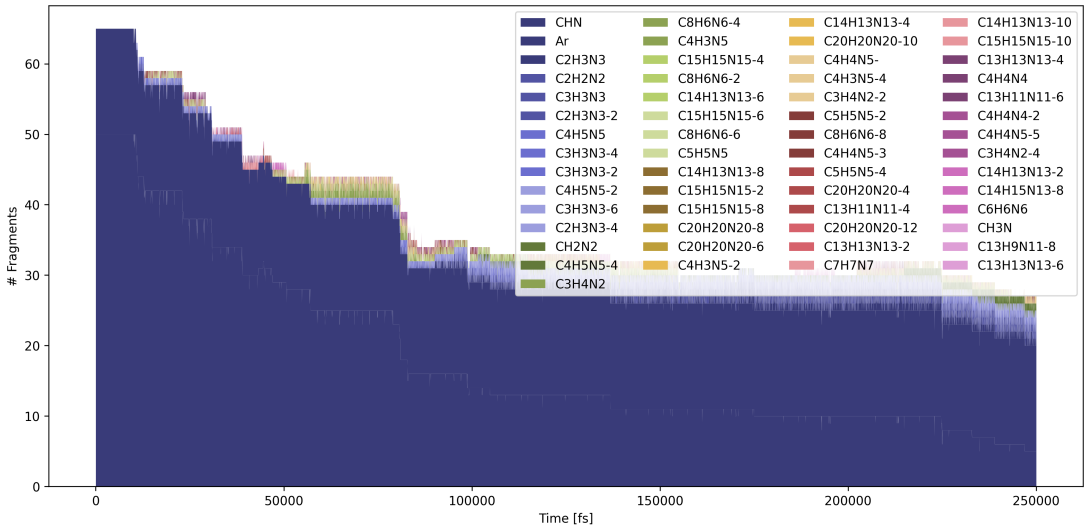


Figure S4: Continuous bar plot for ASHCN4.

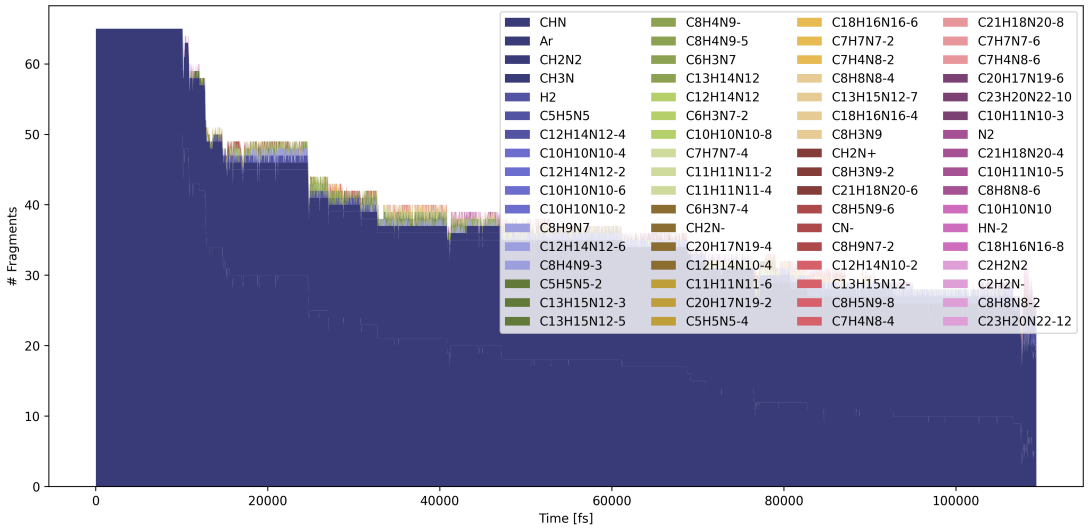


Figure S5: Continuous bar plot for ASHCN5.

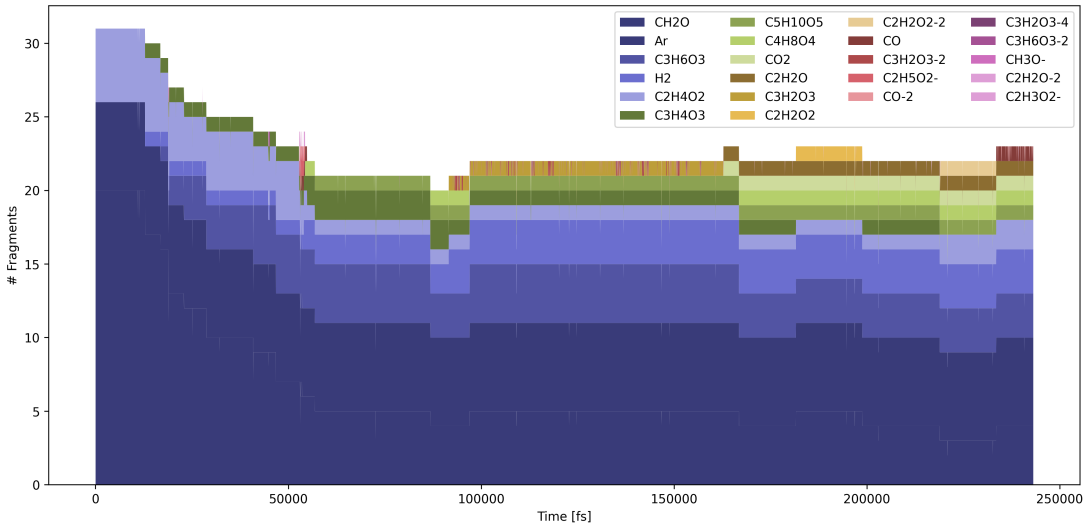


Figure S6: Continuous bar plot for ASForm1.

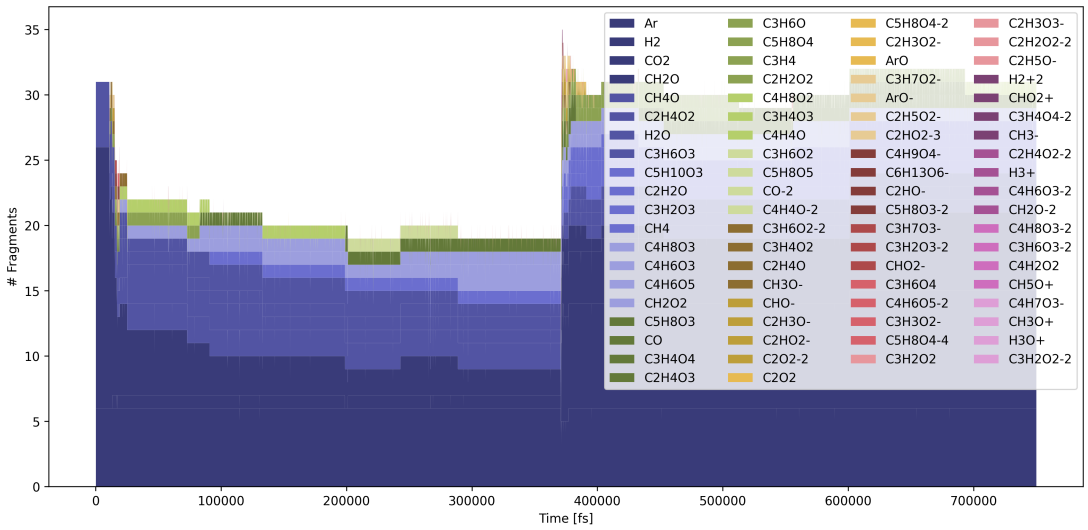


Figure S7: Continuous bar plot for ASForm2.

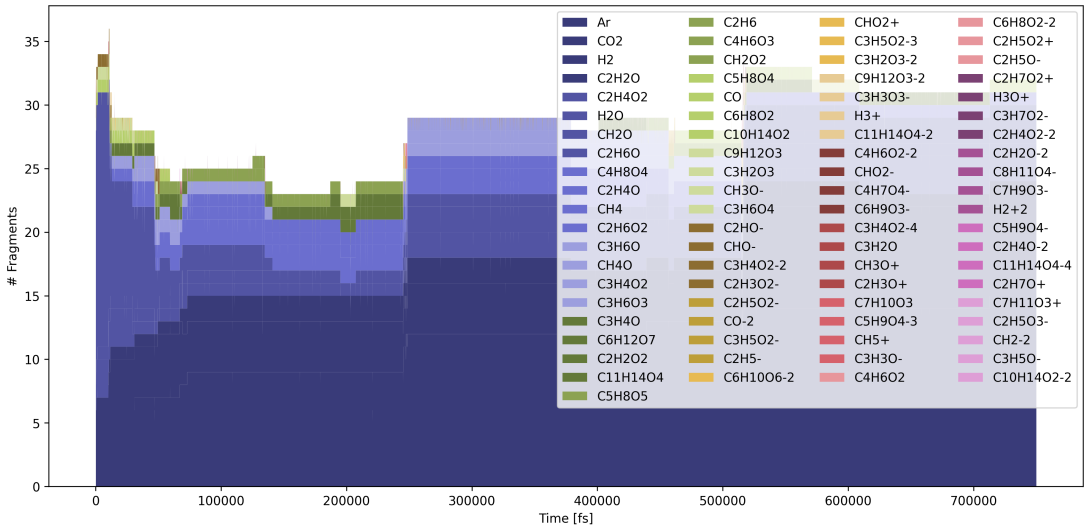


Figure S8: Continuous bar plot for ASForm3.

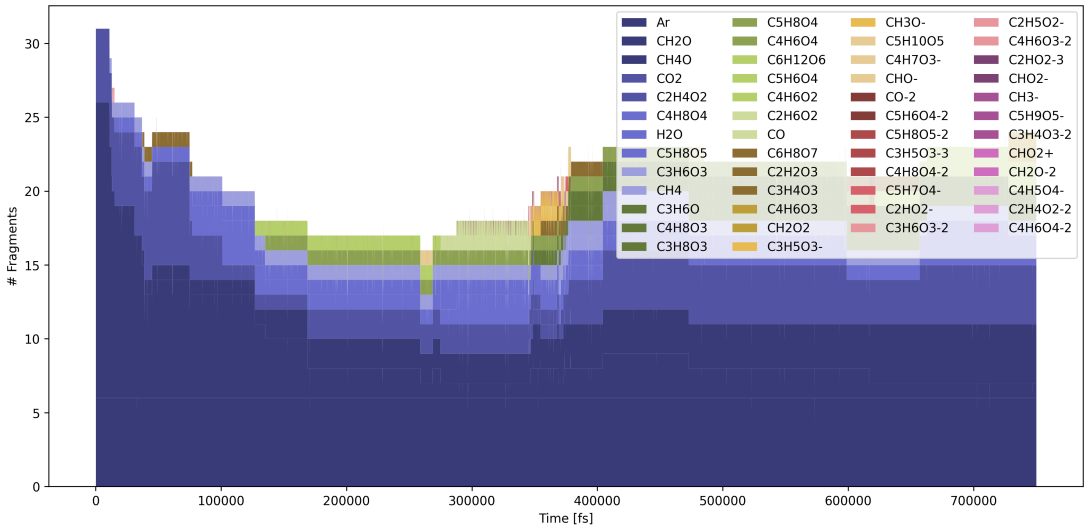


Figure S9: Continuous bar plot for ASForm4.

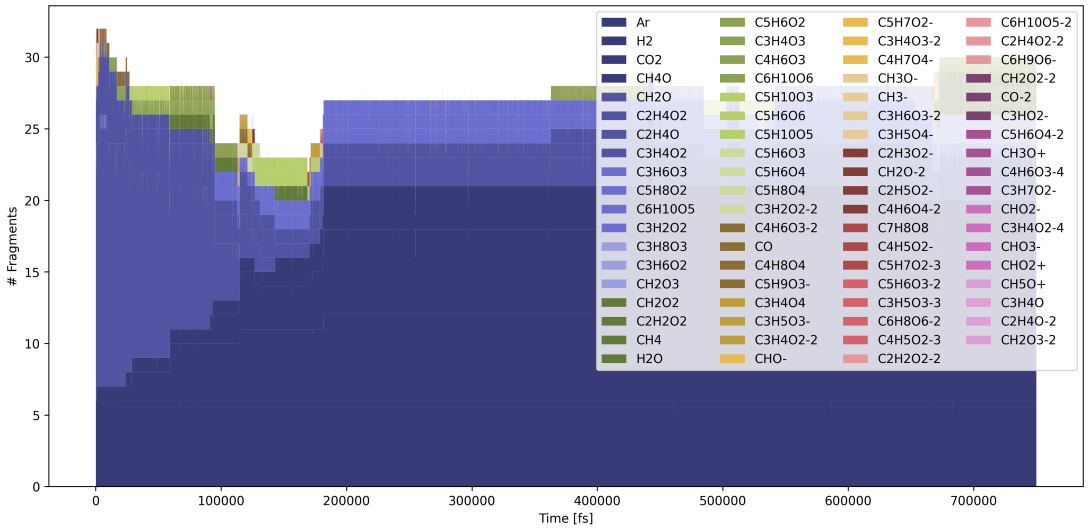


Figure S10: Continuous bar plot for ASForm5.

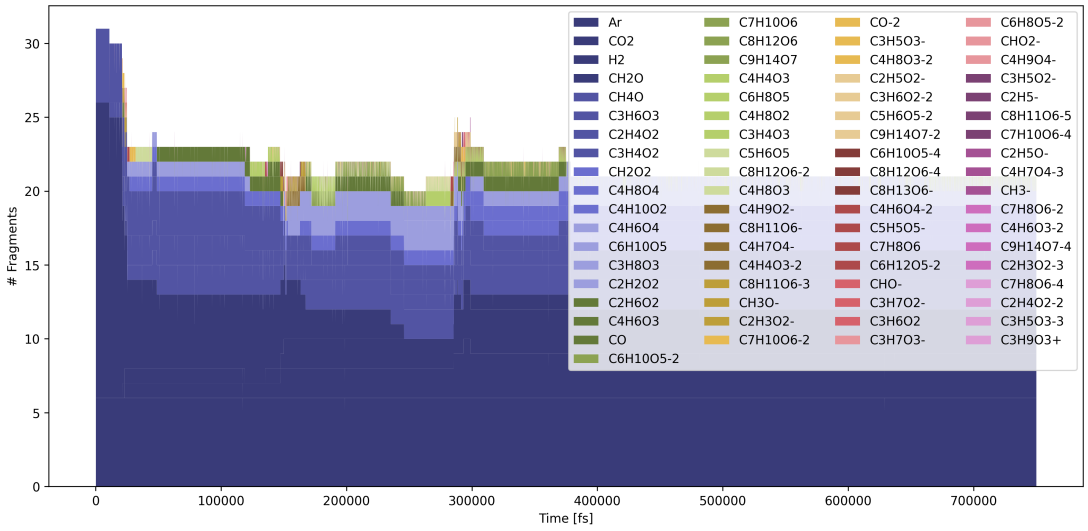


Figure S11: Continuous bar plot for ASForm6.

Algorithms

Algorithm 1 Initialization of Molecules in Nanoreactor Sphere

Require: mols, n_mols, dr
  initialize geom with dummy atom H
  molsArray: initialize mols by amount given in n_mols
  shuffle molsArray
  sumFib = 0
  diff = sumFib - sum(n_mols)
  $r_{\text{nano}} = 0$
  start with 2 samples in the most inner shell: $n = 3$
  while diff < 0 **do**
    compute fibonacci number by index n:
    samples = fib(n)
    r_{nano} += dr
    distribute points on subsphere:
    points = fibonacci_sphere(samples, r_{nano})
    for i in points **do**
      if placedMolecules == $\sum$(nmolecules) **then**
        break
      end if
      add placed molecules at i to geom
      placedMolecules += 1
    end for
    sumFib += samples
    calculate diff
    n += 1
  end while
  delete dummy atom
  return geom, r_{nano}

Algorithm 2 Assign Atoms to Molecular Species Based on Wiberg Bond Orders

Require: wbo_matrix
  initialize frags
  1. Fill up frags:
  for i in atoms **do**
    for j in atoms **do**
      if wbo_matrix < 0.5 **then**
        continue
      end if
      for frag in frags **do**
        check if i already in frag
        check if j already in frag
        append i, j to correct frag or create new frag
      end for
    end for
  end for
  2. Identify missing atoms and assign them based on relative bond length:
  for i in atoms **do**
    exclude buffer atoms from assignment
    for j in atoms **do**
      exclude buffer atoms from assignment
      calculate bond_dist(i, j)
      std$[j]$ = |bond_dist(i, j) − standard_dist(i, j)|
    end for
    atom_partner = argmin(std)
    assign i to correct frag based on atom_partner
  end for
  return frag

Algorithm 3 Build Molecules from Adjacency Matrix

Require: atom_list, adjacency_matrix
  create an empty, editable RDKit mol object rwM
  for i in atom_list **do**
    add i to rwM
    store index of i
  end for
  for row in adjacency_matrix **do**
    for *bond* in row **do**
      add single, double or triple RDKit bonds to rwM based on *bond*
    end for
  end for
  convert rwM to mol object m
  for i_{at} in n_{atoms} **do**
    determine atomic number z and formal charge c_{formal}
    add charges if necessary
  end for
  recalculate valences
  do partial sanitization of m
  return m

Algorithm 4 Create Data Frame and Molecular Grid

Require: mol_species, WBOs File, Trajectory
  Initialize *df*
  add mol_species and traj columns to *df*
  reconstruct bo_matrix
  use Alg. 2 to compute mol_species
  for frag in mol_species **do**
    use Alg. 3 to construct RDKit mols
    compute SMILES
    calculate mol_formulas
  end for
  add SMILES and mol_formulas objects to *df*
  sort mols according to absolute occurence
  plot mols on grid using the RDKit function `MolsToGridImage`
  return *df*, grid

Algorithm 5 Find Reactions Based on Atom Index and Fragment

Require: atom_index, t
  find product based on atom_index
  add product to product SMILES list l_{ps}
  for i_{P} in product **do**
    find reactant(s)
    add reactant to reactant SMILES list l_{rs}
  end for
  sort and merge reactant indices list l_{ri}
  sort and merge product indices list l_{pi}
  while $l_{\mathrm{ri}} \neq l_{\mathrm{pi}}$ **do**
    find products
    find reactants
  end while
  return ($[l_{\mathrm{ri}},l_{\mathrm{pi}}]$,$[l_{\mathrm{rs}},l_{\mathrm{ps}}]$)

Algorithm 6 Create Reaction Network and List of Reactions

Require: *df*
  read *df*
  create empty reaction list r_{list}
  for t in exp_state **do**
    find reactions using Alg. 5
    store reaction in r_{list}
    print t, reaction
  end for
  for i in r_{list} **do**
    add nodes to graph G
    add edges to graph G
  end for
  store r_{list} as JSON
  plot G
  return G, r_{list}

4.4 Publication IV: RNA Oligomerisation without Added Catalyst from 2',3'-Cyclic Nucleotides by Drying at Air-Water Interfaces

Avinash Vicholous Dass , Sreekar Wunnava , Juliette Langlais, **Beatriz von der Esch**, Maik Krusche, Lennard Ufer, Nico Chrisam, Romeo C. A. Dubini, Florian Gartner, Severin Angerpointner, Christina F. Dirscherl, Petra Rovó, Christof B. Mast, Judit Šponer, Christian Ochsenfeld, Erwin Frey, Dieter Braun

"RNA Oligomerisation without Added Catalyst from 2',3'-Cyclic Nucleotides by Drying at Air-Water Interfaces"
ChemSystemsChem **2022**, 5, e202200026.

Abstract:
For the emergence of life, the abiotic synthesis of RNA from its monomers is a central step. We found that in alkaline, drying conditions in bulk and at heated air-water interfaces, 2',3'-cyclic nucleotides oligomerised without additional catalyst, forming up to 10-mers within a day. The oligomerisation proceeded at a pH range of 7–12, at temperatures between 40–80 °C and was marginally enhanced by K^+ ions. Among the canonical ribonucleotides, cGMP oligomerised most efficiently. Quantification was performed using HPLC coupled to ESI-TOF by fitting the isotope distribution to the mass spectra. Our study suggests a oligomerisation mechanism where cGMP aids the incorporation of the relatively unreactive nucleotides C, A and U. The 2',3'-cyclic ribonucleotides are byproducts of prebiotic phosphorylation, nucleotide syntheses and RNA hydrolysis, indicating direct recycling pathways. The simple reaction condition offers a plausible entry point for RNA to the evolution of life on early Earth.

Avinash Vicholous Dass , Sreekar Wunnava , Juliette Langlais, **Beatriz von der Esch**, Maik Krusche, Lennard Ufer, Nico Chrisam, Romeo C. A. Dubini, Florian Gartner, Severin Angerpointner, Christina F. Dirscherl, Petra Rovó, Christof B. Mast, Judit Šponer, Christian Ochsenfeld, Erwin Frey, Dieter Braun
"RNA Oligomerisation without Added Catalyst from 2',3'-Cyclic Nucleotides by Drying at Air-Water Interfaces"
ChemSystemsChem **2022**, 5, e202200026.

ChemSystemsChem
Research Article
doi.org/10.1002/syst.202200026
Chemistry Europe
European Chemical Societies Publishing
www.chemsystemschem.org

RNA Oligomerisation without Added Catalyst from 2′,3′-Cyclic Nucleotides by Drying at Air-Water Interfaces**

Avinash Vicholous Dass[+],[a] Sreekar Wunnava[+],[a] Juliette Langlais[+],[a] Beatriz von der Esch,[b] Maik Krusche,[a] Lennard Ufer,[a] Nico Chrisam,[a] Romeo C. A. Dubini,[d] Florian Gartner,[f] Severin Angerpointner,[f] Christina F. Dirscherl,[a] Petra Rovó,[d, e] Christof B. Mast,[a] Judit E. Šponer,[g] Christian Ochsenfeld,[b, c] Erwin Frey,[f] and Dieter Braun*[a]

For the emergence of life, the abiotic synthesis of RNA from its monomers is a central step. We found that in alkaline, drying conditions in bulk and at heated air-water interfaces, 2′,3′-cyclic nucleotides oligomerised without additional catalyst, forming up to 10-mers within a day. The oligomerisation proceeded at a pH range of 7–12, at temperatures between 40–80 C and was marginally enhanced by K^+ ions. Among the canonical ribonucleotides, cGMP oligomerised most efficiently. Quantification was performed using HPLC coupled to ESI-TOF by fitting the isotope distribution to the mass spectra. Our study suggests a oligomerisation mechanism where cGMP aids the incorporation of the relatively unreactive nucleotides C, A and U. The 2′,3′-cyclic ribonucleotides are byproducts of prebiotic phosphorylation, nucleotide syntheses and RNA hydrolysis, indicating direct recycling pathways. The simple reaction condition offers a plausible entry point for RNA to the evolution of life on early Earth.

Introduction

The central and multifunctional role of RNA within biology points towards RNA as a chief informational biopolymer for the onset of molecular evolution.[1] Polymerisation involving more than a single type of canonical nucleotide, generating a varied pool of RNA strands, has not been achieved under aqueous conditions.[2–6] Chemical activation strategies are deployed to trigger RNA polymerisation[3,7,8] and template-directed primer extension of sequences.[9,10] In the earliest self-replicating systems, the formation of complementary strands for replication and transfer of genetic information by non-enzymatic processes is believed to be important and homopolymers are not considered very useful as genes.[11] Short RNA strands, especially from dimers[11] to tetramers[12,13] have been shown to enhance the copying of mixed-sequence templates in comparison to monomers. Thus, it is necessary to have a oligomerisation mechanism that is able to generate short mixed-sequences that later function as primers and templates for copying of longer sequences.

We base this study on 2′,3′-cyclic mononucleotides (cNMP) which (a) possess an intrinsically activated phosphate; (b) are products of several prebiotic phosphorylation and nucleotide syntheses;[14–18] and (c) are products of neutral to alkaline chemical and enzymatic hydrolyses of RNA.[19–23] In comparison, the dry oligomerisation of 3′,5′-cGMP[24–26] did not foster the oligomerisation of the other ribonucleotides.[26] Orgel and coworkers, reported conditions for 2′,3′-cAMP oligomerisation by drying for 40 days with a 5-fold excess of ethane-1,2-diamine and yields up to 0.67 % of 14-mers.[4,6] Other catalysts such as imidazole or urea required temperatures up to 85 C and offered lower yields.[4,6]

[a] Dr. A. V. Dass,[+] S. Wunnava,[+] J. Langlais,[+] M. Krusche, L. Ufer, N. Chrisam, C. F. Dirscherl, Dr. C. B. Mast, Prof. Dr. D. Braun
Faculty of Physics, Systems Biophysics
Ludwig-Maximilians-Universität München
Amalienstraße 54, 80799, Munich, Germany
E-mail: dieter.braun@lmu.de
[b] B. von der Esch, Prof. Dr. C. Ochsenfeld
Chair of Theoretical Chemistry, Department of Chemistry,
Ludwig-Maximilians-Universität München
Butentandtstraße 5–13, 81377, Munich, Germany
[c] Prof. Dr. C. Ochsenfeld
Max Planck Institute for Solid State Research
Heisenbergstr. 1, 70569 Stuttgart, Germany
[d] R. C. A. Dubini, Dr. P. Rovó
Faculty of Chemistry and Pharmacy,
Ludwig-Maximilians-Universität München
Butentandtstraße 5–13, 81377, Munich, Germany
[e] Dr. P. Rovó
Institute of Science and Technology Austria
c/o NMR Facility, Am Campus 1, 3400 Klosterneuburg, Austria
[f] Dr. F. Gartner, S. Angerpointner, Prof. Dr. E. Frey
Faculty of Physics, Statistical and biological physics,
Ludwig-Maximilians-Universität München,
Theresienstraße. 37, D-80333, Munich, Germany
[g] Dr. J. E. Šponer
Institute of Biophysics Academy of Sciences of the Czech Republic
Královopolská 135, 61265 Brno, Czech Republic

[+] These authors contributed equally to this work.

[**] A previous version of this manuscript has been deposited on a preprint server (http://doi.org/10.26434/chemrxiv-2022-zwh2 t-v2).

Supporting information for this article is available on the WWW under https://doi.org/10.1002/syst.202200026

ChemSystemsChem
Research Article
doi.org/10.1002/syst.202200026
Chemistry Europe
European Chemical Societies Publishing

We found that 2′,3′-cGMP oligomerised spontaneously under alkaline (pH 7-12) drying conditions (40-80 C), within a day. The other canonical cNMP were relatively inert under similar conditions. Our observations of cAMP and cUMP forming up to trimers are consistent with literature.[27,28] The oligomerisation is demonstrated in the presence of bulk water at the air-water interface, within a microfluidic thermal chamber. The chamber mimics conditions of a heated, water filled volcanic rock pore that includes a gas bubble.

In an oligomerisation mixture of cNMP, we observed oligomers rich in G nucleotides, but with C, A and U incorporated at lower concentrations. Computational and modelling results suggest that the oligomers of cGMP form a self-assembled scaffold in the dry state, which could incorporate the nucleotides C, A and U to form short mixed-sequence oligomers.

Results

Polymerisation of cGMP

An aqueous solution of the sodium salt of 2′,3′-cGMP (20 mM) was dried for 18 hours at 40 C in the presence of 40 mM KCl. Since the monomers are monosodium salts, there was an equal concentration of Na^+ ions when in solution (20 mM). All the reported concentrations throughout the article are calculated for a volume of 100 µL. The total concentration of each n-mer (oligomer) is a sum of oligomers containing the linear-phosphate (-P) and the cyclic-phosphate (-cP) on the n-mer terminus. Both endings are well discriminated by HPLC as the n-mer-cP is eluted before a n-mer-P of the same length (S2d). Typically, about 90% of the n-mers consisted of -P endings (S5d). Due to propensity of purines to form non-covalent aggregates in mass spectrometry detection,[29] a combination of HPLC and ESI-TOF techniques were used for detection of oligonucleotides. The non-covalent stacked n-mers (eg. two 4-mers) are discriminated from covalent n-mers (eg. an 8-mer) due to the higher mass of the stacked n-mers by one H_2O in the MS and the corresponding HPLC retention times of n-mers under denaturing HPLC conditions.[30–32]

The denaturing conditions of the HPLC column at 60 C efficiently resolved synthetic oligoG n-mers without signs of aggregation, as shown in Figure 1c. It must be noted that an n-mer-cP and a cyclised n-mer of the same length would have the same mass, but are unlikely not to be discriminated by the HPLC retention times. The presence of n-mer-cP is established from the ^{31}P NMR peak at ~20 ppm in Figure 1d. Oligomers from 2- to 15-mers (S8a) were detected by HPLC-MS for cGMP oligomerisation. For quantification, only 2- to 10-mers were considered throughout the study.

The error bars can be estimated based on plots of cGMP oligomerisation (5 replicates) in S8a, with a mean standard deviation of 2.95 µM between independent runs of the experiment. The error bars are not indicated in the figures as they would appear insignificant on the log scale. For quantification, the HPLC retention times of the oligomer standards of G were

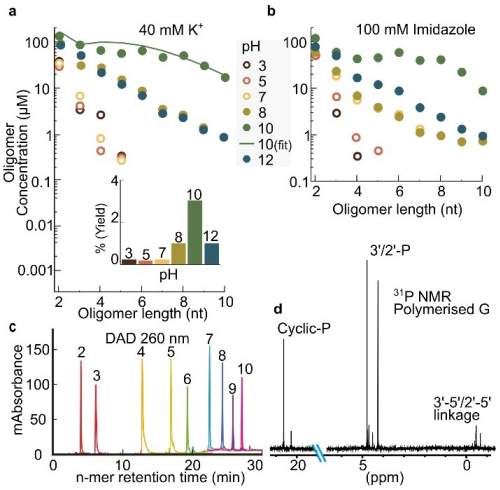


Figure 1. Oligomerisation of Guanosine-2′,3′-cyclic monophosphate (cGMP·Na). A 20 mM cGMP·Na solution was heat-dried with 40 mM K^+ at 40 C for 18 hours, under ambient pressure in 100 µL volume. (a) Polymerisation was screened over a range of pH 3–12. The reported concentrations were the sum of terminal cyclic (-cP) and linear phosphate (-P) containing oligomers. Oligomers without terminal phosphates were not detected. Polymerisation was optimal at pH 10 with total oligomer yields of ~3.5% (inset). The solid line shows results of the polymerisation model based on stacked assembly (S19). (b) pH screen with 100 mM imidazole under similar conditions. No significant increase in oligomerisation was found by adding imidazole. (c) Diode array detector (DAD) absorbance at 260 nm for 50 µM oligoG standards (-P endings) and 100 µM KCl used for confirming HPLC separation and the determination of retention time for quantification with ion counts. (d) ^{31}P proton-decoupled NMR spectrum (10% D_2O, pH 10), of oligomerised G sample: the signals corresponding to phosphodiester linkages for both 3′-5′ and 2′-5′ are between −0.8 and −1.1 ppm.

first optimised on an RP C-18 HPLC column coupled to ESI-TOF. Figure 1c shows the HPLC chromatogram of 2- to 10-mers for oligoG standards (with 1 eqv. of KCl) with their respective retention times. We found efficient separation and no evidence for the formation of aggregates. The ion counts of the n-mer with their HPLC retention times are shown in S2b. By comparing the ion counts, we confirmed the high efficiency of the post-polymerisation ethanol precipitation protocol and its negligible influence (S3). However, the precipitation was used to remove excess monomers which would otherwise saturate the HPLC column, yielding a robust method for the quantification of the complex oligonucleotide mixtures (S2c).

The calculated isotope probabilities of the n-mers in the various charge states were fitted to the raw mass spectra using a self-written LabView program. This allowed us to identify salt adducts formed in the mass spectrometer and to fit overlapping isotope patterns. The retention times of the oligoG standards were used to obtain time-brackets to sum the mass spectra. Further details on the calibration used for the quantification within the program and the functional modes of the program are elaborated in S1–S6. Based on preliminary enzymatic digestion experiments, we estimated that the formed G

ChemSystemsChem
Research Article
doi.org/10.1002/syst.202200026
Chemistry Europe
European Chemical Societies Publishing

oligomers were linked by 2′-5′ and 3′-5′ phosphodiester linkages in about 1:1 ratio (S21–S23). The linkage type in the oligomerisation was also confirmed by ^{31}P NMR (Figure 1d) and the peaks were assigned based on the literature values.[33,34]

Figures 1a and 1b compare the effect of pH on the lengths and concentrations of the n-mers formed by drying with K^+ (Cl^-) and imidazole respectively. We determined the optimal reaction temperature to be 40 C (S5, S8b). Imidazole and its derivatives are used in the literature as nucleotide activation agents for templated primer extension reactions,[9] as a buffering agent and a catalyst for oligomerisation.[4] The addition of imidazole did not enhance the length and concentration of n-mers in comparison to oligomerisation with K^+.

Polymerisation from cNMP

We also tested the polymerisation tendencies of cAMP, cUMP and cCMP under the same heat-drying conditions and found that these monomers did not polymerise to the same lengths and concentrations as cGMP. Figure 2a shows that the polymerisation trend decreases in the order cGMP > cUMP > cAMP > cCMP. The dominance of G-polymerisation prompted us to investigate the copolymerisation of these moderately reactive mononucleotides under the influence of the well oligomerising cGMP. We found that a mixture of two or four different monomers was capable of generating mixed sequence oligomers, where the majority of the mixed oligomers were rich in G. We probed if the oligomerisation of a G and C mixture could reach levels where hybridisation between strands could be possible. Thus, we oligomerised a binary mixture of cGMP and cCMP (20 mM each), under heat-drying conditions (40 C) in the presence of 40 mM KCl. Comparing quantities of C_2 in Figure 2a and 2b, the concentration of C_2 is enhanced 2 fold and C_3 became detectable; besides the fact that mixed GC oligomers are formed (Figure 2b). The detailed sequence composition for GC mixed polymerisation is seen in Figure 2e, showing that the G_2 to G_{10} contribute to the bulk of the oligomers formed in the polymerisation mixture. Up to two C's were incorporated into oligomers $\leq$4-mers, one C is incorporated into 5-mers and none were detectable beyond them. A similar analysis of GA and GU binary mixtures is available in S9a, b.

GC mixed polymerisation was favoured at temperatures ranging from 40 C to 80 C (Figure 2c), similar to cGMP (S5b, S8b). It must be noted that in reactions at 30 C for 18 hours, the drying was incomplete within the polypropylene tubes used for the experiment and the reaction kinetics in the dry state was reduced. Higher temperatures on the other hand possibly contributed to the degradation of the monomers (S4c) and the formed oligomers as seen in the trace comparisons under 80, 60 and 40 C in Figure 2c.

Specific cations also influenced cNMP oligomerisation. We found that K^+ ions yielded higher concentrations and lengths of the oligomers in comparison to Na^+ ions at the same concentrations. The presence of Mg^{2+} ions in the reaction mixture inhibited polymerisation (Figure 2d). The dependence

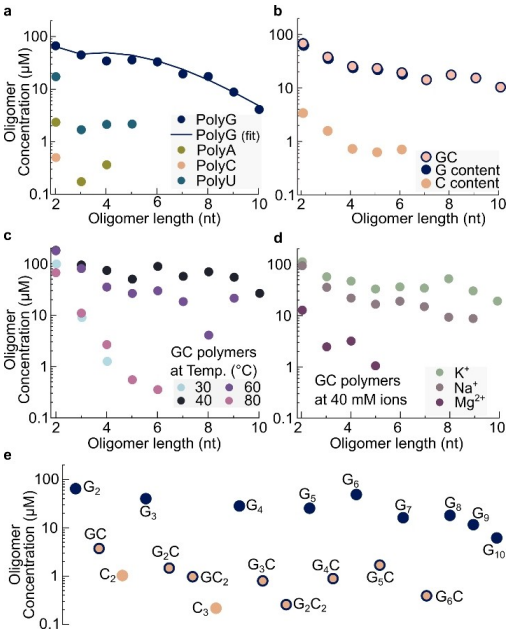


Figure 2. Oligomerisation of mixed Nucleotide 2′,3′-cyclic monophosphate (cNMP). (a) Homooligomers of cGMP·Na, cAMP·Na, cCMP·Na, cUMP·Na were individually produced from a 20 mM solution at 40 C for 18 hours. Oligomers of G were formed in far higher concentrations than polyC, A and U. The solid line shows results of the polymerisation model based on stacked assembly (S18). The approximately 3x lower yield for oligoG compared to Figure 1a is attributed to the lack of K^+ ions. (b) Oligomerisation of cGMP and cCMP at 40 C with 40 mM K^+. The base C is incorporated into the sequences in the presence of cGMP while only dimers were detected without it. (c) Temperature screening over a range of 30–80 C for GC oligomerisation. Reduced concentration of n-mers >3 is observed for 80 C, possibly due to ring opening of the cyclic phosphate monomers (S4c). (d) The presence of 40 mM K^+ increased the concentration of n-mers while added Mg^{2+} quenched polymerisation (S10a). (e) Sequence composition of cGMP and cCMP mixed oligomers at 40 mM K^+. Oligomers show G-rich n-mers and suggest the presence of all possible combinations in trimer sequences.

of polymerisation on K^+, Na^+ and Mg^{2+} salt concentrations is shown in S10, indicating that 1–3 eqv. of the same cation display similar results, but the type of cation affected the efficiency of oligomerisation.

Polymerisation of cNMP in a heated rock pore mimic

Wet-dry cycles in surface-based geological settings are subjected to a drift in salt and pH conditions due to the imbalance caused by the evaporation of pure water and the rehydration of the fluid that contains salt. Wet-dry cycling can also occur in a closed chamber, subjected to a temperature gradient.[35] The water that evaporates on the warm side re-enters the fluid on its cold side. This causes interface shifts and the dew droplet dynamics on the cold side, offering wet-dry cycles under

ChemSystemsChem
Research Article
doi.org/10.1002/syst.202200026
Chemistry Europe
European Chemical Societies Publishing

constant pH and salt conditions. The geological analogues of such a setting would be volcanic rock pores which are partially filled with fluids and are subjected to a thermal gradient. We have previously reported prebiotically important processes such as accumulation, phosphorylation, encapsulation, gelation, strand separation, enzymatic DNA replication and crystallisation within such settings.[35–37]

For the polymerisation within this setting, we started with 20 mM total monomers (5 mM each of cG, cC, cA and cU). After the chamber was loaded with the monomer solution, a thermal gradient was applied which drove continuous wet-dry cycles just above the air-water interface inside the chamber (Figure 3a). Over time, the meniscus of the bulk liquid receded in an oscillatory manner depending on how many dew droplets formed above the interface; and dried material precipitated on the warm side as a consequence (Figure 3b). The dew droplets grew at the cooler side of the chamber by surface-tension driven fusion and made contact with the warm side, rehydrating the dried material and transporting it back into the bulk.[37] This phenomenon was allowed to continue for 18 hours, after which the setup was dismounted and the remaining bulk liquid and the dried flakes (after dissolution) were sampled for analysis.

The pH of the samples at the end the reaction was found to be lowered by a pH unit, indicating the formation of acidic species in the reaction mixture. A likely cause of the pH drop is the acidification by the ring opening of the cyclic phosphate in the mononucleotides and the oligomers (S4e, S5c, d). At higher temperatures, the pH drop was 1.5 to 2 pH units (S4e).

Despite the presence of bulk water, the oligomerisation inside the simulated volcanic-rock pore showed comparable yields as that of the heat-dried conditions. This indicates that the heated interface can access conditions favourable for polymerisation similar to bulk dried polymerisation conditions. The constant feeding of monomers from the bulk fluid could also be an important factor. A length-selective enzymatic DNA replication was reported recently within this setting, indicating the possible continuity of prebiotic chemistry in such a setting.[37]

We observed all the dimer sequence combinations and most of the trimers (Figure 3c). However, the tetramers and pentamers are predominantly sequences rich in G. The length selectivity of the HPLC allowed the detection of longer sequences. However, the isotopic fit to the raw mass spectra provided by our LabView-based analysis showed that longer species with concentrations lower than 0.2 μM were lost in the background noise of the mass spectra. Moreover, different oligomers can have similar masses (eg. Table S3 and S4), so to avoid false positives, sequences with mass overlaps were not included here. This is in addition to the rigid selection criteria, based on fitting of the isotopic distribution (S12) and only considering mass spectra within the optimised n-mer retention times of the HPLC. A full sequence composition analysis for GC and GCAU mixtures with comparison between dry polymerisation and simulated rock-pore polymerisation is provided in S11. In comparison, CAU reaction mixture yielded only dimers (S9c), indicating again the central role of G in the copolymerisation process.

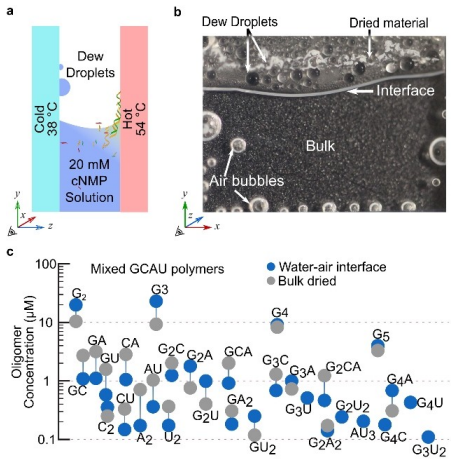


Figure 3. RNA oligomerisation in the vicinity of air inclusions in a heated simulated rock pore. (a) Side view. The chamber is 500 μm in depth and subjected to a heat flow with a temperature gradient of 38–54 C. (b) Front view. The thermal gradient drives continuous evaporation and recondensation in the air inclusions, triggering accumulation and wet-dry cycles. Molecules accumulated at the interface are dried from a receding interface due to evaporation. Rehydration is provided by dew droplets on the cold side which merge with the bulk solution due to surface tension. (c) Oligomerisation of four canonical monomers: cGMP, cCMP, cAMP and cUMP, 5 mM each, 40 mM KCl at pH 10 for 18 hours. Especially for the longer strands, the oligomerisation in the simulated rock pore shows improved yields over the dry reaction. The physically triggered wet-dry cycling and length selectivity in this environment has been shown to drive efficient replication and selection cycles,[37] making the finding of oligomerisation to provide the raw material for templated ligation very interesting. Moreover, this shows that oligomerisation under simulated geological conditions is possible without the need for arid conditions on early Earth. The trends show a rich set of mixed short sequences when all four nucleotides are mixed together for oligomerisation.

Computational study of the proposed intercalated stacked arrangement

Based on the hypothesis that a stack-assisted geometry is triggering the oligomerisation of 3′,5′-cGMP,[26] we studied the suitability of intercalated stack arrangements for the oligomerisation of 2′,3′-cNMP. We explored the stability of the stack arrangements, and the incorporation of cNMP monomers into polymerised cGMP scaffold, based on minimum energy structures and molecular dynamics simulations (Figure 4a–c and S24–34).

To investigate the suggested intercalated stack arrangement for several possible species, we have computed the stacking interaction energies and evaluated the minimum energy geometries obtained at ωB97 M-V/def2-TZVPD level of theory.[38–41] All systems were studied in the gas phase as well as with implicit solvation (C-PCM).[42] The quantum mechanical

ChemSystemsChem
Research Article
doi.org/10.1002/syst.202200026
Chemistry Europe
European Chemical Societies Publishing

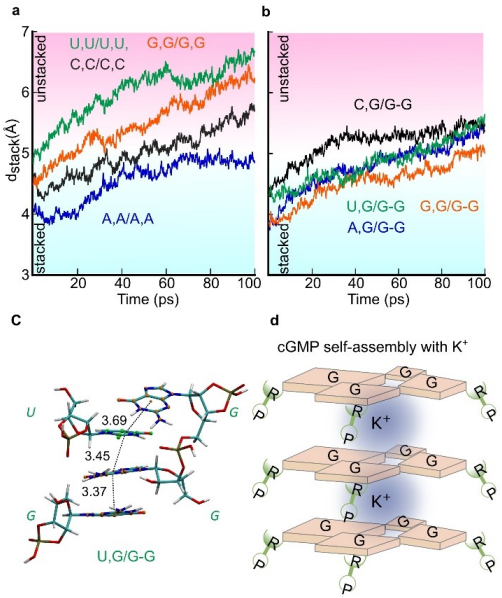


Figure 4. Possible supramolecular assemblies facilitating polymerisation of cNMP. Molecular dynamics simulations suggest polymerisation of A, U and C by intercalating into stacks of oligoG. (a) Distances between bases in the complex of unpolymerised nucleotides show that polymerisation is disfavoured due to drifting away of the complex. (b) The assembly formed when the bases are templated by covalently linked G–G (2′-5′) dimers, forms the most stable complexes and make possible the incorporation of the G, C, A, and U within the n-mers observed in our experiments. (c) Snapshot after energy minimization of stacking interaction between a oligomerised scaffold of cGMP (G–G) to template the cGMP and cUMP monomers (U,G). The dotted lines mark distances between the bases used to evaluate the stability of the complex. (d) However, the self-oligomerisation of oligoG could also be based on stacks of G-tetrads, stabilised by inner K^+ ions, coinciding with the promotion of oligoG formation by K^+ ions. R denotes the ribose of RNA.

computations were performed using FermiONs++[43–45] in combination with Chemshell.[46]

These computations were complemented by GFN-FF molecular dynamics simulations,[47] for the systems encapsulated in an explicit water sphere using xtb.[48] The stacking of homogeneous monomers were tested (N,N/N,N) with N = A, U, G, C and the incorporation of monomers into a dimer and trimer scaffold of G was probed (N,N/G–G or N,N/G–G–G). The 3′-5′ linked G–G and G–G–G accommodate A, U and G monomers into the scaffold providing a stable arrangement for the initiation of polymerisation. For C an alternate arrangement involving hydrogen bonding with a G within the scaffold is observed (S28). We found that a 2′-5′ oligoG scaffold seemed to enhance the alignment (Figure 4a, b, c), confirmed both by static and dynamic computations (S30, S31).

Theoretical model of cGMP polymerisation

Additional evidence supporting a stacked polymerisation mechanism comes from the observed non-exponential length distribution of the oligomer concentrations. This supports the idea that the formation of dimers is the rate limiting step: the concentration drop from monomers to dimers was most significant. For the cGMP oligomerisation in Figure 1a, the 20 mM monomer concentrations drop to 0.15 mM for G_2, then forming a flat concentration plateau, in contrast to the typical exponential length distribution in homogeneous polymerisation.[49]

To test this idea, we fit the concentration distribution of G homooligomers with a stacked polymerisation model (solid line Figure 1a and 2a). The model assumed a three-step polymerisation reaction: i) a monomer of length i and a oligoG scaffold k can stack together with rate ν, ii) the de-stacking rate $\delta_{k,i}$ decreases exponentially with the number of stacked bases $n_{k,i}$, iii) another monomer of length j can stack to the complex. If the stacks persist long enough, the polymerisation reaction ligates the two monomers with rate ρ (see for details S18–S19). The model fits the experimental data, suggesting a rate limiting step for the formation of short oligomers due to the required mutual alignment. It should be noted that it is difficult to distinguish between inter-base stacking or a plausible G-tetrad arrangement suggested based on the enhanced polymerisation observed with K^+ (Figure 4d).

Discussion

Our data suggests that cGMP oligomerises in dry state at moderate temperatures and pH. The oligomerisation occurs over a range of temperatures (40–80 C) and pH (7–12) and does not require additional catalysts, making this reaction robust. Dissolved gases and salts could adjust the pH of the environment, making RNA formation more probable under early Earth models.[50,51] We also showed polymerisation in the wet-dry cycling environment at a heated air-water interface, adding RNA polymerisation to the pool of prebiotic processes possible within such a setting.[35–37] The tested conditions of wet-dry cycles at an air-water interface or direct drying keep the reaction out of equilibrium. The cyclic monomers undergo polymerisation and ring-opening (Figure 1d), of which the ring-opening is still the dominant product at the tested temperatures (S4). Under the tested conditions, the reaction yielded oligomers up to 15-mers. The formed oligoG incorporated cCMP, cAMP and cUMP monomers, albeit in lower concentration, which did not homooligomerise significantly. As a rough comparison to the yields achieved by Verlander and Orgel with homooligomers of cAMP in the presence of ethane-1,2-diamine, we observed ~0.35% for a 6-mer of oligoG in 18 hours compared to 0.81% for polyA in 40 days.[6]

An important feature of this oligomerisation is that the 2′,3′-cyclic phosphate group, under alkaline pH, is sufficient to trigger oligomerisation without ex-situ or in-situ activation mechanism or added catalysts, and under low salt conditions.

ChemSystemsChem
Research Article
doi.org/10.1002/syst.202200026
Chemistry Europe
European Chemical Societies Publishing

The finding that the oligomerisation starts without added catalysts – and that the reaction site is not yet blocked by a catalyst – is a very good starting point for Darwinian evolution to speed up this reaction rate. Low salt conditions are interesting for RNA evolution since they notably help strand separation and reduce RNA degradation.[36]

cNMP oligomerisation is found to be a relatively clean reaction under the tested conditions. In comparison, *in situ* EDC activation yields side products, especially at high temperatures.[52] We did not detect any major side products with ESI-TOF, other than the salt adducts of sodium and potassium.

The abiotic formation and recyclability of the cNMP monomers is feasible, as they are known to be produced under several phosphorylation conditions, nucleotide syntheses and are common degradation products of RNA.[19–23] Thus, with the likelihood of finding catalytic boosts for this found reaction mode, a cycle of reactions involving polymerisation, oligomer extension, polymer hydrolysis and reactivation of monomers under early Earth conditions becomes conceivable. Furthermore, recombination and templated ligation involving 2′,3′-cyclic ending oligomers[53] have been observed.

For our studies, we compared two monovalent ions (K^+, Na^+) and one divalent cation (Mg^{2+}). They were chosen for their relevance in contemporary cytosolic media, their abundance on the early Earth[54] and for the role of Mg^{2+} in ribozyme activity.[55] Polymerisation is enhanced in the presence of K^+ in comparison to Na^+ ions. The inhibition by Mg^{2+} ions possibly occurs by a combination of base catalysis mechanism, the deactivation of -cP ends of the reactant, products and enhanced oligomer hydrolysis. Despite its role in ribozyme functionality, at high concentrations Mg^{2+} inhibits RNA replication by creating strong RNA duplexes, limiting thermal denaturation and enhancing temperature dependent hydrolysis.[56] It is also known that the presence of ~1.5 mM Mg^{2+} is sufficient to inhibit the membrane self-assembly of fatty acids and this has been considered an incompatible aspect for the co-emergence of RNA and fatty acid membranes.[57,58] However, under the discussed reaction conditions of cNMP oligomerisation, RNA formation and encapsulation with fatty acids might be conceivable within freshwater locations on the primordial Earth. Moreover, we have shown that efficient strand separation can be achieved by low sodium concentrations, triggered by microscale water cycles within heated rock pores.[36,37]

Our very preliminary digestion studies and ^{31}P NMR results suggest a considerable backbone heterogeneity (2′-5′ and 3′-5′) within the oligomers. However, a full quantitative treatment is beyond the scope of this study. It has been demonstrated that the presence of 2′-5′ linkages allow efficient strand separation by reducing the melting temperature (*Tm*) of oligomers, which is pertinent in the case of G-rich sequences that are observed in this oligomerisation.[59] Lowering of *Tm* is critical to replication of sequences.[59,60] These studies also show that the presence of 2′-5′ linkages allow the folding of RNA into three-dimensional structures, similar to native linkages and do not hinder the evolution of functional RNAs, such as ribozymes. The susceptibility towards enhanced hydrolysis of the 2′-5′ over the 3′-5′ linkages could select the latter in wet-dry cycling conditions, similar to the reported backbone selection of RNA and DNA.[59,61,62]

Mechanistically, molecular dynamics studies indicated that cGMP oligomerisation could be due to the formation of intercalated stacks of cGMP as a consequence of hydrophobic interactions between the guanine bases. On attaining a stable intermolecular arrangement, the 5′-OH of a nucleotide can attack the cyclic phosphate of the neighbouring nucleotide. This could allow the formation of oligomeric G-scaffolds (Figure 4c, d, S25). However, the formation of tetrad stacks over one another with a central K^+ ion between the stacks could also promote oligomerisation (Figure 4d).

The notion of multi-molecular assemblies is supported by the presence of slow-diffusing species observed in ^{1}H, ^{31}P diffusion ordered spectroscopy (DOSY) of cGMP-KCl solution (S15, S16). Reports in literature point to self-assembly of 5′-GMP and 3′-GMP into helical stacks.[63] The presence of several slow-diffusing species indicate a range of molecular environments, making it impossible to identify a single type of self-assembly by NMR. It has also been reported that G-quadruplex structures could be stable up to a pH of ~10.8 at ambient temperatures.[64,65]

The formation of dimers appears to be a limiting step in the oligomerisation. Such a threshold behaviour is known to be an optimal control strategy for self-assembly processes.[66] With this, monomers remain available in high concentration, leading to long-tailed, non-exponential polymer distributions. This limits the total efficiency of the polymerisation but favours the formation of the oligomers, important for downstream reactions such as templated replication.

It should be noted that an efficient generation of very long and random RNA sequences would make hybridisation and replication inconceivable. At this point, the generated G-rich sequences might not seem optimal for hybridisation and replication. However, a biased pool of short oligomers (10- to 15-mers) further constrains the sequence space, favouring selectivity and making templated replication plausible.[11,67,68] We think that the findings are a first step to provide oligonucleotides for templated ligation and the emergence of an evolutionary dynamics with RNA.

Conclusion

We report the oligomerisation of canonical nucleotides that produced RNA of mixed sequences under drying conditions in bulk and at heated air-water interface. A wide range of temperatures (40–80 C) and pH (7–12) promoted oligomerisation. Best yields were reported by mild heating (40 C) of monomers at low salt concentrations and under alkaline drying conditions (pH 10). The reaction proceeded best at 1–2 equivalents of K^+ and Na^+, while Mg^{2+} ions inhibited it. In an equal mixture of four nucleotides, equal incorporation of all four was not observed and the mixed sequences were dominated by G. However, 2′,3′-cGMP fostered the incorporation of the otherwise scarcely reactive C, A, and U, generating short, mixed

ChemSystemsChem | Research Article doi.org/10.1002/syst.202200026 | Chemistry Europe European Chemical Societies Publishing

sequences. This reaction under the tested temperatures, pH and salt conditions provide a novel route to fresh water oligomerisation towards short RNA strands, an important intermediate step towards providing the raw materials for an RNA-based emergence of life.

Acknowledgements

We would like to thank Ulrich Gerland, Tobias Göppel, Joachim Rosenberger and Bernhard Altaner for their helpful remarks and discussions; Thomas Matreux, Alexandra Kühnlein, Noël Yeh Martin and Maximilian Weingart for comments on the manuscript. The authors thank J. Kussmann (LMU Munich) for providing a development version of the FermiONs++ program package. Financial support was provided by the European Research Council (ERC Evotrap, grant no. 787356, the Simons Foundation (grant no. 327125), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 364653263 – TRR 235 (CRC 235), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311, and the Center for NanoScience. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords: RNA · Polymerisation · Prebiotic chemistry · Non-equilibrium · Air-water interfaces

[1] H. S. Bernhardt, *Biol. Direct* **2012**, *7*, 1–10.
[2] G. Costanzo, S. Pino, A. M. Timperio, J. E. Šponer, J. Šponer, O. Nováková, O. Šedo, Z. Zdráhal, E. Di Mauro, *PLoS One* **2016**, *11*, 1–14.
[3] J. P. Ferris, A. R. Hill, R. Liu, L. E. Orgel, *Nature* **1996**, *381*, 59–61.
[4] M. S. Verlander, R. Lohrmann, L. E. Orgel, *J. Mol. Evol.* **1973**, *2*, 303–316.
[5] C. V. Mungi, N. V. Bapat, Y. Hongo, S. Rajamani, *Life* **2019**, *9*, 1–11.
[6] M. S. Verlander, L. E. Orgel, *J. Mol. Evol.* **1974**, *3*, 115–120.
[7] A. Luther, R. Brandsch, G. Von Kiedrowski, *Nature* **1998**, *396*, 245–248.
[8] S. J. Zhang, D. Duzdevich, D. Ding, J. W. Szostak, *Proc. Nat. Acad. Sci.* **2022**, *119*, 2021.09.07.459201.
[9] T. Walton, W. Zhang, L. Li, C. P. Tam, J. W. Szostak, *Angew. Chem. Int. Ed.* **2019**, *58*, 10812–10819; *Angew. Chem.* **2019**, *131*, 10926–10933.
[10] E. Kervio, M. Sosson, C. Richert, *Nucleic Acids Res.* **2016**, *44*, 5504–5514.
[11] C. Richert, G. Leveau, D. Pfeffer, B. Altaner, E. Kervio, U. Gerland, F. Welsch, *Angew. Chem. Int. Ed.* **2022**, *202203067*, 1–6.
[12] L. Li, N. Prywes, C. P. Tam, D. K. Oflaherty, V. S. Lelyveld, E. C. Izgu, A. Pal, J. W. Szostak, *J. Am. Chem. Soc.* **2017**, *139*, 1810–1813.
[13] D. K. O'Flaherty, N. P. Kamat, F. N. Mirza, L. Li, N. Prywes, J. W. Szostak, P. Sheeringa, *J. Am. Chem. Soc.* **1997**, *140*, 1–5.
[14] E. I. Jiménez, C. Gibard, R. Krishnamurthy, *Angew. Chem. Int. Ed.* **2020**, *60, 19, 1077–10783*, DOI 10.1002/anie.202015910.
[15] Z. Liu, L. F. Wu, J. Xu, C. Bonfio, D. A. Russell, J. D. Sutherland, *Nat. Chem.* **2020**, *12*, 3–10.
[16] H. J. Kim, S. A. Benner, *Astrobiology* **2021**, *21*, 298–306.
[17] M. W. Powner, B. Gerland, J. D. Sutherland, *Nature* **2009**, *459*, 239–242.
[18] Y. Yamagata, H. Inoue, K. Inomata, *Origins Life Evol. Biospheres* **1995**, *25*, 47–52.
[19] R. Breslow, *Acc. Chem. Res.* **1991**, *24*, 317–324.
[20] Y. Li, R. R. Breaker, *J. Am. Chem. Soc.* **1999**, *121*, 5364–5372.
[21] H. Peng, B. Latifi, S. Müller, A. Lupták, I. A. Chen, *RSC Chem. Biol.* **2021**, *2*, 1370–1383.
[22] A. M. Pyle, *Science* **1993**, *261*, 709–714.
[23] S. I. Nakano, D. M. Chadalavada, P. C. Bevilacqua, *Science* **2000**, *287*, 1493–1497.
[24] M. Morasch, C. B. Mast, J. K. Langer, P. Schilcher, D. Braun, *ChemBioChem* **2014**, *15*, 879–883.
[25] J. E. Šponer, J. Šponer, A. Giorgi, E. Di Mauro, S. Pino, G. Costanzo, *J. Phys. Chem. B* **2015**, *119*, 2979–2989.
[26] S. Wunnava, C. F. Dirscherl, J. Výravský, A. Kovařík, R. Matyášek, J. Šponer, D. Braun, J. E. Šponer, *Chem. A Eur. J.* **2021**, *27, 70, 17581–17585*, DOI 10.1002/chem.202103672.
[27] C. M. Tapiero, J. Nagyvary, *Nature* **1971**, *231*, 42–43.
[28] S. Dagar, S. Sarkar, S. Rajamani, *RNA* **2020**, *26*, 756–769.
[29] J. E. Šponer, J. Šponer, A. Kovařík, O. Šedo, Z. Zdráhal, G. Costanzo, E. Di Mauro, *Life* **2021**, *11*, 1–13.
[30] A. Premstaller, P. J. Oefner, *LCGC Eur.* **2002**, *15, 7, 410–422.*
[31] A. Premstaller, P. J. Oefner, *Denaturing High-Performance Liquid Chromatography*, Humana Press, New Jersey, **n.d.**
[32] P. B. Danielson, R. Kristinsson, R. J. Shelton, G. S. LaBerge, *Expert Rev. Mol. Diagn.* **2005**, *5*, 53–63.
[33] H. R. Palmer, J. J. Bedford, J. P. Leader, R. A. J. Smith, *J. Biol. Chem.* **2000**, *275*, 27708–27711.
[34] S. Motsch, D. Pfeffer, C. Richert, *ChemBioChem* **2020**, *21, 2013–2018*.
[35] M. Morasch, J. Liu, C. F. Dirscherl, A. Ianeselli, A. Kühnlein, K. Le Vay, P. Schwintek, S. Islam, M. K. Corpinot, B. Scheu, D. B. Dingwell, P. Schwille, H. Mutschler, M. W. Powner, C. B. Mast, D. Braun, *Nat. Chem.* **2019**, *11*, 779–788.
[36] A. Ianeselli, C. B. Mast, D. Braun, *Angew. Chem. Int. Ed.* **2019**, *58*, 13155–13160; *Angew. Chem.* **2019**, *131*, 13289–13294.
[37] A. Ianeselli, M. Atienza, P. W. Kudella, U. Gerland, C. B. Mast, D. Braun, *Nat. Phys.* **2022**, *18, 579–585*, DOI 10.1038/s41567-022-01516-z.
[38] N. Mardirossian, M. Head-Gordon, *J. Chem. Phys.* **2016**, *144*, DOI 10.1063/1.4952647.
[39] O. A. Vydrov, T. Van Voorhis, *J. Chem. Phys.* **2010**, *133*, DOI 10.1063/1.3521275.
[40] F. Weigend, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.
[41] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
[42] M. Cossi, N. Rega, G. Scalmani, V. Barone, *J. Comput. Chem.* **2003**, *24*, 669–681.
[43] J. Kussmann, C. Ochsenfeld, *J. Chem. Phys.* **2013**, *138*, DOI 10.1063/1.4796441.
[44] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2015**, *11*, 918–922.
[45] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2017**, *13*, 3153–3159.
[46] P. Sherwood, A. H. De Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjøvoll, A. Fahmi, A. Schäfer, C. Lennartz, *J. Mol. Struct.* **2003**, *632*, 1–28.
[47] S. Spicher, S. Grimme, *Angew. Chem.* **2020**, *132*, 15795–15803; *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.
[48] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, 1–49.
[49] C. B. Mast, S. Schink, U. Gerland, D. Braun, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 8030–8035.
[50] J. F. Kasting, M. T. Howard, *Philos. Trans. R. Soc. London Ser. B* **2006**, *361*, 1733–1741.
[51] J. Krissansen-Totton, G. N. Arney, D. C. Catling, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4105–4110.
[52] E. Edeleva, A. Salditt, J. Stamp, P. Schwintek, J. Boekhoven, D. Braun, *Chem. Sci.* **2019**, *10*, 5807–5814.
[53] A. V. Lutay, E. L. Chernolovskaya, M. A. Zenkova, V. V. Vlasov, *Dokl. Biochem. Biophys.* **2005**, *401*, 163–166.
[54] S. Maurer, *Life* **2017**, *7*, DOI 10.3390/life7040044.
[55] T. S. Lee, C. S. López, G. M. Giambaşu, M. Martick, W. G. Scott, D. M. York, *J. Am. Chem. Soc.* **2008**, *130*, 3053–3064.

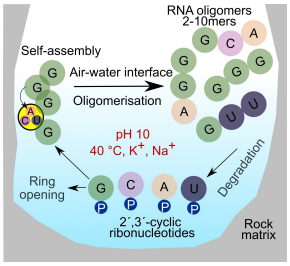
ChemSystemsChem
Research Article
doi.org/10.1002/syst.202200026
Chemistry Europe
European Chemical Societies Publishing

[56] A. Salditt, L. M. R. Keil, D. P. Horning, C. B. Mast, G. F. Joyce, D. Braun, *Phys. Rev. Lett.* **2020**, *125*, 48104.
[57] I. A. Chen, K. Salehi-Ashtiani, J. W. Szostak, *J. Am. Chem. Soc.* **2005**, *127*, 13213–13219.
[58] K. Adamala, J. W. Szostak, *Science* **2013**, *342*, 1098–1100.
[59] A. E. Engelhart, M. W. Powner, J. W. Szostak, *Nat. Chem.* **2013**, *5*, 390–394.
[60] J. Sheng, L. Li, A. E. Engelhart, J. Gan, J. Wang, J. W. Szostak, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3050–3055.
[61] S. Bhowmik, R. Krishnamurthy, *Nat. Chem.* **2019**, *11*, 1009–1018.
[62] A. Mariani, J. D. Sutherland, *Angew. Chem. Int. Ed.* **2017**, *56*, 6563–6566; *Angew. Chem.* **2017**, *129*, 6663–6666.
[63] M. Gellert, M. N. Lipsett, D. R. Davies, *Proc. Natl. Acad. Sci. USA* **1962**, *48*, 2013–2018.
[64] Y. Y. Yan, J. H. Tan, Y. J. Lu, S. C. Yan, K. Y. Wong, D. Li, L. Q. Gu, Z. S. Huang, *Biochim. Biophys. Acta Gen. Subj.* **2013**, *1830*, 4935–4942.
[65] R. Del Villar-Guerra, R. D. Gray, J. B. Chaires, *Curr. Protoc. Nucleic Acid Chem.* **2017**, *2017*, 17.8.1–17.8.16.
[66] F. M. Gartner, I. R. Graf, E. Frey, *Proc. Natl. Acad. Sci. USA* **2022**, *119*, DOI 10.1073/pnas.2116373119.
[67] P. W. Kudella, A. V. Tkachenko, A. Salditt, S. Maslov, D. Braun, *Proc. Natl. Acad. Sci. USA* **2021**, *118*, DOI 10.1073/pnas.2018830118.
[68] S. Toyabe, D. Braun, *Phys. Rev. X* **2019**, *9*, 11056.

Manuscript received: July 29, 2022
Accepted manuscript online: September 19, 2022
Version of record online: ■■, ■■■

RESEARCH ARTICLE

Without catalyst! A non-enzymatic, spontaneous route to the formation of short, mixed-sequence RNA strands from 2′,3′-cyclic ribonucleotides is shown. The reaction occurs by drying at the air–water interfaces.



*Dr. A. V. Dass, S. Wunnava, J. Langlais, B. von der Esch, M. Krusche, L. Ufer, N. Chrisam, R. C. A. Dubini, Dr. F. Gartner, S. Angerpointner, C. F. Dirscherl, Dr. P. Rovó, Dr. C. B. Mast, Dr. J. E. Šponer, Prof. Dr. C. Ochsenfeld, Prof. Dr. E. Frey, Prof. Dr. D. Braun**

1 – 9

RNA Oligomerisation without Added Catalyst from 2′,3′-Cyclic Nucleotides by Drying at Air-Water Interfaces

ChemSystemsChem

Supporting Information

RNA Oligomerisation without Added Catalyst from 2′,3′-Cyclic Nucleotides by Drying at Air-Water Interfaces**

Avinash Vicholous Dass[+], Sreekar Wunnava[+], Juliette Langlais[+], Beatriz von der Esch, Maik Krusche, Lennard Ufer, Nico Chrisam, Romeo C. A. Dubini, Florian Gartner, Severin Angerpointner, Christina F. Dirscherl, Petra Rovó, Christof B. Mast, Judit E. Šponer, Christian Ochsenfeld, Erwin Frey, and Dieter Braun*

Supplementary Material

Contents

Materials and methods

2´,3´-cyclic nucleotide monophosphates (cNMP's): cGMP (monosodium salt), cCMP (monosodium salt), cAMP (monosodium salt), cUMP (monosodium salt). cGMP and cUMP were purchased from Biolog Life Science Institute GmbH & Co. KG. cAMP and cCMP were purchased from Sigma Aldrich. KOH, NaOH, HCl, $MgCl_2$, KCl and Imidazole were purchased from Carl Roth. Oligomer standards for HPLC-MassSpec calibration were purchased from biomers.net GmbH. For the oligo linkage analysis, the enzyme Nuclease P_1 from *Penicillium citrinum* was purchased from Sigma Aldrich. The 2´- 5´ linked 5mers were purchased from biomers.net GmbH.

Reaction

From stock solutions of 200mM cNMP's, 20mM of each cNMP is prepared and the volume is made up to 100µL, including pH adjustments carried out using KOH and HCl. When reactions are carried out with KCl salts, then KOH is used to adjust the pH, NaOH for NaCl. For MgCl2, the pH was adjusted using KOH. The 100µL sample is then heated on a heat block for 18 hours, unless the time is specifically mentioned. After drying the sample, the dried pellet is rehydrated with 20µL of nuclease-free water. To this volume, 2µL of 10mg/mL of glycogen is added. Then 2µL of 5M ammonium acetate is added. To this volume 100µL of cold 100% ethanol is added and kept overnight at 4 °C. Then the sample is then centrifuged at 4 °C for 30 minutes at 21000 rpm. The supernatant is then decanted. To the pellet 100µL of 70% ethanol is added, tapped gently and then centrifuged again at 4°C, 30 minutes, 21000 rcf/15000 rpm and finally decanted. To the pellet, 100µL of nuclease-free water is added, vortexed and the same volume is transferred into HPLC vials for anaylsis. For pH determination of the reactions, bulk dried samples were rehydrated with 100µL of nuclease-free water. After the pH of these samples were determined, the samples were subjected to ethanol precipitation and subsequently, used for HPLC-MS analyses. In the case of simulated rock pore reactions, the bulk of the liquid (90%) remained post-reaction and this was extracted and used directly for pH determination before further analyses.

Determination of the phosphodiester linkage by Nuclease P1

Nuclease P1 digestion assay was done to assess the type of the phosphodiester linkages in the oligomerised samples. 3'-5' phosphodiester linkages are susceptible to Nuclease P1 while the 2'-5' linkages are not. This had been used previously to determine the 2'-5' linkages in ligation site of Peach Latent Mosaic Viroid.[1] For the assay, 100 µL of 20 mM 2',3'-cGMP, pH 10 was oligomerised by drying at 40 °C for 24 hours. Oligomerised samples were ethanol precipitated and re-dissolved to contain 100 µM equivalents of phosphodiester bond.

Nuclease P1 from *Penicillium citrinum* (Sigma-Aldrich N8630) was dissolved in P1 storage buffer (25 mM Tris-HCl, 50 mM NaCl, 1 mM $ZnCl_2$, 50% Glycerol, pH 7.2 at 25 °C) at a concentration >0.5 U/µL. The digestion reaction was done in P1 reaction buffer (50 mM Sodium Acetate, pH 5.5) with oligomerised 2',3'-cGMP sample concentration of 1 µM phosphodiester bond equivalents and 0.5 U enzyme in a 100 µL volume. For the control, 1 µL of P1 Storage buffer without the Nuclease P1 was used. The samples were incubated at 37 °C for 10, 30 or 240 minutes following which the enzyme was inactivated by heating at 75 °C for 10 minutes. Phenol-chloroform-isoamyl alcohol (Carl-Roth A156.3) extraction was done by adding equal volume to the sample followed by vigorous mixing and centrifugation at 15000 rpm. The top aqueous phase was removed carefully and was used for LCMS analysis.

For the LCMS analysis, the EIC (extracted ion chromatogram) for the oligomers (Linear-P, Cyclic-P and with no-P ends) were extracted and the concentration equivalent of the phosphodiester linkages was calculated for each sample and normalized to one of the no enzyme control (Figure S14, S15 and S16). For Figure S14 and Figure S15 control with 10-minute incubation (C10) and for Figure S16 control with 30-minutes of incubation (C30) was used for normalisation.

Figure 14a, 15a and 16a show the UV 260 nm chromatogram of the control samples as an example for retention times with annotations made to mark the peaks shown in the corresponding b-i figures. 100 µL of 20 mM 2',3'-cGMP, pH 10 was oligomerised by drying at 40 °C for 24 hours. Oligomerised samples were ethanol precipitated and re-dissolved to 100 µL. The digestion assay was done three times and the data for each is presented in Figures 21, 22 and 23 respectively.

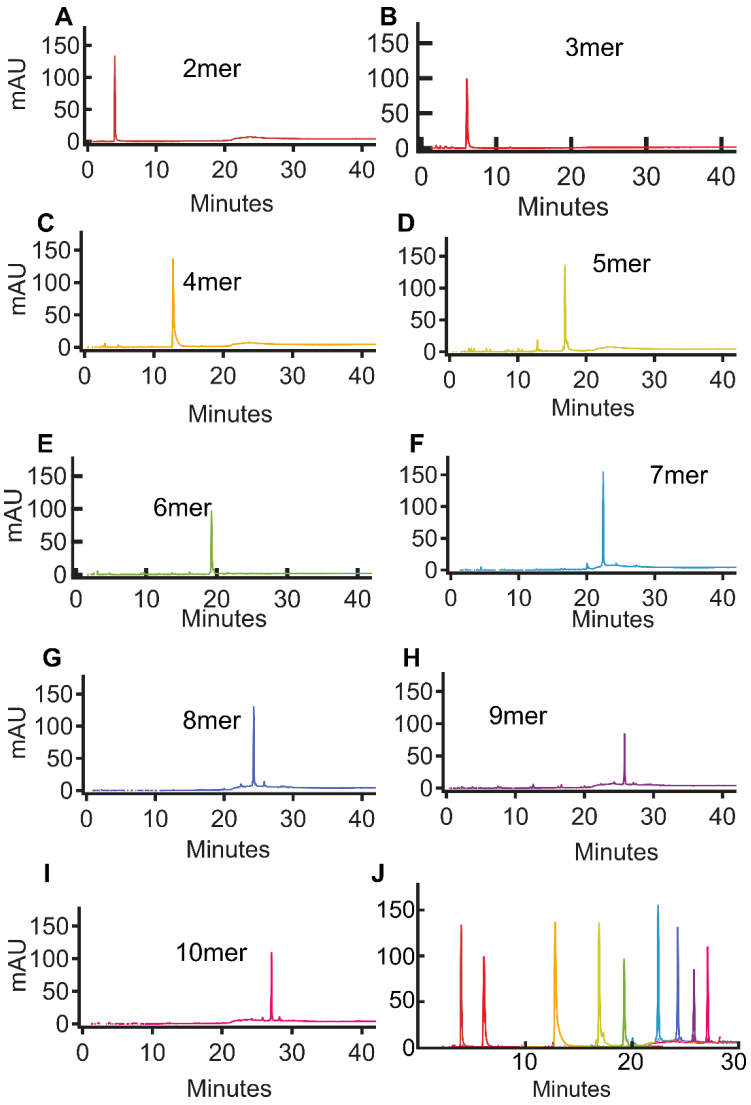
HPLC protocol

The measurements were performed on Agilent G6230BA LC/TOF with G5654A 1260 InfinityII bioinert HPLC with a G7115A 1260 Infinity diode array detector. For the HPLC analysis, we utilised Agilent AdvanceBio Oligonucleotide (4.6 x 150 mm, 2.7 µm) column with a pressure rating of 600 bar, inner diameter (ID) of 4.6 mm, length150 mm and particle size of 2.7 µm.

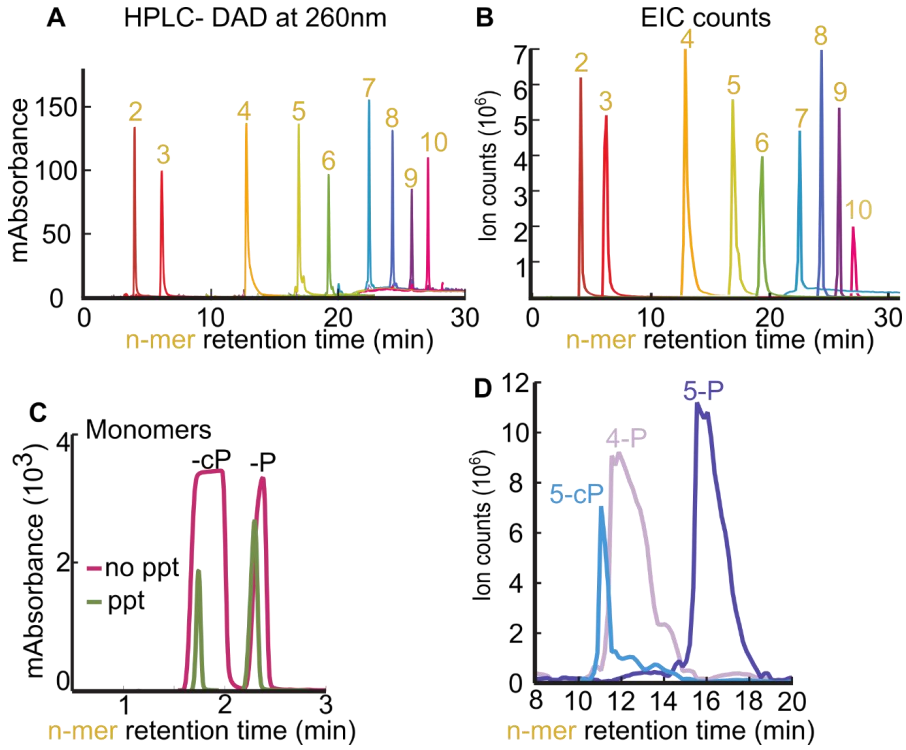
1) The standards and oligomerisation samples were analysed in water-methanol solvent system. The solvent composition of bottle A was 200mM HFIP and 8mM of TEA in water and bottle B was 50:50 methanol-water with 200mM HFIP and 8mM TEA. The samples were analysed in step gradient flow with an initial isocratic step of 1%B flushed for first 5 minutes, followed by an isocratic flow of 1% up to 30% B in 22.5 minutes and a final isocratic flow of 30% up to 40% B in the next 15 minutes at a constant flow rate of 1 mL/minute. The multisampler G5668A was set at sampling speed of 100µL/min, ejection speed of 400µL/min and the needle height position set at 0.3mm. The column temperature was set at 60 °C for denaturation HPLC conditions as recommended in literature.[2–5] The columns are always equilibrated at 60 °C prior to loading of the samples to the column. The separation efficiency of the oligomers is clearly seen in Fig 1C of the main manuscript and in figure S1. The separation efficiency of each n-mer on the HPLC and the corresponding n-mer mass counts are displayed in figure S2.

2) For the -cP and –P monomer analysis, the samples were analysed in water-methanol solvent system. The solvent composition of bottle A was 200mM HFIP and 8mM of TEA in water and bottle B was 50:50 methanol-water with 200mM HFIP and 8mM TEA. The samples were analysed in step gradient flow with an isocratic flow of 1%B flushed for first 5 minutes, followed by upto 4% B in 3 minutes, 3.3 minutes at up to 8%B, next 4.8 minutes at up to 35%B at a constant flow rate of 1 mL/minute. The multisampler G5668A was set at sampling speed of 100µL/min, ejection speed of 400µL/min and the needle height position set at 0.3mm. The column temperature was set at 30 °C for HPLC conditions in this case.

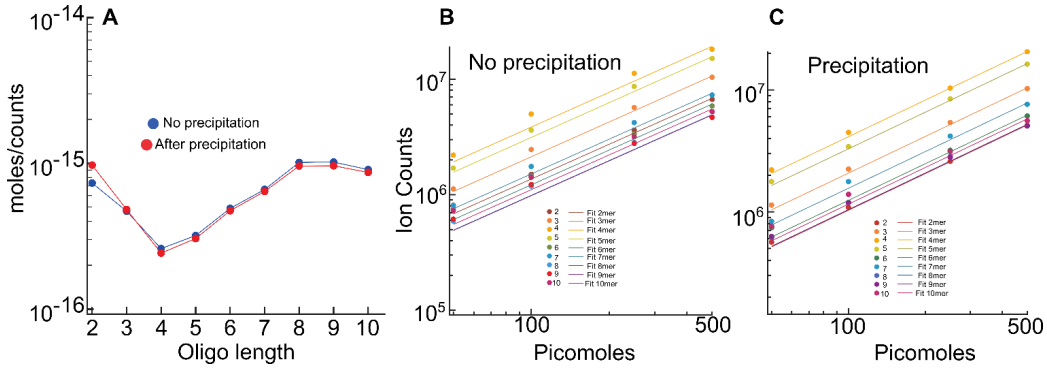
The Mass spectrometer-TOF is equipped with a dual AJS ESI ion source. The instrument parameters of the TOF were set to a gas temperature of 325 °C, gas flow at 8 l/min, nebulizer at 45 psig, sheath gas temperature at 400 °C and sheath gas flow at 11 l/min. The samples were anaylsed in the negative ion mode. The scan source parameters were as follows: Vcap 3000, nozzole voltage at 2000V, fragmentor at 175V, Skimmer at 65 and octapoleRFpeak value of 750. Reference masses were run in parallel to the sample run using standard reference/calibration and tuning mix recommended by Agilent (product number G1969-85000).



***S1*:** 50 µM of each polyG n-mer (2 to 10-mer) was mixed with 50 µM KCl, incubated for 2 hours and then injected into the ESI HPLC-MS. The oligomers in this test were not precipitated and injected directly into the HPLC-MS to demonstrate the separation efficiency of our HPLC protocol even without the ethanol precipitation.
A-I) Efficient separation is observed in the HPLC-DAD (diode array detector) at 260nm. They are simultaneously injected into the ESI-MS after their elution from the HPLC column. **J**) Shows the collective HPLC-DAD chromatograms of the 2-10mers. The colours indicated in figures A-I correspond to 2 to 10mers and the same colours correspond to 2-10mers in the collective chromatogram.



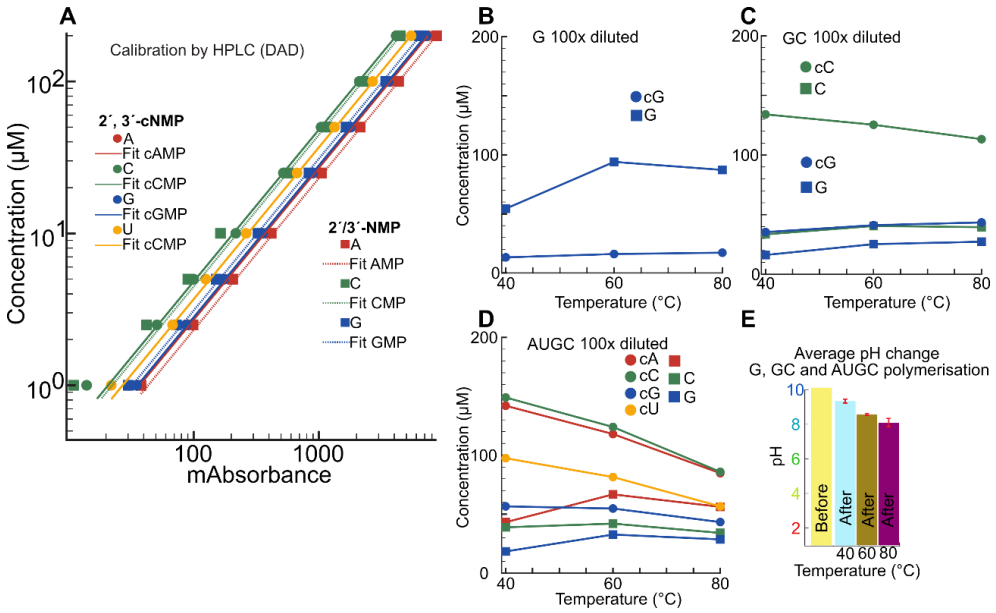
***S2*:** 50 µM of each polyG n-mer (2 to 10-mer) were mixed with 50µM KCl, incubated for 2 hours and then injected into the ESI HPLC-MS. The oligomers in this test were not precipitated.
A) Efficient separation is observed with the HPLC-DAD (diode array detector) at 260nm. They are simultaneously injected into the ESI-MS after their elution from the HPLC column. **B)** The eluted peaks of each n-mer (2-10mer) are then detected based on their masses and is confirmed by the MS peaks. The Y-axis displays the ion counts plotted against the corresponding HPLC retention times (x-aixs) of the n-mers. **C)** 20 mM of monomer was injected to compare the amount of monomers removed after the ethanol precipitation process and compared to an non-precipitated sample. This was analysed by HPLC with diode array detector (DAD) with 260 nm UV detection. The results show that precipitation of the reaction mixture was favourable to prevent the saturation of the DAD with monomers and more than 50% of the monomers were removed by the ethanol precipitation protocol. **D)** The plot shows an example of the separation of the 5-mer-P from a 5-mer-cP with the respective retention time from the oligomerisation mixture of 20 mM cGMP, 40 mM KCl, at pH 10, 40 °C drying for 18 hours. The 5-mer-cP is retained on the column before the 4-mer-P and 5-mer-P and is identified based on the mass in the MS and eluted with their respective retention times. The quantification of n-mer-cP was done similarly for 2 to 10-mers by using identification of their masses from their retention times and n-mer-cP always preceded the (n-mer -1)-P (read as n-mer minus 1 with P ending).



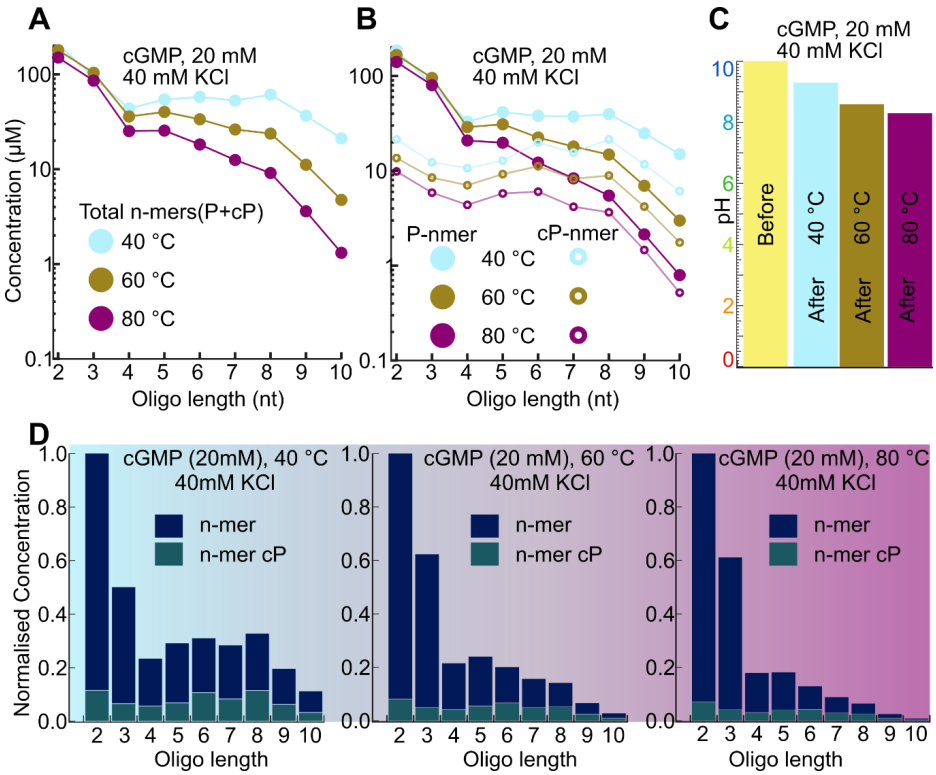
S3. Calibration of 2 to 10-mers polyG standards.
A) We compared 2-10mer standards by 1) without ethanol precipitation protocol and 2) with ethanol precipitation protocol. For the analysis of oligomerisation samples, only the ethanol precipitated polyG n-mer standards calibration values were used as all the oligomerisation reaction samples were subjected to ethanol precipitation for HPLC-MS analysis. The calibration was performed for implementation of the concentration calculator on the customized SpectralBrowser program. This allowed for correlating the number of counts on the TOF for an oligomer of defined length, to its concentration. This was done by injection and analysis of 2 to 10mers of polyG oligomers at concentrations ranging from 1 to 500 pmoles. **B, C)** The raw counts obtained from the MS analysis were then plotted against the pmols for each length of the oligomers (S3A and S3B). The values for the curves were linearly fit and their slopes (values in Table S1) were obtained. The moles/counts for each n-mer oligo were calculated and the average was determined as the value in calibration (moles/ions) (S6, denoted as 7). The ratio of the average value to slope of each n-mer was used as the length dependent factors (LDF) in the SpectralBrowser (S6). The values used for oligo concentrations of 2-10mers was for a 100µL injection volume in the spectralBrowser.

Table S1

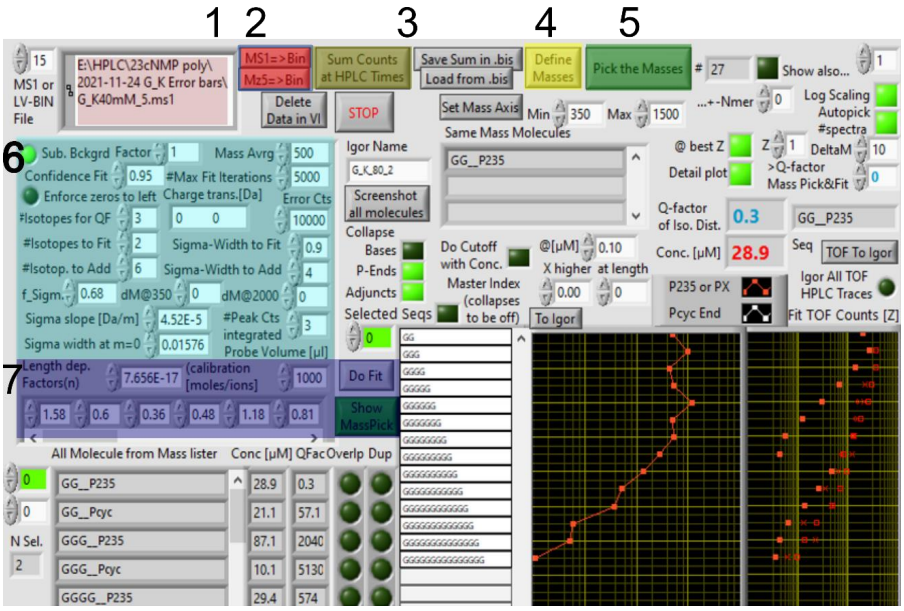
Oligo Length (Linear Fit)	**Coefficient Values ± one standard deviation (*counts/moles*)**	
	Slope un-precipitated	***Slope Precipitated***
2	13656 ± 259	10288 ± 87.7
3	21307 ± 453	20768 ± 276
4	38472 ± 1.78e+003	41287 ± 307
5	31293 ± 896	32831 ± 292
6	12173 ± 408	12407 ± 242
7	15117 ± 444	15601 ± 313
8	9866.8 ± 403	10429 ± 275
9	9806 ± 400	10404 ± 270
10	11048 ± 509	11551 ± 366



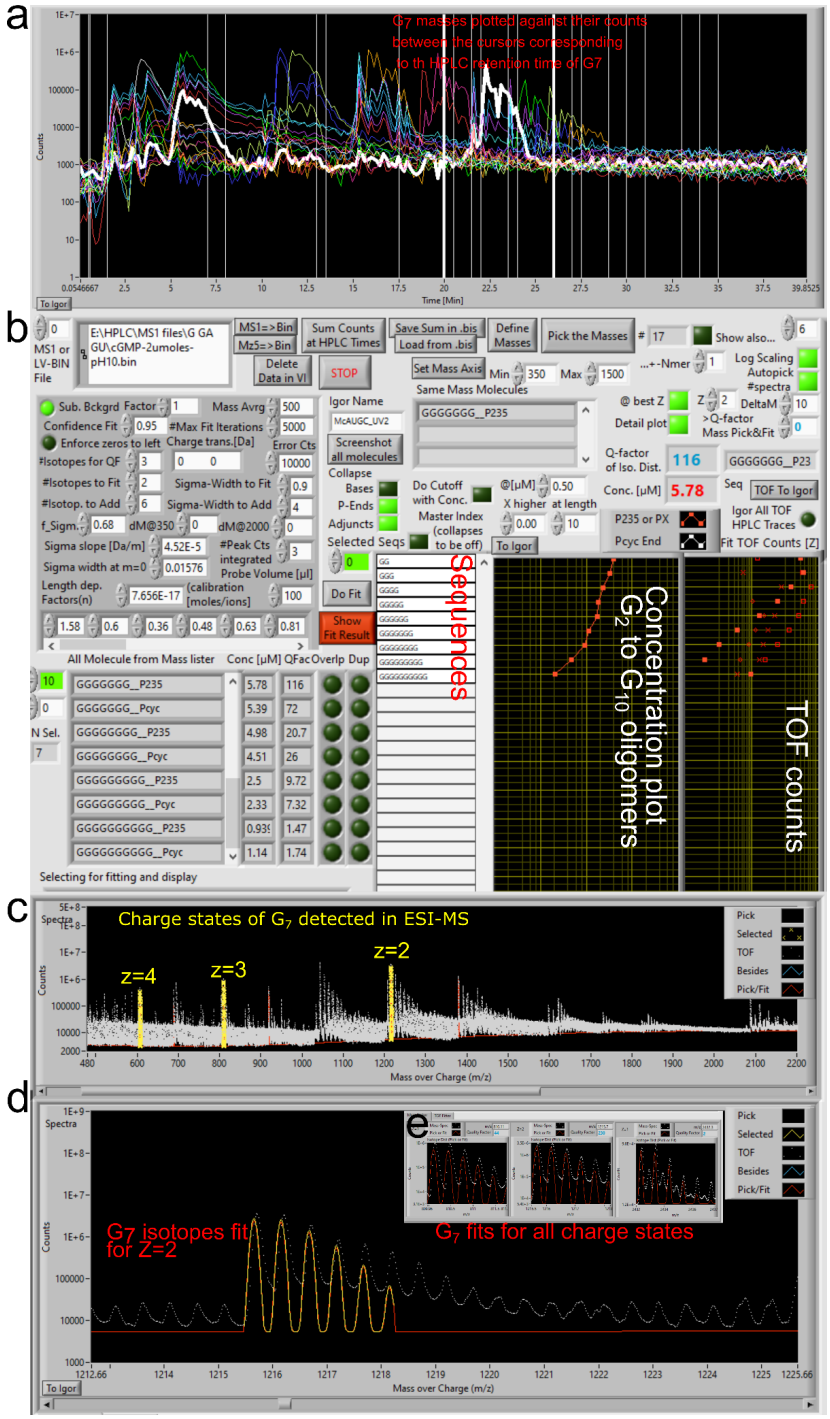
S4. Calibration of cyclic phosphate monomers and linear phosphate monomers standards was performed to determine the concentration of monomers formed in the oligomerisation reaction.
A) A calibration was performed at the following concentrations of cAMP, cUMP, cGMP, cCMP and 2'/3' linear monomers AMP, GMP, CMP: 200, 100, 50, 25, 10, 5, 2.5 and 1 uM concentrations. S4A shows the calibration curve for concentration in uM vs absorbance at 260 nM recorded by the DAD (diode array detector) coupled to the HPLC. The G only oligomerisation samples were conducted with a starting concentration of 20 mM cGMP, the GC samples with 10 mM each for cGMP and cCMP and the AUGC samples with 5 mM for cGMP, cCMP, cAMP and cUMP. G only samples were diluted 200 times, the GC samples 100 times and the AUGC 50 times, so that the theoretical concentration of reactants in each samples was 0.1mM, in order to achieve a good separation of the compounds in the HPLC and to be in the linear range of the calibration. **B**, **C** and **D)** show the final concentration of cyclic and linear phosphate monomers as a function of the drying temperature (40, 60, 80°C). Overall, the concentration of cyclic monomers decreases with the temperature. Nonetheless the drop remains limited and even at 80°C after one day we still have a high concentration of non-hydrolysed cyclic reactants. For the GC and AUGC oligomerisation samples, the final concentration of cyclic monomers is still higher that the concentration of the hydrolysed product. Only the G oligomerisation shows a different behavior. The concentrations of linear product seem also to increase globally with the temperature. In conclusion, maybe the hydrolysis of the reactants plays a role in the drop of efficiency of the oligomerisation at higher temperature. **E)** Shows the effect of temperature on the overall pH of post-oligomerisation mixture. The dried samples are rehydrated with 100 µL nuclease-free water and the pH were measure. The results are displayed as an average pH drop for G, GC and GCAU mixtures. All the reaction mixtures were initially adjusted to pH 10 and allowed to dry at 40, 60 and 80 °C. The results Show that at 40 °C, the pH drop is by 1 pH unit (Avg). The pH drops are 2 pH units after 80 °C reaction, indicating the formation of acidic products due to ring opening of monomers and n-mer-cP (see S5 for n-mers results of G oligomerisation). We could not point to any other simultaneous processes leading to pH drop.



S5. Quantification of n-mer-cP and n-mer-P (2 to 10mers; for monomers quantification see S4) in cGMP oligomerisation. The reactions were carried out for 18 hours in 100μL reaction volume. The reaction mixture was adjusted to pH 10 using KOH solution before the start of the reaction. The total concentration of the reaction mixture was adjusted to 40 mM using KCl solution. The total concentration of K+ ions included the K+ from KOH for pH adjustment and the remaining solution was adjusted using KCl to reach a total of 40 mM concentration. **A)** shows the total concentration n-mer-cP and n-mer-P and their dependence on temperature. Best oligomerisation is obtained at 40 °C with relative decrease in oligomerisation at 80 °C. However, the oligomerisation is favoured from 40 °C to 80 °C to varying extent and thus showing the prebiotic relevance of this reaction. **B)** shows the break-up of the concentrations of n-mer-cP from n-mer-P in the oligomerisation mixture. **C)** shows the post-oligomerisation pH drops as a function of varying temperature. We attribute this pH drop to the hydrolysis of the of cP n-mers and cP-monomers. As it can be seen in S5A and S5B that the amount of oligomers drops significantly especially from 4-mers to 10-mers with increasing reaction temperature. **D)** plots show normalised concentrations (between 0-1) for n-mers (2 to 10). In each plot the concentrations are normalised to 2-mers. Thus, indicating the relative concentrations of cP-n-mers in comparison to P-n-mers. The bar plots clearly indicate that the amounts of cP-oligomers are reduced significantly as the reaction are carried out at higher temperatures. For pH determination of the reactions, bulk dried samples were rehydrated with 100μL of nuclease-free water. After the pH was determined, the samples were subjected to ethanol precipitation and subsequently, used for HPLC-MS analyses. In the case of simulated rock pore reactions, the bulk of the liquid (90%) remained post-reaction and this was extracted and used directly for pH determination before further analyses.



S6. Representative screenshot of the workflow in the SpectralBrowser. The numbering above the arrows indicate the direction of the workflow starting from 1 and moving upwards. The MS1 data file generated from the MSConvert is dropped into 1. The MS1 file is converted into a binary file that is read by the SpectralBrowser by function 2. The counts at defined cursor positions are accumulated by function 3. Oligomeric masses from 2-10mers with the cyclic phosphate or linear phosphate ends are defined by 4. Those masses are then processed, plotted and displayed in the program. The 'Do Fit' function in the program is applied to remove overlapping masses and the concentrations are calculated based on the fit qualities. Qantification using SpectralBrowser. The mass spectrometry data was handled by using a freely available program called MSConvert from ProteoWizard.[6] The program allowed the conversion of '.d' file format generated by the Agilent MS system to a MS1 or MZ5 file formats. Once the MS1 or MZ5 file was generated, it is now ready to be used in the SpectralBrowser. The MS1/MZ5 file can be dragged and dropped into the program. The workflow within the SpectralBrowser is shown in S6. The MS1 /MZ5 file is then converted into binary file by loading all the spectra from 0-40 minutes and saving it into a subVI within the program. The subVI generates a '.bin' file with the mass arrival times, the m/z values arriving at those times and corresponding counts of the masses. The masses are then accumulated, by clicking on 'sum counts at HPLC times', within the defined set of cursors assigned based on the retention times of the oligomers eluting in the HPLC-DAD. Then the masses are defined within a subVI using 'Define masses' key, where the masses are calculated for cyclic ended phosphates and linear phosphate oligomers. If required, the salt adduct masses are also calculated. Following this, the masses are processed, plotted and displayed in the program. The length dependency factor, moles/ion counts values for 2-10mers are obtained from the calibration values of standards, and the injection volumes are used to calculate the oligomer concentrations for the analysis. The program further uses a fitting algorithm that plots and overlays the fit on to the experimental isotope distributions obtained from the raw data. The fit is calculated after several rounds of iterative overlays to obtain a final average fit. A cubic spline fit is used for the interpolation between the counts of the isotope masses and overlaid on the empirical isotope distribution. The fit parameters such as the number of isotopes to be fit, the standard deviation of the iterative fit values, the width of the fit over the empirical isotope distribution can be varied based on the data being analysed. The concentrations of the oligomers are calculated and based on the fit values and they are exported from the program.


a
G7 masses plotted against their counts between the cursors corresponding to th HPLC retention time of G7
Counts
Time [Min]
b
Sequences
Concentration plot G2 to G10 oligomers
TOF counts
c
Charge states of G7 detected in ESI-MS
z=4
z=3
z=2
Mass over Charge (m/z)
d
e
G7 isotopes fit for Z=2
G7 fits for all charge states
Mass over Charge (m/z)

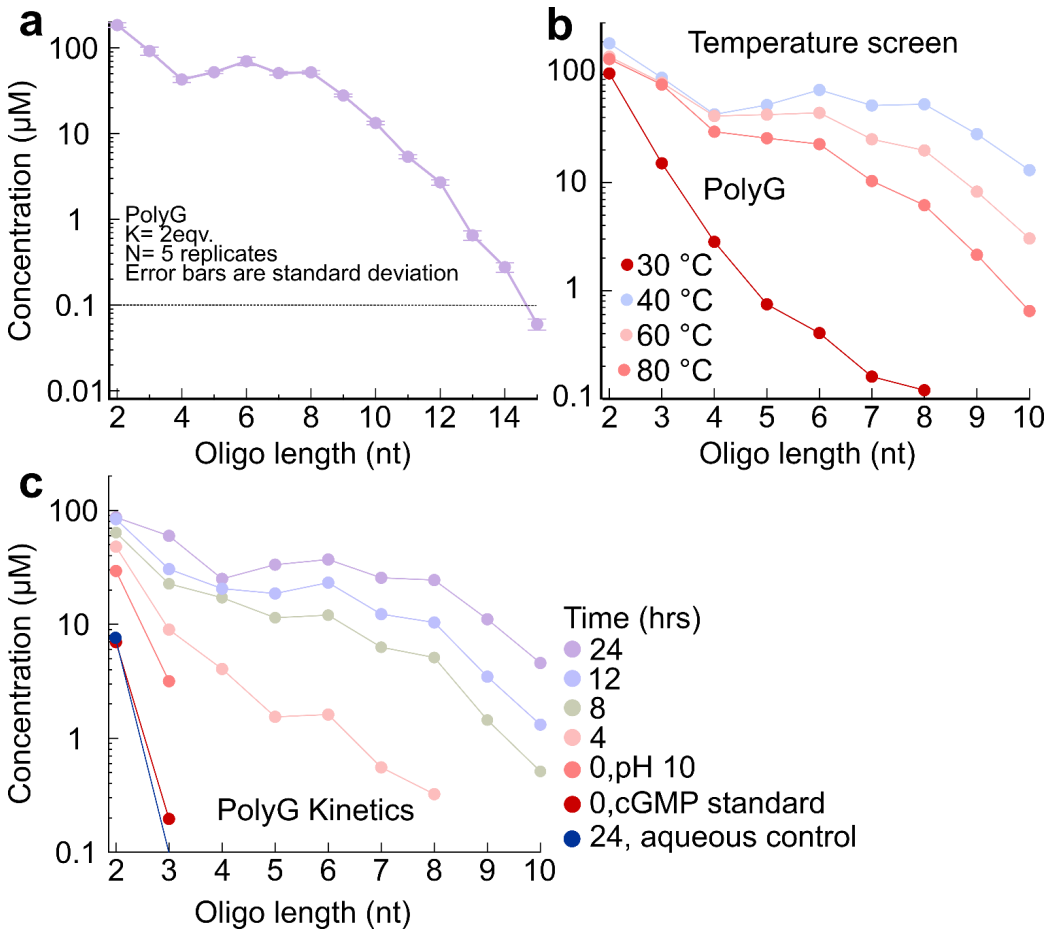
S7. Example of a routine analysis of G7 oligomer using the SpectralBrowser program. a) The masses of G_7 (sum of the counts of all the charge states) plotted against time on the x-axis. The G_7 white trace which corresponds to the HPLC-DAD retention time of G_7 with linear phosphate terminus is between the two bold cursors. b) Shows the in-program plot of all the oligomers of G on the x-axis vs their corresponding concentrations. On the right of concentration plot is the raw counts of the corresponding G oligomers. c) The full spectrum of all the mass accumulated from between the defined cursor positions of the of 2-10mers in grey background. The yellow traces correspond to all the charge states (z=2,3 and 4) detected in the ESI-MS for a G_7 oligo. z=1 is a dormant charge state for a G_7 oligo and hence not observed. d) The fit results for z=2, i.e. the charge state with most counts and a m/z of 1215.7. e) Inset are the fits for all the charge states of a G_7 oligomer.

Table S2. Example of mass overlap and selection criteria

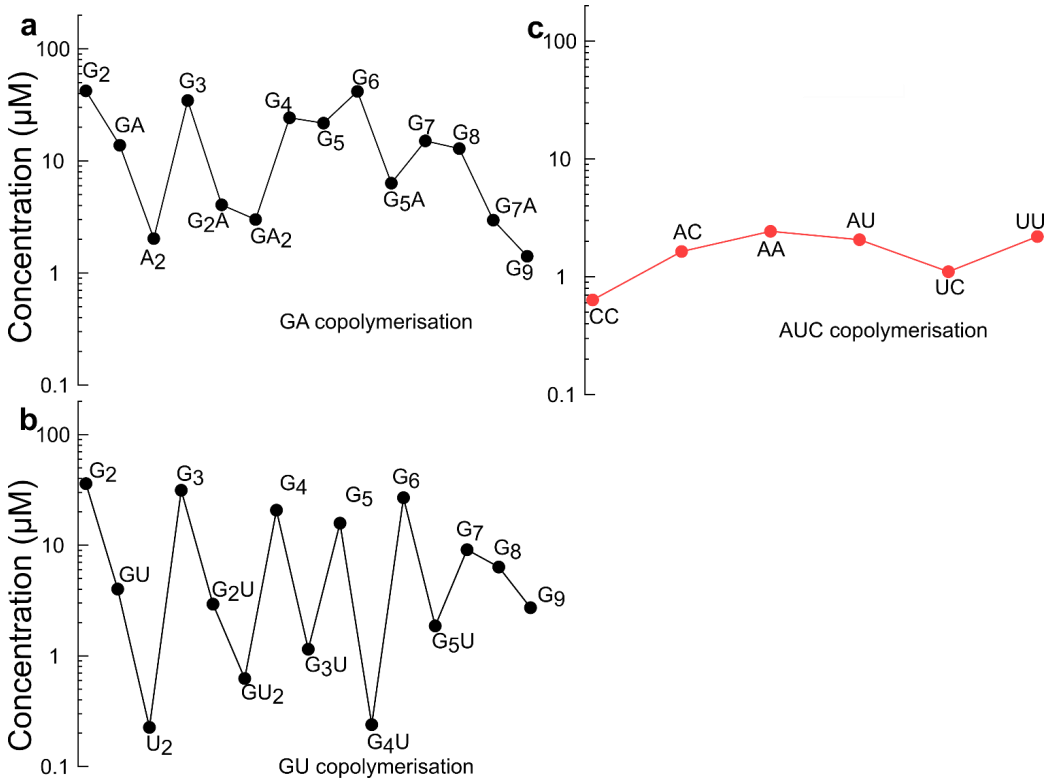
Sequence	*Mass Overlap sequences*	*Charge state*				*Other overlaps*	*Comment*
		1	*2*	*3*	*4*		
GGG-P 1052.1456 (1) 525.5692 (2)	*AAA_cP_Na3*	*1052.0961*					*The overlaps fall within the cursor ranges of trimers.*
	CCA_cP_K3	*1051.9955*					
	GCC_cP_NaK2	*1052.0164*					
	GGGGAA_cP_Na3				*525.0529*		*Fall outside the cursor range and thus, the overlap is not directly relevant*
	GAAAAA_cP_K3				*525.0371*		
	GGGGCC_cP_K3				*525.0277*		
	GGAAAA_cP_NaK2				*525.0424*		
	GGGAAA_cP_Na2K				*525.0476*		
	GCU_P_K2					*524.0109*	*Within Cursor range, however the major isotopes do not overlap.*
	UUU_P_K3					*523.9698*	
	CAU_cP_K3					*525.9861*	
	GCU_cP_NaK2					*525.9966*	

Table S3. Example of mass overlap and selection criteria

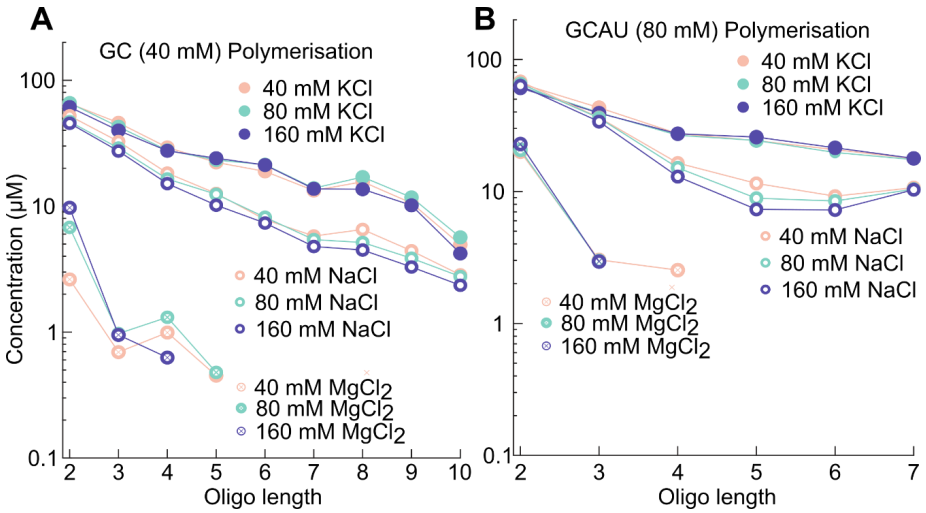
Sequence	*Mass Overlap sequences*	*Charge state*				*Comment*
		1	*2*	*3*	*4*	
GGG_cP 1034.135 516.5639	*GGC_P_Na*	*1034.1214*	*516.5571*			*All the overlapping sequences fall within the assigned cursors range for the cyclic trimers. The sequences selected here (indicated in green shading) are based upon the quality of the fit algorithm (which takes into consideration the most abundant isotope and at least three isotopes were considered), the presence of the corresponding sequence without the salt adduct and the isotopic distribution pattern. More over the observations from homooligomerisation data of G, C, A, U and two monomers cooligomerisation from GC, GA, GU and AUC also aids in the decision process. This does not necessarily indicate the absence of the sequences with overlapping masses.*
	GCA_P_K	*1034.1004*	*516.5466*			
	AUU_P_K2	*1034.0182*	*516.5054*			
	GUU_P_NaK	*1034.0392*	*516.5159*			
	GGU_cP_K	*1033.0688*	*516.0307*			
	CAU_P_K2	*1033.0342*	*516.0134*			
	GCU_P_NaK	*1033.0551*	*516.0239*			
	UUU_p_NaK 2	*1032.9729*	*515.9828*			



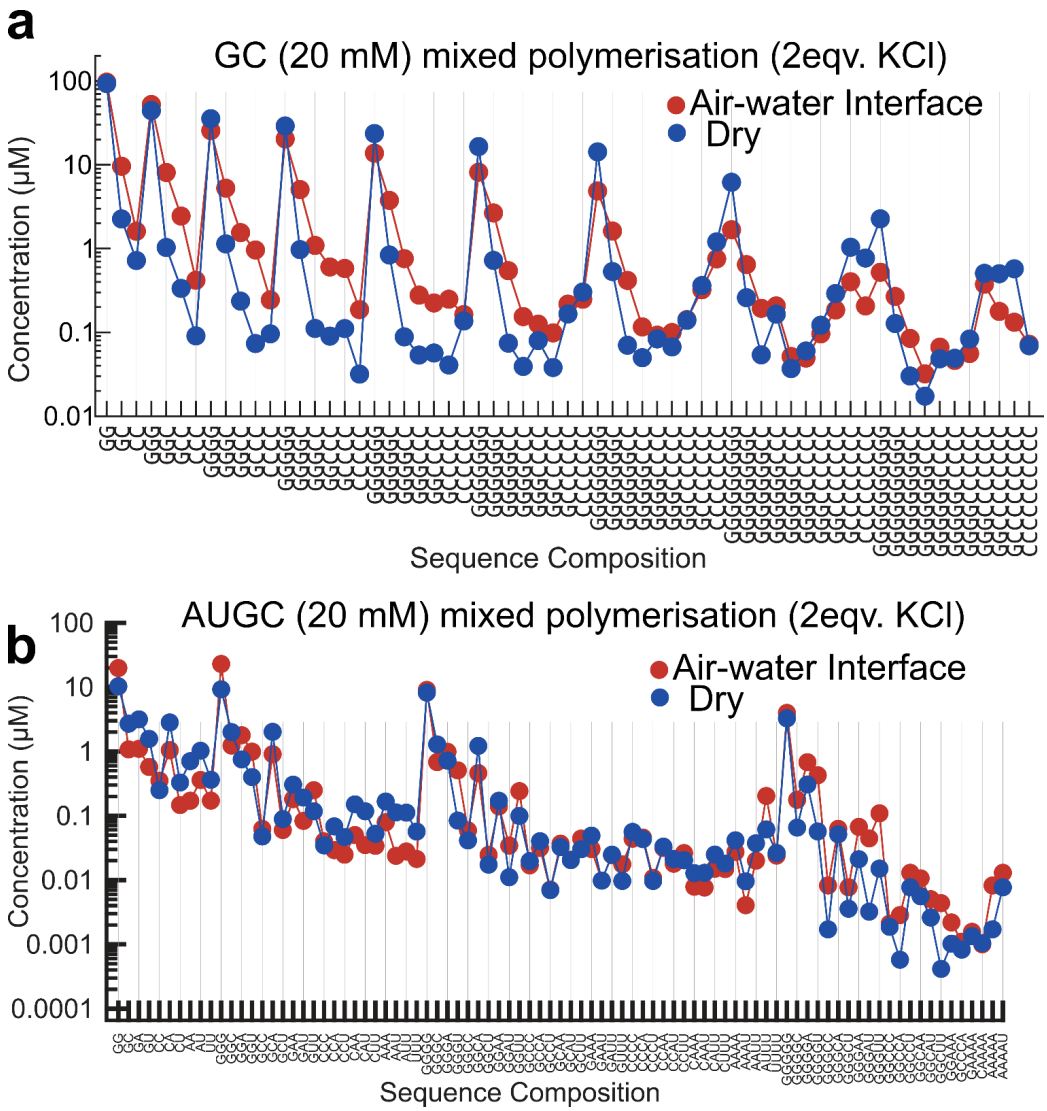
S8. **a)** Error bars calculated as standard deviation for 5 replicates. The error bars are smaller than the markers. The values beyond the 10mers are not quantified, but oligomers were detected clearly in the mass spectrometry analysis and the values shown here for n-mers >10 are using moles/ions values at 1. The dotted horizontal line indicates the quantification limit. **b)** Temperature screen for cGMP oligomerisation at 30, 40, 60 and 80 °C. The reaction worked best at 40 °C for 24 hrs reaction time. **c)** The kinetics of the cGMP reaction are from 0 to 24 hours are shown. A progression in the buildup of oligomers is observed from 4 hours to 24 hours and beyond which the concentrations of oligomers are not observed to increase significantly. The aqueous control was carried using an oil layer on the top of the reaction mixture to prevent evaporation.



S9. a) Cooligomerisation of cGMP and cAMP (at 40 °C) with 2 eqv. K^+, shows G and A differential contribution in GA cooligomers. Sequence composition cGMP and cAMP cooligomerisation at 2 eqv. K+. b) Cooligomerisation of cGMP and cUMP and the sequence composition of the same. c) cAMP, cUMP, cCMP cooligomerisation and its sequence composition. The reaction shows very low concentration of mostly dimers formed and thus showing that cGMP played a major role in forming mixed sequence oligomers of >3mers in the rested conditions. All the reactions were carried out at 40 °C, 2eqv. K^+ ions for 18 hours.

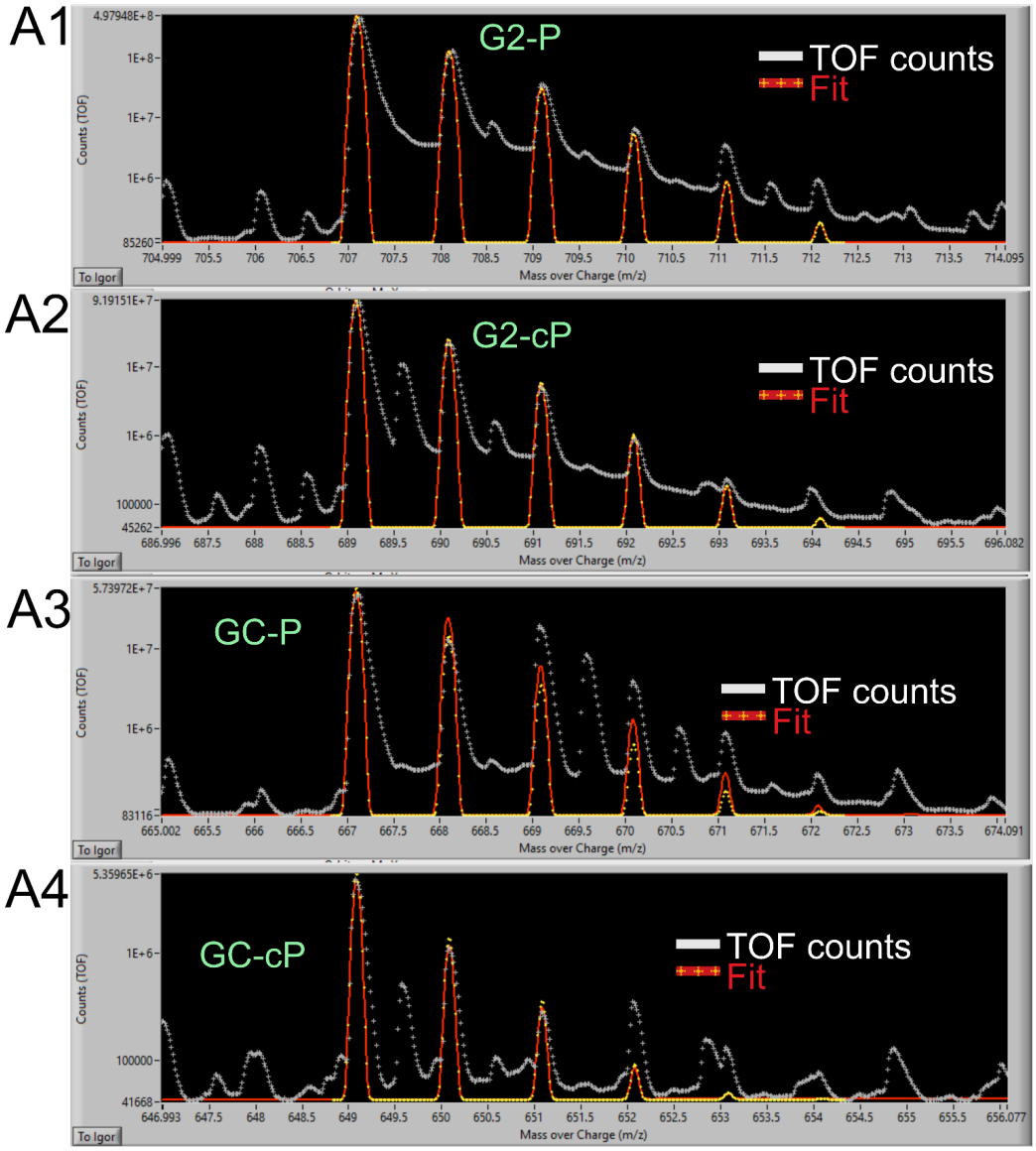


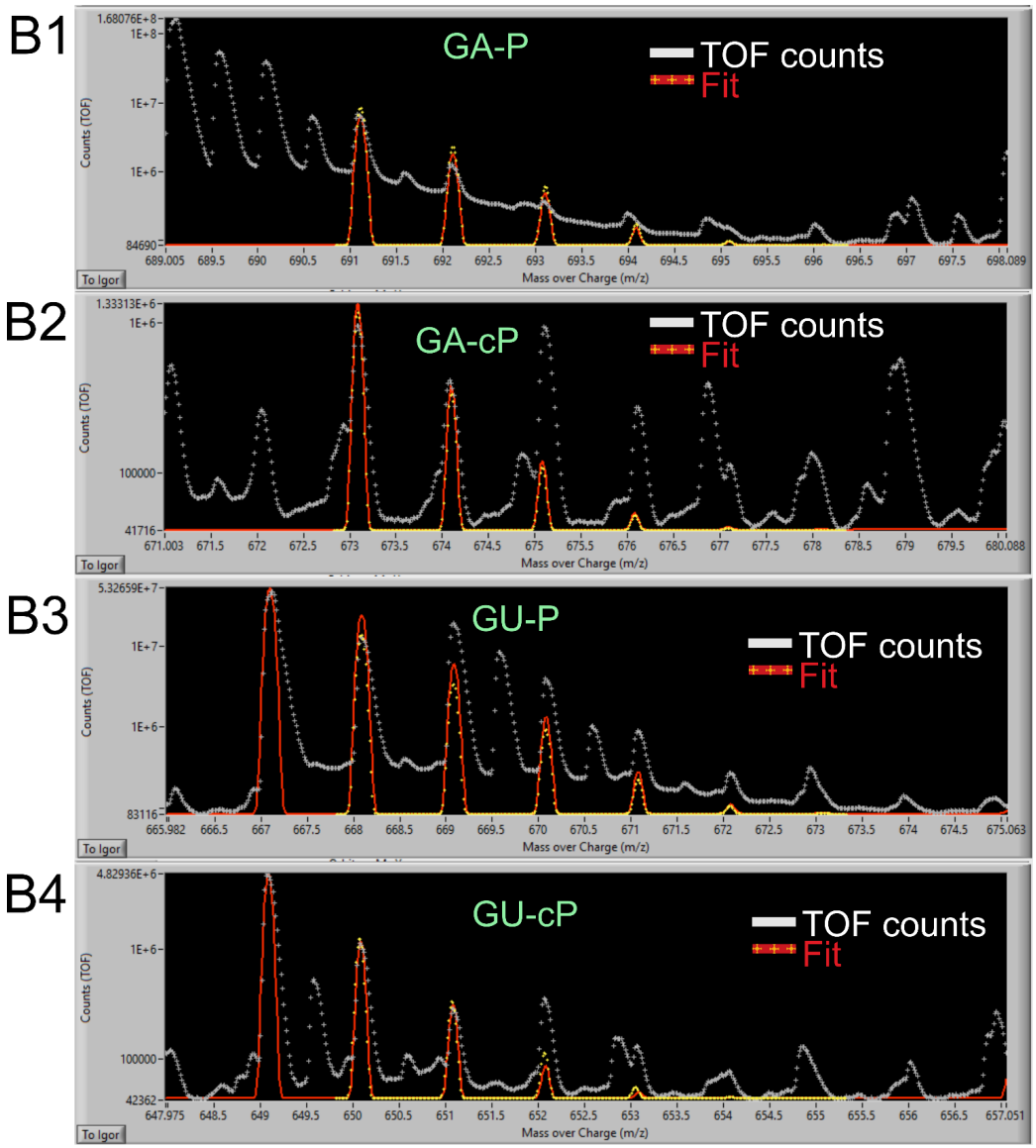
S10. Quantification of GC oligomerisation as a function of varying salt concentrations. The reactions were carried out for 18 hours in 100µL reaction volume. The reaction mixture was adjusted to pH 10 using KOH solution before the start of the reaction. The total concentration of the reaction mixture was adjusted to the desired salt concentrations for each reaction using KCl solution. The total concentration of K+ ions includes the K+ from KOH for pH adjustment and the remaining solution was adjusted using KCl. S10A and S10B clearly show that the GC and GCAU oligomerisations are not affected by the amounts of salt. However, we observe that the type of salt definitely plays a major role in the yields of the reactions. K+ shows best oligomerisation assistance in comparison to Na+. Mg2+ drastically inhibits oligomerisation.

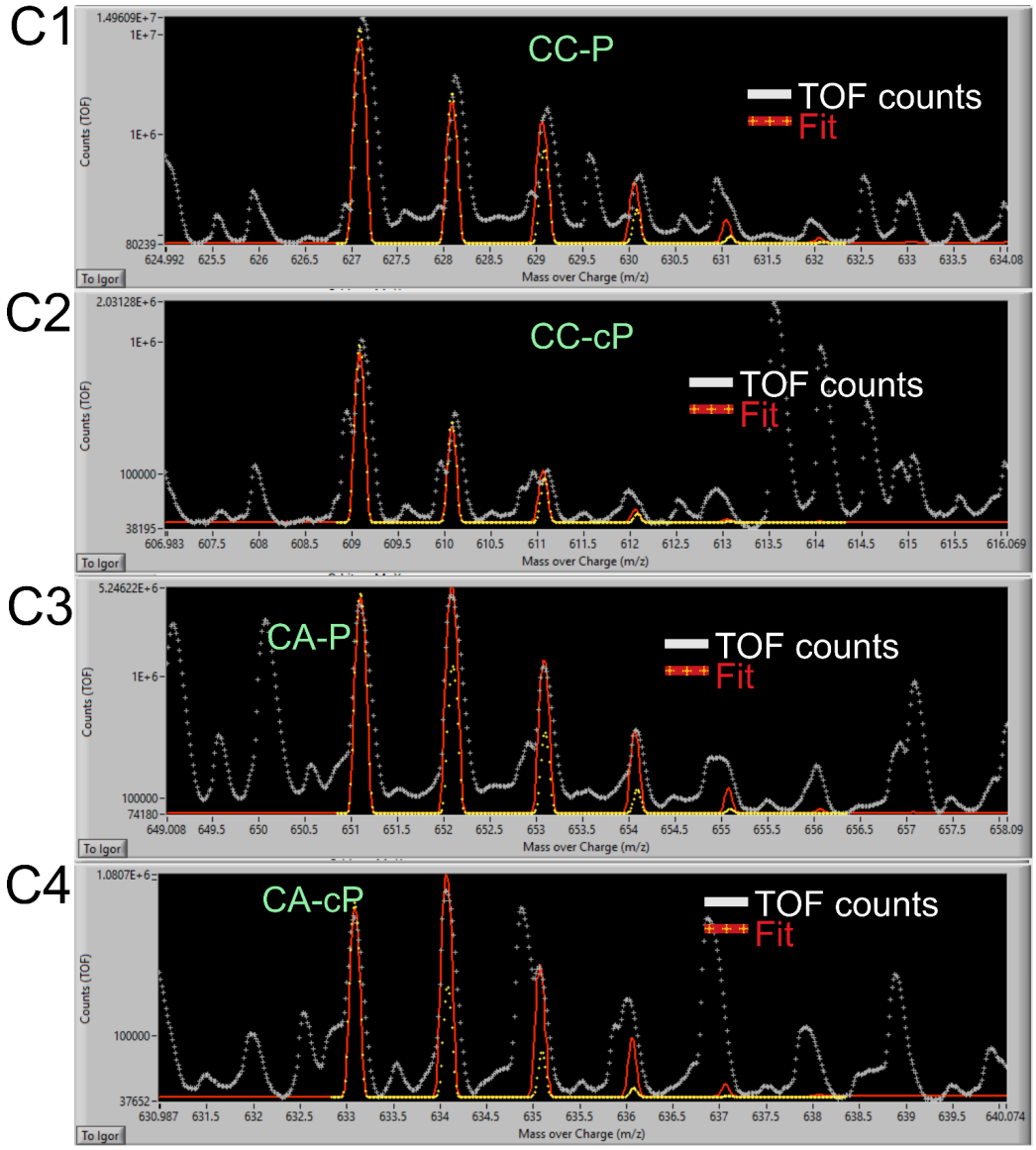


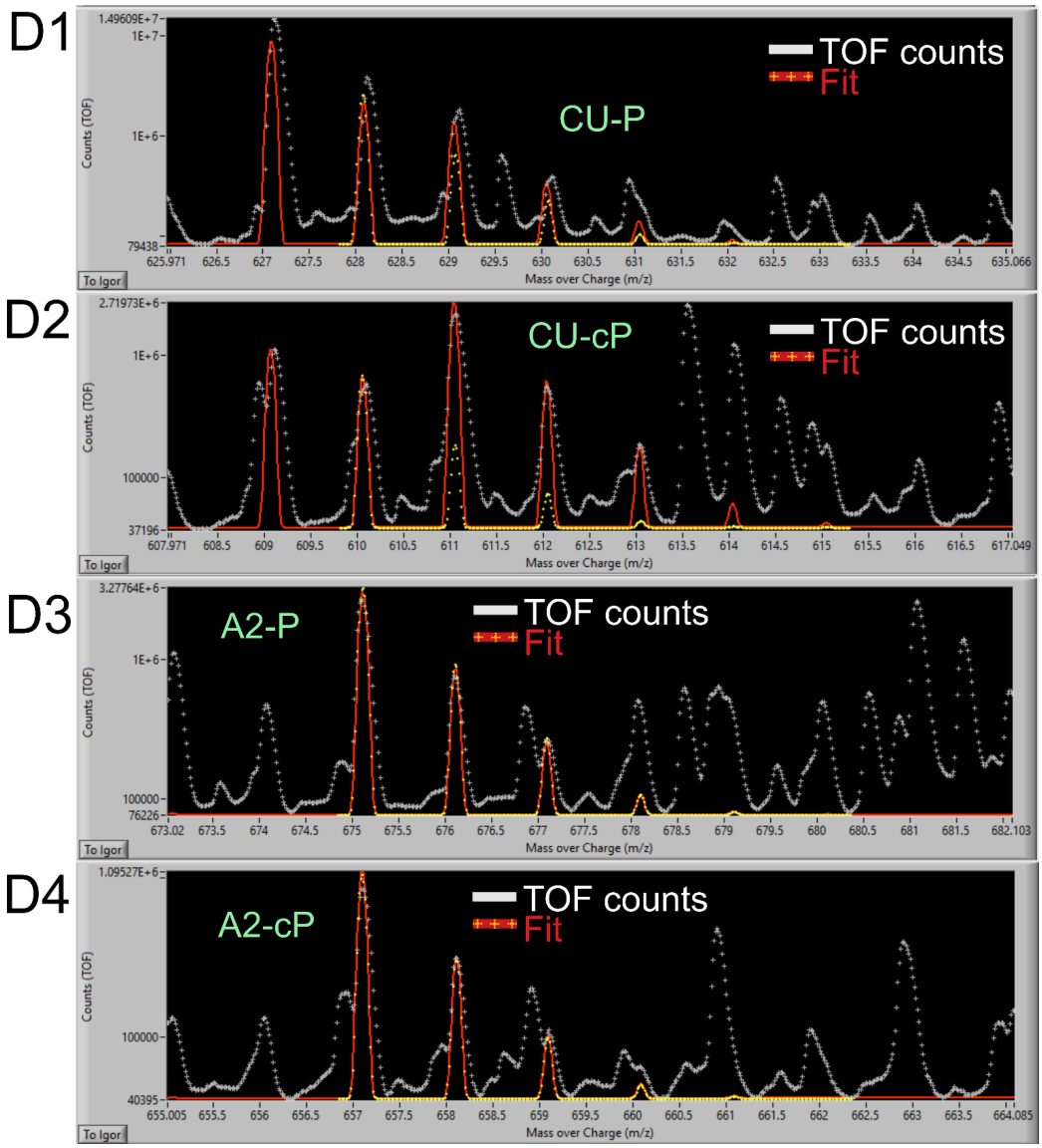
S11. Comparison of GC and AUGC oligomerisation due to bulk drying vs within a simulated rock pore environment. The reactions were carried out for 18 hours in 100μL reaction volume. The reaction mixture was adjusted to pH 10 using KOH solution before the start of the reaction. The total concentration of the K+ was fixed at 40mM and was adjusted to the desired concentrations for the reactions using KCl solution. The total concentration of K+ ions includes the K+ from KOH for pH adjustment and the remaining solution was adjusted using KCl. The sequence composition shown include the mass overlaps of species which are fit by the program. The Corresponding figures in Figure 2e and Figure 3c of the manuscript are compiled after a more rigorous selection criteria based on the retention times, isotope distribution, mass overlaps and a good fit quality. GC oligomers were fit for 2-10mers, however, for the AUGC oligomers we attempted to fit only 2-7mery due excessive mass overlaps with salt adducts and due high noise beyond 7mers for detection of cooligomers.

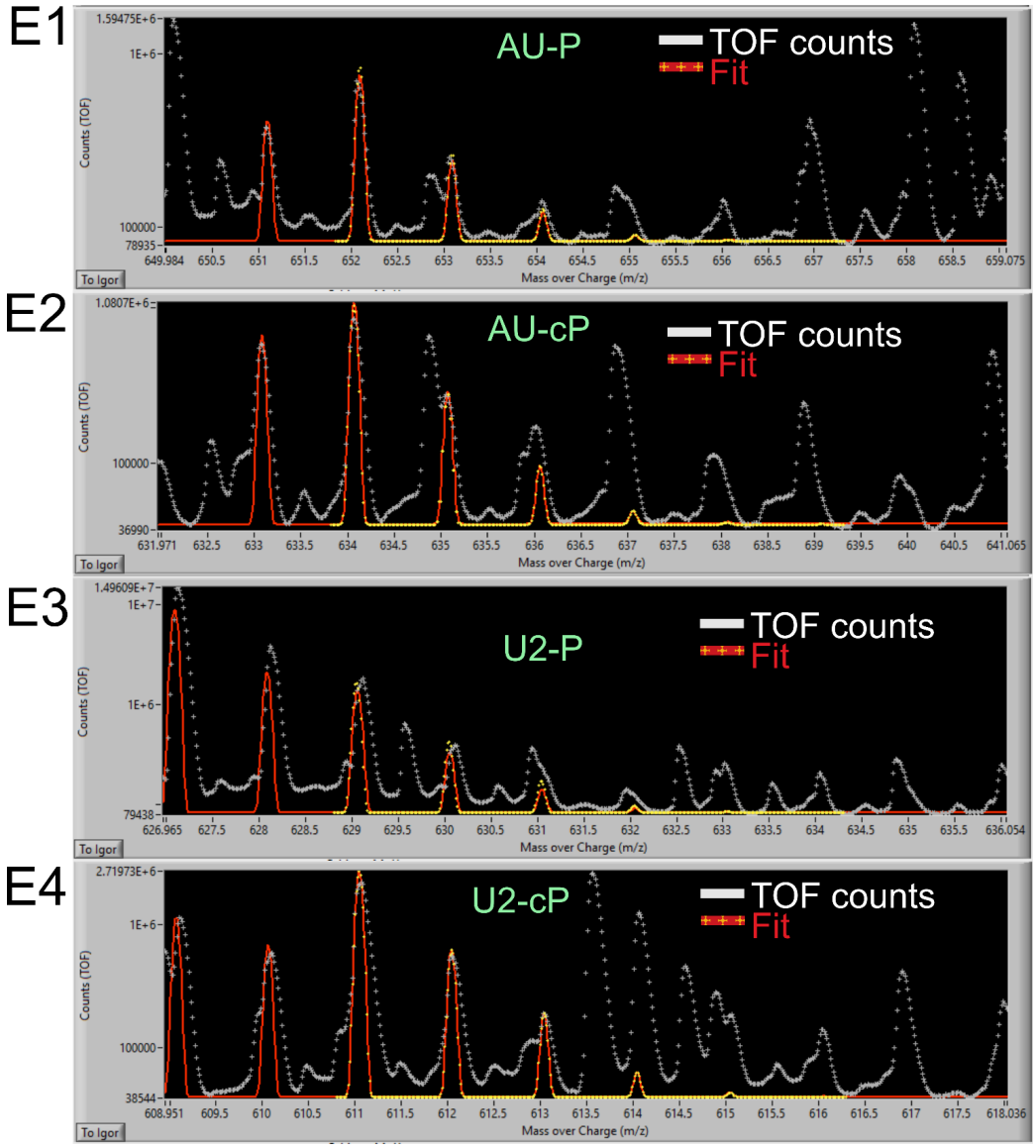
S12 The following are GCAU analyses images within the SpectralBrowser LabView program showing the the isotopic distributions, fits.

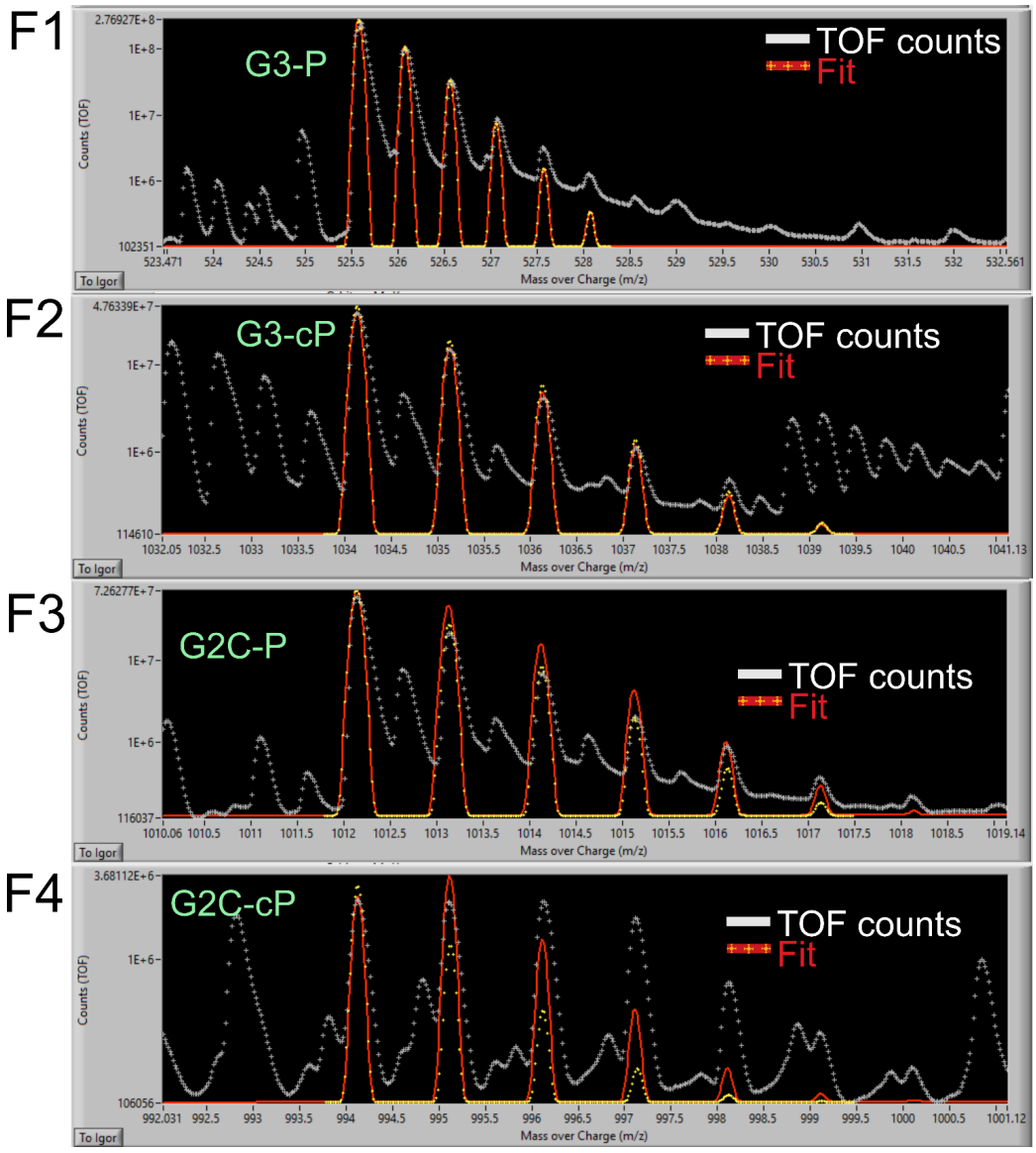


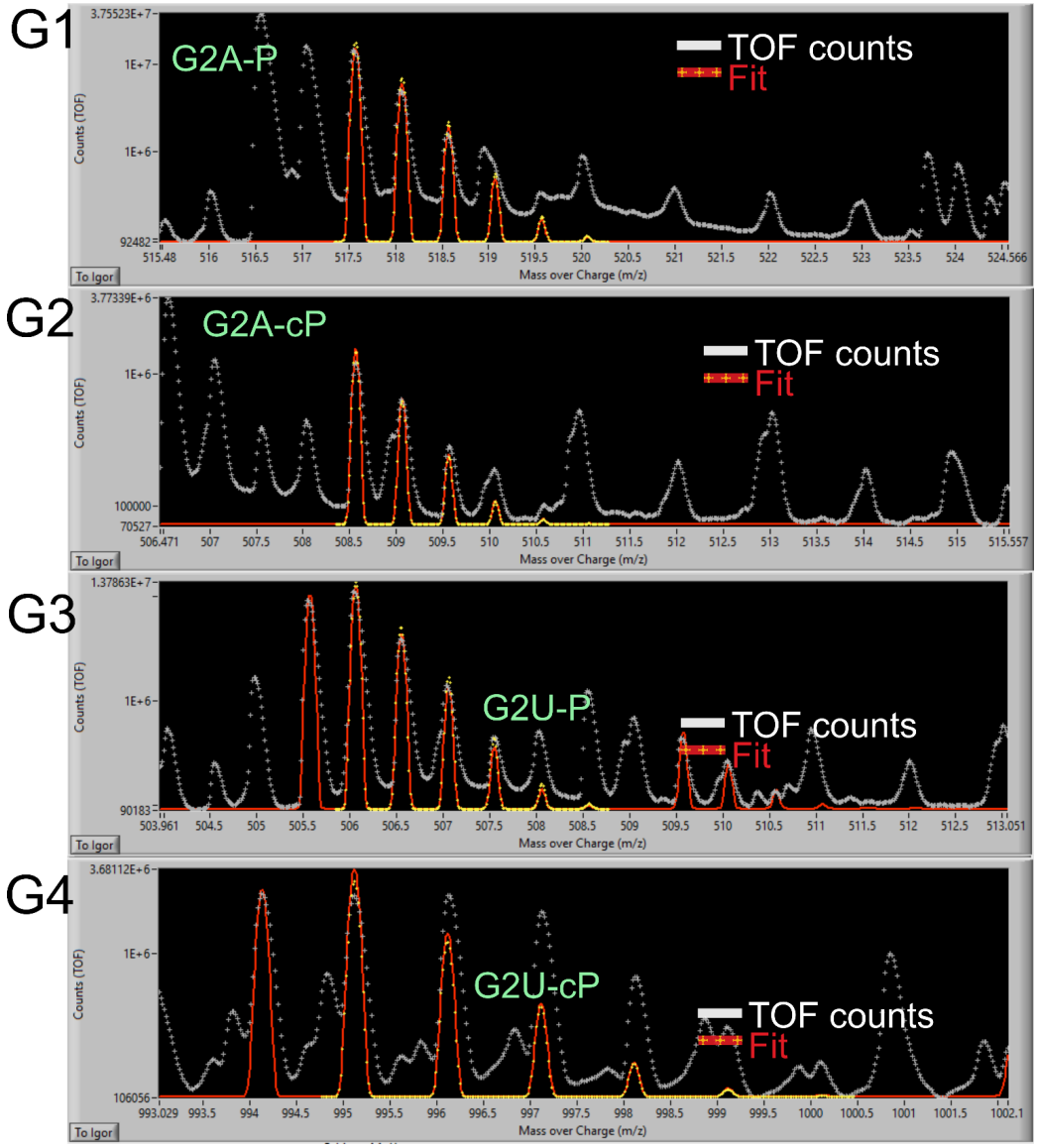

B1
GA-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
B2
GA-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
B3
GU-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
B4
GU-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor

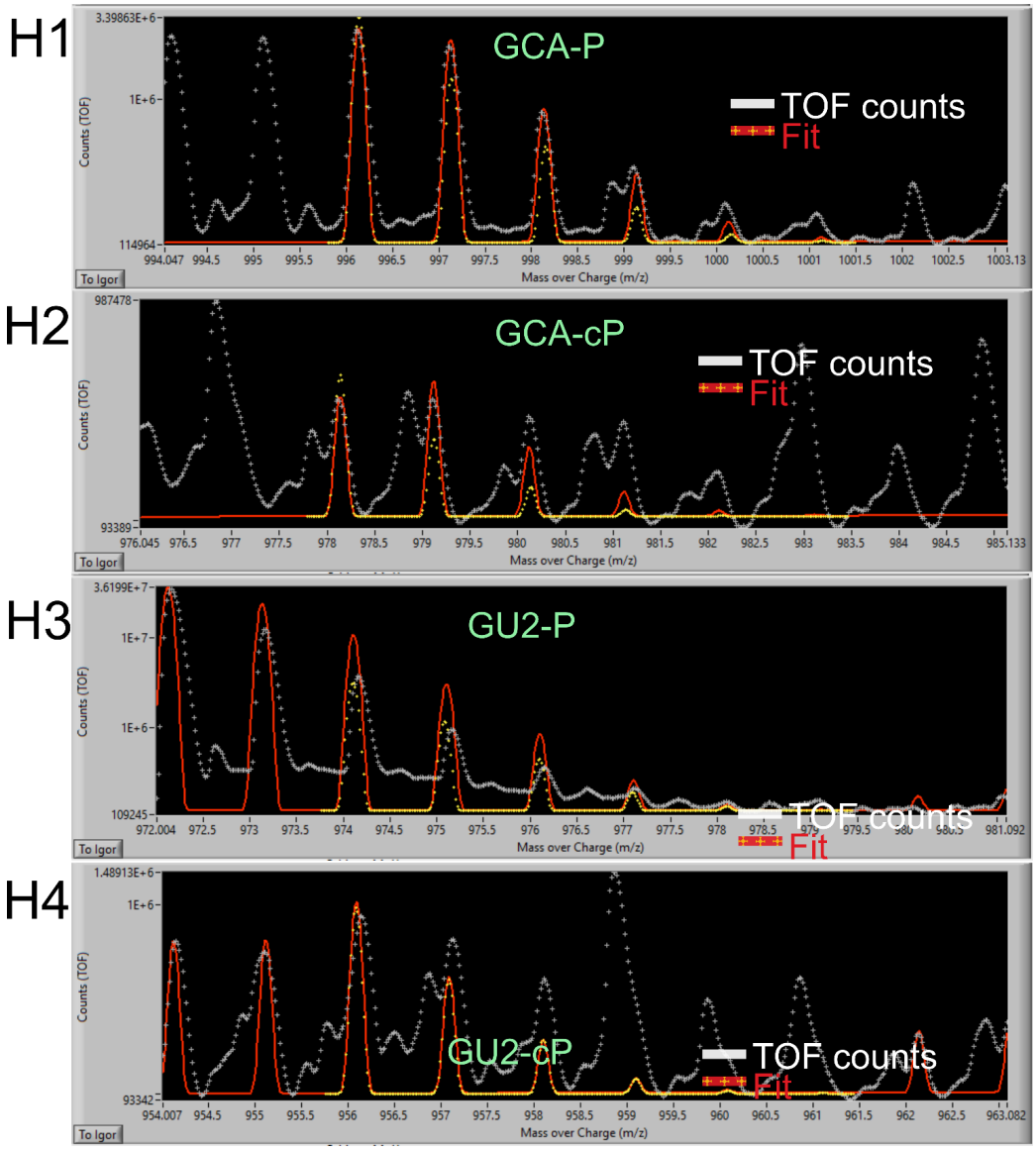

C1
CC-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
C2
CC-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
C3
CA-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
C4
CA-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)

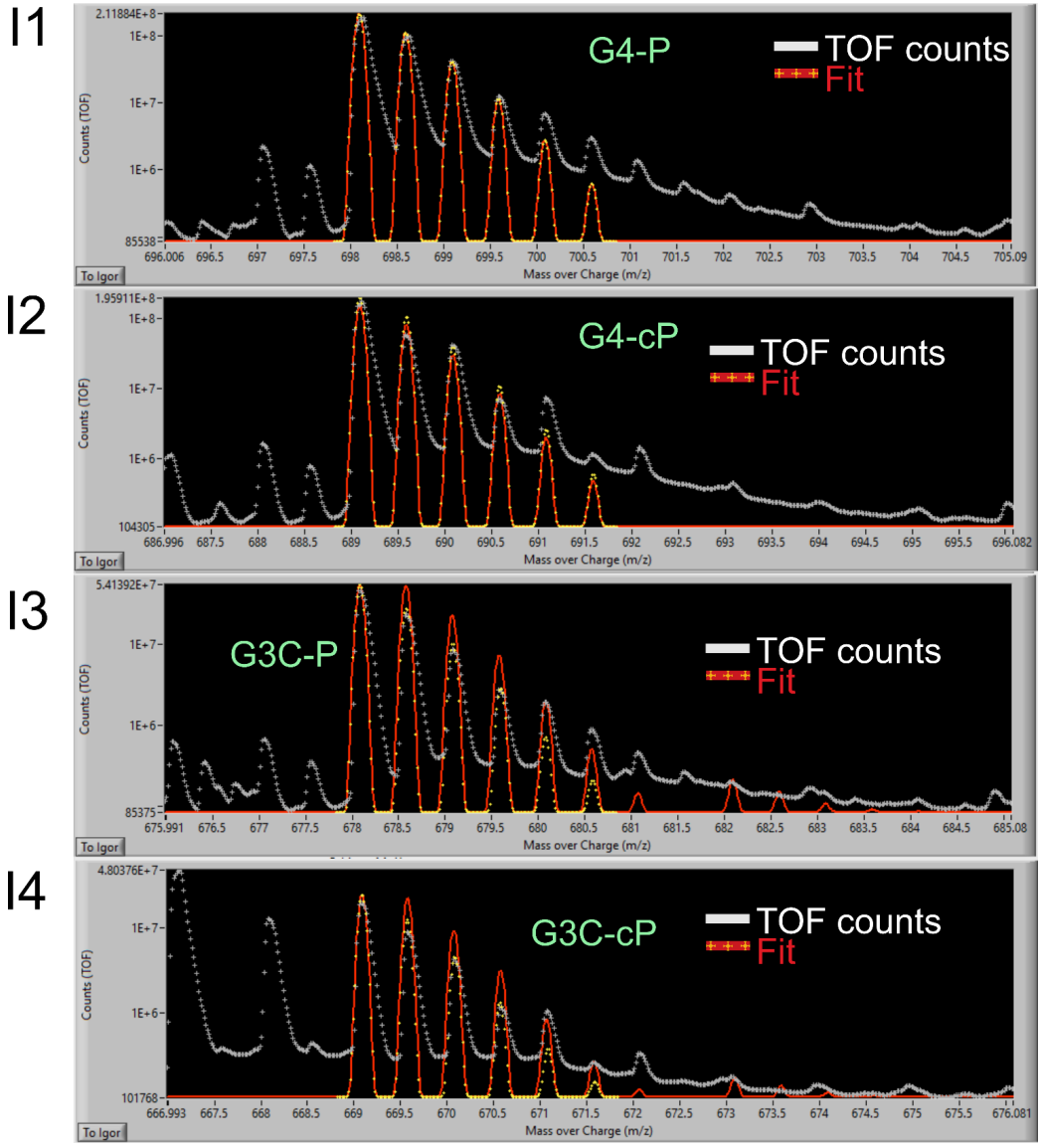

D1
CU-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
D2
CU-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
D3
A2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
D4
A2-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor

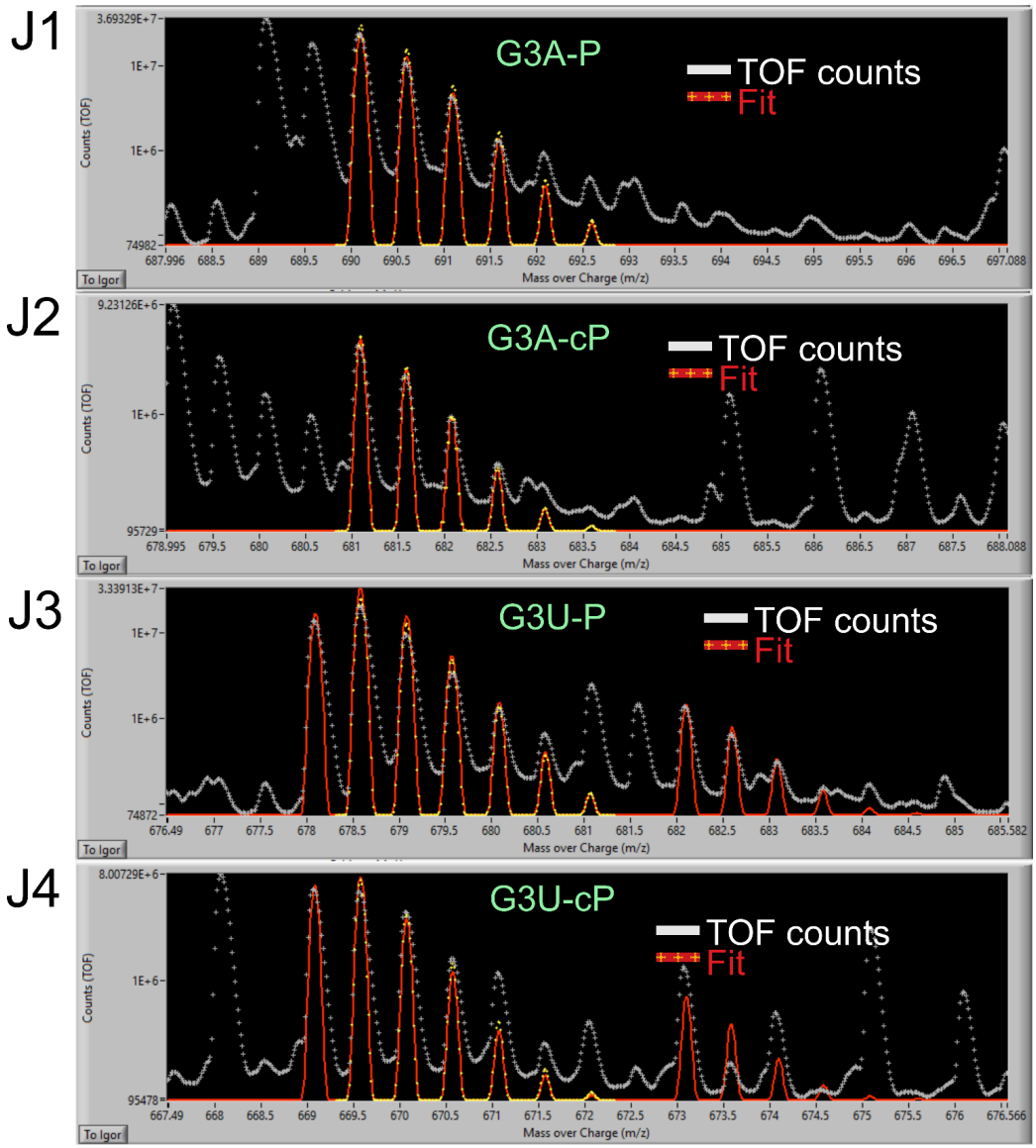

E1
AU-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
E2
AU-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
E3
U2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
E4
U2-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor

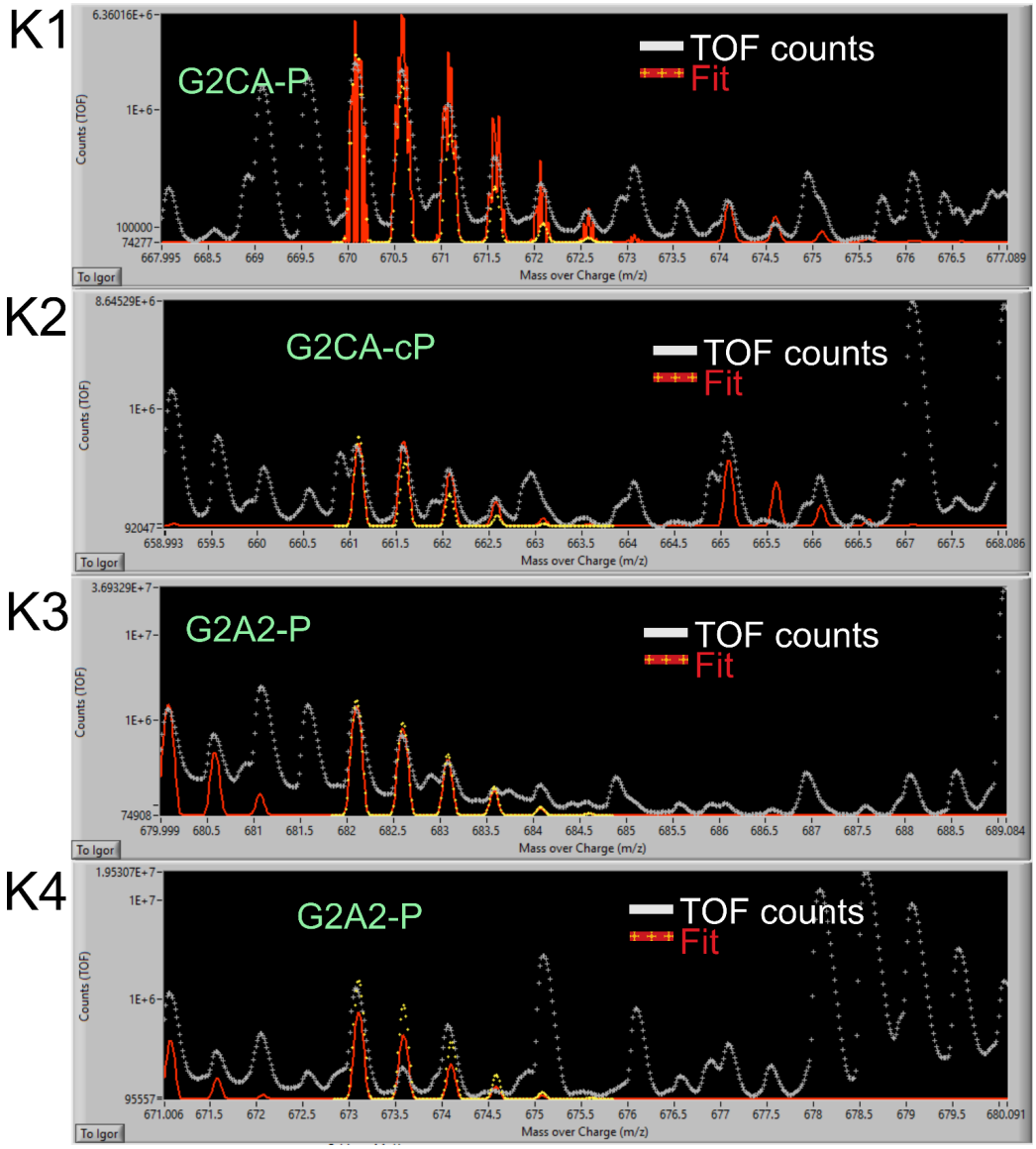

F1
G3-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
F2
G3-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
F3
G2C-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
F4
G2C-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor

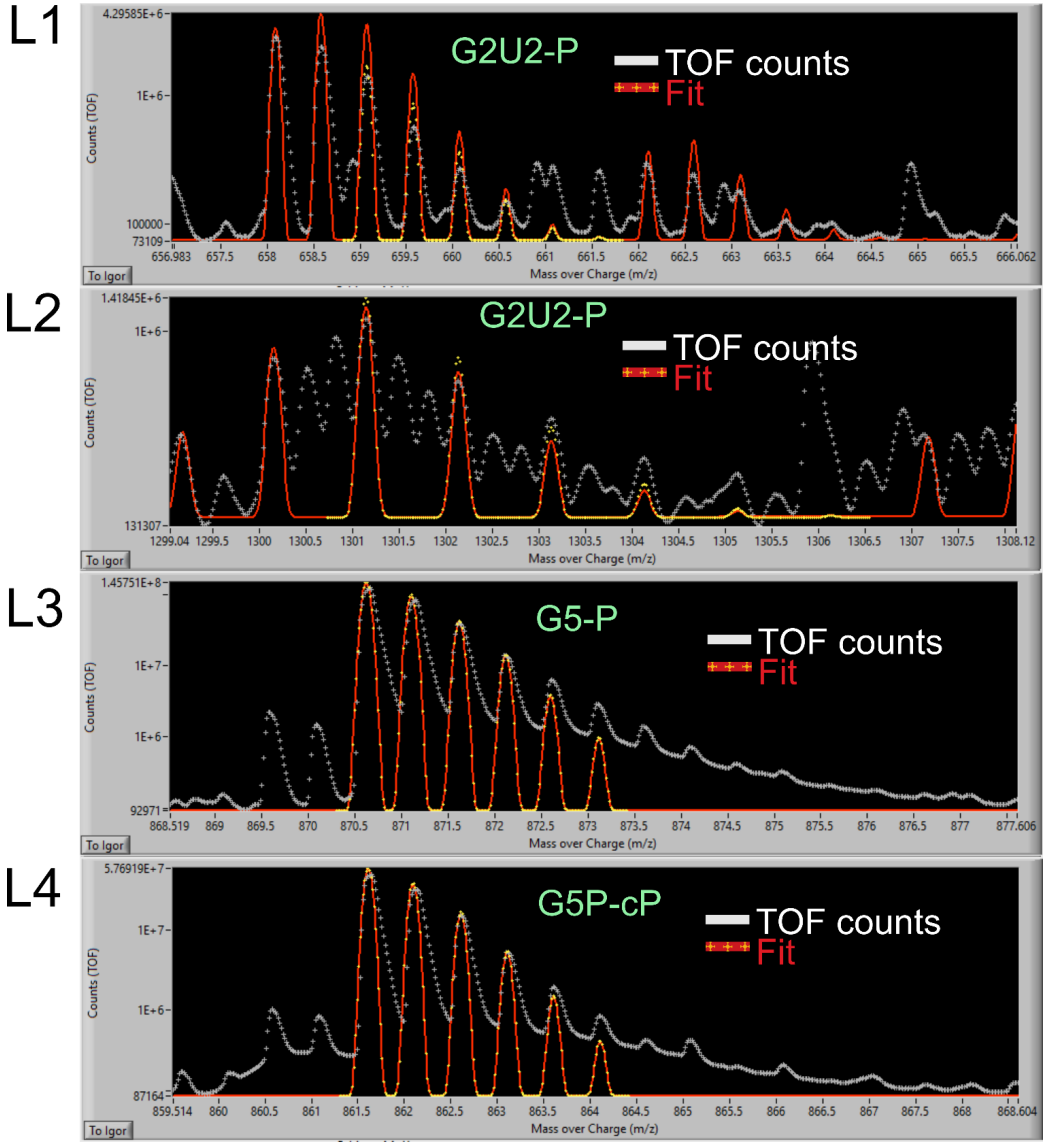

G1
G2A-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
G2
G2A-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
G3
G2U-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
G4
G2U-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor

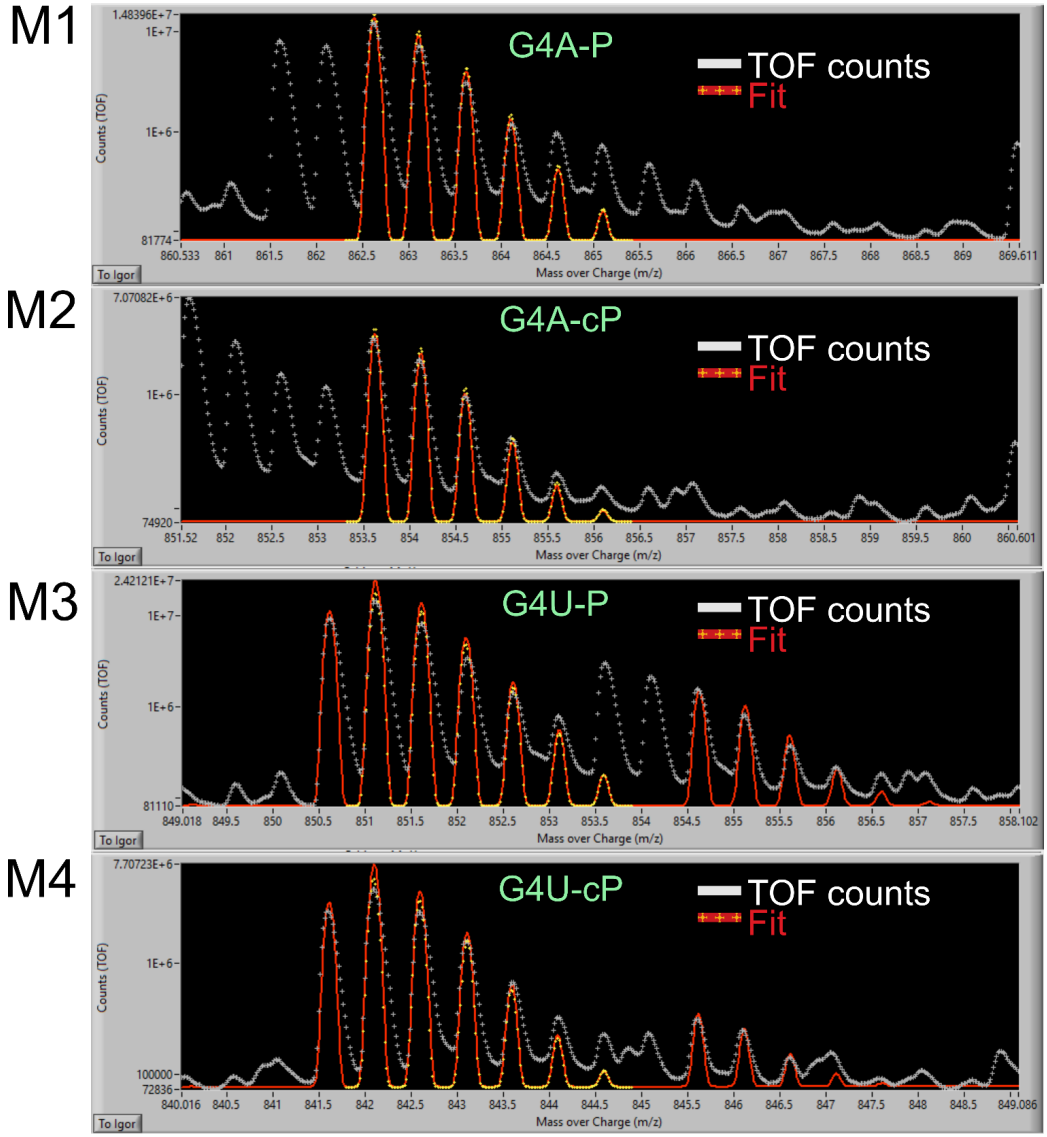

H1
GCA-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
H2
GCA-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
H3
GU2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
H4
GU2-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor

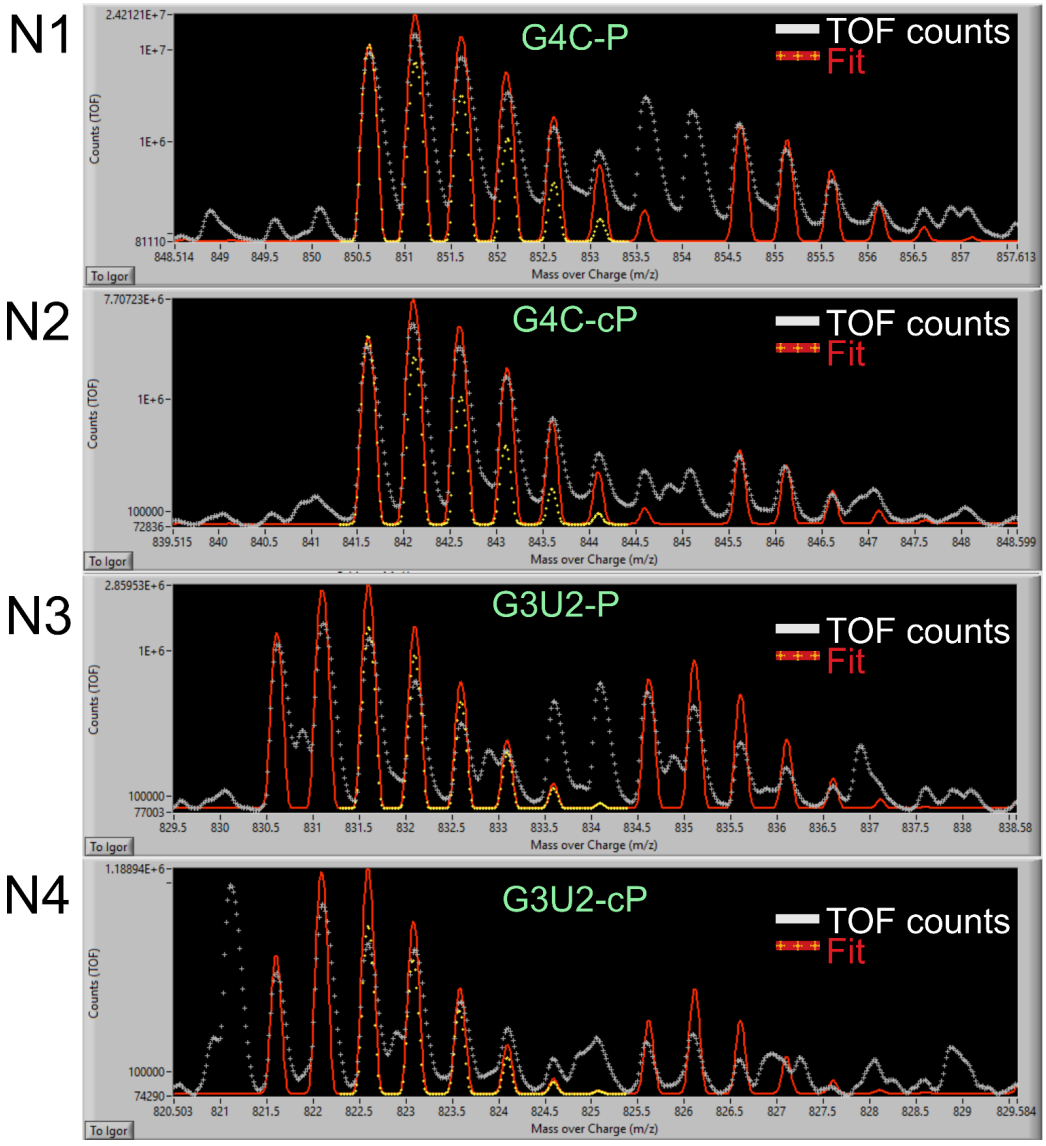

I1
G4-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
I2
G4-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
I3
G3C-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
I4
G3C-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)


J1
G3A-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
J2
G3A-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
J3
G3U-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
J4
G3U-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor


K1
G2CA-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
K2
G2CA-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
K3
G2A2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
K4
G2A2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor


L1
G2U2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
L2
G2U2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
L3
G5-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
L4
G5P-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor


M1
G4A-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
M2
G4A-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
M3
G4U-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
M4
G4U-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor


N1
G4C-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
N2
G4C-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
N3
G3U2-P
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor
N4
G3U2-cP
TOF counts
Fit
Counts (TOF)
Mass over Charge (m/z)
To Igor

NMR Results

NMR experimental section:

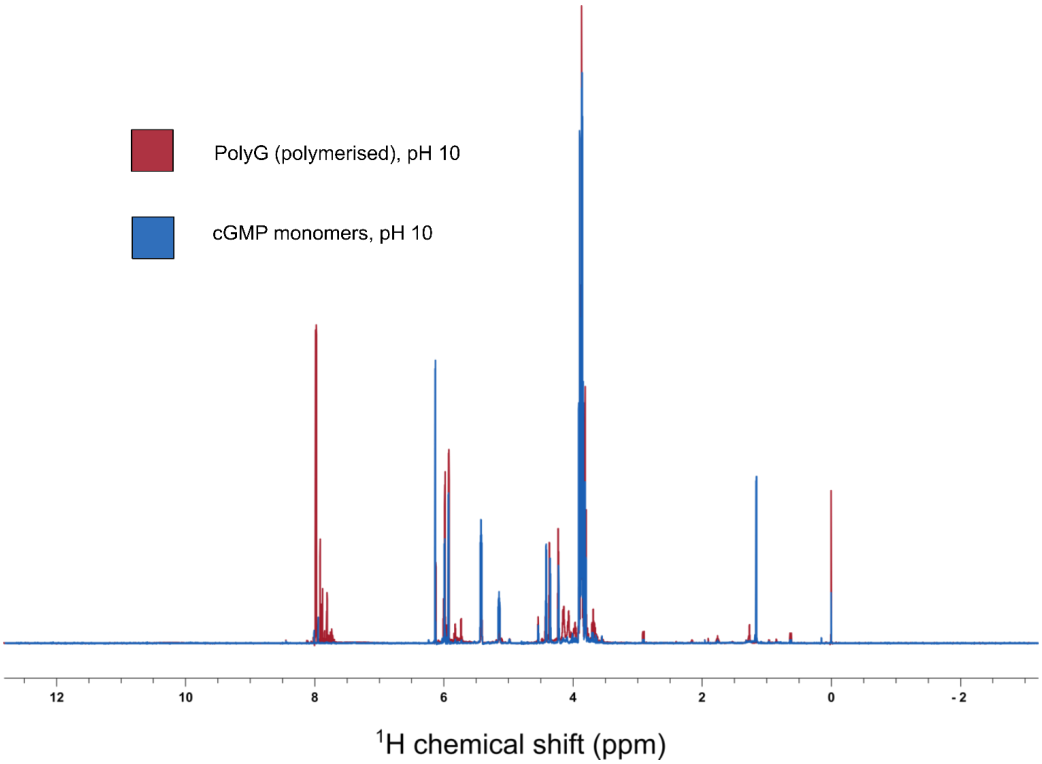
^{1}H (^{31}P) NMR spectra were recorded at 25 °C with a Bruker Avance III spectrometer operating at a ^{1}H (^{31}P) Larmor frequency of 800 MHz (162 MHz). Deuterium oxide (D_2O) was used as a solvent and chemical shift values are given in ppm. DSS (sodium trimethylsilylpropanesulfonate) was used as an internal standard. ^{1}H chemical shifts are reported in δ units relative to DSS methyl peak (appearing as a singlet at $\delta_H = 0.00$).

Compounds *Oligomerised sample of polyG (maroon) and cG monomers (blue)*… ^{1}H Diffusion ordered spectroscopy (DOSY) spectra were recorded using the pulse sequence stebpgp1s19 (available in the standard Bruker library), while ^{31}P DOSY spectra were recorded by using an analogous pulse sequence exchanging the relevant irradiation channel. ^{1}H (^{31}P) DOSY spectrum was recorded with 8k (4k) direct points, 8 (2k) scans and an acquisition time of 510 (310) ms. In each DOSY measurement, 64 spectra were recorded with linearly increasing gradient strength between 5 and 95%. Parameters Δ (diffusion period) and δ (length of the gradient pulses) were optimized for the specific sample and each spectrometer used and were ultimately set to 50 (250) and 1.2 (2) ms, respectively. All spectra were processed with zero-filling to 16k direct points. Subsequent data analysis was conducted using commercially available softwares Bruker TopSpin version 4.0.5 and Bruker Dynamics Center version 2.7.2. Peak intensities were extracted from the DOSY spectrum and the signal decays for each signal were fitted to Equation 1:

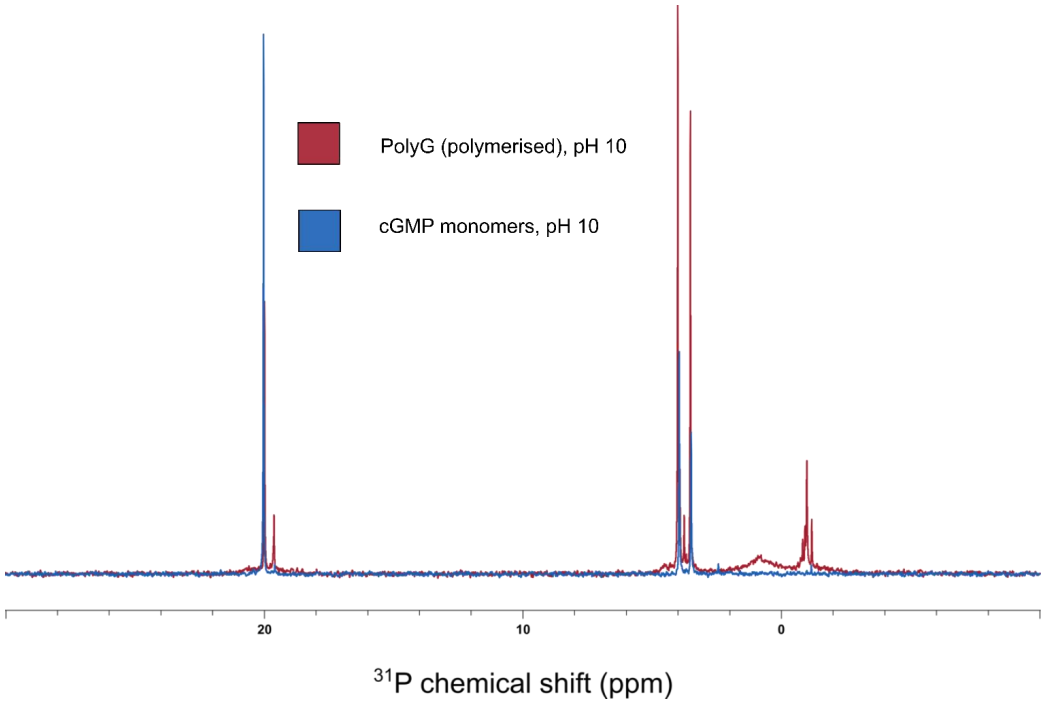
$$I_G = I_0 e^{(-(\gamma\delta G)^2 D\left(\Delta - \frac{\delta}{3}\right))}$$

where I is the normalized signal intensity, γ is the gyromagnetic ratio, G is the gradient strength, and D is the diffusion coefficient. Fitted profiles and the relevant experimental errors are available in Supplementary Tables S4-S7.

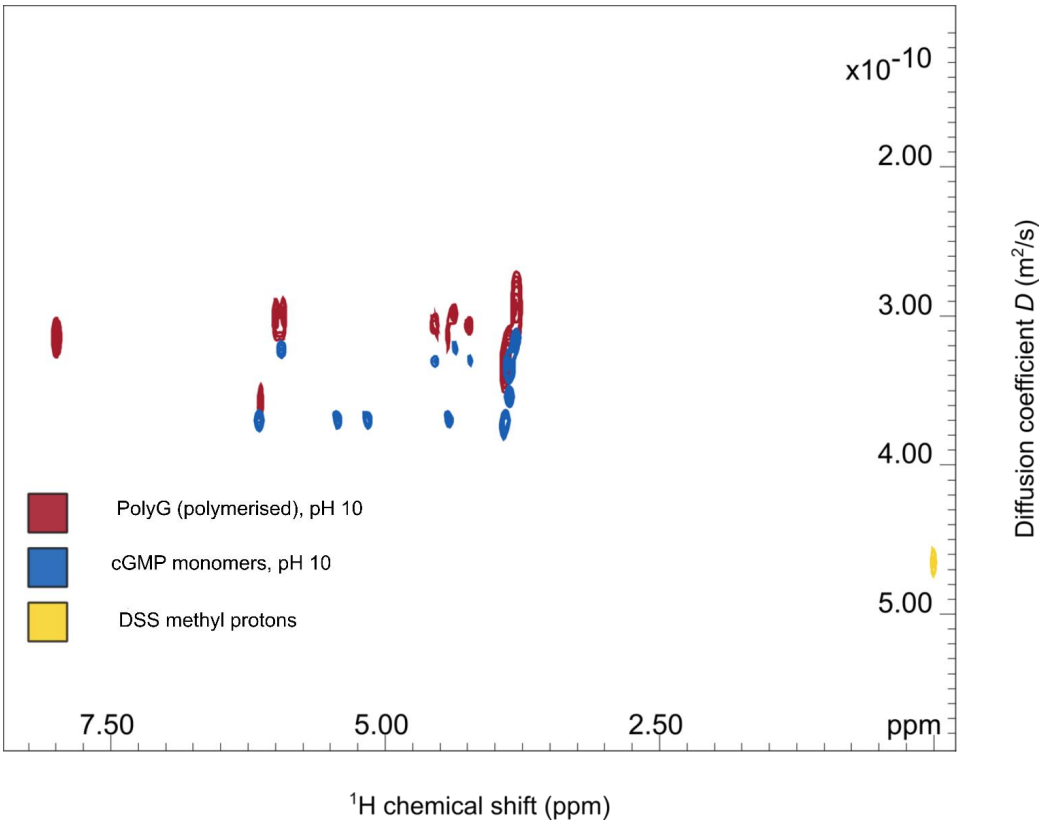
NMR supplementary figures:



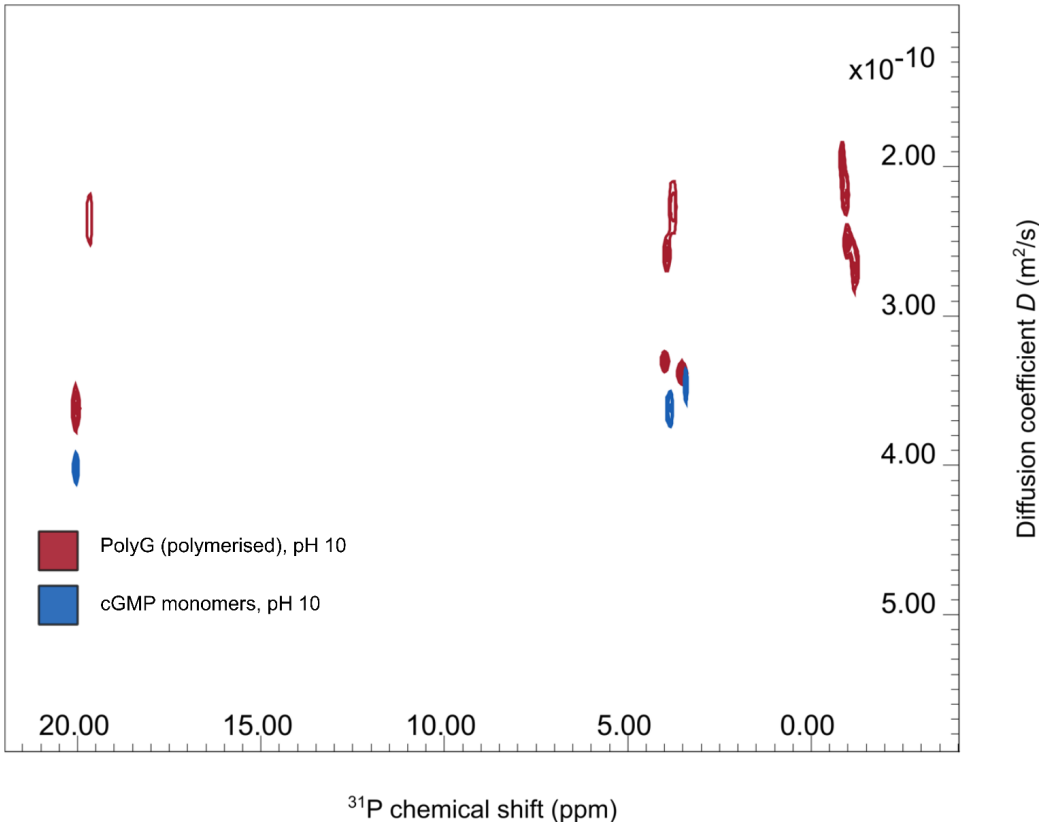
S13 ^{1}H NMR spectrum of Oligomerised sample of PolyG (maroon) and cGMP monomers (blue) each at 60 mM concentration in 600 µL volume. Relaxation delay was set at 1 s. Spectrum is referenced to the signals at 0.00 ppm belonging to DSS methyl protons (internal standard).



S14. *^{31}P NMR spectrum of* Oligomerised sample of polyG (maroon) and cG monomers (blue) each at 60 mM concentration in 600 µL volume. *Relaxation delay was set at 10 s.*



S15. Overlap of pseudo-2D ^{1}H NMR DOSY spectra of Oligomerised sample of PolyG (maroon) and cG monomers (blue), each at 60 mM concentration in 600 µL volume samples in D2O at 25 °C. The internal standard DSS is highlighted in yellow. The fitted value of 4.63x10 -10 m2s-1 is in good agreement with previously reported values.[7]



S16. Overlap of pseudo-2D 31P NMR DOSY spectra of Oligomerised sample of polyG (maroon) and cG monomers (blue) each at 60 mM concentration in 600 µL volume in D2O at 25 °C.

NMR supplementary tables:

Table S4: Fitted values for ^{1}H DOSY experiment of the cGMP monomers sample.

Peak name	*F2 [ppm]*	*D [m2/s]*	*error*
1	*6.133*	*3.737e-10*	*1.1817e-12*
2	*6.131*	*3.735e-10*	*1.1596e-12*
3	*5.930*	*3.243e-10*	*1.9195e-12*
4	*5.923*	*3.197e-10*	*1.8430e-12*
5	*5.422*	*3.727e-10*	*1.2362e-12*
6	*5.143*	*3.709e-10*	*1.5689e-12*
7	*4.795*	*1.479e-09*	*4.8166e-10*
8	*4.535*	*3.318e-10*	*5.3356e-12*
9	*4.410*	*3.697e-10*	*1.4674e-12*
10	*4.349*	*3.252e-10*	*1.4682e-12*
11	*4.219*	*3.290e-10*	*2.1606e-12*
12	*3.907*	*3.798e-10*	*2.1546e-12*
13	*3.891*	*3.724e-10*	*1.0567e-12*
14	*3.864*	*3.320e-10*	*8.8253e-13*
15	*3.860*	*3.367e-10*	*2.2188e-12*
16	*3.852*	*3.537e-10*	*1.7368e-12*
17	*3.842*	*3.543e-10*	*2.3051e-12*
18	*3.836*	*3.549e-10*	*2.2265e-12*
19	*3.815*	*3.221e-10*	*1.4775e-12*
20	*3.811*	*3.185e-10*	*1.0990e-12*
21	*3.800*	*3.182e-10*	*2.1016e-12*
22	*3.795*	*3.147e-10*	*2.2578e-12*

Table S5: Fitted values for ^{1}H DOSY experiment of the oligomerised polyG sample. Data highlighted in red indicates the DSS internal standard diffusion coefficient.

Peak name	*F2 [ppm]*	*D [m2/s]*	*error*
1	*7.984*	*3.153e-10*	*4.2449e-12*
2	*7.971*	*3.142e-10*	*4.7841e-12*
3	*6.123*	*3.592e-10*	*4.0076e-12*
4	*6.120*	*3.568e-10*	*4.0470e-12*
5	*5.982*	*3.097e-10*	*3.6368e-12*
6	*5.977*	*3.024e-10*	*3.7378e-12*
7	*5.924*	*3.029e-10*	*3.2028e-12*
8	*5.917*	*2.999e-10*	*4.4590e-12*
9	*4.536*	*3.053e-10*	*3.0396e-12*
10	*4.420*	*3.116e-10*	*4.0481e-12*
11	*4.362*	*2.959e-10*	*2.3600e-12*
12	*4.223*	*3.041e-10*	*2.6135e-12*
13	*3.908*	*3.370e-10*	*6.0910e-12*
14	*3.892*	*3.323e-10*	*5.6738e-12*
15	*3.864*	*3.145e-10*	*1.2053e-12*
16	*3.807*	*2.910e-10*	*1.7754e-12*
17	*3.795*	*2.945e-10*	*6.0192e-12*
18	*3.792*	*2.902e-10*	*6.8443e-12*
19	*-0.001*	*4.632e-10*	*4.5269e-12*

Table S6: Fitted values for ^{31}P DOSY experiment of the cGMP monomers sample.

Peak name	*F2 [ppm]*	*D [m2/s]*	*error*
1	*20.024*	*4.026e-10*	*8.7340e-12*
2	*3.941*	*3.581e-10*	*1.3170e-11*
3	*3.476*	*3.499e-10*	*1.5036e-11*

Table S7: Fitted values for ^{31}P DOSY experiment of the oligomersied polgG sample.

Peak name	*F2 [ppm]*	*D [m2/s]*	*error*
1	*19.991*	*3.608e-10*	*1.3481e-11*
2	*19.619*	*2.320e-10*	*2.9944e-11*
3	*4.003*	*3.276e-10*	*4.1305e-12*
4	*3.923*	*2.600e-10*	*1.0986e-11*
5	*3.756*	*2.293e-10*	*2.3110e-11*
6	*3.517*	*3.402e-10*	*5.4489e-12*
7	*-0.829*	*1.978e-10*	*1.8610e-11*
8	*-0.919*	*2.154e-10*	*1.5313e-11*
9	*-0.988*	*2.493e-10*	*7.2111e-12*
10	*-1.179*	*2.691e-10*	*1.7769e-11*

Numerical Polymerisation Model

Fit of the data with an effective polymerisation model

In this section we show how to derive an effective polymerisation model and use it to fit the length distribution of polymers measured from the experiments in the trap. To evaluate the basic assumptions of the model we only fit the length distribution of homopolymers (consisting purely of G, A, U, or C, respectively), but do not consider mixed polymers. For these homopolymers, with the help of suitable approximations, the total number of fit parameters can be reduced to only three to four parameters per species and good fit results can be achieved.

In the following, we first derive the effective model that describes ligation of two polymers as an effective three-particle reaction of the two polymers together with a template polymer. To derive these effective rates, we consider only leading order contributions, i.e., we consider only those configurations of polymers bound with templates that contribute most to the overall effective ligation rate. We also discuss why we expect that this approximation via an effective rate adequately describes the dynamics of the system. Finally, we analyze the fits of the model to the data.

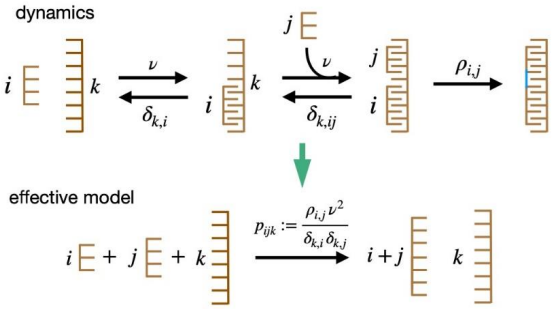
Derivation of the effective model

First, we consider only one species (G) which forms the scaffolds and does not seem to be significantly affected by the presence of the other, so that it can be analyzed in isolation. The concentration of polymers of length i is denoted by c_i. In the following, by using the term 'polymer' we typically also include the monomers (polymers of length 1) and denote the concentration of monomers by c_1. Ligation of polymers is only possible via stacking with a 'template' polymer. Therefore, it can be assumed that ligation takes place in three steps (S10).

In the first step, a polymer (ligand) of length i stacks with a template polymer (template) of length k ($i = 4$ and $k = 8$ in the example in (S10). The stacked state is unstable, because stacking is typically weak and thus highly reversible. We denote the total concentration of stacked pairs of two polymers of size k and i by $c_{k,i}$ and their decay rate (destacking rate) by $\delta_{k,i}$. We assume that the decay rate decreases exponentially with the number of stacking bonds $n_{k,i}$ between the template of size k and ligand of size i,

$$\delta_{k,i} = Ae^{-n_{k,i}h}. \qquad (1)$$

Here, h can be interpreted as the free energy difference per stacking bond relative to the thermal energy scale k_BT. To determine the numbers $n_{k,i}$ of bonds between the polymers, we only account for those configurations that have a maximum number of bonds and, therefore, will make the largest contribution to the ligation rate. For example, if a polymer of size i binds with a template of larger size k, as depicted in figure S17, the (maximum) number of stacking bonds is $2i$ (each base of the ligand forms two stacks with the template). As we will see below, however, only the total number of bonds between the ligands and the template is relevant for our model, and for this total number we can find a rather simple expression.



S17. Effective model for the polymerisation kinetics. We assume that two RNA polymers of lengths i and k can stack together with rate ν. The de-stacking rate $\delta_{k,i}$ decreases exponentially with the number of stacked bases $n_{k,i}$ (here 2i), see Eq. (1). Subsequently, another polymer of length j can stack to the complex and if the stacks persist long enough, the polymerisation reaction ligates the two polymer strands with rate ρ. Since for polymers up to a certain length the de-stacking rate is large compared to the stacking rate, we can derive an effective model that describes ligation of the two strands of length i and j as an effective three-particle reaction with effective rate constant p_{ijk}. In total, the effective model thus depends on only four fit parameters, two of which can be well estimated from the literature.

In the second step, a second ligand of length j stacks next to the first ligand. We denote the concentration of two ligands of length i and j stacked with a template of length k by $c_{k,ij}$ and the corresponding decay rate by $\delta_{k,ij}$. Again, we only consider those configurations that are most stable and hence contribute most to the total ligation rate. To leading order, the decay rate $\delta_{k,ij}$ will thus be determined by the number of bonds between the template k with the smaller of the two ligands i or j, which we assume to be j (the reaction where the smaller polymer binds first has only a subleading contribution on the total ligation rate). Hence,

$$\delta_{k,ij} \approx \delta_{k,j} = Ae^{-n_{k,j}h}. \tag{2}$$

Lastly, in the third step, a polymerisation reaction ligates the strands i and j. It can be assumed that polymerisation is slow and irreversible on the time scale of the experiment. Furthermore, it is likely that the polymerisation rate depends on the lengths of the ligands. We denote by ρ_{ij} the polymerisation rate of two ligands of lengths i and j. As a first order approximation it is sufficient to set all ligation rates equal and distinguish only the case when both ligands are monomers. However, we found that the fits can be improved if we additionally distinguish the case when one ligand is a dimer and the other one a monomer. Hence, we define

$$\rho_{ij} = \begin{cases} \mu_1 \rho & \text{if } i = j = 1 \\ \mu_2 \rho & \text{if } i = 2, j = 1 \\ \rho & \text{else}, \end{cases} \tag{3}$$

where μ_1 and μ_2 are dimensionless fit parameters that quantify how much slower polymerisation proceeds for two monomers or a monomer and a dimer, respectively, as compared to when larger polymers are involved. A possible reason for the slower ligation of monomers is that they are not as tightly constrained in the template as larger polymers.[8] It should take longer, on average, for a successful ligation

to occur if the orientation and spacing between monomers changes frequently due to thermal fluctuations.

Denoting the stacking rate ν, the dynamics of the concentrations of the intermediate states $c_{k,i}$ and $c_{k,ij}$ and the final product c_{i+j} can now be described by the following set of rate equations:

$$\begin{aligned}\frac{d}{dt}c_{k,i} &= \nu c_i c_k - \delta_{k,i}c_{k,i} - \nu c_{k,i}c_j + \delta_{k,ij}c_{k,ij}\\ \frac{d}{dt}c_{k,ij} &= \nu c_{k,i}c_j - \delta_{k,ij}c_{k,ij} - \rho_{ij}c_{k,ij}\\ \frac{\mathrm{d}}{dt}c_{i+j} &= \rho_{ij}c_{k,ij}\end{aligned} \tag{4}$$

Assuming that the concentrations of the intermediate states are stationary, the rate of production of the final product is determined by equating the first and the second equation with 0 and eliminating $c_{k,i}$,

$$\frac{d}{dt}c_{i+j} = \rho_{ij}\frac{\nu^2 c_i c_j c_k}{\delta_{k,i}\delta_{k,ij}+\rho_{ij}(\delta_{k,i}+\nu c_j)} \approx \frac{\rho_{ij}\nu^2}{\delta_{k,i}\delta_{k,ij}}c_i c_j c_k =: p_{ijk}\, c_i c_j c_k\,, \tag{5}$$

where in the second step we assumed that $\delta_{k,i}\delta_{k,ij} \gg \rho_{ij}\left(\delta_{k,i}+\nu c_j\right)$.

Let us halt here for a moment to evaluate what these approximations mean and whether they are indeed justified for the system. The stationarity assumption for the concentrations of the intermediate states is usually justified if the off-rates for the intermediate reactions are much larger than the corresponding on-rates, or, in other words if

$$c_i \nu \ll \delta_{k,i} \qquad for\ all\ i\,, \tag{6}$$

where Eq. (2) was used to simplify the condition for the second intermediate reaction.

Note that a polymer i can have at most $2i$ bonds with another polymer k and thus a more stringent condition is obtained by replacing $\delta_{k,i}$ with Ae^{-2ih} in the above inequality. Typical values for the rate constants found in the literature are $\nu \approx 1\ (\mu Ms)^{-1}$, $h = 0.5$ and $A \approx 10^5\ \mathrm{s}^{-1}$ at 40° Celsius.[9–12] Evaluation of Eq. (6) with the final concentrations from the experiment (and the initial concentration $c_1 = 20\,mM$ for the monomers) shows that the condition is most stringent for the monomers (because the concentration is large) and the largest polymers observed in the experiment (because the exponential factor is dominant). Only for the largest polymers (size > 10) observed in the final distribution could the stationarity condition be violated, since the left-hand side in Eq. (6) may slightly exceed the right-hand side if the above-mentioned values for the parameters are assumed. However, since the concentration of large polymers is rather low, their influence on the dynamics as catalysts for ligation is probably very low anyway. Furthermore, at earlier times during the experiment, their concentration is even significantly lower so that until a certain time Eq. (6) is approximately fulfilled for all types of polymers. Under these considerations, stationarity in Eq. (4) is a reasonable assumption. Assuming that Eq. (6) is fulfilled, the approximation in the derivation of Eq. (5) is justified if both $\delta_{k,i}$ and $\delta_{k,ij}$ are larger than ρ_{ij}, which is again very likely.[13]

Taken together, we expect that ligation in the experiment is well described by the effective rate in Eq. (5). Note, however, that the approximation would break down if polymers grew to a significantly larger size or if the concentrations were strongly increased.
Equation (5) states that ligation is an effective three-particle reaction of two ligands with a template. Together with Eqs. (1) and (2), the effective ligation rate can be transformed to

$$p_{ijk} = \frac{\rho_{ij} v^2}{\delta_{k,i}\delta_{k,j}} = \frac{\rho_{ij} v^2}{A^2} \frac{1}{e^{-(n_{k,i}+n_{k,j})h}}, \tag{7}$$

where $(n_{k,i} + n_{k,j})$ is the total number of bonds between the template k and the two ligands i and j. The largest contribution to the overall rate comes from those configurations that have a maximum number of bonds. The maximum number of bonds is limited by the smaller of either the length of the template or the combined lengths of the ligands. Hence, $(n_{k,i} + n_{k,j}) \approx 2\min(k, i+j)$ which further simplifies the effective rate constant:

$$p_{ijk} = \frac{\rho_{ij} v^2}{A^2} \frac{1}{e^{-2\min(k,i+j)h}} =: \alpha_{ij}\, e^{2\min(k,i+j)h} \tag{8}$$

For the overall pre-factor α_{ij} we distinguish only the case when both ligands are monomers and when one is a dimer and the other a monomer (cf. Eq. (3)):

$$\alpha_{ij} = \begin{cases} \mu_1 \frac{\rho v^2}{A^2} =: \mu_1\, \alpha & \text{if } i = j = 1 \\ \mu_2\, \alpha & \text{if } i = 2,\ j = 1 \\ \alpha & \text{else} \end{cases} \tag{9}$$

Therefore, we fit the model with only four parameters $(\alpha, \mu_1, \mu_2, h)$ for the species G. To fit the other species as well, we use the same model but with independent fitting parameters $(\alpha_\sigma, \mu_{1,\sigma}, \mu_{2,\sigma}, h_\sigma)$ for each species $\sigma = \mathrm{U}, \mathrm{A}, \mathrm{C}$ assuming that only G-polymers can act as template and catalyze ligation. This assumption is reasonable because, first, the yield for G is significantly larger than the yield for the other species and, second, stacking is less stable between A, U and C with themselves than with G.

To formulate the rate equations for the full dynamics of G (the other species are included analogously), it is useful to define the effective ligation rate for two ligands of length i and j

$$\tilde{p}_{ij} \coloneqq \sum_k p_{ijk}\, c_k\,, \tag{10}$$

by performing the weighted sum over the length distribution of templates (G-polymers).

The full model for G then reads

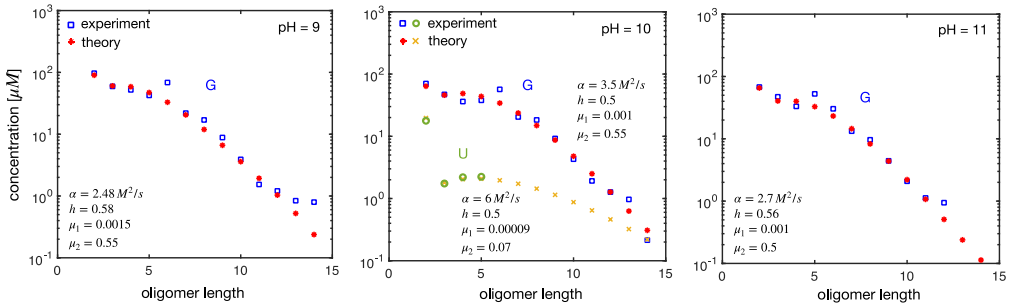
$$\begin{aligned} \frac{d}{dt} c_1 &= -\sum_{j\neq 1} \tilde{p}_{1j}\ c_1 c_j - 2\tilde{p}_{11}\, c_1 c_1\,, \\ \frac{d}{dt} c_\ell &= \frac{1}{2} \sum \sum_{i+j=\ell} \tilde{p}_{ij}\, c_i c_j - \sum_{j\neq \ell} \tilde{p}_{\ell j}\ c_\ell c_j - 2\tilde{p}_{\ell\ell}\, c_\ell c_\ell\,. \end{aligned} \tag{11}$$

The factor of $1/2$ in the second line of Eq. (11) avoids double counting in the sum and the two factors of 2 in the first and second line are stoichiometric factors (two copies of a polymer of length ℓ are lost if they are ligated).

Equation (11) formally describes the concentrations of polymers of arbitrary lengths. However, simulations of the rate equations are only reasonable for polymers up to a finite cut-off length. Introducing such a cut-off length is necessary because the approximation breaks down for large polymers. This also matches the experimental observations, where only polymers with a length of up to 14 bases are observed with significant statistics (S18). Furthermore, a cut-off length slightly larger than 10 significantly speeds up the simulation. We chose a cut-off length of 15.

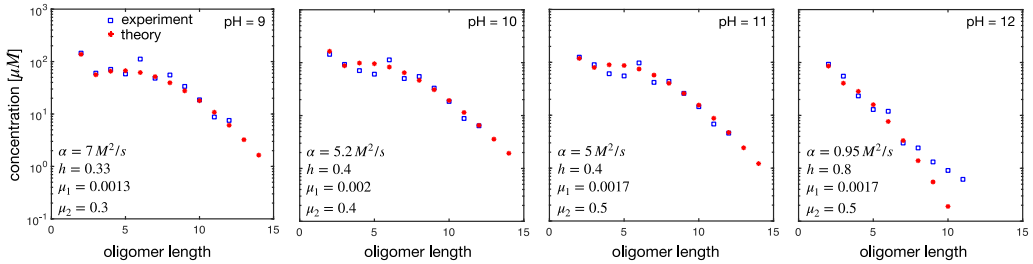
Fitting the model to the data

To fit the parameters of our model, we simulated Eq. (11) numerically from time $t_0 = 0$ until $t_{end} = 18\,h$, which corresponds to the runtime of the experiments. We fitted the resulting final polymer-length distribution to the length distribution obtained from the experiments, using standard least-squares optimization to determine the model parameters. Obvious outliers in the experimental data (often the data point for the hexamers) have been ignored in the fitting procedure. Since the model has only four parameters, it was most effective to sample large regions of the parameter space systematically by varying the parameters on a dense grid. To this end, we first sampled physically reasonable regions of the parameter space with a rough grid of $10^4 - 10^5$ equidistant grid points and subsequently refined the grid around the optimum to finely adjust the parameters. In this way, we were able to ensure that we actually found a global optimum and not just a local one.



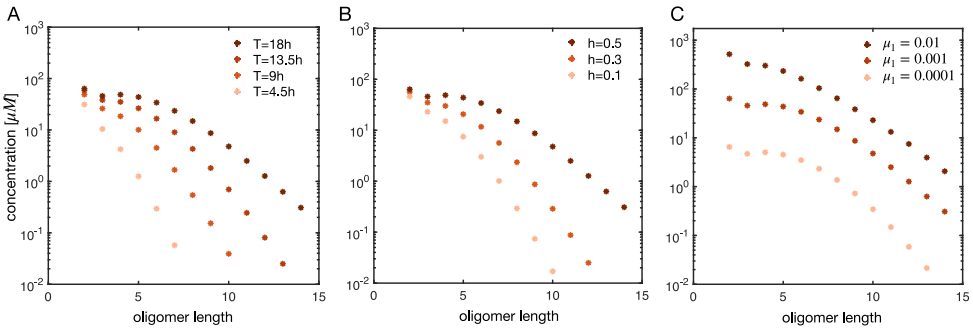
S18: Fit of the final length distribution of homo-G polymers in the experiment after 18 hours in 2 eqv. Na+ solution at different pH levels (blue squares) with the theoretical model (red stars). Fits were obtained with the method of least squares by sampling the complete (physically meaningful) parameter space and numerically integrating the model until time $\mathrm{t} = 18\,\mathrm{h}$. The initial monomer concentration in the experiment and in the simulation was $20\,\mathrm{mM}$. The final monomer concentration is omitted for clarity. The parameter values for which the optimal fits were obtained are indicated in each plot. For $\mathrm{pH} = 10$ we additionally plot the distribution of homo-U polymers (green circles) and the corresponding fit by the model (yellow crosses).

S18 shows the fit to the experimental polymer length distribution obtained at different pH levels in 40 mM Na+ salt solution with an initial monomer concentration of 20 mM. The parameter values that generated the best fit are indicated in the plots. For the system with pH=10 we also show the best fit to the length distribution of homo-U polymers. However, since for U polymers only four data points from the experiment are available, the fitted distribution and the induced parameter values are not fully reliable in this case. The same problem applies to A and C polymers, for which even less data points are available, so that it does not make sense to fit the four-parameter model to this data. S12 analogously shows the fit of the model to the polymer length distribution obtained in 40 mM solution of K+ at different pH levels.



S19. Fit of the final length distribution of homo G polymers in the experiment after 18 hours in $40\ \mathrm{mM}$ K+ solution at different pH levels (blue squares) with the theoretical model (red stars). Fits were obtained with the method of least squares by sampling the complete (physically meaningful) parameter space and numerically integrating the model until time $\mathrm{t} = 18\ \mathrm{h}$. The initial monomer concentration in the experiment and in the simulation was $20\ \mathrm{mM}$. The final monomer concentration is omitted for clarity. The parameter values for which the optimal fits were obtained are indicated in each plot.

In all cases, the experimental distribution can be described well by our theoretical model, apart from a few outliers. In most cases, the parameter h at the optimal fit is rather close to the presumed value of 0.5. The parameter α is fitted best, on average, with a value of approximately $4\ M^2 s^{-1}$, which can be reduced (for example) to $\nu = 1\ (\mu M s)^{-1}$, $A = 10^5\ \mathrm{s}^{-1}$, $\rho = 0.04\ \mathrm{s}^{-1}$, all of which appear as plausible values for the rate constants according to estimates obtained from the literature.[9–12] Interestingly, to fit the curves accurately, the parameter μ_1 must be chosen rather small of the order 10^{-3}. Larger values for μ_1, according to the model, would lead to a greatly enhanced concentration of dimers and subsequently also enhanced concentrations for the larger polymers (see S20C).

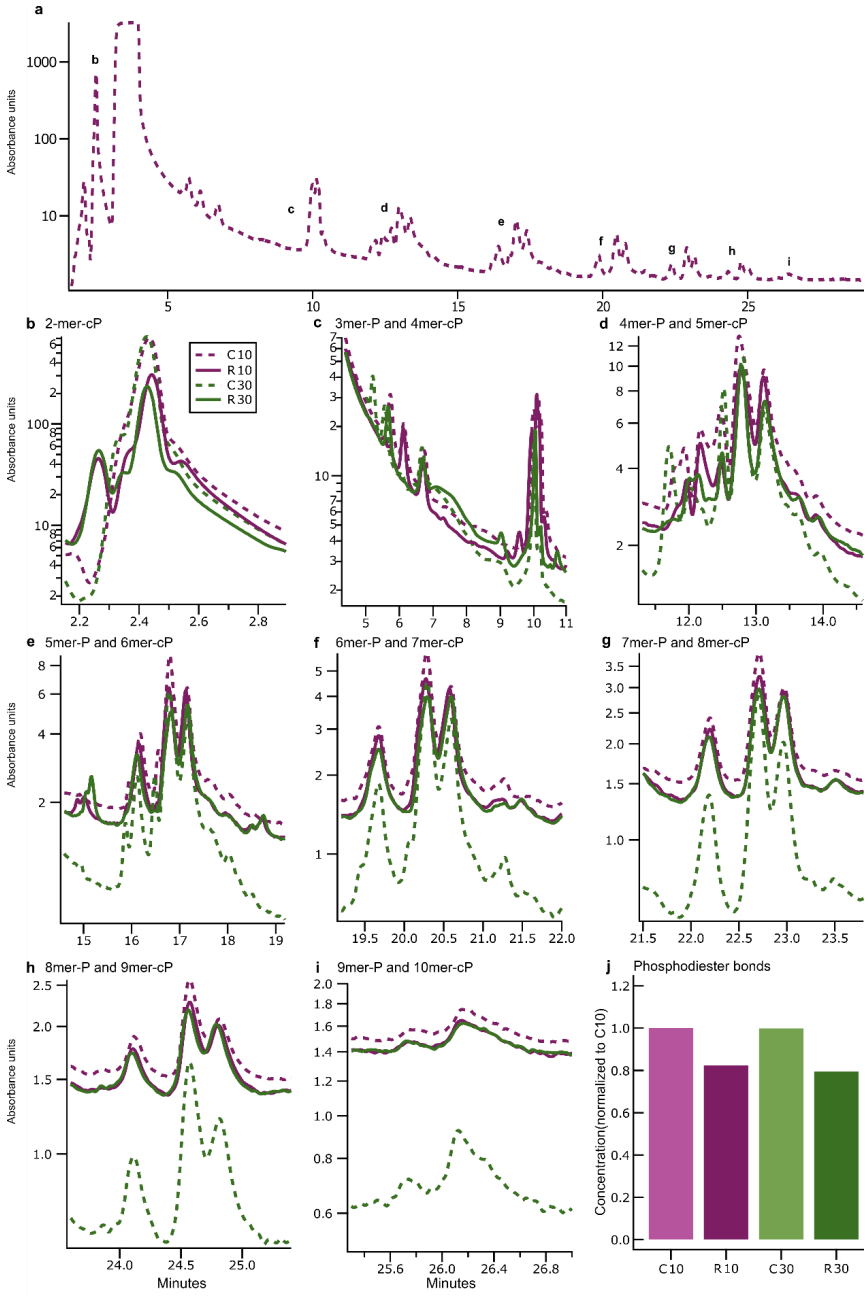


S20. Effect of the fitting parameters. In the optimal fit shown for the system in the middle panel ($\mathrm{pH} = 10$), we vary single fitting parameters of the theory to demonstrate their effect on the final polymer distribution. **A** Variation of the runtime of the simulation, which is equivalent to a variation in the parameter ν that sets the overall time scale. Primarily, the concentration of larger polymers is strongly suppressed at earlier times and the distribution is steeper. **B** Variation of the exponential factor h. Smaller values of h imply that the stacks are less stable and thus templated reactions are less frequent, which decreases the concentration of larger polymers. **C** Variation of the dimerisation barrier. The dimerisation barrier mainly affects the distribution as a whole, shifting it up or down, but only slightly affects the slope of the curve.

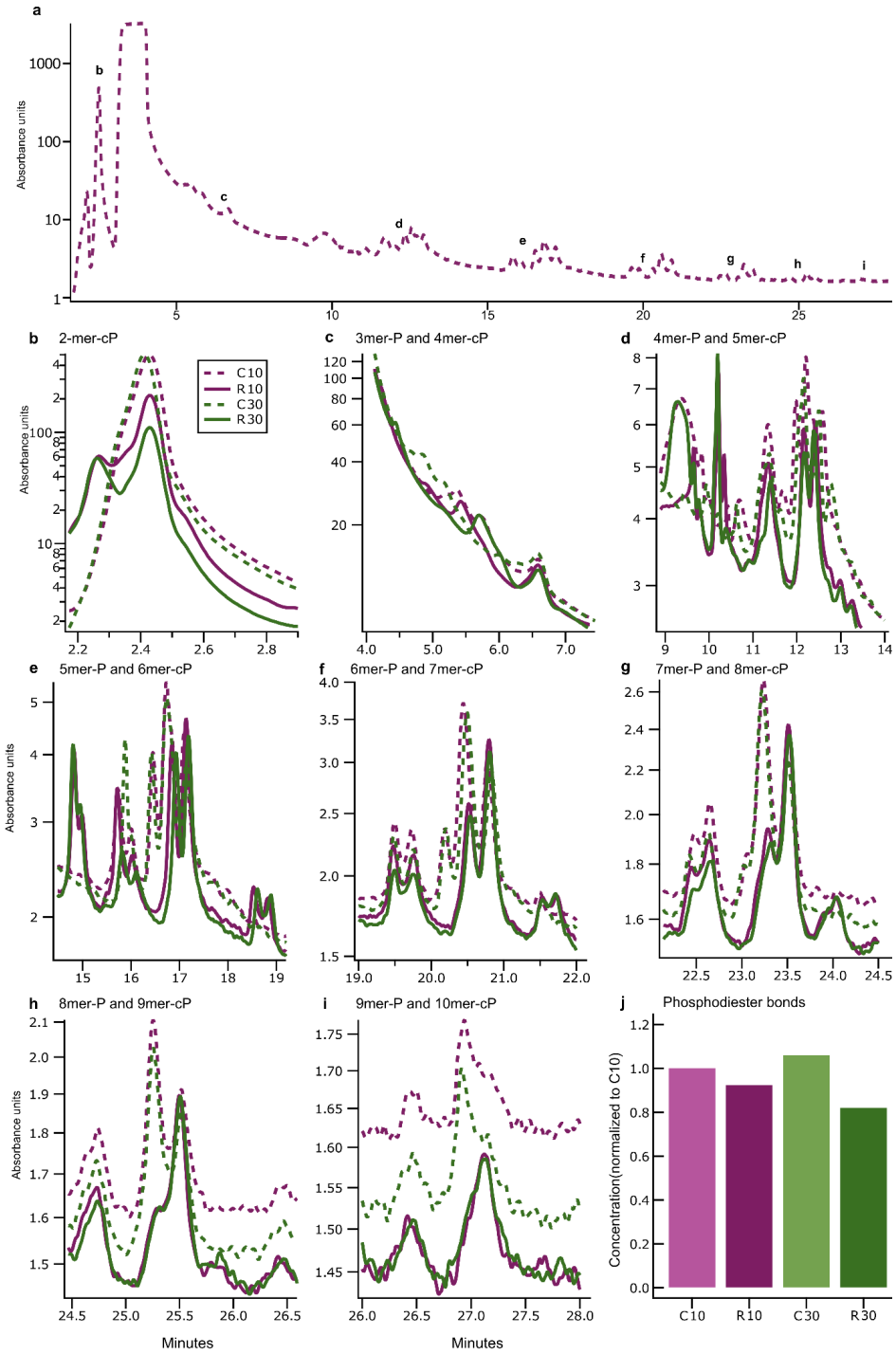
The model – with appropriately fitted parameters – furthermore allows us to study the dynamics of the system and the specific effect of the fitting parameters on the final polymer size distribution. S20A shows the polymer size distribution at four equidistant intermediate time points obtained with the fitting parameters for G in S18 (middle panel, pH=10). Note that, since the parameter ν determines the time scale, varying the runtime t_{end} of the simulation is equivalent to varying the parameter ν. Moreover, S20B and S20C show the effects of variations of the parameters h and μ_1, which either change the steepness of the distribution or roughly shift the curve in the vertical direction.

In total, the experimental finding of a flat size distribution of polymers up to a size of approximately 8 bases (cf. S18 and S19) at a rather small overall yield of 2-3 % of the monomer concentration suggests the existence of a significant dimerisation barrier. Such a dimerisation barrier could be induced by a reduced ligation rate for monomers as we discussed above (cf. Eq. (3)). Alternatively, the same behavior could be explained by assuming that the stacking rate between monomers is reduced compared to the stacking rate of larger polymers or, equivalently, that the specific detachment rate of monomers from the template is strongly underestimated by the model. We assume an exponential increase of the detachment rate δ with decreasing number of bonds (see Eqs. (1) and (2)), which might not be reasonable for single monomers. Both of these explanations (slower ligation and faster detachment) would affect the effective ligation rate between monomers and dimers as well, but to a much lesser extent. This is also reflected in the parameter μ_2 which is fitted by a value of the order 10^{-1}, about two orders of magnitude larger than μ_1.

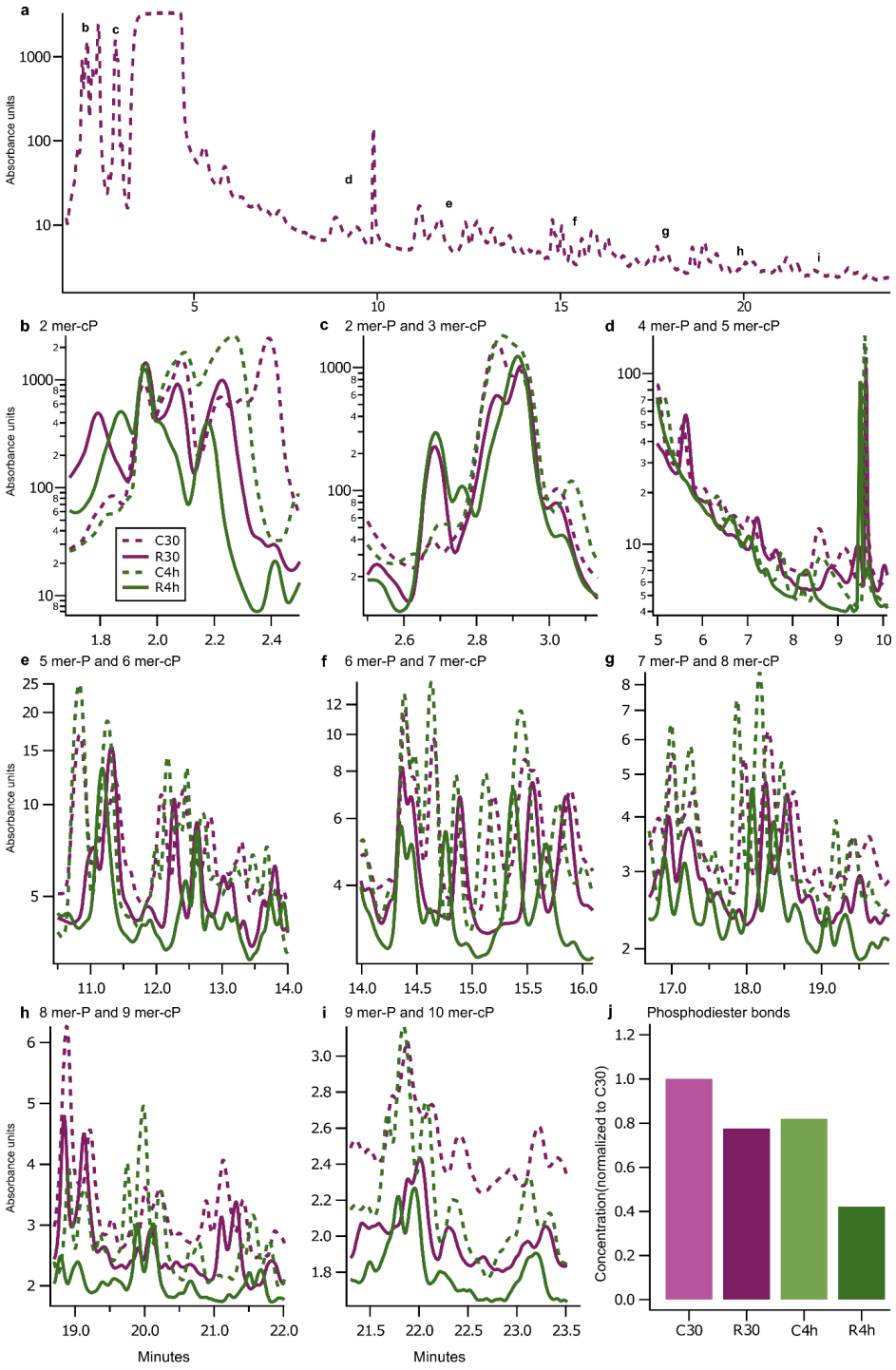
Determination of the phosphodiester linkage by Nuclease P1



S21. The Nuclease P1 digestion was conducted for polymerised 2',3'-cGMP. **a.** UV 260 nm chromatogram of no-enzyme control with 10-minute incubation (C10). Chromatogram shows retention times of different length oligonucleotides (annotated b-i). Saturated peak at 4 minute is due to traces of phenol in the extraction masking some oligonucleotide peaks. **b-i.** UV chromatograms for the samples C10 (no-enzyme control, 10-minute incubation, purple-dashed), R10 (enzyme digest, 10-minute incubation, purple-solid), C30 (no-enzyme control, 30-minute incubation, green-dashed) and R30 (enzyme digest, 30-minute incubation, green-solid). Intensity of some of the peaks are lowered in enzyme digest samples. **b.** 2 mer-cP, **c.** 3 mer-P and 4 mer-cP, **d.** 4 mer-P and 5 mer-cP, **e.** 5 mer-P and 6 mer-cP, **f.** 6 mer-P and 7 mer-cP, **g.** 7 mer-P and 8 mer-cP, **h.** 8 mer-P and 9 mer-cP, **i.** 9 mer-P and 10 mer-cP. **j.** Concentration equivalent of phosphodiester linkages were calculated for each sample using the integrated EIC counts and multiplying it by the number of phosphodiester linkages. Concentrations of the samples R10, C30 and R30 were normalized to that of C10. Both R10 and R30 show ~20% enzymatic hydrolysis of the oligonucleotides. Background hydrolysis is low for 30 minutes of incubation as can be seen on comparing C10 and C30.


a
Absorbance units
b
c
d
e
f
g
h
i
b 2-mer-cP
C10
R10
C30
R30
c 3mer-P and 4mer-cP
d 4mer-P and 5mer-cP
e 5mer-P and 6mer-cP
f 6mer-P and 7mer-cP
g 7mer-P and 8mer-cP
h 8mer-P and 9mer-cP
i 9mer-P and 10mer-cP
j Phosphodiester bonds
Concentration(normalized to C10)
Minutes
Minutes

***S22*.** The Nuclease P1 digestion was conducted for polymerised 2',3'-cGMP. **a.** UV 260 nm chromatogram of no-enzyme control with 10-minute incubation (C10). Chromatogram shows retention times of different length oligonucleotides (annotated b-i). Saturated peak at 4 minute is due to traces of phenol in the extraction masking some oligonucleotide peaks. **b-i.** UV chromatograms for the samples C10 (no-enzyme control, 10-minute incubation, purple-dashed), R10 (enzyme digest, 10-minute incubation, purple-solid), C30 (no-enzyme control, 30-minute incubation, green-dashed) and R30 (enzyme digest, 30-minute incubation, green-solid). Intensity of some of the peaks are lowered in the enzymatic digest samples. **b.** 2 mer-cP, **c.** 3 mer-P and 4 mer-cP, **d.** 4 mer-P and 5 mer-cP, **e.** 5 mer-P and 6 mer-cP, **f.** 6 mer-P and 7 mer-cP, **g.** 7 mer-P and 8 mer-cP, **h.** 8 mer-P and 9 mer-cP, **i.** 9 mer-P and 10 mer-cP. **j.** Concentration equivalent of phosphodiester linkages were calculated for each sample using the integrated EIC counts and multiplying it by the number of phosphodiester linkages. Concentrations of the samples R10, C30 and R30 were normalized to that of C10. R10 shows ~10% and R30 shows ~20% enzymatic hydrolysis. Background hydrolysis is low for 30 minutes of incubation as can be seen on comparing C10 and C30.


a
b
c
d
e
f
g
h
i
Absorbance units
b 2 mer-cP
c 2 mer-P and 3 mer-cP
d 4 mer-P and 5 mer-cP
C30
R30
C4h
R4h
e 5 mer-P and 6 mer-cP
f 6 mer-P and 7 mer-cP
g 7 mer-P and 8 mer-cP
h 8 mer-P and 9 mer-cP
i 9 mer-P and 10 mer-cP
j Phosphodiester bonds
Concentration(normalized to C30)
Minutes
Minutes

***S23*.** The Nuclease P1 digestion was conducted for polymerised 2',3'-cGMP. **a.** UV 260 nm chromatogram of no-enzyme control with 30-minute incubation (C30). Chromatogram shows retention times of different length oligonucleotides (annotated b-i). Saturated peak at 4 minute is due to traces of phenol in the extraction masking some oligonucleotide peaks. **b-i.** UV chromatograms for the samples C30 (no-enzyme control, 30-minute incubation, purple-dashed), R30 (enzyme digest, 30-minute incubation, purple-solid), C4h (no-enzyme control, 4-hour incubation, green-dashed) and R4h (enzyme digest, 4-hour incubation, green-solid). Intensity of some of the peaks are lowered in the enzymatic digest samples. **b.** 2 mer-cP, **c.** 2 mer-P and 3 mer-cP, **d.** 4 mer-P and 5 mer-cP, **e.** 5 mer-P and 6 mer-cP, **f.** 6 mer-P and 7 mer-cP, **g.** 7 mer-P and 8 mer-cP, **h.** 8 mer-P and 9 mer-cP, **i.** 9 mer-P and 10 mer-cP. **j.** Concentration equivalent of phosphodiester linkages were calculated for each sample using the integrated EIC counts and multiplying it by the number of phosphodiester linkages. Concentrations of the samples R30, C4h and R4h were normalized to that of C30 (Comparing C10 and C30 in Figures 1 and 2 shows limited background hydrolysis). R30 shows ~23% and R4h shows ~60% enzymatic hydrolysis. Considering the background hydrolysis of C4h (~20%) and comparing the C4h to R4h gives an estimate of ~49% enzymatic digest, suggesting that ~49% of the phosphodiester linkages are 3'-5'.

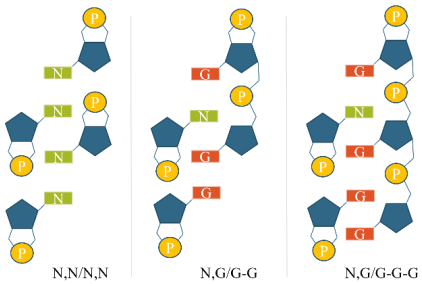
Investigation of Intercalated-Stack Arrangement and its Suitability as Starting Point for Oligomerisation by Computational Simulations

1. Computational Methods

1.1 System Preparation

To study the possible stacked intercalated arrangement previously proposed by Šponer *et al*.[14] for the oligomerisation reaction starting from a 3'-5'-cyclic-GMP tetramer, a set of analogous systems with 2'-3'-cyclic phosphate groups was created.

In order to investigate why predominantly cGMP oligomerises on its own, non-covalently bonded homogeneous tetramers were prepared for the canonical nucleobases, adenine, cytosine, guanine, and uracil (N, N/N, N). Further we aimed to study the role a poly-G matrix could play in facilitating the formation of strands incorporating adenine, cytosine, and uracil. For this purpose, additional systems were created, which include a GMP matrix: (1) N, G/G-G, where N denotes cAMP, cCMP, cGMP, or cUMP, G labels cGMP and G-G indicates two oligomerised GMPs (2) N, G/G-G-G follows the same notation. G-G-G denotes a GMPs trimer. These heterogeneous systems were investigated both with 2'-5' and 3'-5' linkage, respectively. S24 visualizes the notation.



***S24*.** Summary of investigated stacked systems. Bases that are alternated are indicated by N. Non-covalently linked neighbouring bases are separated by a comma. Oligomerised nucleotides are connected by a hyphen. Using a slash, the two strands are indicated.

1.2 Calculation of Stacking Interactions Energies

For the investigation of the stacking behavior, all phosphate groups were saturated with hydrogens to avoid negative charges. The stacking was studied in a static and dynamic manner. Using the program package FermiONs++[15–17] in combination with Chemshell,[18] all systems were optimized in the gas phase, as well as with implicit solvation (C-PCM)[19] at ωB97M-V[20]/def2TZVPD[21,22] level of theory including the VV10[23] dispersion correction.

Following the structure optimizations, the energies of the subunits were computed at the same theory level and in the same environment. To calculate the stacking interaction energies, the subunit energies were subtracted from the energy of the stacked super-system. In addition, the average distances between the centre of mass (based on non-hydrogen atoms) of the nucleobases in each system and the

P–O5 distance are evaluated. This distance is relevant because oligomerisation of the nucleotides would result in a bond forming between these two atoms.

1.3 Dynamic Study of Intercalated Stacked Arrangement

A major shortcoming of investigating the proposed arrangements statically is that, for such an extended system, it is unclear whether one has found a global representative minimum on the potential energy surface during optimization. Moreover, the inclusion of explicit solvent in the calculation of the stacking interaction energies is nontrivial.

To study the relative stabilization of the aggregates, bypassing the need for a representative minimum energy structure and including explicit solvent, we studied the same stacked species by molecular dynamics within a water sphere, with a thickness of 15 Å.

Five GFN-FF[24] molecular dynamics simulations were performed for each system using the xtb[25] program package. All simulations were propagated for 100 ps with a timestep of 2.0 fs at 298.15K. The SHAKE algorithm[26] was used to constrain the X-H motion. To avoid dissociation of the water shell, the system was confined in a cavity by a logfermi potential.[27] Prior to all production runs, the geometries were optimized at GFN-FF level of theory.

2. Results and Discussion of the Studies of the Nucleotide Assembly

2.1 Evaluation of Optimized Stacked Structures

Strong intermolecular interactions, e.g., π-stacking, can lead to nucleotide self-assembly, which is a prerequisite for the auto-oligomerisation reaction. To investigate why the short oligonucleotides, contain predominantly guanine, we computed the stacking interaction energies and evaluated the minimum energy structures of the various stacked system by calculating the average base separation and the P-O5 distance, relevant to the oligomerisation reaction. These results are summarized in TableS9, figures S25 and 26. The minimum energy structures of all systems optimized in the gas phase are shown in S27-31.

We examined tetramers containing non-covalently linked cAMP, cCMP, cGMP, or cUMP nucleotides (N,N/N,N) as well as heterogeneous stacks with 3'-5' and 2'-5' linked GMP strands (N,G/G-G, N,G/G-G-G). Furthermore, we evaluated the stack stability by GFN-FF molecular dynamics simulations.

Table S9: *Stacking interaction energies calculated at ωB97M-V/def2-TZVPD level of theory, average center of mass distances, and P-O5 distances for the N,N/N,N, N,G/G-G, and N,G/G-G-G systems. Energies are given in kcal/mol and distances in Å.*

System	E_{stack}		d_{stack}		*d(P-O5)*		E^{PCM}_{stack}		d^{PCM}_{stack}		d^{PCM}*(P-O5)*	
A,A/A,A	*-54.8*		*3.93*		*4.77, 4.66*		*-39.1*		*3.85*		*4.64, 4.60*	
C,C/C,C	*-49.3*		*4.28*		*4.75, 4.45*		*-31.9*		*4.24*		*4.82, 4.50*	
G,G/G,G	*-51.8*		*3.89*		*4.70, 4.48*		*-38.6*		*3.74*		*4.72, 4.48*	
U,U/U,U	*-32.9*		*4.68*		*4.62, 4.39*		*-22.2*		*4.58*		*4.58, 4.32*	
Linkage	*2'-5'*	*3'-5'*	*2'-5'*	*3'-5'*	*2'-5'*	*3'-5'*	*2'-5'*	*3'-5'*	*2'-5'*	*3'-5'*	*2'-5'*	*3'-5'*
A,G/G-G	*-53.5*	*-68.9*	*3.74*	*3.34*	*4.64*	*3.69*	*-35.3*	*-41.0*	*3.70*	*3.44*	*4.66*	*3.63*
C,G/G-G	*-54.9*	*-60.0*	*4.10*	*4.50*	*4.43*	*9.00*	*-31.2*	*-32.1*	*4.04*	*4.63*	*4.50*	*8.69*
G,G/G-G	*-51.9*	*-74.5*	*3.80*	*4.02*	*4.73*	*3.72*	*-35.8*	*-44.8*	*3.69*	*3.75*	*4.64*	*3.66*
U,G/G-G	*-50.9*	*-72.0*	*3.93*	*3.50*	*4.29*	*5.29*	*-30.9*	*-36.4*	*3.93*	*3.61*	*4.48*	*5.34*
A,G/G-G-G	*-69.2*	*-134.7*	*3.86*	*3.44*	*4.68*	*3.86*	*-44.8*	*-83.2*	*3.83*	*3.44*	*4.59*	*3.84*
C,G/G-G-G	*-79.5*	*-90.1*	*3.97*	*4.28*	*4.51*	*5.07*	*-39.9*	*-57.5*	*3.93*	*4.41*	*4.38*	*3.73*
G,G/G-G-G	*-77.8*	*-90.4*	*4.04*	*4.31*	*4.66*	*3.68*	*-47.2*	*-62.0*	*3.83*	*3.86*	*4.51*	*3.65*
U,G/G-G-G	*-66.5*	*-93.8*	*3.87*	*3.92*	*3.88*	*3.39*	*-39.9*	*-55.3*	*3.91*	*3.88*	*3.76*	*3.60*

E_{stack}: Stacking interaction energies

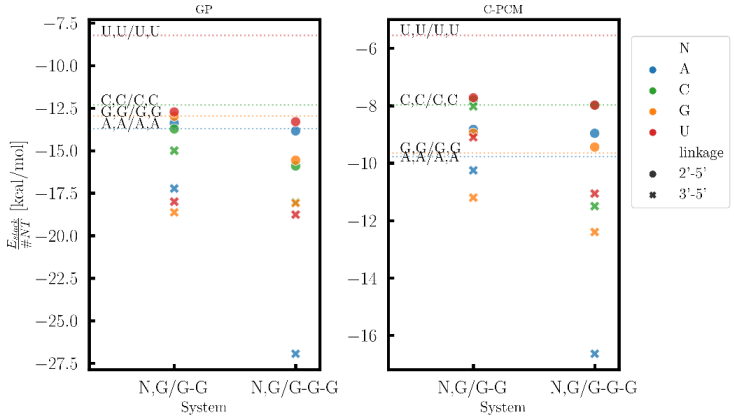
d_{stack}: Average center of mass distance

PCM: Values computed with implicit solvent

The energetic stabilization by stacking, defined as the energy difference of the optimized stacked complex and the single point energies of the subunits, is highest for the adenosine tetramer (gas phase: -54.8 kcal/mol, PCM: -39.1 kcal/mol) closely followed by guanosine and cytosine. The stacking interaction energy for the uracil tetramer is much weaker than for the other nucleobases (gas phase: -32.9 kcal/mol, PCM: -22.2 kcal/mol). S25 shows the stacking interaction energies divided by the number of nucleotides for all investigated systems and reflects the trend described above. It can be seen that this ordering is obtained both in the gas phase as well as incorporating implicit solvent.

Higher stabilization makes the self-assembly more likely. However, other effects such as base solubility, reactivity, and minimum energy geometry affect the rate of oligomerisation.

While the stacking stabilization of the adenosine tetramer is 3 kcal/mol higher than that of the guanosine tetramer in the gas phase, the average interbase distance and both P-O5 distances are larger (S26). This could indicate lower reactivity of the adenosine tetramer than the guanosine species, when stacks have formed. The difference in the stacking interaction energies decreases to 0.5 kcal/mol when implicit solvation is used.



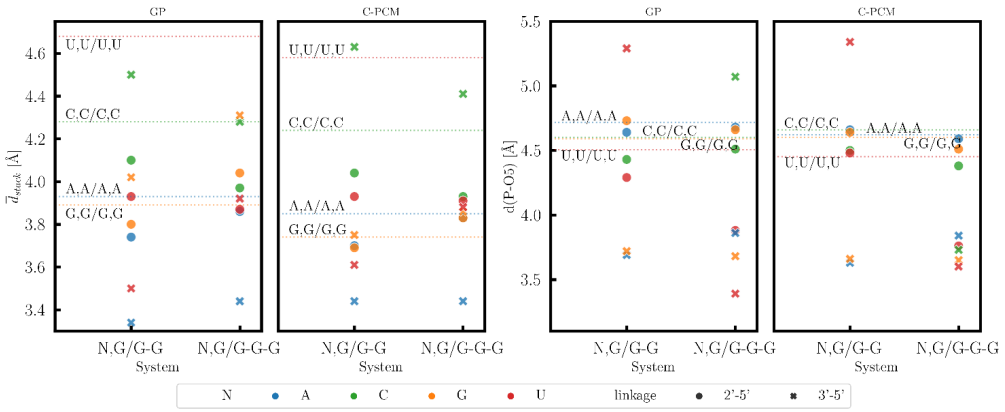
S25. Stacking interaction energies per nucleotide given in kcal/mol. Results from gas-phase calculations are given on the left, and results using implicit solvation are summarized on the right. The values for the N, N/N, N systems are indicated by horizontal lines hereby we aim to ease the comparison to the systems containing a poly-G matrix.

Within the N, N/N, N group the smallest stabilization was determined for U,U/U,U (gas phase: -32.9 kcal/mol, PCM: -22.2 kcal/mol). In the presence of a poly-G matrix, the difference between uracil and the other nucleobases is significantly reduced. This suggests that a poly-G matrix may facilitate the formation of uracil-containing RNA strands.

While a poly-G matrix aids the positioning of cUMP in an intercalated stacked arrangement, it leads to hydrogen-bonded aggregates for cCMP. Even though the obtained hydrogen bonded arrangements are stable, they are presumably less suitable as starting points for oligomerisation. This is reflected in the measured P-O5 distances. For the 3'-5' linked C,G/G-G tetramer they are much greater than all other P-O5 distances (gas phase: 9.00 Å, PCM: 8.69 Å). We hypothesize that cytosine is incorporated into heterogeneous RNA states via an alternative arrangement.

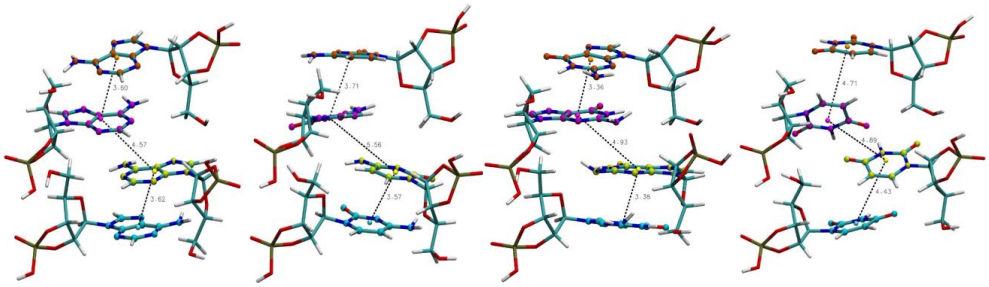
Apart from the 3'-5' linked C,G/G-G system, the strands become more stable and more condensed in presence of a guanine matrix as shown in S26.

Overall, the obtained stacking interaction energies with implicit solvation in water are significantly lower than in the gas phase. The minimum energy structures also differ slightly: the stacks are generally less compact in the gas phase. Nonetheless the same overall trends can be observed using both setups. It should be further noted that the given results are only meant to show trends and are not suitable for direct quantitative comparison to experiment.



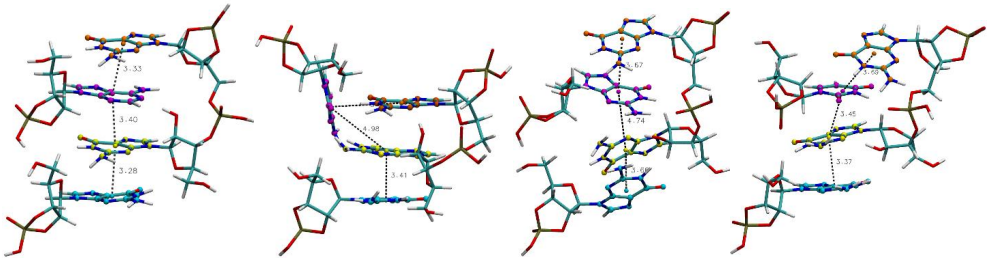
S26. Summary of average center of mass distances and P-O5 Distances of all stacked systems given in Å. Results from gas-phase calculations are given on the left (GP), and results using implicit solvation are summarized on the right (C-PCM). The values for the N,N/N,N systems are indicated by horizontal lines to facilitate comparison with the systems containing a poly-G matrix. The measured P-O5 distances for the 3'-5' linked C,G/G-G tetramer are much greater than all other P-O5 distances (gas phase: 9.00 Å, PCM: 8.69 Å) so that they were omitted.

S27-31 show the obtained minimum energy structures of the stacked species in the gas phase. S27 shows the N,N/N,N set. While A,A/A,A and G,G/G,G are more evenly stacked, more disordered minimum energy structures were obtained for C,C/C,C and U,U/U,U. This matches the lower stability of the stacks.



(a) A,A/A,A. (b) C,C/C,C. (c) G,G/G,G. (d) U,U/U,U.

S27. Visualization of pure-stacked-tetramers (N,N/N,N) optimized at the ωB97M-V/def2TZVPD level of theory. Indicated are the computed center of masses and the associated atoms. Additionally, the center of mass distances is given.



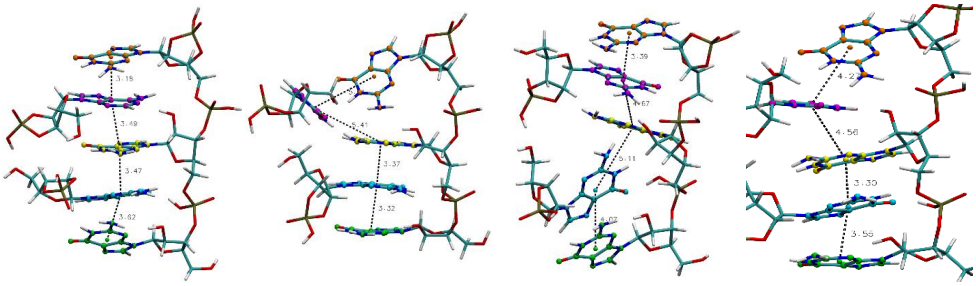
(a) A,G/G-G. (b) C,G/G-G. (c) G,G/G-G. (b) U,G/G-G.

S28. 3'-5' linked heterogeneous-stacked-tetramers (N,G/G-G) optimized at the ωB97M-V/def2-TZVPD level of theory. Indicated are the computed center of masses and the associated atoms. Additionally, the center of mass distances is given.

S28 shows the N,G/G-G set with a 3'-5' linked GMP dimer. A,G/G-G forms a very compact structure, all interbase distances are below 3.5 Å. Baulin *et al.*[28] highlighted the tendency of adenine to intercalate in a recent study where they analyzed the distribution of different motifs in RNA. In the C,G/G-G system the three cGMP nucleotides are *π*-stacked while cCMP is base paired to the inner cGMP nucleotide. As can be seen in S28b the 5'-OH group and the 2'-3' cyclic phosphate group are spatially separated, which makes this assembly unsuitable as starting point for oligomerisation. Other than for the U,U/U,U system, a compact and evenly stacked structure was obtained for U,G/G-G, which indicates that cUMP is suitable for embedding into a poly-G scaffold, where it can be positioned for oligomerisation.

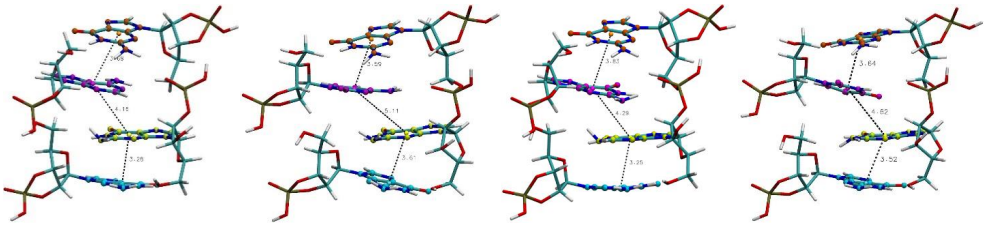
For the heterogeneous-stacked-pentamers (N,G/G-G-G) including 3'-5' cGMP trimers shown in S29, similar observations are made as for the 3'-'5 linked heterogeneous-stacked-tetramers (N,G/G-G) previously discussed.

For the 2'-5' linked heterogeneous stacked systems, shown in S30 and 31, stable stacked structures were obtained for all systems. Therefore, the stacked arrangement might play a bigger role in the formation of 2'-5' linked strands than for 3'-5' linked RNA.



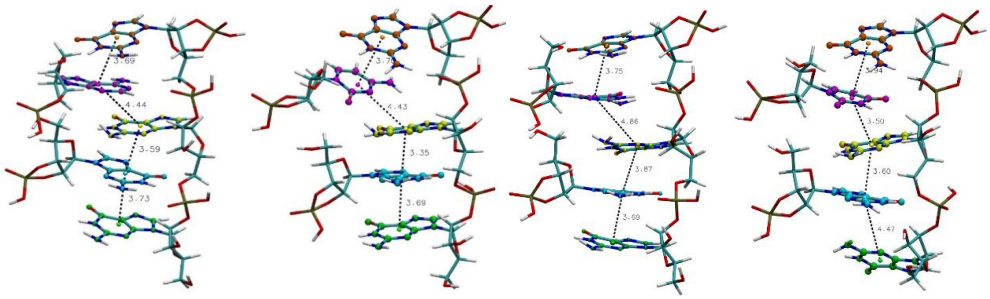
(a) A,G/G-G-G. (b) C,G/G-G-G. (c) G,G/G-G-G. (d) U,G/G-G-G.

S29. 3'-5' linked heterogeneous-stacked-pentamer (N,G/G-G-G) optimized at the ωB97M-V/def2-TZVPD level of theory. Indicated are the computed center of masses and the associated atoms. Additionally, the center of mass distances is given.



(a) A,G/G-G. (b) C,G/G-G. (c) G,G/G-G. (b) U,G/G-G.

***S30*.** 2'-5' linked heterogeneous-stacked-tetramers (N,G/G-G) optimized at the ωB97M-V/def2-TZVPD level of theory. Indicated are the computed center of masses and the associated atoms. Additionally, the center of mass distances is given.

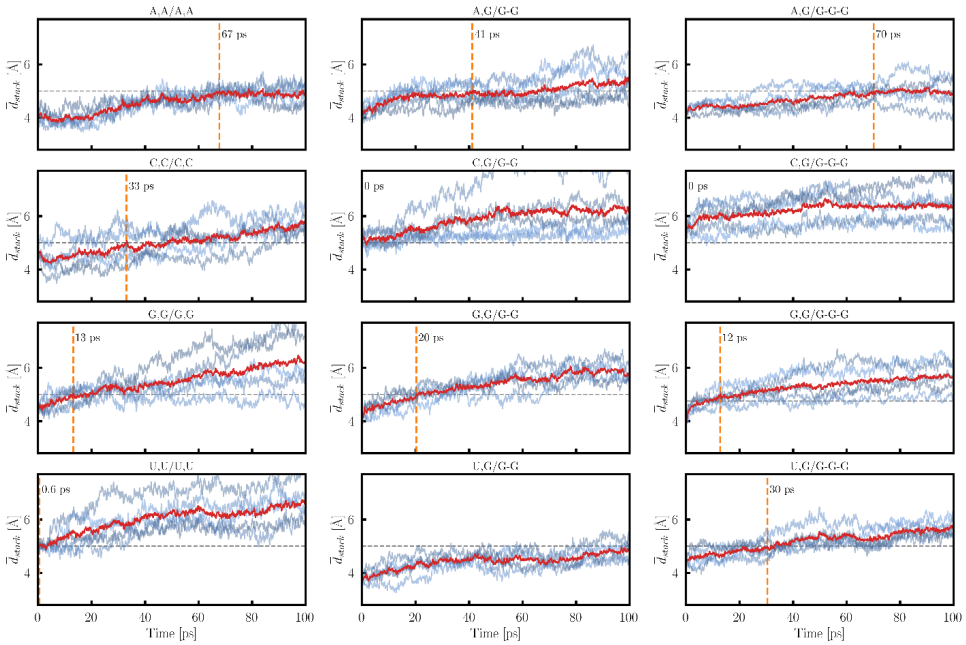


(a) A,G/G-G-G. (b) C,G/G-G-G. (c) G,G/G-G-G. (d) U,G/G-G-G.

S31. 2'-5' linked heterogeneous-stacked-pentamer (N, G/G-G-G) optimized at the ωB97M-V/def2-TZVPD level of theory. Indicated are the computed center of masses and the associated atoms. Additionally, the center of mass distances is given.

2.2 Results of Dynamic Study of the Intercalated Stacked Arrangements

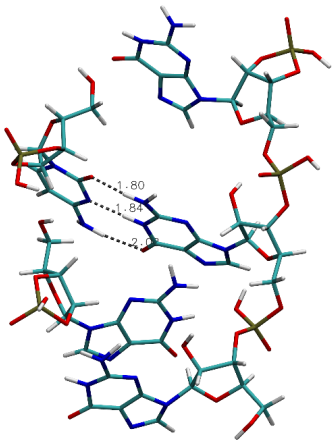
In order to study the relative stability of the stacked systems independent of a representative minimum energy structure, we conducted five 100 ps GFN-FF dynamics simulations for each system. Initially, all systems were closely stacked. To compare the relative stability of the stacks based on molecular dynamics simulations, we compare the time until the stacks disassemble. Here we defined the time of destacking as the point where the mean of the measured average interbase distances of the five simulation runs becomes larger than 5.00 Å.



S32. Change in the average base distance during a GFN-FF dynamics run, for the N,N/N,N (left), 3'-5' linked N,G/G-G (middle) and N,G/G-G-G (right) systems. The five simulation runs are shown in blue, and their average is given in red. The time at which the mean crosses the 5.0 Å destacking threshold is marked in orange.

S32 and 34 show how the average base distances change during the course of five simulation runs (blue). In addition, the mean of the five simulations is given (red) and the destacking time is given and indicated by an orange vertical line. On average, the A,A/A,A species remains stacked significantly longer than the analogous N,N/N,N systems, indicating that the adenine-only system is the most stable among this group. This goes hand in hand with the result of the study of the stacking interaction energies based on an optimized minimum energy structure. The C,C/C,C and G,G/G,G systems reach the destacking threshold of 5 Å after 33 ps and 13 ps, respectively. Unlike the static study, the molecular dynamics simulations suggest that the C,C/C,C stack may be more stable than the G,G/G,G stack. However, this could be an artefact of the low sample size (n=5) and the different level of theory. Similar

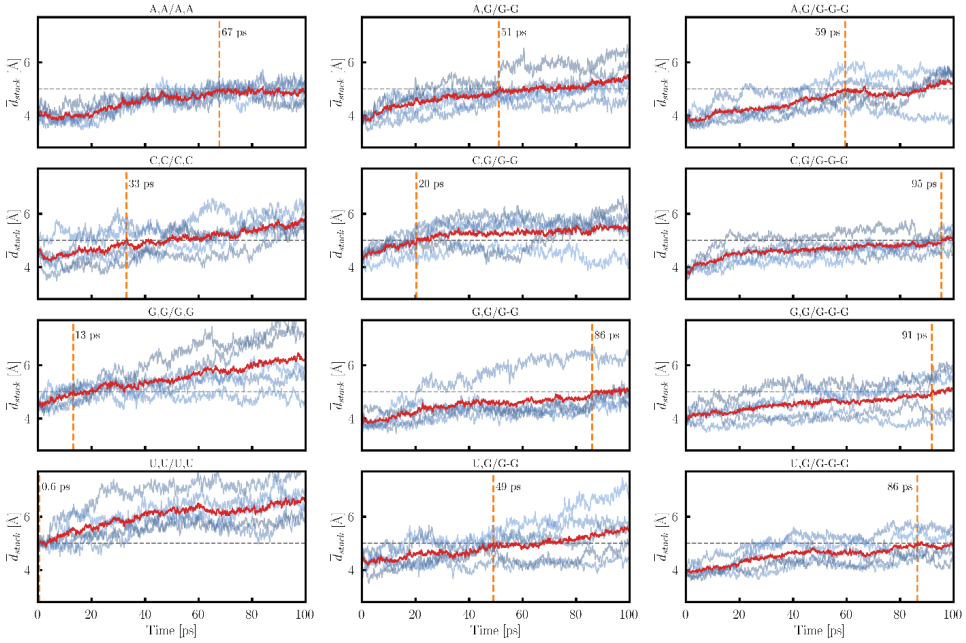
as in the static study, where the stabilization was significantly lower, the U,U/U,U stack appears to be the least stable. On average U,U/U,U disassembles after 0.6 ps. Therefore, uracil containing strands are not expected at short reaction timescales, as the stacked arrangement seems to be very unstable. S28 and 29 shows the results for the 3'-5' linked heterogeneous tetramers and pentamers. It can be seen that the poly-G matrices significantly stabilize the formation of intercalated stacked aggregates including uracil. The species including cytosine loose this arrangement directly and form a base paired alternative structure. This hydrogen bonded assembly then remains stable, which is indicated by the convergence of the measured interbase distances. The base paired arrangement is shown in S33.



S33. Alternative arrangements observed for the 3'-5' linked C, G/G-G-G group, where cytosine and guanine are base paired.

S27 shows the results for the 2'-5' linked heterogeneous systems in comparison to the N, N/N, N species. Other than for the 3'-5' linked analogues, all heterogeneous tetramers and pentamers form stable intercalated aggregates. Stabilization by a poly-G matrix, which might favor the formation of heterogeneous strands, is observed for C, G, and U. A, G/G-G and A, G/G-G-G on average exceed the unstacking threshold earlier than the A, A/A, A system. The A, A/A, A system remains stacked for 67 ps, A, G/G-G for 51 ps and A,G/G-G-G for 59 ps. Thus, although the time the systems remains in a compact stacked arrangement is not extended, all aggregates including A are relatively stable in comparison to the stacks including C, G, and U.

Overall, it appears that a poly-G matrix could enable the co-oligomerisation of cAMP, cCMP, and cUMP by stabilizing them in a stacked intercalated assembly, thus increasing the chances for oligomerisation reactions.



S34. Change in the average base distance during a GFN-FF dynamics run, for the N,N/N,N (left), 2'-5' linked N,G/G-G (middle) and N,G/G-G-G (right) systems. The five simulation runs are shown in blue, and their average is given in red. The time at which the mean crosses the 5.0 Å destacking threshold is marked in orange.

References

[1] F. Côté, D. Lévesque, J.-P. Perreault, *J. Virol.* **2001**, *75*, 19–25.

[2] A. Premstaller, P. J. Oefner, *LCGC Eur.* **2002**, 2–10.

[3] P. B. Danielson, R. Kristinsson, R. J. Shelton, G. S. LaBerge, *Expert Rev. Mol. Diagn.* **2005**, *5*, 53–63.

[4] A. Premstaller, P. J. Oefner, *Denaturing High-Performance Liquid Chromatography*, Humana Press, New Jersey, **n.d.**

[5] C. L. Wysoczynski, S. C. Roemer, V. Dostal, R. M. Barkley, M. E. A. Churchill, C. S. Malarkey, *Nucleic Acids Res.* **2013**, *41*, 1–10.

[6] M. C. Chambers, B. MacLean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz,

J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, P. Mallick, *Nat. Biotechnol.* **2012**, *30*, 918–920.

[7] R. Horst, A. L. Horwich, K. Wüthrich, *J. Am. Chem. Soc.* **2011**, *133*, 16354–16357.

[8] P. W. Kudella, A. V. Tkachenko, A. Salditt, S. Maslov, D. Braun, *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, DOI 10.1073/pnas.2018830118.

[9] T. E. Ouldridge, P. Šulc, F. Romano, J. P. K. Doye, A. A. Louis, *Nucleic Acids Res.* **2013**, *41*, 8886–8895.

[10] N. Srinivas, T. E. Ouldridge, P. Sulc, J. M. Schaeffer, B. Yurke, A. A. Louis, J. P. K. Doye, E. Winfree, *Nucleic Acids Res.* **2013**, *41*, 10641–10658.

[11] D. Y. Zhang, E. Winfree, *J. Am. Chem. Soc.* **2009**, *131*, 17303–17314.

[12] S. John, *Proc. Natl. Acad. Sci.* **1998**, *95*, 1460–1465.

[13] E. Edeleva, A. Salditt, J. Stamp, P. Schwintek, J. Boekhoven, D. Braun, *Chem. Sci.* **2019**, *10*, 5807–5814.

[14] J. E. Šponer, J. Šponer, A. Giorgi, E. Di Mauro, S. Pino, G. Costanzo, *J. Phys. Chem. B* **2015**, *119*, 2979–2989.

[15] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2015**, *11*, 918–922.

[16] J. Kussmann, C. Ochsenfeld, *J. Chem. Phys.* **2013**, *138*, DOI 10.1063/1.4796441.

[17] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2017**, *13*, 3153–3159.

[18] P. Sherwood, A. H. De Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjøvoll, A. Fahmi, A. Schäfer, C. Lennartz, *J. Mol. Struct. THEOCHEM* **2003**, *632*, 1–28.

[19] M. Cossi, N. Rega, G. Scalmani, V. Barone, *J. Comput. Chem.* **2003**, *24*, 669–681.

[20] N. Mardirossian, M. Head-Gordon, *J. Chem. Phys.* **2016**, *144*, DOI 10.1063/1.4952647.

[21] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[22] F. Weigend, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

[23] O. A. Vydrov, T. Van Voorhis, *J. Chem. Phys.* **2010**, *133*, DOI 10.1063/1.3521275.

[24] S. Spicher, S. Grimme, *Angew. Chemie* **2020**, *132*, 15795–15803.

[25] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*, 1–49.

[26] J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327–341.

[27] M. Shiga, M. Masia, *J. Chem. Phys.* **2013**, *139*, 44120.

[28] E. Baulin, V. Metelev, A. Bogdanov, *Nucleic Acids Res.* **2020**, *48*, 8675–8685.

4.5 Publication V: Quantitative Comparison of Experimental and Computed IR-Spectra Extracted from Ab Initio Molecular Dynamics

Beatriz von der Esch, Lena Sauerland, Laurens D. M. Peters, Christian Ochsenfeld
"Quantitative Comparison of Experimental and Computed IR-Spectra Extracted from Ab Initio Molecular Dynamics"
J. Chem. Theory Comput. **2021**, *17*, 985–995.

Abstract: Experimentally measured infrared spectra are often compared to their computed equivalents. However, the accordance is typically characterized by visual inspection, which is prone to subjective judgment. The primary challenge for a similarity-based analysis is that the artifacts introduced by each approach are very different and, therefore, may require pre-processing steps to determine and correct impeding irregularities. To allow for automated objective assessment, we propose a practical and comprehensive workflow involving scaling factors, a novel baseline correction scheme, and peak smoothing. The resulting spectra can then easily be compared quantitatively using similarity measures, for which we found the Pearson correlation coefficient to be the most suitable. The proposed procedure is then applied to compare the agreement of the experimental infrared spectra from the NIST Chemistry Web book with the calculated spectra using standard harmonic frequency analysis and spectra extracted from ab initio molecular dynamics simulations at different levels of theory. We conclude that the direct, quantitative comparison of calculated and measured IR spectra might become a novel, sophisticated approach to benchmark quantum-chemical methods. In the present benchmark, simulated spectra based on ab initio molecular dynamics show in general better agreement with the experiment than static calculations.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

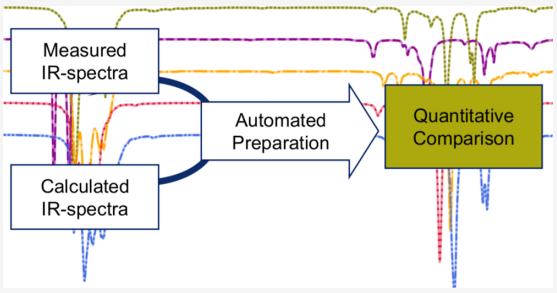
Quantitative Comparison of Experimental and Computed IR-Spectra Extracted from Ab Initio Molecular Dynamics

Beatriz von der Esch, Laurens D. M. Peters, Lena Sauerland, and Christian Ochsenfeld*

Cite This: J. Chem. Theory Comput. 2021, 17, 985−995 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Experimentally measured infrared spectra are often compared to their computed equivalents. However, the accordance is typically characterized by visual inspection, which is prone to subjective judgment. The primary challenge for a similarity-based analysis is that the artifacts introduced by each approach are very different and, therefore, may require preprocessing steps to determine and correct impeding irregularities. To allow for automated objective assessment, we propose a practical and comprehensive workflow involving scaling factors, a novel baseline correction scheme, and peak smoothing. The resulting spectra can then easily be compared quantitatively using similarity measures, for which we found the Pearson correlation coefficient to be the most suitable. The proposed procedure is then applied to compare the agreement of the experimental infrared spectra from the NIST Chemistry Web book with the calculated spectra using standard harmonic frequency analysis and spectra extracted from ab initio molecular dynamics simulations at different levels of theory. We conclude that the direct, quantitative comparison of calculated and measured IR spectra might become a novel, sophisticated approach to benchmark quantum-chemical methods. In the present benchmark, simulated spectra based on ab initio molecular dynamics show in general better agreement with the experiment than static calculations.



1. INTRODUCTION

Infrared (IR) spectroscopy is one of the most widely used analytical techniques for the qualitative and quantitative analyses of various materials, such as organic compounds, polymers, fibers, biomolecules, and even human tissue.[1−4] It is a powerful tool for the chemical characterization of substances, structure elucidation, characterization of surface processes, or monitoring chemical reactions.[2,3,5,6] In general, substances are IR-active if molecular vibrations, caused by irradiation with infrared light, lead to a change of the dipole moment. The resulting distinctive vibrations can be classified into valence and deformation vibrations.[7]

The prediction of IR spectra is an established field of research in theoretical chemistry and an important link to experimental work.[8−24] The quality of the calculated spectra is directly linked to the quality of the theoretical description, making their computation challenging. An adequate description of the entire potential energy surface (not only in direct proximity of the energy minimum) is decisive for reproducing the positions of the peaks in the absorbance spectrum. Additionally, a proper description of the electron density is needed to accurately predict peak intensities from the dipole moments.

Vibrational spectra are traditionally, and today still regularly, calculated from the second derivative of the energy with respect to the nuclear coordinates at a minimum energy structure employing the harmonic approximation. Several (more sophisticated) alternative approaches have been derived and applied in modern quantum chemistry.[8−26] One example is the calculation of IR spectra from ab initio molecular dynamics, which has several advantages over the static method, for example, the accounting for anharmonicity.[27−29] Besides the IR spectrum, Raman[28,30] and Vibrational Circular Dichroism spectra[31] can be extracted from ab initio molecular dynamics.

Even though a lot of effort has been made to increase the quality of the predicted spectra, there is no established technique to quantitatively compare the experimentally recorded and computed IR spectra. So far, calculated spectra are generally compared to measured spectra upon subjective visual agreement, or by the shift of fundamental frequencies. Only a few recent studies[32,33] performed a quantitative comparison of the computed and experimental spectra. The reason for this is that comparing experimental and theoretical

Received: December 10, 2020
Published: January 29, 2021

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

vibrational spectra is not straightforward. In experiments, vibrations can couple and overtones as well as difference and combination vibrations can appear in the spectrum.[7] Additionally, the sample preparation and possible contamination or degradation can influence the resulting experimental spectrum.[7,34,35]

Here, we propose a chemometric procedure to objectively compare the IR spectra. The procedure involves various preprocessing steps for both experimental and calculated spectra to remove artifacts and enable the following quantification of the similarity by the Pearson correlation coefficient. The main ideas of this approach are based on the established concepts within the field of analytical chemistry used for library searches.[34−37] The goal of a library search is to find the matching spectra within a library of known reference spectra.

We start with a brief introduction of IR spectra calculations, preprocessing procedures, and similarity measures in Section 2, followed by the computational details listed in Section 3. In Section 4, we test various similarity indicators using model functions and our preprocessing ansatz using the calculated and experimental IR-spectra of 16 representative molecules. The latter were obtained from the NIST Chemistry Web book.[38] Having validated our approach, we use it to compare the calculated spectra obtained from extended Lagrangian Born−Oppenheimer ab initio molecular dynamics simulations at different levels of theory and from static harmonic frequency calculations. Our conclusions are given in Section 5.

2. THEORY AND METHODS

2.1. Calculation of IR Spectra. Vibrational spectra can be computationally predicted using different approaches. The standard approach is to calculate vibrational harmonic normal modes.[16,28,39] This static method is built upon the assumption that the potential energy surface at an energy minimum can be approximated by a harmonic potential. To calculate harmonic frequencies, first the molecular structure has to be optimized. Subsequently, the second-order derivative of the potential energy (E) with respect to the nuclear coordinates (R) is calculated (eq 1) and mass weighted.

$$\boldsymbol{H} = \left(\frac{1}{\sqrt{m_A m_B}} \frac{\partial^2 E}{\partial R_A \partial R_B} \right)_{A,B=1,\ldots,N} \tag{1}$$

N denotes the number of atoms. Through diagonalization of the resulting mass-weighted Hessian ($\boldsymbol{H}$), the eigenvalues (ϵ_k) are obtained. From ϵ_k, the vibrational frequencies (ν_k) can be calculated

$$\nu_k = \frac{1}{2\pi c}\sqrt{\epsilon_k} \tag{2}$$

where c is the speed of light. The IR intensities are obtained from the change of the dipole moment along the respective normal mode vector.

Although this established harmonic approximation method is commonly applied, it neglects anharmonic effects, which can lead to large deviations between the calculated and experimental spectra.[40] Additionally, its application is limited: (1) because of the harmonic approximation, it cannot be used to monitor reactions and (2) no conditions such as temperature or specific solvent interaction by the inclusion of explicit solvent molecules can be accounted for straightforwardly. Furthermore, a major issue is the immense computational cost, which restricts the applicability to relatively small system sizes. Not only calculating the second derivative for extended systems becomes expensive (see, e.g., ref 41 for a reduced scaling approach), also the effort for finding minimum energy structures and the total number of minima increases significantly. To correctly determine the spectra for large systems, a frequency analysis for each minimum energy structure and subsequent scaling by Boltzmann weights becomes necessary. Finally, this approach only produces a discrete spectrum, therefore peak areas that might contain interesting information are missing.

A more sophisticated technique is to extract the IR-spectra from ab initio molecular dynamics (AIMD).[27−29,42,43] To do so, the autocorrelation function of the time derivative of the dipole moment (μ) has to be calculated and converted from the time domain to the frequency domain by Fourier transformation, which leads to the following proportionality of the IR intensity (I).

$$I(\omega) \propto \int \langle \dot{\mu}(\tau)\dot{\mu}(t+\tau)\rangle_\tau e^{-i\omega t} dt \tag{3}$$

This approach can be accelerated by using a fast Fourier transform (FFT) algorithm, which provides a significant enhancement especially for many time steps.[42] The calculation of IR spectra from AIMD simulations has some clear advantages. In contrast to vibrational frequency calculations, it requires only the first derivative of the dipole moment. Moreover, the harmonicity of the potential energy surface is not assumed, which leads to the occurrence of certain anharmonic effects.[29,40] However, not all anharmonic effects occurring during experimental IR spectroscopy, such as mode coupling or Fermi resonance, are observed.[28] Other than using the previously described static method, systems can be studied in the bulk phase and chemical reactions can be monitored. A further advantage is the continuity of the obtained spectra, which enables the analysis of the vibrational peak widths.

2.2. Spectra Preprocessing. To enable a comparison of experimental and theoretical spectra, artifacts need to be corrected in the experimental spectra, while the calculated spectra are usually rescaled. Additionally, all spectra need to be continuous and normalized with identical range and resolution. Common features found in the experimental spectra are low signal to noise ratios and elevated baselines. Therefore, baseline correction and smoothing algorithms may help to improve spectral quality and the agreement between theory and experiment. All aspects are briefly discussed in this section. If several experimental spectra are available, we suggest selecting the spectrum with the higher resolution or the least disturbances. The latter can be identified by the similarity of the original and preprocessed spectrum. This selection could, in principle, also be automatized.

2.2.1. Scaling Factors. Vibrational frequencies, computed using many quantum mechanical techniques, have to be scaled in order to minimize systematic errors because of an inaccurate description of electron correlation.[44−49] In this work, we determine scaling factors by applying all factors from 0.82 to 1.05 to our test set and calculate the mean of the similarity measure. The optimum scaling factor maximizes the mean score. Scaling factors are not only necessary to overcome theoretical limitations but can also be exploited as plausibility check. Similar to previous studies,[44−50] we use it to verify and validate our procedure (preprocessing and similarity measure) by comparing our determined scaling factors to previously

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

published ones. It has been suggested that multiple scaling factors[51,52] or mass-weighted scaling factors[33,53] can be employed to enhance the accordance of the experimental and computed spectra. We, therefore, also tested the effect of two scaling factors (Table 3). Two scaling factors were optimized for the high- and low-frequency domains. The regions were separated at 2200 cm^{-1} for practical reasons. In the area around 2200 cm^{-1}, fewer peaks are observed, thereby we avoid inconsistencies by the two scaling factors.

2.2.2. Generation of Continuous Spectra. From a normal mode analysis, one can extract peak positions and intensities. To compare the generated spectroscopic data to the measured spectra, either peaks from the experimental spectra have to be picked or the computed data have to be used to construct a continuous spectrum. Both methodologies introduce a bias, either by the choice of peaks or the estimated peak widths. Because generating a continuous spectrum can be more easily automated and subsequently subjected to the same comparison scheme, we apply Lorentzian curves (eq 4) with the corresponding amplitudes A, and with γ = 20 cm^{-1} at the computed peak positions $\tilde{\nu}_k$.

$$I(\tilde{\nu}) = \frac{1}{2\pi} \frac{A\gamma}{(\tilde{\nu} - \tilde{\nu}_k)^2 + (0.5\gamma)^2} \tag{4}$$

Henschel *et al.*[32] have tried to optimize the optimal full width at half-minimum (FWHM) in their recent work. However, they obtain clearly unphysical bandwidths. They reason that wider peaks always lead to higher Pearson correlation coefficients if the peak positions do not match well the experimental results. For one of their test cases, they obtain averaged values between 0.476 and 0.790 when the FWHM is set within a reasonable range.

This problem does not occur when the spectra have been extracted from molecular dynamics, as they are already continuous in this case.

2.2.3. Baseline Correction. Baseline effects occur regularly in the collected IR-spectra and complicate automated treatments. To computationally remove baselines, several methods are available. However, not all can be applied universally, as they require hyper-parameter optimization or supervision by the operator.[35] Because baselines vary greatly and manual removal is time-consuming as well as of limited reproducibility, an automatable methodology has to be found for quantitative analysis.[54,55] Here, we introduce a novel baseline correction involving only two parameters and three steps: (1) determining the areas of interest (where the peaks are located), (2) connect the areas of no interest via linear interpolation, and (3) remove the baseline from the entire spectrum under the condition that the new value is equal or lower than the original intensity. To obtain the region of interest, we calculate the derivative of the spectrum, normalize it, pick the points with values higher than a certain threshold (0.006), and add three neighboring data points (in both directions). The procedure is outlined in Figure 1.

2.2.4. Peak Smoothing. To remove peak splitting, which can arise in gas-phase IR spectroscopy and noise from the baseline corrected spectra, we employ the asymmetric-least-squares algorithm by Boelens et al.[56]

$$F = \sum_i w_i(s_i - z_i) + \lambda \sum_i (\Delta^2 z_i)^2 \tag{5}$$

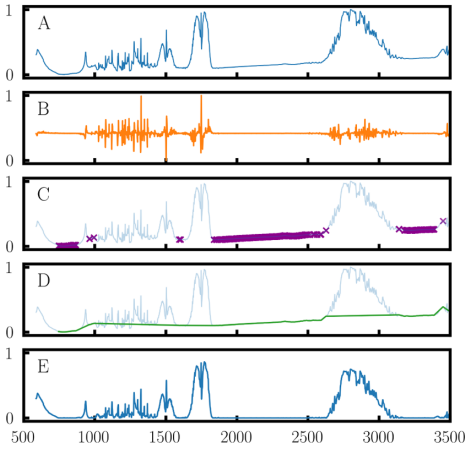


Figure 1. (A) Original IR-spectrum of formaldehyde. (B) First derivative of the spectrum. Data points where the derivative exceeds a certain threshold and points in direct proximity are excluded from the data for the baseline approximation. The remaining points are marked in magenta in plot C. The subsequent plot (D) shows the resulting baseline in green obtained from linear interpolation. Finally, the baseline-corrected spectrum is displayed (E).

where s is the collected spectrum, Δ is a difference operator, z is the baseline-corrected spectrum, w are the weights, which are set asymmetrically to p if $s_i > z_i$ and to $1 - p$ otherwise. The parameters λ and p determine the amount of smoothing and the magnitude of the weights.[54−57] The asymmetric-least-squares algorithm can also be used as the baseline correction algorithm. Here, we use it to smooth the experimental IR-spectra, setting λ and p to 5 and 0.01, respectively.

2.3. Similarity Measures. There are numerous measures for the degree of agreement of vibrational spectra. In the field of analytics, these measures are denoted "Hit Quality Index" (HQI) and are mostly employed for library searches based on the idea that each substance has a highly characteristic vibrational spectrum. They mostly range from zero to one (or 0−100%), where a higher HQI means that the compared samples are closer related.[34−36] For library searches, a similarity measure is needed that has a high detection efficiency and a low processing time because the goal is to identify a sample by comparison to a vast amount of reference spectra. These criteria are only subordinate for our application.[36] For our purpose, the comparison of different computational techniques, a slowly decreasing measure is favorable. This could in contrast lead to false positives during library searches.[58] Furthermore, we want our measure to be symmetric and to produce scores within a fixed range.

The various similarity indicators generally lead to non-identical results as they penalize dissimilarities differently. Widely applied measures are either based on distance metrics, peak picking and matching, or correlation analysis. Common distance functions are the Euclidean distance (ED) (eq 6), the Root Mean Square Deviation (RMSD) (eq 7), the Absolute Difference Value Search (ADV),[36] the Kullback-Leibler Divergence (KL),[59] the Jeffrey Divergence (JD)[60] (eq 8), or the Earth Mover Distance (EMD) (eq 9),[61,62] just to name a few.[34,35,37,63,64] From the field of the correlation analysis, the Pearson correlation coefficient (PCC) (eq 10) is widely used where s and r the two spectra and $\bar{s}$ and $\bar{r}$ their mean values.

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

$$M_{ED} = \left(1 + \sqrt{\sum_{i=1}^{n} |s_i - r_i|^2}\right)^{-1} \quad (6)$$

$$M_{RMSD} = \left(1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n} |s_i - r_i|^2}\right)^{-1} \quad (7)$$

$$M_{JD} = \left(1 + \sum_{i=1}^{n} s_i \log\frac{s_i}{r_i} + r_i \log\frac{r_i}{s_i}\right)^{-1} \quad (8)$$

$$M_{EMD} = 1 - \sqrt{\inf_{\pi\in\Pi(s,r)} \int c(x_0, x_1)\mathrm{d}\pi(x_0, x_1)} \quad (9)$$

$$M_{PCC} = \frac{\sum_{i=1}^{n}(s_i - \bar{s})(r_i - \bar{r})}{\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2 \sum_{i=1}^{n}(r_i - \bar{r})^2}} \quad (10)$$

Recently, the Pearson correlation coefficient and the Spearman correlation coefficient have been used by Henschel et al.[32] and Pracht et al.[33] to quantitatively compare the experimental spectra to the spectra obtained from normal mode analysis.

We apply the given measures to prototypical differences of spectra (peak shifting, peak broadening, etc.,) and compare them regarding the desired properties mentioned above. Many methods and variations of the measures given above used in analytical chemistry were discarded due to their nonsymmetrical character (e.g., Kullback–Leibler Divergence). Since most measures conduct a point by point comparison, the resolution and range of the spectra have to be adapted before comparison, so n is equal for the sample and reference.

3. COMPUTATIONAL DETAILS

If not stated otherwise, the spectra were prepared as described in the theory section, prior to the comparison. All calculations were conducted with the program package FermiONs++[41,65−67] in combination with the LibXC library v4.0.1.[68] For all DFT calculations, the gm5 grid was employed.[69] The SCF convergence criterion was set to 10^{-7} a.u. for the root mean square (RMS) of the FPS-commutator. In the case of GFN2-xTB, the convergence criteria were set to 10^{-6} a.u. for the energy and 10^{-4} a.u. for the RMS of the charges. The electronic temperature was set to 300 K. The experimental gas-phase IR spectra were retrieved from the NIST database.[38]

First, optimized geometries were generated for the molecules listed in Table 1. Each compound was optimized using the following methods: GFN2-xTB,[70,71] HF-3c[50]/minix, PBE[72]/def2-TZVP,[73−75] PBEh-3c[76]/def2-mSVP, and B3LYP[39,77−79]/def2-TZVP. The convergence criteria for the geometry optimization using GFN2-xTB were set to 10^{-6} atomic units (a.u.) for the energy, 3×10^{-4} a.u. for the average RMS force on all atoms, and the threshold for the rms displacement was set to 1.2×10^{-3}. The maximum force had to be lower than 4.5×10^{-4} and the maximum displacement below 1.8×10^{-3}. For all other methods, the convergence criteria for the geometry optimization were set to 10^{-6} a.u. for the energy, 10^{-4} a.u. for the gradient, and 10^{-1} a.u. for the displacement.

Table 1. List of Investigated Substances

acetic acid	acetonitrile	acetylen	ammonia
benzene	carbon dioxide	ethene	formaldehyde
methane	nitrous oxide	phosgene	sulfur dioxide
silicon tetrafluoride	tetrahydrofuran	thiophene	water

Subsequently, five ab initio Born–Oppenheimer molecular dynamics simulations with extended Lagrangian[80,81] were performed for each combination of substance and method varying the initial velocity. For PBE and B3LYP, the def2-SVP[73−75] basis set was employed, other than for the previous optimization. The Velocity Verlet propagator[82,83] was used for propagating the nuclei and a velocity rescaling thermostat[84] (298 K) was applied. This temperature was chosen to enhance the amount of structures sampled within the simulation time; however, it should be noted that the temperature at which the experimental spectra were recorded might deviate and that this can lead to dissimilarities. Each trajectory was 20 ps long using a step size of 0.1 fs. The necessary trajectory length was determined based on the convergence of the resulting spectra. Plots showing the convergence are provided in the Supporting Information. For extended, highly flexible systems, longer simulation times are recommended. IR spectra were calculated based on five trajectories according to eq 3 and the method described in ref 29.

Additional harmonic frequency calculations were carried out at the B3LYP-D3[39,77−79,85,86]/def2-SVP level of theory based on the previously optimized geometries.

4. RESULTS AND DISCUSSION

4.1. Response of Similarity Measures to Manipulations of Prototypical Spectra. A suitable similarity measure for the comparison of experimental and theoretical spectra, computational benchmarks, or convergence studies should (1) be resistant to minor irregularities such as noise, (2) correctly reflect changes such as peak shifts, broadening, and relative intensities, (3) decrease slowly so minor differences can be reflected and comparisons are still possible if there are some disagreements, and (4) ideally be insensitive to the normalization technique. "To select a suitable similarity measure, dierent manipulations are carried out on Lorentzian model functions and the resulting eects on the PCC, ED, RMSD, JD, and EMD are investigated (Figure 2). Derivative based methods are not considered because these react even more sensitively toward dissimilarities.

The JD is not a suitable similarity indicator as it does not fulfill any of the criteria mentioned above. Only the signal broadening with subsequent normalization by the standard deviation affects the measure.

The distance-based methods reflect all changes. They only decrease slightly when the entire function is translated (Figure 2c), in all other cases, the ED decreases very rapidly. Overall, the RMSD decreases more slowly than the ED and, therefore, seems to be a more suitable similarity measure.

The EMD reacts quite similar to the RMSD when functions are broadened and intensities are manipulated (Figure 2a,b,e,f). Its response to peak shifts is too moderate, both, when the entire spectrum is translated (Figure 2c) as well as when only a single peak is manipulated (Figure 2d).

The PCC is indifferent to changes of magnitude when the entire "spectrum" is affected (Figure 2a), therefore it is not influenced by the normalization method. When only one of the peak intensities is changed, the measure decreases slowly (Figure 2b). The homogeneous shift (Figure 2c) also leads to a

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

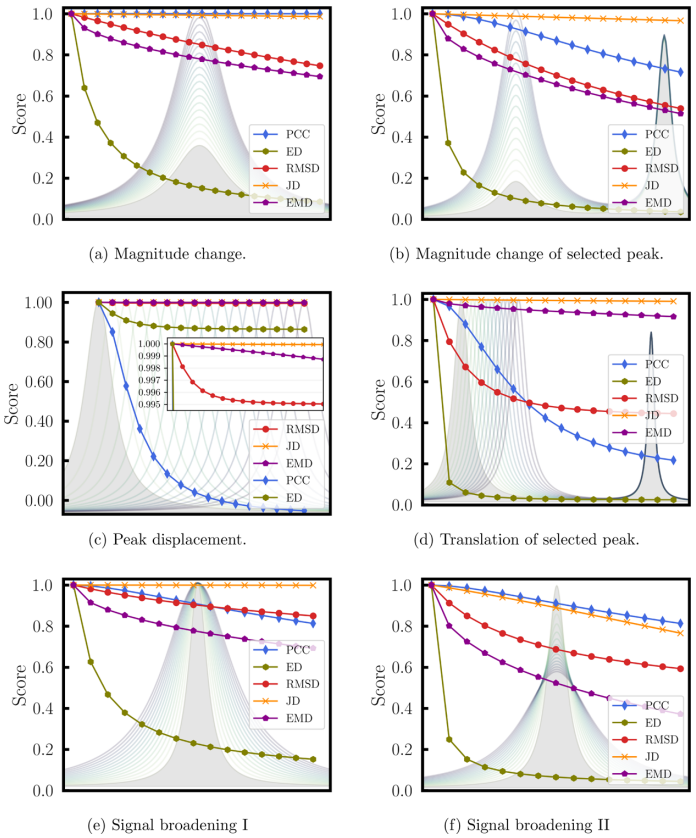


Figure 2. Visualization of various manipulations and the response of the PCC, ED, RMSD, and EMD. (a,b) Effect of intensity change of the entire reference function and a selected peak, both without subsequent normalization. (c,d) Effect of gradually increasing peak displacement on the similarity measures. (d) Only one of the peaks is shifted. (e,f) Effect of peak broadening while normalizing the functions from zero to one and normalizing by the standard deviation, respectively.

slow decrease. In contrast to all other measures, the PCC reaches zero. The translation of a selected peak reduced the PCC (Figure 2d), however, it does not reach a plateau as fast as the RMSD and does not reach zero as the ED and, therefore, represents the displacement best. Both changes of the peak areas are reflected by the PCC (Figure 2e,f) as the measure decreases steadily with enhancing disagreement.

Because the PCC is additionally robust toward noise, as the peaks have a greater influence on the resulting score, we conclude that all criteria are met. However, it should be noted that negative values can be obtained. These indicate an anticorrelation, for spectra this can be interpreted as very dissimilar.

4.2. Validation. First, we investigate the effect of scaling factors on the mean similarity score M_{PCC} for the different computational methods, obtained when comparing the computed spectra to the experimental spectra. Figure 3 shows their dependence, while the scaling factors that maximize M_{PCC} are listed in Table 2 together with factors determined in previous studies.[44,47,50,87,88]

The good agreement of the determined factors with those obtained by previous studies (relying on other methodologies) is evidence that the proposed workflow produces reasonable results. Another interesting finding is that while for B3LYP and

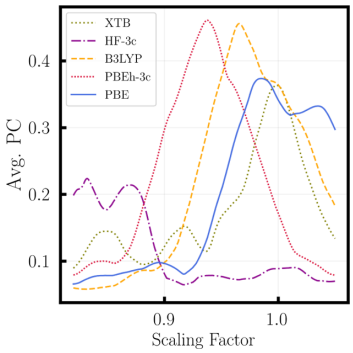


Figure 3. Visualization of the scaling factor search, where values from 0.82 and 1.05 were tested and the mean similarity indicator were evaluated.

PBEh-3c, sharp peaks are obtained, for PBE, GFN2-xTB, and HF-3c, multiple maxima are found (see Figure 3). For HF-3c, the two maxima have almost identical mean similarity scores. Multiple maxima might indicate that either the optimal scaling factors differ for high and low frequency vibrations or for the examined compounds. Therefore, not the same systematic

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

Table 2. Method-Dependent Scaling Factors for Vibrational Frequencies Obtained by Maximizing the Mean Similarity Score

	B3LYP	PBEh-3c	PBE	HF-3c	GFN2-xTB
scaling factors	0.966	0.938	0.985	0.832	0.999
reference values	0.961[47]	0.950[76]	0.990[47]	0.860[50]	1.000[71]

error is effecting all peaks. The method-dependent scaling factors for the high- and low-frequency vibrations obtained by maximizing the mean similarity score are given in Table 3. Overall, a higher accordance between the experimental and computational spectra was obtained when using two factors, therefore we decided to proceed with two scaling factors.

Table 3. Method-Dependent Scaling Factors for the High- and Low-Frequency Vibrations Obtained by Maximizing the Mean Similarity Score

scaling factors (cm^{-1})	B3LYP	PBEh-3c	PBE	HF-3c	GFN2-xTB
>2200	0.97	0.94	0.99	0.83	1.00
<2200	1.00	0.94	1.04	0.85	0.99

As further validation of the introduced similarity measure, a cross check is conducted. The experimental spectra of acetic acid, acetonitrile, acetylene, ammonia, benzene, carbon dioxide, ethene, formaldehyde, and nitrous oxide are compared to each other. Figure 4 (left) shows the cross correlation that is obtained if only the resolution of the two spectra is adapted, while Figure 4 (middle) is determined from spectra, which were additionally baseline corrected and smoothed. Figure 4 (right) shows the differences between the two plots.

Figure 4 shows that two identical spectra (diagonal elements) yield a similarity score of one. The off diagonal elements show the determined similarity scores for the comparison of the spectra of different compounds, which is symmetric. The cross correlation with preprocessing exhibits less negative scores and more values close to zero than the one without preprocessing. On the other hand, some off-diagonal elements increase (up to 0.49). All elevated scores can, however, be explained by comparing the spectra visually. Acetic acid (1) and formaldehyde (8) feature a very prominent C−O stretch vibration, which is perfectly aligned, as well as similar C−H vibrations and an overlapping signal at about 1200 cm^{-1}. The IR spectra of ethene (7) and benzene (5), acetonitrile (2), and ammonia (4) also show a significant overlap. The corresponding spectra are compared in the Supporting Information.

The elevated scores of off-diagonal elements, however, highlight that the rather tolerant PCC might produce many false positives when applied for library searches. Obtained matches should, therefore, always be reviewed. This, in our opinion, however, does not affect the suitability of the PCC measure for benchmarks of computational methods because only gradual changes are expected.

To complete the validation of the proposed workflow, we now check the agreement between the visual rating and the similarity indicator. In Figure 5, spectra calculated from a harmonic frequency analysis, extracted from AIMD-simulations, and experimental equivalents are compared. Shown are the spectra of acetylene, benzene, thiophene, and tetrahydrofuran. The remaining spectra can be found in the Supporting Information. The determined similarity scores are given in Table 4.

For acetylene (Figure 5a), scores between −0.047 and 0.699 are determined. PBE scores highest with 0.699, closely followed by B3LYP and PBEh-3c with 0.681 and 0.664, respectively. The spectra include all recorded signals with their position being in very good agreement with the experiment. The spectra calculated with HF-3c and GFN2-xTB show less agreement. The peak positions as well as the relative intensities are not reproduced accurately. The harmonic frequencies ($B3LYP^h$ in Figure 5a) are partly matching. The prominent C−H vibration is represented well, however, the vibration at around 1300 cm^{-1} is not found and the signal at 3400 cm^{-1} is shifted significantly, while its relative intensity is overestimated. For acetylene, the determined similarity scores (see Table 4) are in good agreement with the visual rating.

For benzene (Figure 5b), all spectra extracted from dynamics agree relatively well with the experimental work. This is adequately represented by the similarity measures. All values lie above 0.492. GFN2-xTB has the highest similarity score (0.764), however, it does not produce the peaks of low intensities between 1000 and 2000 cm^{-1}. This insufficiency only has a minor impact on the Pearson correlation coefficient. The DFT-functionals produce very similar spectra, they only vary in the peak areas of the minor signals, which are missing in the GFN2-xTB spectrum. In general, the peak positions of the spectra calculated from dynamics are in remarkably good agreement with the experimental spectrum. However, some of

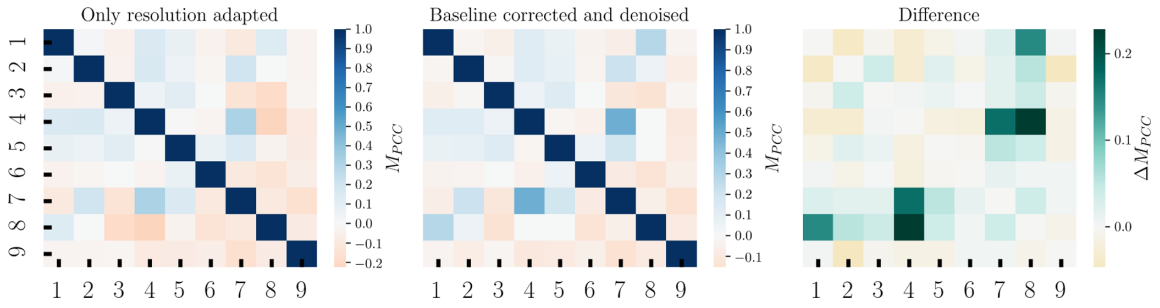


Figure 4. Cross check of the similarity score M_{PCC}, based on the experimental spectra of CH_3COOH (1), CH_3CN (2), C_2H_2 (3), NH_3 (4), C_6H_6 (5), CO_2 (6), C_2H_4 (7), CH_2O (8), and N_2O (9). On the left, the experimental spectra were only resolution adapted, while in the middle the spectra were additionally baseline corrected and denoised. The right heat-map highlights the change in scores. It should be noted that the color scale differs from the other two plots.

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

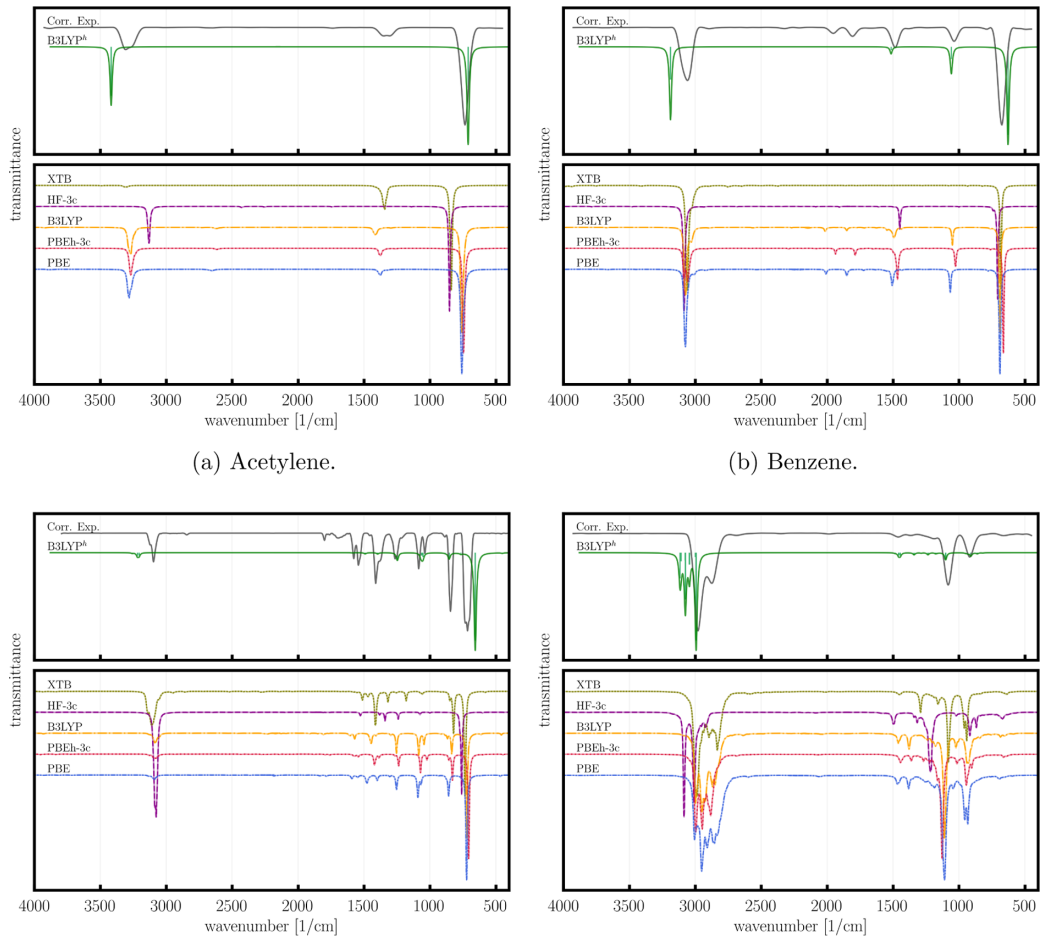


Figure 5. Comparison of the calculated spectra by means of harmonic frequency analysis (denoted B3LYP^h, upper panel), from AIMD-simulations (lower panel), and experimental reference spectra retrieved from the NIST Chemistry Web book.[38] As examples, the spectra of (a) acetylene, (b) benzene, (c) thiophene, and (d) tetrahydrofuran are shown.

Table 4. M_{PCC} Scores Obtained by Comparison of the Calculated Spectra of Aetylene, Benzene, Thiophene, and Tetrahydrofuran With Their Measured Equivalents[a]

	C_2H_2	C_6H_6	C_4H_4S	C_4H_8O
GFN2-xTB	0.014	0.764	0.592	0.876
HF-3c	−0.047	0.492	0.118	0.276
B3LYP	0.681	0.736	0.658	0.860
PBEh-3c	0.664	0.738	0.684	0.769
PBE	0.699	0.712	0.653	0.855
B3LYP^h	0.293	−0.055	0.010	0.396

[a]B3LYP^h denotes the results obtained from harmonic frequency calculations at the B3LYP-D3/def2-SVP level of theory.

the minor signals are not always reproduced. The harmonic frequency analysis gives four IR-active frequencies. The two most prominent peaks are slightly shifted. Again the visual judgment and rating go hand in hand.

The highest similarity score obtained for thiophene is 0.684 with PBEh-3c. As can be seen in Figure 5c, most of the peak positions, as well as peak areas, are represented accurately. Nevertheless, the signal around 3100 cm^{-1} is too weak. This inadequacy is shared with the spectra calculated using PBE and B3LYP. PBE and B3LYP only score marginally lower than PBEh-3c. GFN2-xTB has a score of 0.592. Among the dynamics-based spectra, HF-3c matched the experimental results least, as indicated by the similarity score of 0.118. HF-3c overestimates the relative intensity of the characteristic C−H stretch vibration. All methods underestimate the intensities of the peaks between 1500 and 1800 cm^{-1}. The harmonic frequencies (again) have the poorest agreement with the experimental IR-spectrum, with a low score of 0.010.

Finally, for tetrahydrofuran (Figure 5d), all spectra calculated from dynamics using DFT are in good agreement with the experimental spectra. GFN2-xTB yields the highest score, 0.860. Especially, the peak widths are reflected accurately. These observations agree with the corresponding scores. The agreement of HF-3c and B3LYP^h is still acceptable. For tetrahydrofuran, a slightly higher correspondence between

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

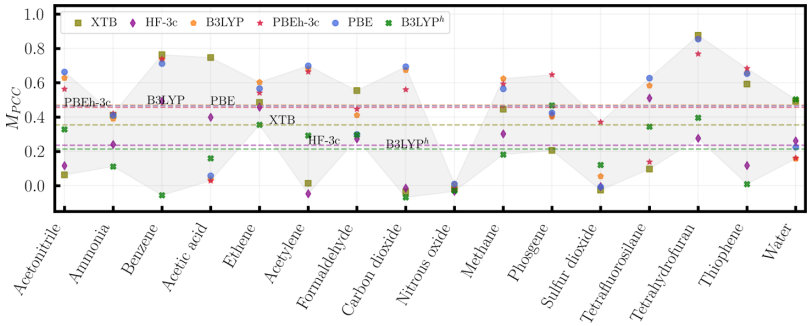


Figure 6. M_{PCC} scores per compound colored using the computational method. The mean scores for each method are marked by a horizontal line. B3LYP^h denotes the results obtained from harmonic frequency calculations at the B3LYP-D3/def2-SVP level of theory. All results were scaled using two scaling factors.

B3LYP^h and the experimental spectrum is indicated than for HF-3c.

In summary, the applied similarity scores reflect the relation between spectra correctly. However, it should be noted that especially the displacement of peaks leads to a strong lowering of the score. The comparison of visual judgment and quantitative rating also shows that scores around 0.5 already indicate good agreement.

4.3. Method Comparison. Figure 6 shows the obtained M_{PCC} scores for each combination of method and compound, and the corresponding averages. The plot clearly shows that the tested DFT functionals perform best, the accordance between the computed spectra and all three methods is fairly equal. The average score of B3LYP is 0.470, for PBE 0.467, and for PBEh-3c 0.459. However, the differences in the similarity indicator are too little to rank these methods. Surprisingly, HF-3c scores far worse and is even surpassed by the much less time-consuming, semi-empirical tight binding method, GFN2-xTB. Furthermore, we were not able to obtain a stable HF-3c dynamics simulation for N_2O, which is probably because of a poor description of the relatively challenging electronic structure. The spectra obtained from harmonic frequency analysis (B3LYP-D3/def2-SVP) are (on average) slightly worse than those extracted from dynamics at the HF-3c/minix level of theory. This might be due to the lack of anharmonicity in the second derivative-based ansatz and clearly shows the superiority of dynamics, when it comes to reproducing the spectroscopic data.

In cases where collectively all methods achieve low scores, this can also indicate a low quality of the reference spectrum or the occurrence of several irregularities. For example, in the experimental spectrum of N_2O, several overtones appear. However, automated detection and elimination have not yet been resolved. These so far unavoidable experimental effects also limit the possible agreement of the calculated and measured spectra. Hence, scores above 0.75 are very unlikely and scores above 0.5 already indicate a very good agreement.

An equivalent of Figure 6 where only one scaling factor is used is given in the Supporting Information.

Figure 7 shows the similarity scores grouped by methods. The plot visualizes the minimum and maximum values, the upper and lower quartile, the median (orange line), and the arithmetic mean (red dotted line) achieved by each method. Additionally, the underlying data are shown by blue dots. The

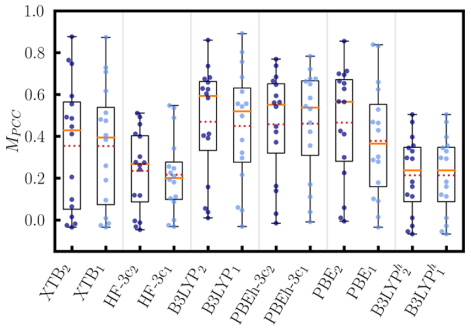


Figure 7. Box-plot that summarizes the achieved M_{PCC} grouped by methods. The medians are marked by an orange line and the arithmetic means by red-dashed lines. B3LYP^h denotes the results obtained from harmonic frequency calculations at the B3LYP-D3/def2-SVP level of theory. The subscripts "1" and "2" denote whether one or two scaling factors were employed. The underlying data are shown by additional blue dots.

plot shows the results obtained when using one or two scaling factors, this is denoted by the subscript.

The use of two different scaling factors for the low and high-frequency domain has the greatest effect on the spectra calculated with HF-3c and PBE. For GFN2-xTB, PBEh-3c, and the normal modes (B3LYP^h), the effect is negligible. When only one scaling factor is employed, PBEh-3c/def2-mSVP produces the qualitatively best spectra out of the tested methods followed by B3LYP/def2-SVP and PBE/def2-SVP. This order changes and the differences are diminished when a second scaling factor is introduced. PBE and B3LYP surpass PBEh-3c. However, the scores of PBEh-3c are less spread.

GFN2-xTB shows the highest spread of values, which is probably because of its semi-empirical, simplistic nature. As mentioned above, on average, GFN2-xTB produces much more accurate spectra than HF-3c, which makes it an attractive time-efficient alternative to the tested DFT methods. All methods produce at least one spectrum that deviates significantly from the experimental reference spectrum. This means that all of the methods are far away from being black-box approaches to reproduce the spectroscopic data.

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

5. CONCLUSIONS

Strong links between theoretical and experimental studies are essential for both fields. Here, the simulation of vibrational frequencies is an important bridge. However, the agreement of computational and experimental spectra has so far mainly been assessed visually. Therefore, we examined the application of the Pearson correlation coefficient (PCC) as an indicator of similarity to enable an objective, quantitative evaluation. Furthermore, quantitative similarity measures can be used to assess the performance of different theoretical methods, for convergence studies and compound verification. In the context of automated comparison of spectra, we presented a preprocessing procedure to reduce impeding effects, which comprises range and resolution adaption, a novel baseline correction, noise reduction, and scaling of spectra. The proposed methods were successfully used to illustratively compare the measured spectra of 16 chemical compounds and their computed equivalents obtained from either AIMD simulations or harmonic frequency analysis calculations at different levels of theory. It was shown that spectra calculated from AIMD simulations are significantly closer to the experimental data. However, the high cost of ab initio dynamics is currently still limiting the applicability. While the performance of ab initio dynamics is continuously increasing, also other approaches[89−92] like machine learning-assisted dynamics and force field developments might be helpful, so that we are convinced that the computation of spectroscopic properties from dynamics will become more and more feasible in the near future.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c01279.

Study of the convergence of spectra; underlying data of cross check; visualization of IR-spectra overlap of acetonitrile, ammonia, benzene, and ethene; all calculated and experimental reference spectra; and M_{PCC} scores per compound using one scaling factor (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Christian Ochsenfeld − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0002-4189-6558; Email: christian.ochsenfeld@uni-muenchen.de

Authors

Beatriz von der Esch − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0002-8366-5272

Laurens D. M. Peters − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0001-6572-8738

Lena Sauerland − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c01279

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 364653263—TRR 235 "Emergence of Life". C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

■ REFERENCES

(1) Lang, P. L.; Katon, J. E.; O'Keefe, J. F.; Schiering, D. W. The identification of fibers by infrared and Raman microspectroscopy. *Microchem. J.* **1986**, *34*, 319−331.

(2) Steiner, G.; Tunc, S.; Maitz, M.; Salzer, R. Conformational Changes during Protein Adsorption. FT-IR Spectroscopic Imaging of Adsorbed Fibrinogen Layers. *Anal. Chem.* **2007**, *79*, 1311−1316.

(3) Mudunkotuwa, I. A.; Minshid, A. A.; Grassian, V. H. ATR-FTIR spectroscopy as a tool to probe surface adsorption on nanoparticles at the liquid-solid interface in environmentally and biologically relevant media. *Analyst* **2014**, *139*, 870−881.

(4) Diem, M.; Ergin, A.; Remiszewski, S.; Mu, X.; Akalin, A.; Raz, D. Infrared micro-spectroscopy of human tissue: principles and future promises. *Faraday Discuss.* **2016**, *187*, 9−42.

(5) Brudler, R.; Rammelsberg, R.; Woo, T. T.; Getzoff, E. D.; Gerwert, K. Structure of the I_1 early intermediate of photoactive yellow protein by FTIR spectroscopy. *Nat. Struct. Biol.* **2001**, *8*, 265−270.

(6) Violet, L.; Mifleur, A.; Vanoye, L.; Nguyen, D. H.; Favre-Réguillon, A.; Philippe, R.; Gauvin, R. M.; Fongarland, P. Online monitoring by infrared spectroscopy using multivariate analysis - background theory and application to catalytic dehydrogenative coupling of butanol to butyl butyrate. *React. Chem. Eng.* **2019**, *4*, 909−918.

(7) Fleming, I.; Williams, D. *Spectroscopic Methods in Organic Chemistry*; Springer: Cham, 2019; pp 85−121.

(8) Pawłowski, F.; Halkier, A.; Jørgensen, P.; Bak, K. L.; Helgaker, T.; Klopper, W. Accuracy of spectroscopic constants of diatomic molecules from ab initio calculations. *J. Chem. Phys.* **2003**, *118*, 2539−2549.

(9) Neugebauer, J.; Hess, B. A. Fundamental vibrational frequencies of small polyatomic molecules from density-functional calculations and vibrational perturbation theory. *J. Chem. Phys.* **2003**, *118*, 7215−7225.

(10) Neugebauer, J.; Reiher, M. Mode tracking of preselected vibrations of one-dimensional molecular wires. *J. Phys. Chem. A* **2004**, *108*, 2053−2061.

(11) Dierksen, M.; Grimme, S. Density functional calculations of the vibronic structure of electronic absorption spectra. *J. Chem. Phys.* **2004**, *120*, 3544−3554.

(12) Dierksen, M.; Grimme, S. The vibronic structure of electronic absorption spectra of large molecules: A time-dependent density functional study on the influence of "Exact" Hartree-Fock exchange. *J. Phys. Chem. A* **2004**, *108*, 10225−10237.

(13) Mroginski, M. A.; Mark, F.; Thiel, W.; Hildebrandt, P. Quantum mechanics/molecular mechanics calculation of the Raman spectra of the phycocyanobilin chromophore in α-C-phycocyanin. *Biophys. J.* **2007**, *93*, 1885−1894.

(14) Hrenar, T.; Werner, H.-J.; Rauhut, G. Accurate calculation of anharmonic vibrational frequencies of medium sized molecules using local coupled cluster methods. *J. Chem. Phys.* **2007**, *126*, 134108.

(15) Petrenko, T.; Neese, F. Analysis and prediction of absorption band shapes, fluorescence band shapes, resonance Raman intensities, and excitation profiles using the time-dependent theory of electronic spectroscopy. *J. Chem. Phys.* **2007**, *127*, 164319.

(16) Rauhut, G.; Knizia, G.; Werner, H.-J. Accurate calculation of vibrational frequencies using explicitly correlated coupled-cluster theory. *J. Chem. Phys.* **2009**, *130*, 054105.

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

(17) Luber, S.; Neugebauer, J.; Reiher, M. Intensity tracking for theoretical infrared spectroscopy of large molecules. *J. Chem. Phys.* **2009**, *130*, 064105.
(18) Neese, F. Prediction of molecular properties and molecular spectroscopy with density functional theory: From fundamental theory to exchange-coupling. *Coord. Chem. Rev.* **2009**, *253*, 526−563.
(19) Heislbetz, S.; Rauhut, G. Vibrational multiconfiguration self-consistent field theory: Implementation and test calculations. *J. Chem. Phys.* **2010**, *132*, 124102.
(20) Rodriguez-Betancourtt, V.-M.; Quezada-Navarro, V.-M.; Neff, M.; Rauhut, G. Anharmonic frequencies of [F,C,N,X] isomers (X=O,S) obtained from explicitly correlated coupled-cluster calculations. *Chem. Phys.* **2011**, *387*, 1−4.
(21) Pfeiffer, F.; Rauhut, G. Anharmonic Frequencies of CX2Y2 (X, Y = O, N, F, H, D) Isomers and Related Systems Obtained from Vibrational Multiconfiguration Self-Consistent Field Theory. *J. Phys. Chem. A* **2011**, *115*, 11050−11056.
(22) Bieler, N. S.; Haag, M. P.; Jacob, C. R.; Reiher, M. Analysis of the Cartesian Tensor Transfer Method for Calculating Vibrational Spectra of Polypeptides. *J. Chem. Theory Comput.* **2011**, *7*, 1867−1881.
(23) Weymuth, T.; Haag, M. P.; Kiewisch, K.; Luber, S.; Schenk, S.; Jacob, C. R.; Herrmann, C.; Neugebauer, J.; Reiher, M. MOVIPAC: Vibrational spectroscopy with a robust meta-program for massively parallel standard and inverse calculations. *J. Comput. Chem.* **2012**, *33*, 2186−2198.
(24) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.
(25) Bloino, J.; Barone, V. A second-order perturbation theory route to vibrational averages and transition properties of molecules: General formulation and application to infrared and vibrational circular dichroism spectroscopies. *J. Chem. Phys.* **2012**, *136*, 124108.
(26) Carnimeo, I.; Puzzarini, C.; Tasinato, N.; Stoppa, P.; Charmet, A. P.; Biczysko, M.; Cappelli, C.; Barone, V. Anharmonic theoretical simulations of infrared spectra of halogenated organic compounds. *J. Chem. Phys.* **2013**, *139*, 074310.
(27) Gaigeot, M.-P.; Sprik, M. Ab Initio Molecular Dynamics Computation of the Infrared Spectrum of Aqueous Uracil. *J. Phys. Chem. B* **2003**, *107*, 10344−10358.
(28) Thomas, M.; Brehm, M.; Fligg, R.; Vöhringer, P.; Kirchner, B. Computing vibrational spectra from ab initio molecular dynamics. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6608−6622.
(29) Peters, L. D. M.; Kussmann, J.; Ochsenfeld, C. Efficient and Accurate Born-Oppenheimer Molecular Dynamics for Large Molecular Systems. *J. Chem. Theory Comput.* **2017**, *13*, 5479−5485.
(30) Thomas, M.; Brehm, M.; Kirchner, B. Voronoi dipole moments for the simulation of bulk phase vibrational spectra. *Phys. Chem. Chem. Phys.* **2015**, *17*, 3207−3213.
(31) Thomas, M.; Kirchner, B. Classical Magnetic Dipole Moments for the Simulation of Vibrational Circular Dichroism by ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2016**, *7*, 509−513.
(32) Henschel, H.; Andersson, A. T.; Jespers, W.; Mehdi Ghahremanpour, M.; van der Spoel, D. Theoretical Infrared Spectra: Quantitative Similarity Measures and Force Fields. *J. Chem. Theory Comput.* **2020**, *16*, 3307−3315.
(33) Pracht, P.; Grant, D. F.; Grimme, S. Comprehensive Assessment of GFN Tight-Binding and Composite Density Functional Theory Methods for Calculating Gas-Phase Infrared Spectra. *J. Chem. Theory Comput.* **2020**, *16*, 7044−7060.
(34) Renner, G.; Schmidt, T. C.; Schram, J. A New Chemometric Approach for Automatic Identification of Microplastics from Environmental Compartments Based on FT-IR Spectroscopy. *Anal. Chem.* **2017**, *89*, 12045−12053.
(35) Renner, G.; Nellessen, A.; Schwiers, A.; Wenzel, M.; Schmidt, T. C.; Schram, J. Data preprocessing & evaluation used in the microplastics identification process: A critical review & practical guide. *TrAC, Trends Anal. Chem.* **2019**, *111*, 229−238.
(36) Kwiatkowski, A.; Smulko, J.; Gnyba, M.; Wierzba, P. Algorithms of Chemicals Detection Using Raman Spectra. *Metrol. Meas. Syst.* **2010**, *17*, 549−559.
(37) Gautam, R.; Vanga, S.; Ariese, F.; Umapathy, S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *Eur. Phys. J. Tech. Instrum.* **2015**, *2*, 2195−7045.
(38) Wallace, W. E. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg MD; Vol. *20899*, (retrieved July 27, 2020); Chapter "Infrared Spectra" by NIST Mass Spectrometry Data Center.
(39) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(40) Kumarasiri, M.; Swalina, C.; Hammes-Schiffer, S. Anharmonic effects in ammonium nitrate and hydroxylammonium nitrate clusters. *J. Phys. Chem. B* **2007**, *111*, 4653−4658.
(41) Kussmann, J.; Luenser, A.; Beer, M.; Ochsenfeld, C. A reduced-scaling density matrix-based method for the computation of the vibrational Hessian matrix at the self-consistent field level. *J. Chem. Phys.* **2015**, *142*, 094101.
(42) Futrelle, R. P.; McGinty, D. J. Calculation of spectra and correlation functions from molecular dynamics data using the fast Fourier transform. *Chem. Phys. Lett.* **1971**, *12*, 285−287.
(43) Kirchner, B.; di Dio, P. J.; Hutter, J. *Multiscale Molecular Methods in Applied Chemistry*; Springer: Berlin, Heidelberg, 2011; Vol. *307*; pp 109−154.
(44) Scott, A. P.; Radom, L. Harmonic vibrational frequencies: An evaluation of Hartree-Fock, Møller-Plesset, quadratic configuration interaction, density functional theory, and semiempirical scale factors. *J. Phys. Chem.* **1996**, *100*, 16502−16513.
(45) Irikura, K. K.; Johnson, R. D.; Kacker, R. N. Uncertainties in scaling factors for ab initio vibrational frequencies. *J. Phys. Chem. A* **2005**, *109*, 8430−8437.
(46) Tantirungrotechai, Y.; Phanasant, K.; Roddecha, S.; Surawatanawong, P.; Sutthikhum, V.; Limtrakul, J. Scaling factors for vibrational frequencies and zero-point vibrational energies of some recently developed exchange-correlation functionals. *J. Mol. Struct.* **2006**, *760*, 189−192.
(47) Merrick, J. P.; Moran, D.; Radom, L. An evaluation of harmonic vibrational frequency scale factors. *J. Phys. Chem. A* **2007**, *111*, 11683−11700.
(48) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational thermochemistry: Scale factor databases and scale factors for vibrational frequencies obtained from electronic model chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872−2887.
(49) Borowski, P. An evaluation of scaling factors for multiparameter scaling procedures based on DFT force fields. *J. Phys. Chem. A* **2012**, *116*, 3866−3880.
(50) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672−1685.
(51) Laury, M. L.; Carlson, M. J.; Wilson, A. K. Vibrational frequency scale factors for density functional theory and the polarization consistent basis sets. *J. Comput. Chem.* **2012**, *33*, 2380−2387.
(52) Bouteiller, Y.; Gillet, J.-C.; Grégoire, G.; Schermann, J. P. Transferable Specific Scaling Factors for Interpretation of Infrared Spectra of Biomolecules from Density Functional Theory. *J. Phys. Chem. A* **2008**, *112*, 11656−11660.
(53) Irikura, K. K. Mass scaling for vibrational frequencies from ab initio calculations. *Chem. Phys. Lett.* **2005**, *403*, 275−279.
(54) Liland, K. H.; Almøy, T.; Mevik, B.-H. Optimal choice of baseline correction for multivariate calibration of spectra. *Appl. Spectrosc.* **2010**, *64*, 1007−1016.
(55) Peng, J.; Peng, S.; Jiang, A.; Wei, J.; Li, C.; Tan, J. Asymmetric least squares for multiple spectra baseline correction. *Anal. Chim. Acta* **2010**, *683*, 63−68.

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

(56) Boelens, H. F. M.; Dijkstra, R. J.; Eilers, P. H. C.; Fitzpatrick, F.; Westerhuis, J. A. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *J. Chromatogr. A* **2004**, *1057*, 21−30.
(57) Eilers, P. H. C. A perfect smoother. *Anal. Chem.* **2003**, *75*, 3631−3636.
(58) Henschel, H.; van der Spoel, D. An Intuitively Understandable Quality Measure for Theoretical Vibrational Spectra. *J. Phys. Chem. Lett.* **2020**, *11*, 5471−5475.
(59) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(60) Harold, J. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. London, Ser. A* **1946**, *186*, 453−461.
(61) Rubner, Y.; Tomasi, C.; Guibas, L. J. Earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99−121.
(62) Villani, C. *Optimal Transport: Old and New*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp 93−111.
(63) Deborah, H.; Richard, N.; Hardeberg, J. Y. A Comprehensive Evaluation of Spectral Distance Functions and Metrics for Hyperspectral Image Processing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3224−3234.
(64) Liu, H.; Song, D.; Rüger, S.; Hu, R.; Uren, V. *Information Retrieval Technology*; Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 44−50.
(65) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, 134114.
(66) Kussmann, J.; Ochsenfeld, C. Preselective Screening for Linear-Scaling Exact Exchange-Gradient Calculations for Graphics Processing Units and General Strong-Scaling Massively Parallel Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 918−922.
(67) Kussmann, J.; Ochsenfeld, C. Hybrid CPU/GPU Integral Engine for Strong-Scaling Ab Initio Methods. *J. Chem. Theory Comput.* **2017**, *13*, 3153−3159.
(68) Lehtola, S.; Steigemann, C.; Oliveira, M. J. T.; Marques, M. A. L. Recent developments in libxc - A comprehensive library of functionals for density functional theory. *SoftwareX* **2018**, *7*, 1−5.
(69) Laqua, H.; Kussmann, J.; Ochsenfeld, C. An improved molecular partitioning scheme for numerical quadratures in density functional theory. *J. Chem. Phys.* **2018**, *149*, 204111.
(70) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.
(71) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(72) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(73) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571−2577.
(74) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(75) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057−1065.
(76) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*, 054107.
(77) Lee, C.; Yang, W.; Parr, R. G. Development of the Colic-Salvetti correlation-energy into a functional of the electron density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.
(78) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.
(79) Becke, A. D. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372−1377.
(80) Niklasson, A. M. N.; Steneteg, P.; Odell, A.; Bock, N.; Challacombe, M.; Tymczak, C. J.; Holmström, E.; Zheng, G.; Weber, V. Extended Lagrangian Born-Oppenheimer molecular dynamics with dissipation. *J. Chem. Phys.* **2009**, *130*, 214109.
(81) Souvatzis, P.; Niklasson, A. M. N. Extended Lagrangian Born-Oppenheimer molecular dynamics in the limit of vanishing self-consistent field optimization. *J. Chem. Phys.* **2013**, *139*, 214102.
(82) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev. A: At., Mol., Opt. Phys.* **1967**, *159*, 98−103.
(83) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76*, 637−649.
(84) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(85) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(86) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(87) Sinha, P.; Boesch, S. E.; Gu, C.; Wheeler, R. A.; Wilson, A. K. Harmonic vibrational frequencies: Scaling factors for HF, B3LYP, and MP2 methods in combination with correlation consistent basis sets. *J. Phys. Chem. A* **2004**, *108*, 9213−9217.
(88) Kesharwani, M. K.; Brauer, B.; Martin, J. M. L. Frequency and zero-point vibrational energy scale factors for double-hybrid density functionals (and other selected methods): Can anharmonic force fields be avoided? *J. Phys. Chem. A* **2015**, *119*, 1701−1714.
(89) Botu, V.; Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074−1083.
(90) Wang, J.; Li, C.; Shin, S.; Qi, H. Accelerated Atomic Data Production in Ab Initio Molecular Dynamics with Recurrent Neural Network for Materials Research. *J. Phys. Chem. C* **2020**, *124*, 14838−14846.
(91) Han, R.; Luber, S. Trajectory-based machine learning method and its application to molecular dynamics. *Mol. Phys.* **2020**, *118*, No. e1788189.
(92) Jia, W.; Wang, H.; Chen, M.; Lu, D.; Lin, L.; Car, R.; Zhang, L. Pushing the Limit of Molecular Dynamics with Ab Initio Accuracy to 100 Million Atoms with Machine Learning. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*; IEEE Press: Atlanta, USA, November 9−19, 2020.

https://dx.doi.org/10.1021/acs.jctc.0c01279
J. Chem. Theory Comput. 2021, 17, 985−995

Supporting Information: "Quantitative comparison of experimental and computed IR-Spectra extracted from ab initio molecular dynamics"

Beatriz von der Esch, Laurens D. M. Peters, Lena Sauerland, and Christian Ochsenfeld*

Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany

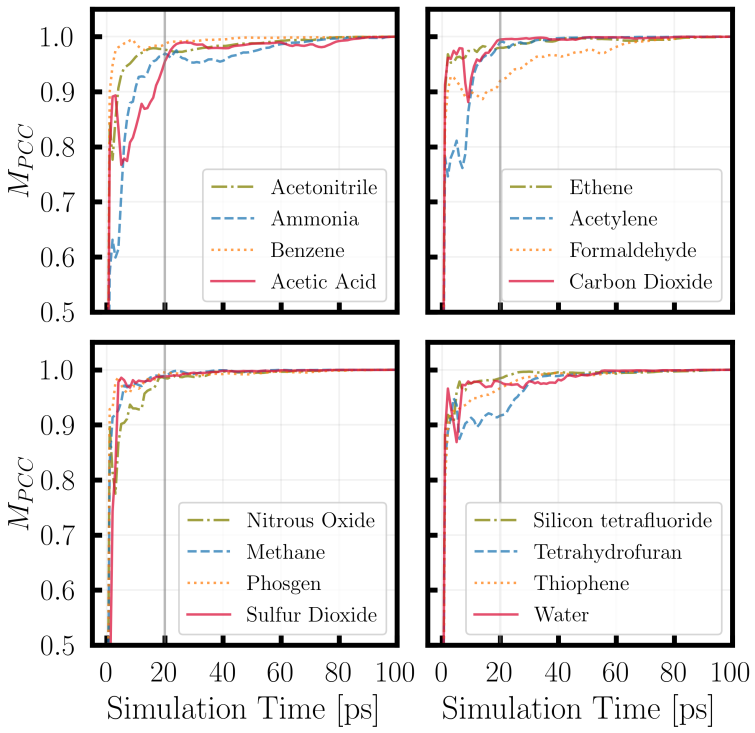
E-mail: christian.ochsenfeld@uni-muenchen.de

Visualisations

All plots were produced using the python-packages *matplotlib*[1] and *seaborn*.[2]

Spectra convergence study

The PCC similarity measure was used to asses the convergence of IR-spectra with respect to the simulation time based on a 100 ps GFN2-xTB[3,4] dynamics calculation. SFigure 1 shows that after 20 ps only minor changes are observed. To reduce the simulation length and increase the explored extent of phase space five 20 ps simulations were run in parallel.



SFigure 1: Convergence of computed IR-spectra from dynamics.

Underlying data of Cross Check

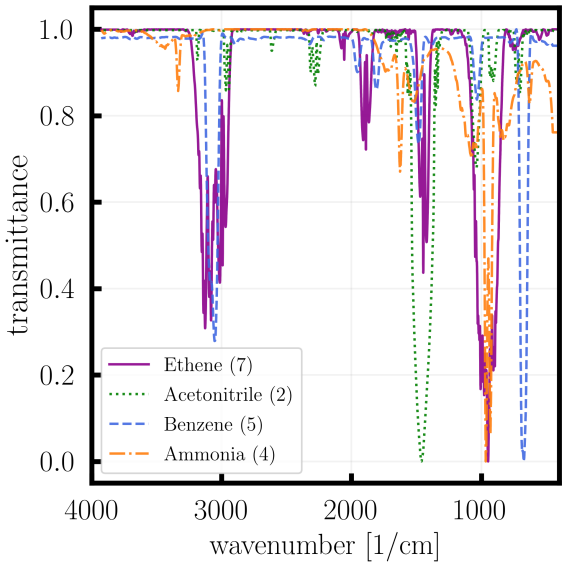
STable 1: Cross check based on experimental spectra.

	CH_3COOH	CH_3CN	C_2H_2	NH_3	C_6H_6	CO_2	C_2H_4	CH_2O	N_2O
CH_3COOH	1.	0.023	-0.052	0.158	0.076	-0.041	-0.093	0.136	-0.026
CH_3CN	0.0236	1.	-0.047	0.158	0.062	-0.018	0.202	0.002	-0.027
C_2H_2	-0.052	-0.047	1.	0.06	0.116	0.006	-0.134	-0.187	-0.016
NH_3	0.158	0.158	0.06	1.	-0.	-0.028	0.315	-0.224	-0.09
C_6H_6	0.076	0.062	0.116	-0.	1.	0.066	0.139	-0.037	-0.077
CO_2	-0.041	-0.018	0.006	-0.028	0.066	1.	-0.094	-0.135	-0.066
C_2H_4	-0.093	0.202	-0.134	0.315	0.139	-0.094	1.	-0.097	-0.157
CH_2O	0.136	0.002	-0.187	-0.224	-0.037	-0.135	-0.097	1.	-0.077
N_2O	-0.026	-0.027	-0.016	-0.09	-0.077	-0.066	-0.157	-0.077	1.

STable 2: Cross check based on pre-processed experimental spectra with baseline correction and smoothing.

	CH_3COOH	CH_3CN	C_2H_2	NH_3	C_6H_6	CO_2	C_2H_4	CH_2O	N_2O
CH_3COOH	1.	-0.021	-0.058	0.125	0.064	-0.038	-0.066	0.285	-0.021
CH_3CN	-0.021	1.	-0.009	0.125	0.081	-0.028	0.221	0.055	-0.072
C_2H_2	-0.058	-0.009	1.	0.062	0.128	0.01	-0.115	-0.146	-0.021
NH_3	0.125	0.125	0.062	1.	-0.019	-0.048	0.489	0.004	-0.102
C_6H_6	0.064	0.08	0.128	-0.019	1.	0.062	0.192	0.003	-0.092
CO_2	-0.038	-0.028	0.01	-0.048	0.062	1.	-0.078	-0.130	-0.062
C_2H_4	-0.066	0.22	-0.115	0.489	0.192	-0.078	1.	-0.060	-0.149
CH_2O	0.285	0.055	-0.146	0.004	0.003	-0.130	-0.060	1.	-0.088
N_2O	-0.021	-0.072	-0.021	-0.102	-0.092	-0.062	-0.149	-0.088	1.

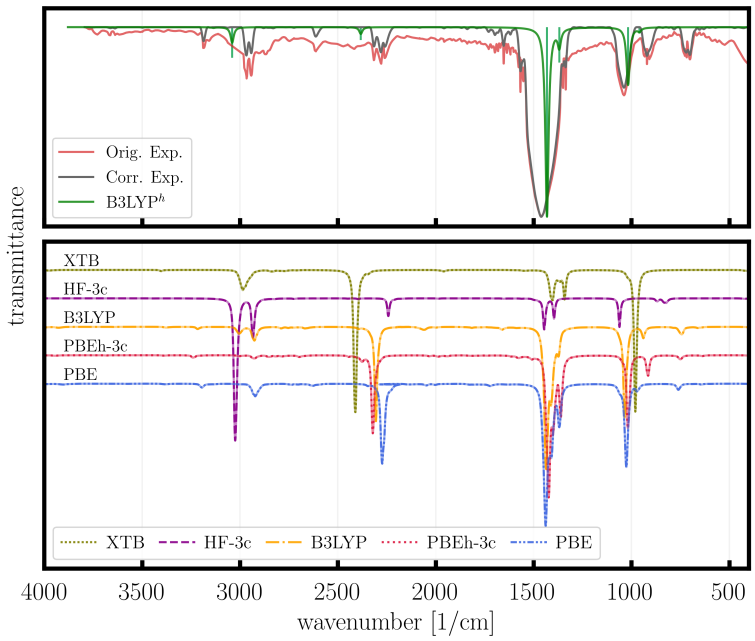
Visualisation of IR-spectra overlap of acetonitrile, ammonia, benzene, and ethene



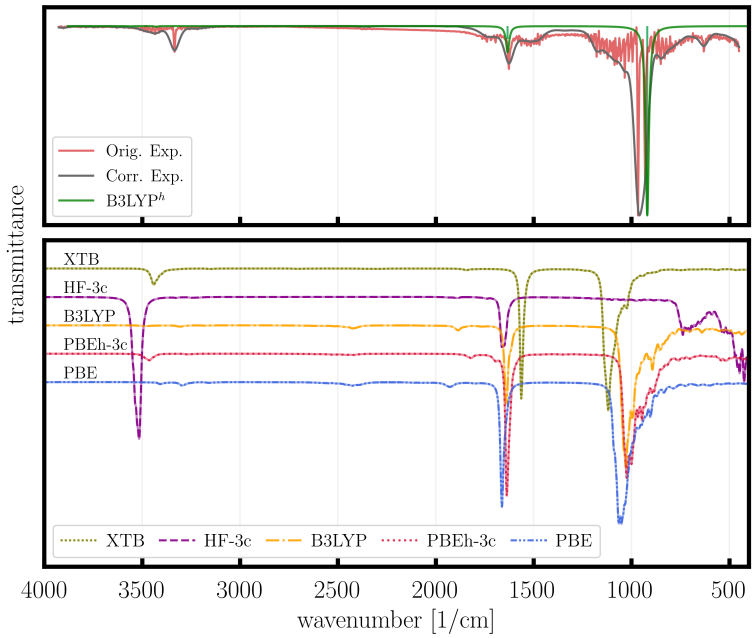
SFigure 2: Pre-processed experimental spectra of acetonitrile, ammonia, benzene, and ethen.

All calculated and experimental reference spectra

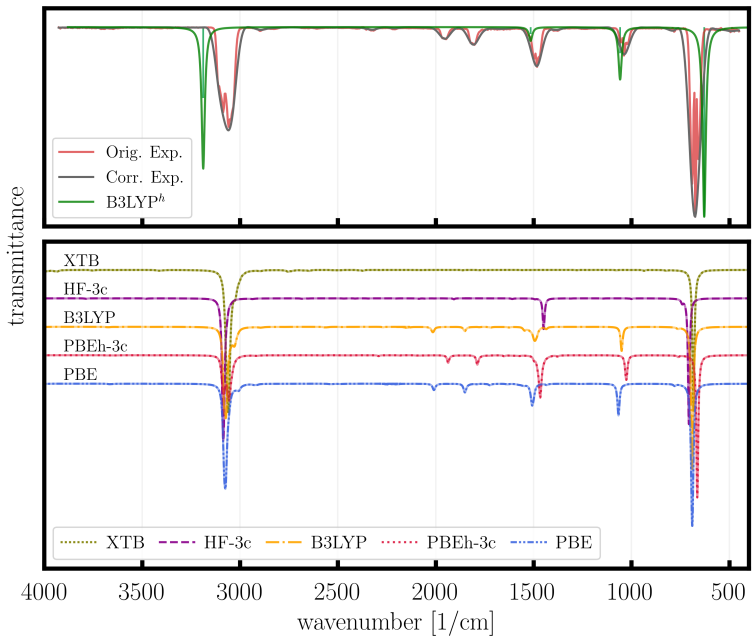
Below all original experimental spectra are displayed (red) and the resulting corrected spectra (grey). Additionally, all spectra obtained from dynamics and the static harmonic frequency analysis (green) are shown.



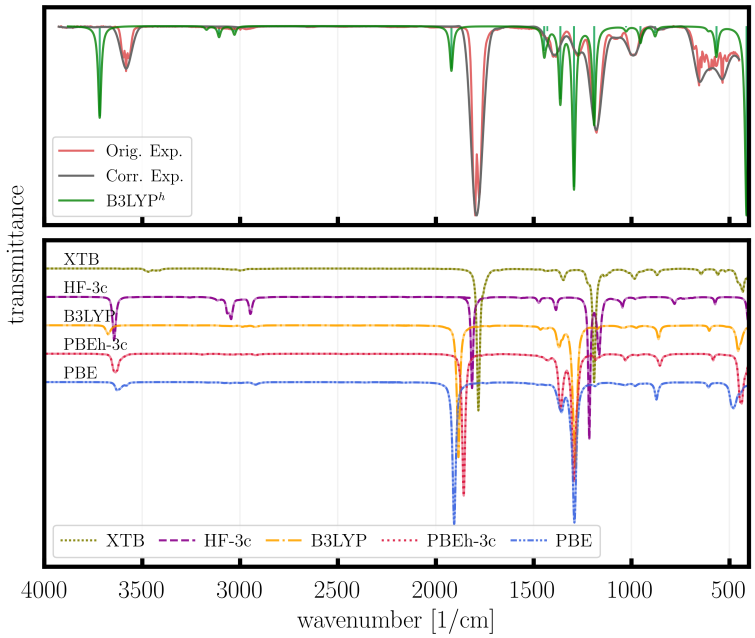
SFigure 3: Acetonitrile[5]



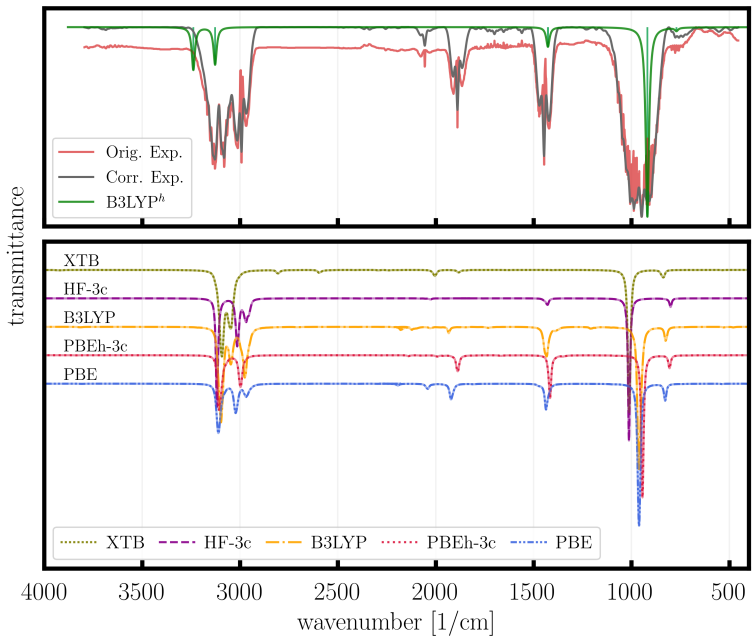
SFigure 4: Ammonia[5]



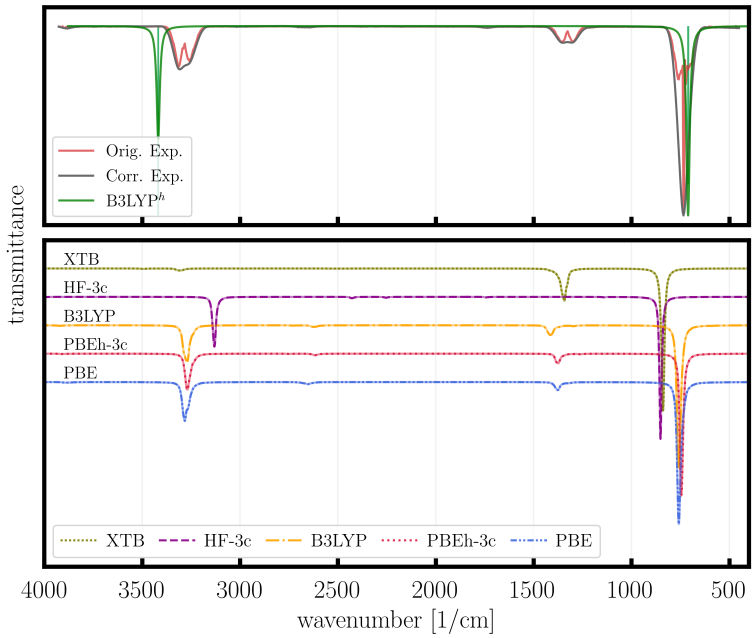
SFigure 5: Benzene[5]



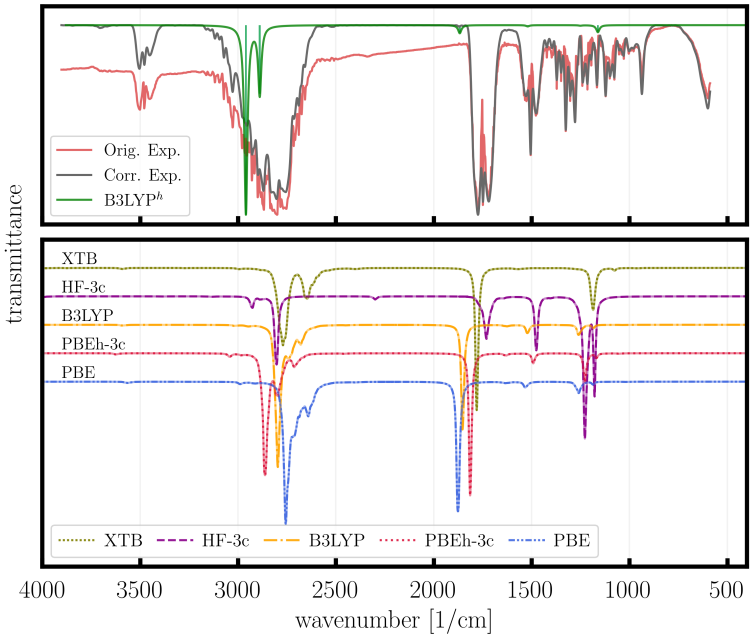
SFigure 6: Acetic Acid[5]



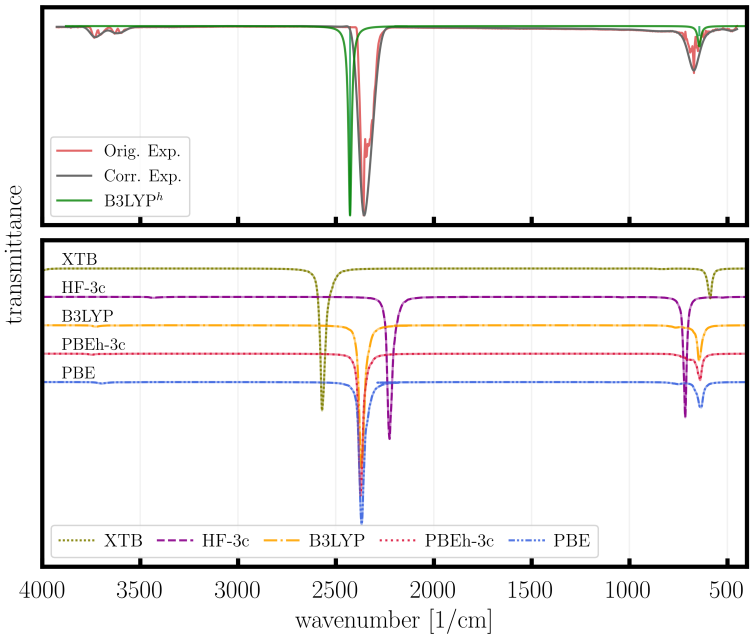
SFigure 7: Ethene[5]



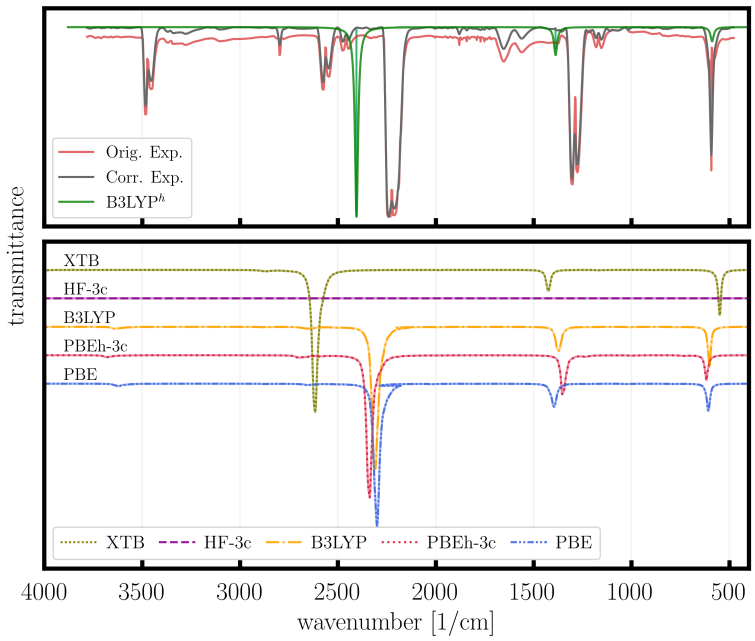
SFigure 8: Acetylene[5]



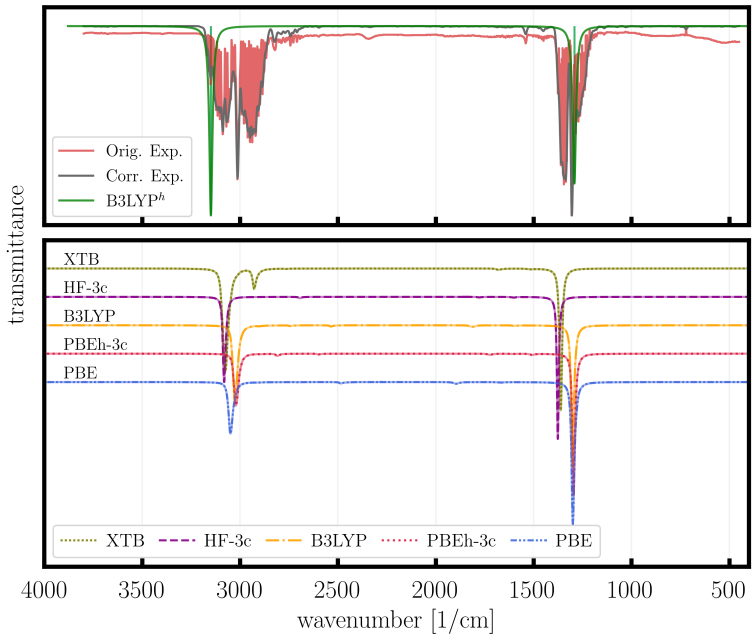
SFigure 9: Formaldehyde[5]



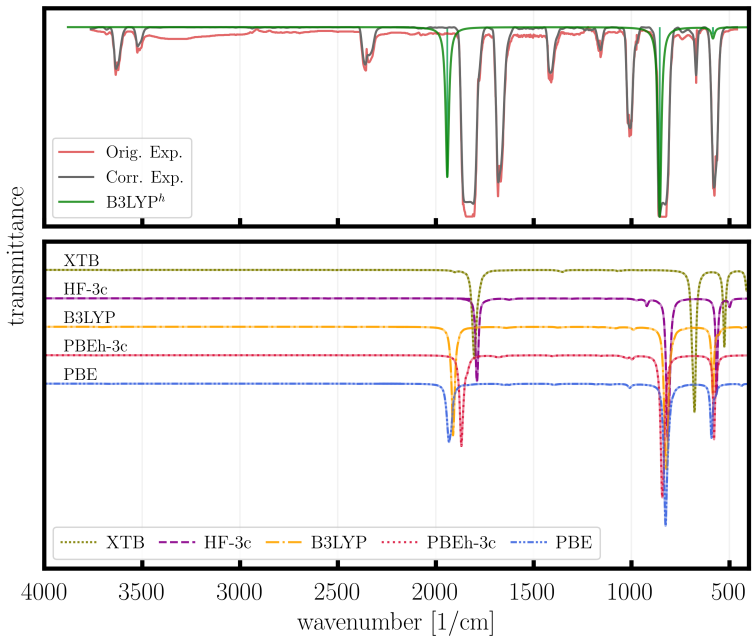
SFigure 10: Carbon dioxide[5]



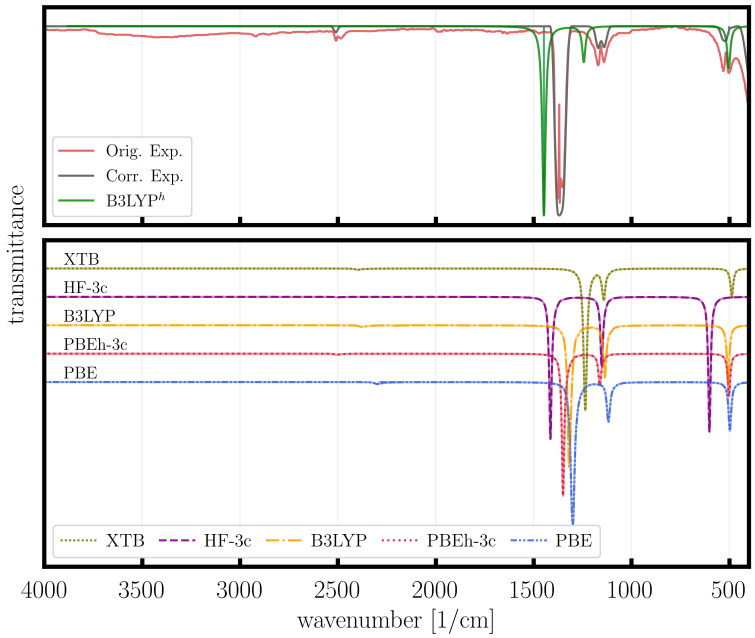
SFigure 11: Nitrous oxide[5]



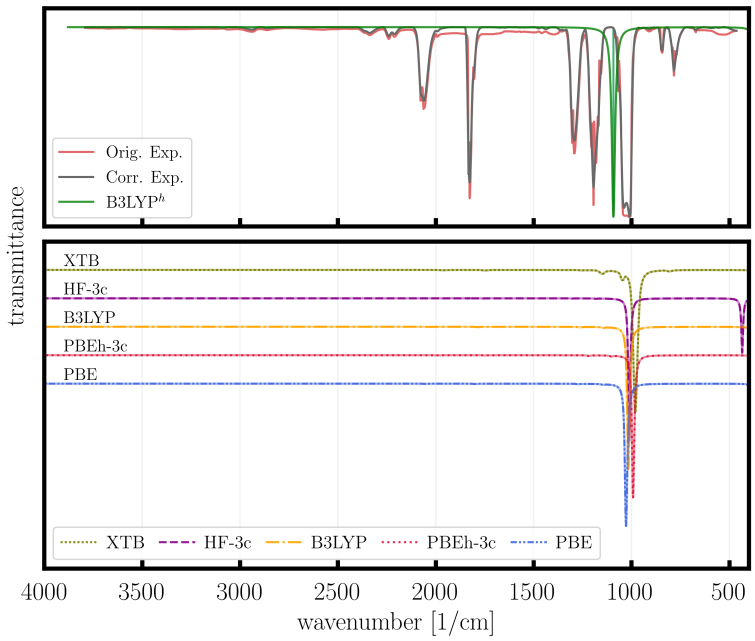
SFigure 12: Methane[5]



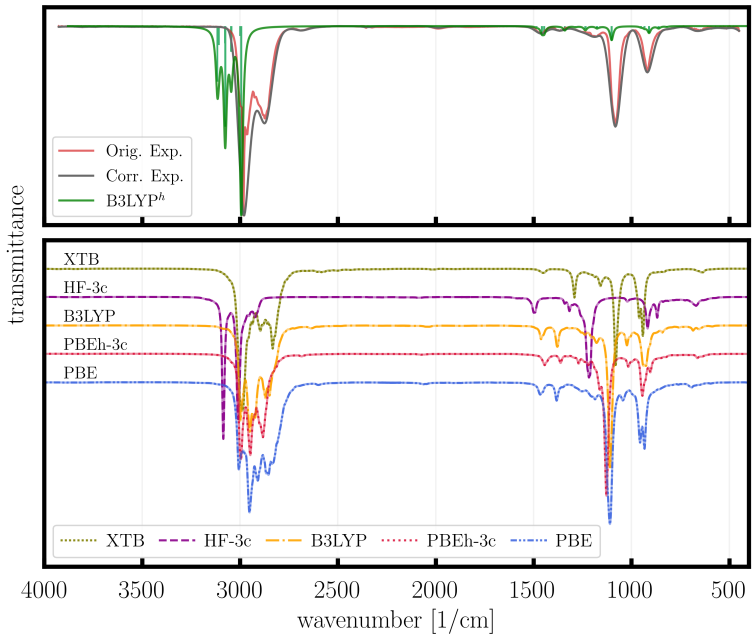
SFigure 13: Phosgene[5]



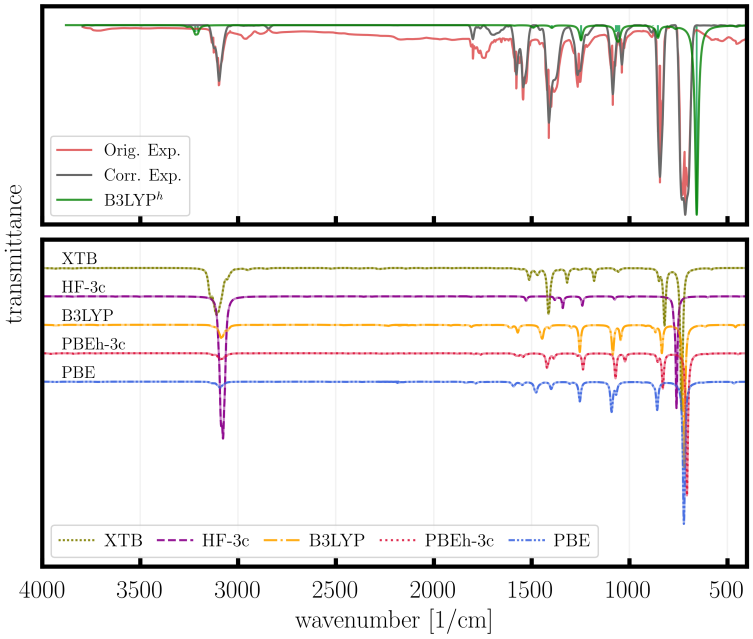
SFigure 14: Sulfur dioxide[5]



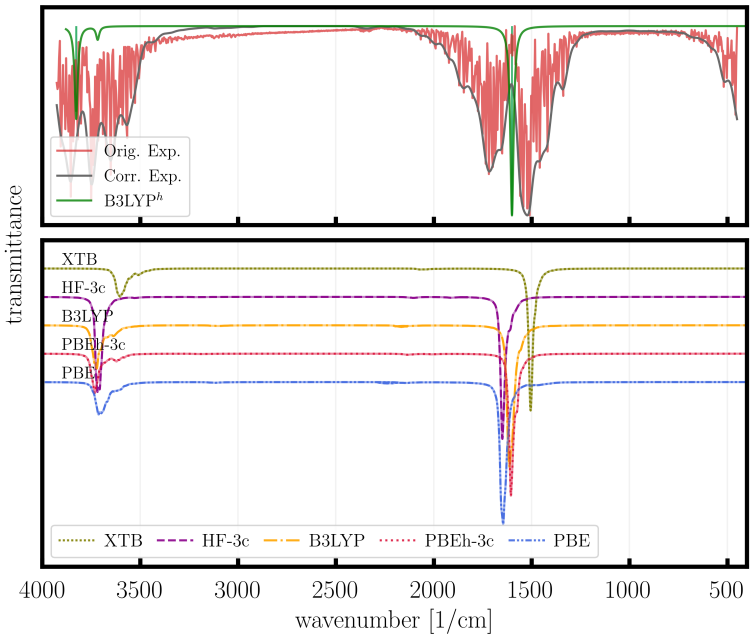
SFigure 15: Silicon tetrafluoride[5]



SFigure 16: Tetrahydrofuran[5]

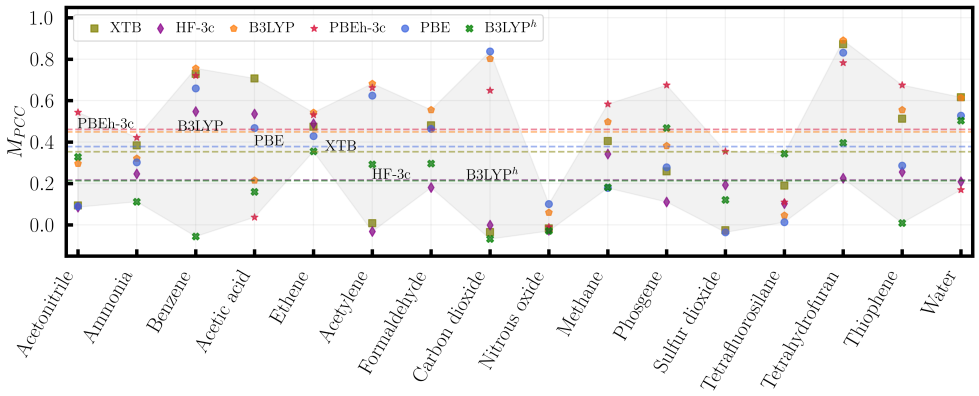


SFigure 17: Thiophene[5]



SFigure 18: Water[5]

M_{PCC} scores per compound using one scaling factor



SFigure 19: M_{PCC} scores per compound colored by computational method. The mean scores for each method are marked by a horizontal line. B3LYP^h denotes the results obtained from harmonic frequency calculations at the B3LYP-D3/def2-SVP level of theory. All results were scaled using one scaling factor.

References

(1) Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

(2) Waskom, M.; Botvinnik, O.; O'Kane, D.; Hobson, P.; Lukauskas, S.; Gemperline, D. C.; Augspurger, T.; Halchenko, Y.; Cole, J. B.; Warmenhoven, J. et al. mwaskom/seaborn: v0.8.1 (September 2017). 2017; `https://doi.org/10.5281/zenodo.883859`.

(3) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

(4) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multi-

pole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(5) Wallace, W. E. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology, Gaithersburg MD, 20899, Eds. P.J. Linstrom and W.G. Mallard, (retrieved July 27, 2020); Chapter "Infrared Spectra" by NIST Mass Spectrometry Data Center.

Chapter 5

Conclusion and Outlook

In the scope of this thesis, six sampling-driven studies are presented. In the first three published studies, and the outlined project (see chapter 3), reaction paths are characterized, each focusing on a different challenge. In publication **I**, the selection of an appropriate starting configuration that significantly influences the outcome of QM/MM reaction studies was addressed with machine learning. Using a simple linear regression model for the prediction of transition barriers, reactive periods were identified within an MD trajectory and further knowledge was gained about structural factors that govern the reactivity. The proposed approach was applied to the initial step of the desuccinylation catalyzed by sirtuin 5, but is transferable to any extended molecular system.
In publication **II**, the same reaction step was further studied using QM/MM-MD simulations resulting in the free energy surface allowing for the direct comparison of the 'true' free energy barrier and the barrier extrapolated from the minimum energy barriers computed in the previous study. This comparison highlights the necessity for sampling to accurately represent complex reactions and calls into question the reliability of transition barriers obtained from a single or very few minimum energy paths.
In the study summarized in chapter 3, the WTM-eABF method was used to characterize a synthetic route toward deoxyribonucleosides under prebiotic conditions from the canonical nucleobases, acetaldehyde, and glyceraldehyde as proposed by Teichert *et al.*[11] The route involves two steps, the formation of a vinylated nucleobase followed by the formation of the sugar ring. The latter was experimentally observed to be highly regio- and stereo-selective. This selectivity was computationally examined. So far, the results remain inconclusive. However, the optimization of the chosen collective variables and investigating the solvent influence could lead to interesting insights in future work.
Publication **III**, other than the before-mentioned studies aimed to accelerate the discovery of novel reaction paths. The molecular nanoreactor approach pioneered by Wang *et al.*[14] was optimized and alternate reactivity-enhancing spherical constraint functions as well as the use of buffer atoms was introduced. Furthermore, a fully automated post-processing routine is presented. Using the developed approach, prebiotically relevant primary and secondary precursors were obtained from a collection of HCN molecules. Aldotrioses, aldotetroses, as well as other reactive compounds were observed when reproducing the

formose network.[117]
Following the emergence of simple organic building blocks that constitute the biomolecules of living organisms, these must have aggregated and polymerized to lipids, carbohydrates, proteins, and the information-carrying polymers DNA and RNA. In publication **IV** the non-enzymatic polymerization of 2',3'-cNMPs is proposed under drying conditions or at a heated air-water interface at moderate temperature, low salt concentration in an alkaline environment. Experimentally, copolymerization of all four canonical nucleobases was observed, where cGMP polymerizes first. By evaluating the stability of homogeneous and heterogeneous stacked intercalated tetramers, we were able to show that a GMP scaffold could stabilize the otherwise unreactive nucleotides in an assembly suitable for polymerization. The experimental and computational findings were complemented by spectroscopic data.
Combining spectroscopic and computational findings enables the assessment of experimental hypothesises. In publication **V**, different quantitative indicators are tested to compare recorded and computed IR-spectra, thereby bridging both fields. So far the agreement between these was mainly assessed visually. Instead, we examined using the Pearson Correlation Coefficient and other quantitative measures to objectively compare spectra. The measure was further used to assess the performance of various electronic structure methods. Furthermore, in the scope of this project, several pre-procressing procedures were discussed to remove impeding effects such as background noise. The study clearly shows that extracting spectra from AIMD simulations yields results that are in better agreement with measured IR-spectra than spectra obtained from normal mode analysis.
All presented studies point towards a superiority of sampling-based studies over computational routines relying on a single or very few molecular configurations with rising importance of adequate sampling with the extent and flexibility of the system.
As quantum-chemical calculations become more and more efficient, larger data sets can be created opening up new fields in computational chemistry. The sheer amount of quantum chemical data that can be generated today allows us to design new, more accurate approaches, enables better, more general fitting to QM-level results, and allows for the interplay of machine learning and quantum chemistry. The latter being a true game-changer as machine learning, in turn, can accelerate the characterization of chemical system. For example, machine learning is already used to approximate *ab initio* dynamics.[118,119]
As more and more data becomes available to computational chemists, efficient routines have to be established that fully harness the rising amount of simulation data.

Bibliography

[1] M. Schiedel, D. Robaa, T. Rumpf, W. Sippl and M. Jung, *Med. Res. Rev.*, 2018, **38**, 147–200.

[2] J. Schemies, U. Uciechowska, W. Sippl and M. Jung, *Med. Res. Rev.*, 2010, **30**, 861–889.

[3] S. Michan and D. Sinclair, *Biochem. J.*, 2007, **404**, 1–13.

[4] A. Chalkiadaki and L. Guarente, *Nat. Rev. Cancer*, 2015, **15**, 608–624.

[5] J. Du, Y. Zhou, X. Su, J. J. Yu, S. Khan, H. Jiang, J. Kim, J. Woo, J. H. Kim, B. H. Choi, B. He, W. Chen, S. Zhang, R. A. Cerione, J. Auwerx, Q. Hao and H. Lin, *Science*, 2011, **334**, 806–809.

[6] R. H. Houtkooper, E. Pirinen and J. Auwerx, *Nat. Rev. Mol. Cell Biol.*, 2012, **13**, 225–238.

[7] R. Sure and S. Grimme, *J. Comput. Chem.*, 2013, **34**, 1672–1685.

[8] U. Ryde, *J. Chem. Theory Comput.*, 2017, **13**, 5745–5752.

[9] G. Torrie and J. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.

[10] M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 124105.

[11] J. S. Teichert, F. M. Kruse and O. Trapp, *Angew. Chemie Int. Ed.*, 2019, **58**, 9944–9947.

[12] H. Fu, H. Zhang, H. Chen, X. Shao, C. Chipot and W. Cai, *J. Phys. Chem. Lett.*, 2018, **9**, 4738–4745.

[13] H. Fu, X. Shao, W. Cai and C. Chipot, *Acc. Chem. Res.*, 2019, **52**, 3254–3264.

[14] L. P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nat. Chem.*, 2014, **6**, 1044–1048.

[15] G. F. Joyce, *Nature*, 1989, **338**, 217–224.

[16] I. V. Delidovich, A. N. Simonov, O. P. Taran and V. N. Parmon, *ChemSusChem*, 2014, **7**, 1833–1846.

[17] J. D. Sutherland, *Angew. Chemie*, 2016, **128**, 108–126.

[18] N. Kitadai and S. Maruyama, *Geosci. Front.*, 2018, **9**, 1117–1153.

[19] M. Thomas, M. Brehm, R. Fligg, P. Vöhringer and B. Kirchner, *Phys. Chem. Chem. Phys.*, 2013, **15**, 6608–6622.

[20] F. Jensen, *Introd. to Comput. Chem.*, John Wiley & Sons, Incorporated, 3rd edn, 2017, pp. 418–431.

[21] K. E. Ranaghan and A. J. Mulholland, *Int. Rev. Phys. Chem.*, 2010, **29**, 65–133.

[22] W. Li and A. Ma, *Mol. Simul.*, 2014, **40**, 784–793.

[23] P. V. Banushkina and S. V. Krivov, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2016, **6**, 748–763.

[24] K. Zinovjev and I. Tuñón, *J. Phys. Chem. A*, 2017, **121**, 9764–9772.

[25] K. Zinovjev and I. Tuñón, *WIREs Comput. Mol. Sci.*, 2018, **8**, e1329.

[26] H. JÓNSSON, G. MILLS and K. W. JACOBSEN, Class. Quantum Dyn. Condens. Phase Simulations, 1998, pp. 385–404.

[27] G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978–9985.

[28] J. Kästner, J. M. Carr, T. W. Keal, W. Thiel, A. Wander and P. Sherwood, *J. Phys. Chem. A*, 2009, **113**, 11856–11865.

[29] G. Henkelman, B. P. Uberuaga and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9901–9904.

[30] D. Sheppard, R. Terrell and G. Henkelman, *J. Chem. Phys.*, 2008, **128**, 134106.

[31] E. L. Kolsbjerg, M. N. Groves and B. Hammer, *J. Chem. Phys.*, 2016, **145**, 094107.

[32] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816–9872.

[33] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko and K. R. Müller, *J. Chem. Theory Comput.*, 2013, **9**, 3404–3419.

[34] S. Geman, E. Bienenstock and R. Doursat, *Neural Comput.*, 1992, **4**, 1–58.

[35] H. Abdi and L. J. Williams, *Wiley Interdiscip. Rev. Comput. Stat.*, 2010, **2**, 433–459.

[36] L. J. Broadbelt, S. M. Stark and M. T. Klein, *Ind. Eng. Chem. Res.*, 1994, **33**, 790–799.

[37] M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

[38] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.

[39] D. Weininger, *J. Chem. Inf. Model.*, 1988, **28**, 31–36.

[40] D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 97–101.

[41] J. Behler, *J. Chem. Phys.*, 2011, **134**, 074106.

[42] A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.

[43] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller and E. K. U. Gross, *Phys. Rev. B*, 2014, **89**, 205118.

[44] A. S. Christensen, L. A. Bratholm, F. A. Faber and O. Anatole von Lilienfeld, *J. Chem. Phys.*, 2020, **152**, 044107.

[45] M. F. Langer, A. Goeßmann and M. Rupp, *npj Comput. Mater.*, 2022, **8**, 41.

[46] F. Jensen, *Introd. to Comput. Chem.*, John Wiley & Sons, Incorporated, 3rd edn, 2017, pp. 447–468.

[47] D. G. Truhlar, B. C. Garrett and S. J. Klippenstein, *J. Phys. Chem.*, 1996, **100**, 12771–12800.

[48] D. Flaig, M. Beer and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2012, **8**, 2260–2271.

[49] S. Das, K. Nam and D. T. Major, *J. Chem. Theory Comput.*, 2018, **14**, 1695–1705.

[50] J. Chen, J. Kato, J. B. Harper, Y. Shao and J. Ho, *J. Phys. Chem. B*, 2021, **125**, 9304–9316.

[51] A. M. Cooper and J. Kästner, *ChemPhysChem*, 2014, **15**, 3264–3269.

[52] B. von der Esch, J. C. B. Dietschreit, L. D. M. Peters and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2019, **15**, 6660–6667.

[53] J. C. B. Dietschreit, B. von der Esch and C. Ochsenfeld, *Phys. Chem. Chem. Phys.*, 2022, **24**, 7723–7731.

[54] B. Roux, *Comput. Phys. Commun.*, 1995, **91**, 275–282.

[55] J. Hénin, T. Lelièvre, M. R. Shirts, O. Valsson and L. Delemotte, *arXiv:2202.04164v2*, 2022.

[56] J. Schlitter, M. Engels and P. Krüger, *J. Mol. Graph.*, 1994, **12**, 84–89.

[57] A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci.*, 2002, **99**, 12562–12566.

[58] E. Darve and A. Pohorille, *J. Chem. Phys.*, 2001, **115**, 9169–9183.

[59] J. Comer, J. C. Gumbart, J. Hénin, T. Lelievre, A. Pohorille and C. Chipot, *J. Phys. Chem. B*, 2015, **119**, 1129–1151.

[60] A. Lesage, T. Lelièvre, G. Stoltz and J. Hénin, *J. Phys. Chem. B*, 2017, **121**, 3676–3685.

[61] A. Barducci, G. Bussi and M. Parrinello, *Phys. Rev. Lett.*, 2008, **100**, 020603.

[62] C. H. Bennett, *J. Comput. Phys.*, 1976, **22**, 245–268.

[63] N. Goldman, *Nat. Chem.*, 2014, **6**, 1033–1034.

[64] T. J. Martínez, *Acc. Chem. Res.*, 2017, **50**, 652–656.

[65] J. Meisner, X. Zhu and T. J. Martínez, *ACS Cent. Sci.*, 2019, **5**, 1493–1495.

[66] T. Das, S. Ghule and K. Vanka, *ACS Cent. Sci.*, 2019, **5**, 1532–1540.

[67] D. M. Matheu, A. M. Dean, J. M. Grenda and W. H. Green, *J. Phys. Chem. A*, 2003, **107**, 8552–8565.

[68] D. Rappoport, C. J. Galvin, D. Y. Zubarev and A. Aspuru-Guzik, *J. Chem. Theory Comput.*, 2014, **10**, 897–907.

[69] C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Comput. Phys. Commun.*, 2016, **203**, 212–225.

[70] A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2018, **8**, 1–20.

[71] G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A*, 2019, **123**, 385–399.

[72] A. Puliyanda, K. Srinivasan, K. Sivaramakrishnan and V. Prasad, *Digit. Chem. Eng.*, 2022, **2**, 100009.

[73] A. Barducci, M. Bonomi and M. Parrinello, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2011, **1**, 826–843.

[74] S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 2847–2862.

[75] T. Lei, W. Guo, Q. Liu, H. Jiao, D. B. Cao, B. Teng, Y. W. Li, X. Liu and X. D. Wen, *J. Chem. Theory Comput.*, 2019, **15**, 3654–3665.

[76] K. B. Wiberg, *Tetrahedron*, 1968, **24**, 1083–1096.

[77] G. Landrum, P. Tosco, B. Kelley, Sriniker, Ric, Gedeck, R. Vianello, N. Schneider, A. Dalke, E. Kawashima, D. N, B. Cole, M. Swain, S. Turk, D. Cosgrove, AlexanderSavelyev, A. Vaucher, G. Jones, M. Wójcikowski, D. Probst, G. Godin, V. F. Scalfani, A. Pahl, F. Berenger, JLVarjo, Strets123, JP, DoliathGavid, G. Sforna and J. H. Jensen, *rdkit/rdkit: 2021_03_4 (Q1 2021) Release*, 2021, `https://zenodo.org/record/5085999`.

[78] J. Šponer, P. Jurečka, I. Marchan, F. J. Luque, M. Orozco and P. Hobza, *Chem. - A Eur. J.*, 2006, **12**, 2854–2865.

[79] J. Šponer, J. Leszczyński and P. Hobza, *J. Phys. Chem.*, 1996, **100**, 5590–5596.

[80] B. Guillot, *J. Chem. Phys.*, 1991, **95**, 1543–1551.

[81] N. Wiener, *Acta Math.*, 1930, **55**, 117–258.

[82] A. Khintchine, *Math. Ann.*, 1934, **109**, 604–615.

[83] A. P. Scott and L. Radom, *J. Phys. Chem.*, 1996, **100**, 16502–16513.

[84] K. K. Irikura, R. D. Johnson and R. N. Kacker, *J. Phys. Chem. A*, 2005, **109**, 8430–8437.

[85] Y. Tantirungrotechai, K. Phanasant, S. Roddecha, P. Surawatanawong, V. Sutthikhum and J. Limtrakul, *J. Mol. Struct. THEOCHEM*, 2006, **760**, 189–192.

[86] J. P. Merrick, D. Moran and L. Radom, *J. Phys. Chem. A*, 2007, **111**, 11683–11700.

[87] I. M. Alecu, J. Zheng, Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2010, **6**, 2872–2887.

[88] P. Borowski, *J. Phys. Chem. A*, 2012, **116**, 3866–3880.

[89] P. H. Eilers, *Anal. Chem.*, 2003, **75**, 3631–3636.

[90] H. F. Boelens, R. J. Dijkstra, P. H. Eilers, F. Fitzpatrick and J. A. Westerhuis, *J. Chromatogr. A*, 2004, **1057**, 21–30.

[91] K. H. Liland, T. Almøy and B. H. Mevik, *Appl. Spectrosc.*, 2010, **64**, 1007–1016.

[92] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li and J. Tan, *Anal. Chim. Acta*, 2010, **683**, 63–68.

[93] A. Kwiatkowski, M. Gnyba, J. Smulko and P. Wierzba, *Metrol. Meas. Syst.*, 2010, **17**, 549–560.

[94] G. Renner, T. C. Schmidt and J. Schram, *Anal. Chem.*, 2017, **89**, 12045–12053.

[95] G. Renner, A. Nellessen, A. Schwiers, M. Wenzel, T. C. Schmidt and J. Schram, *TrAC - Trends Anal. Chem.*, 2019, **111**, 229–238.

[96] R. A. K. S. and Leibler, *Ann. Math. Stat.*, 1951, **22**, 79–86.

[97] H. Jeffreys, *Proc. R. Soc. Lond. A. Math. Phys. Sci.*, 1946, **186**, 453–461.

[98] Y. Rubner, C. Tomasi and L. J. Guibas, *Int. J. Comput. Vis.*, 2000, **40**, 99–121.

[99] B. von der Esch, L. D. M. Peters, L. Sauerland and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2021, **17**, 985–995.

[100] A. Eschenmoser, *Science*, 1999, **284**, 2118–2124.

[101] G. Springsteen and G. F. Joyce, *J. Am. Chem. Soc.*, 2004, **126**, 9578–9583.

[102] N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110.

[103] A. Schäfer, H. Horn and R. Ahlrichs, *J. Chem. Phys.*, 1992, **97**, 2571–2577.

[104] F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.

[105] F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.

[106] S. Grimme, J. G. Brandenburg, C. Bannwarth and A. Hansen, *J. Chem. Phys.*, 2015, **143**, 054107.

[107] O. A. Vydrov, T. V. Voorhis and T. Van Voorhis, *J. Chem. Phys.*, 2010, **133**, 244103.

[108] J. Kussmann, M. Beer and C. Ochsenfeld, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2013, **3**, 614–636.

[109] J. Kussmann, A. Luenser, M. Beer and C. Ochsenfeld, *J. Chem. Phys.*, 2015, **142**, 094101.

[110] H. Laqua, T. H. Thompson, J. Kussmann and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2020, **16**, 1456–1468.

[111] A. M. N. Niklasson, P. Steneteg, A. Odell, N. Bock, M. Challacombe, C. J. Tymczak, E. Holmström, G. Zheng and V. Weber, *J. Chem. Phys.*, 2009, **130**, 214109.

[112] P. Souvatzis and A. M. N. Niklasson, *J. Chem. Phys.*, 2013, **139**, 214102.

[113] A. Klamt and G. Schüürmann, *J. Chem. Soc. Perkin Trans. 2*, 1993, 799–805.

[114] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *J. Chem. Phys.*, 1982, **76**, 637–649.

[115] A. Hulm, J. C. B. Dietschreit and C. Ochsenfeld, *J. Chem. Phys.*, 2022, **157**, 024110.

[116] J. C. B. Dietschreit, D. J. Diestler and C. Ochsenfeld, *J. Chem. Phys.*, 2022, **156**, 114105.

[117] S. A. Benner, E. A. Bell, E. Biondi, R. Brasser, T. Carell, H. Kim, S. J. Mojzsis, A. Omran, M. A. Pasek and D. Trail, *ChemSystemsChem*, 2020, **2**, e1900035.

[118] M. Rupp, *Int. J. Quantum Chem.*, 2015, **115**, 1058–1073.

[119] M. Gastegger, J. Behler and P. Marquetand, *Chem. Sci.*, 2017, **8**, 6924–6935.