Revisiting Hybridization Kinetics with Improved Elementary Step Simulation

Jordan Lovrod
University of British Columbia, Vancouver, Canada

Boyan Beronov
University of British Columbia, Vancouver, Canada

Chenwei Zhang
University of British Columbia, Vancouver, Canada

Erik Winfree
California Institute of Technology, Pasadena, CA, USA

Anne Condon
University of British Columbia, Vancouver, Canada

Abstract

Nucleic acid strands, which react by forming and breaking Watson-Crick base pairs, can be designed to form complex nanoscale structures or devices. Controlling such systems requires accurate predictions of the reaction rate and of the folding pathways of interacting strands. Simulators such as *Multistrand* model these kinetic properties using continuous-time Markov chains (CTMCs), whose states and transitions correspond to secondary structures and elementary base pair changes, respectively. The transient dynamics of a CTMC are determined by a kinetic model, which assigns transition rates to pairs of states, and the rate of a reaction can be estimated using the mean first passage time (MFPT) of its CTMC. However, use of *Multistrand* is limited by its slow runtime, particularly on rare events, and the quality of its rate predictions is compromised by a poorly-calibrated and simplistic kinetic model. The former limitation can be addressed by constructing truncated CTMCs, which only include a small subset of states and transitions, selected either manually or through simulation. As a first step to address the latter limitation, Bayesian posterior inference in an Arrhenius-type kinetic model was performed in earlier work, using a small experimental dataset of DNA reaction rates and a fixed set of manually truncated CTMCs, which we refer to as *Assumed Pathway* (AP) state spaces. In this work we extend this approach, by introducing a new prior model that is directly motivated by the physical meaning of the parameters and that is compatible with experimental measurements of elementary rates, and by using a larger dataset of 1105 reactions as well as larger truncated state spaces obtained from the recently introduced stochastic *Pathway Elaboration* (PE) method. We assess the quality of the resulting posterior distribution over kinetic parameters, as well as the quality of the posterior reaction rates predicted using AP and PE state spaces. Finally, we use the newly parameterised PE state spaces and *Multistrand* simulations to investigate the strong variation of helix hybridization reaction rates in a dataset of Hata et al. While we find strong evidence for the nucleation-zippering model of hybridization, in the classical sense that the rate-limiting phase is composed of elementary steps reaching a small "nucleus" of critical stability, the strongly sequence-dependent structure of the trajectory ensemble up to nucleation appears to be much richer than assumed in the model by Hata et al. In particular, rather than being dominated by the collision probability of nucleation sites, the trajectory segment between first binding and nucleation tends to visit numerous secondary structures involving misnucleation and hairpins, and has a sizeable effect on the probability of overcoming the nucleation barrier.

2012 ACM Subject Classification Applied computing → Chemistry

Keywords and phrases DNA reaction kinetics, kinetic model calibration, simulation-based Bayesian inference, continuous-time Markov chains

Digital Object Identifier 10.4230/LIPIcs.DNA.29.5

29th International Conference on DNA Computing and Molecular Programming (DNA 29).
Editors: Ho-Lin Chen and Constantine G. Evans; Article No. 5; pp. 5:1–5:24
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Nucleic acid strands, which can react *in vitro* or *in vivo* by forming and breaking Watson-Crick base pairs, can be designed to fold into specific two and three-dimensional nanostructures [19, 41] through methods such as tile and brick assembly [64, 13, 36] and DNA origami [54, 21]. These nanostructures can execute various chemical [65, 11], mechanical [45, 5, 6], computational [4, 63, 75], and biomedical tasks [62, 20, 12, 74, 80, 26, 40]. To design and debug such systems of interacting nucleic acid strands within their environment, it is valuable to predict both their thermodynamic behaviour (such as energetically stable structures) and their kinetic behaviour (such as folding pathways or reaction rates).

Efficient, general-purpose methods are available for predicting thermodynamic properties of interacting nucleic acid strands [34, 85, 76, 25]. These methods leverage "nearest-neighbor" models of nucleic acid thermodynamics that have been well calibrated from experimental data over many decades [46, 58, 57]. In contrast, general-purpose methods for predicting kinetic properties can be slow and inaccurate. One such simulation model, *Multistrand* [59, 60], samples folding trajectories from initial to final states through the space of all possible secondary structures of the nucleic acid system. Each step along the trajectory is an elementary transition [23], in which a base pair forms or breaks and in which a holding time is consumed. Such simulations can be prohibitively slow when the reaction is a rare event, e.g., when the initial and final states are separated by a high energy barrier. Moreover, *Multistrand*'s elementary rates are determined by a combination of nearest neighbor thermodynamics and a 2-parameter Metropolis kinetic model, which is too simplistic to produce reliable rate predictions.

Two important improvements address these limitations of *Multistrand*. First, the *Pathway Elaboration* (PE) method uses *Multistrand* to build truncated state spaces of reaction kinetics [83, 81]. These smaller state spaces are amenable to matrix methods for rate computation, which are efficient even on rare events. Second, a 15-parameter Arrhenius kinetic model takes into account the local context around each elementary step, which in principle should improve reaction rate predictions [82]. This Arrhenius model was calibrated using a dataset of a few hundred reactions, on small, customised state spaces, that reflect assumptions about likely pathways. We refer to these as *Assumed Pathway* (AP) state spaces.

1.1 Improved parameter inference

However, previous work stopped short of the challenging computational task of calibrating the Arrhenius model using the PE state spaces. In this work, we describe a Bayesian inference approach to this problem. Section 3.1 introduces a new prior distribution over the kinetic parameters, which is directly motivated by their physical meaning and which is compatible with experimental measurements of elementary rates. We use a dataset of 1105 reactions, as will be described in Section 4.2, that had previously been sourced, but only a small subset of which had been used for inference in past work. For each reaction in this dataset, we generated a truncated state space with AP, and also using the existing PE implementation when it completed within 7 days and with 10GB of guaranteed RAM on a single CPU core[1].

Despite a significantly higher computational cost, we find that the larger PE state spaces do not always lead to more accurate rate predictions than the small, manually designed AP state spaces. For posterior approximation, we apply the standard *random walk Metropolis*

[1] The majority of the runtime and memory footprint were caused by the unoptimised Python implementation of PE, rather than by the *Multistrand* stochastic simulations in its inner loops.

(RWM) algorithm. The resulting posterior approximations in Section 4.3, which are often multimodal, recover an expected correlation structure among the kinetic parameters. However, we also uncover severe numerical instability in the linear equation systems required for rate prediction. Due to numerous design limitations in the legacy software, a significant refactoring effort was required to implement the above extensions. In Section 4.4, we discuss several modifications to the kinetic model and inference methods that could improve inference quality. Full details of this work are described in Lovrod's MSc Thesis [43].

1.2 Case study

In Section 5, we use the mode of the new posterior to parameterise the *Multistrand* model, and present a case study based on the helix association data of Hata et al. [33]. Although the examined sequences have equal length and similar melting temperatures, the reported hybridization rates spread over more than two orders of magnitude. Our thermodynamic simulations confirm a correlation between the experimental rate and the expected number of free bases in the Boltzmann ensemble of unbound strands. However, our kinetic simulations suggest that the process of nucleation, in the non-probabilistic sense of reaching three consecutive correct inter-strand base pairs, is nontrivial and insufficiently explained by the time to first binding. We then employ *Multistrand*'s "first step mode" (FSM) to analyse the probabilistic and temporal behaviour of trajectories that start at the moment of initial binding and end either with dissociation or hybridization. In particular, we find a positive correlation between the proportion of successfully associating trajectories and the experimental rate. Moreover, the two-stranded complexes often spend extended periods of time exploring conformations with misnucleation and/or hairpins, and visualisations of the reactive FSM trajectory ensemble indicate a rich and strongly sequence-dependent structure, including a multimodal first passage time distribution for some reactions.

2 Background and related work

The DNA reactions we consider occur in systems with fixed experimental conditions (solution volume V, temperature T, and concentrations of Na^+ and Mg^{2+} ions). We often use the inverse temperature $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant. A nucleic acid reaction, in which DNA or RNA strands fold from one three-dimensional structure into another by forming and breaking base pairs, can be described at the secondary structure level with an initial microstate (or initial distribution over microstates) representing the reactants, and a final microstate (or final region of microstates) representing the products. The number of microstates may in general scale exponentially in the total strand length l.

A thermodynamic model defines the Gibbs free energy $\Delta G(x)$ relative to some reference state, at all allowed states $x \in \mathcal{X}$ of a system, and gives rise to the Gibbs-Boltzmann distribution π and its partition function Z_β at inverse temperature β,

$$\pi(x \mid \beta) = \frac{1}{Z_\beta} \cdot e^{-\beta \cdot \Delta G(x)}, \qquad Z_\beta = \int e^{-\beta \cdot \Delta G(x)} dx, \tag{1}$$

which can be used to compute all quantities of interest at thermodynamic equilibrium.

2.1 Kinetic models of nucleic acid elementary steps

We focus in this work on elementary step models of nucleic acid reactions [23, 15, 59, 60, 22, 82], which offer a relatively fine-grained view of the system, with states and transitions corresponding to secondary structures and isolated changes in base pairs, respectively. These simulators model the reaction kinetics using a *continuous-time Markov chain* (CTMC) on this state space, which will be defined in the next section. Notably, *Multistrand* can model multi-stranded reactions, but like other currently available elementary step models, it only allows pseudoknot-free secondary structures.

2.1.1 CTMCs and their mean first passage times (MFPTs)

A finite CTMC is characterised by an initial probability distribution π_0 over a finite set of allowed states $\mathcal{X}$ and a *transition rate matrix* $K : \mathcal{X}^2 \to \mathbb{R}$, such that for all pairs of distinct states x and x', $K(x, x')$ is the instantaneous transition rate from x to x', and $K(x, x) = -\sum_{x' \in \mathcal{X} \setminus x} K(x, x')$ [69]. We refer to the subset of states $I \subset \mathcal{X}$ with non-zero initial probability π_0 as the *initial region.* The *transition probability matrix* $P : \mathcal{X}^2 \to [0, 1]$ is the normalised rate matrix, $P(x, x') = -\frac{K(x,x')}{K(x,x)}$, whereas the *transition matrices* (or *propagators*) $Q_t = e^{tK}$ for $t \in \mathbb{R}_{\geq 0}$ determine the transient dynamics of a CTMC. A stochastic process $\{X(t)\}_{t \in \mathbb{R}_{\geq 0}}$ can then be identified with this CTMC if and only if

$$\mathbb{P}\left(X(t_0) = x_0, \ldots, X(t_n) = x_n\right) = \pi_0(x_0) \prod_{m \in [0, n-1]} Q_{t_{m+1} - t_m}(x_m, x_{m+1}) \tag{2}$$

holds for any $n \in \mathbb{N}_0$, $t_0 < \cdots < t_n \in \mathbb{R}_{\geq 0}$, and $x_0, \ldots, x_n \in \mathcal{X}$. In other words, trajectory probabilities can be arbitrarily decomposed into segment probabilities due to Markovianity, and the probability of any segment is determined solely by its spatial endpoints (x_m, x_{m+1}), its temporal endpoints (t_m, t_{m+1}) and the transition rate matrix.

For a fixed final region $F \subset \mathcal{X}$, the *mean first passage time* (MFPT) $\tau_F : \mathcal{X} \to \mathbb{R}_{\geq 0}$ denotes the expected time to reach F for the first time from each state. It satisfies

$$-\tau_F(x) \cdot K(x, x) = 1 + \sum_{x' \in \mathcal{X} \setminus x} \tau_F(x') \cdot K(x, x') \quad \text{for all} \quad x \in \mathcal{X} \setminus \{F\}\ , \tag{3}$$

which is a numerically solvable matrix equation for sufficiently small CTMCs in which all states are connected to F [69], and this is the approach taken in this work. The MFPT from I to F is then defined by taking the expectation over the initial state distribution. When the state space of a CTMC is large, it can be infeasible to use matrix methods to exactly compute quantities such as the MFPT. One can instead resort to Monte Carlo estimation, e.g., via the *stochastic simulation algorithm* (SSA) [31], although SSA can be prohibitively inefficient for rare event simulation such as in systems with several metastable regions.

2.1.2 Functional form of CTMC rates

A kinetic model with free parameters θ describes the non-equilibrium dynamics of a CTMC via transition rates. We classify elementary transitions into three distinct types: *association*, in which a base pair is formed between two previously separate complexes, *dissociation*, in which a base pair breaks and causes a complex with multiple strands to separate into two distinct complexes, and *isomerisation*, in which a base pair is formed/broken within a complex. The kinetic model in *Multistrand* [59, 60] can be expressed as

$$\ln K(x, x' | \beta, \theta) = \ln \bar{K}(x, x' | \beta, \theta) + \begin{cases} -\beta \cdot \mathbb{1}_{\Delta G(x') \geq \Delta G(x)} (\Delta G(x') - \Delta G(x)), & \text{isomerisation} \\ \ln(u \cdot \alpha_\theta), & \text{association} \\ \ln(u \cdot \alpha_\theta) - \beta \cdot (\Delta G(x') - \Delta G(x)), & \text{dissociation} \end{cases} \tag{4}$$

where the base transition rate function $\bar{K} : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$ and the bimolecular scaling parameter α_θ are parametric choices, x and x' are adjacent microstates, and u is the initial strand concentration. A functional form of $\bar{K}(x, x'|\beta, \theta)$ which is symmetric with respect to x and x' can easily be shown to satisfy *detailed balance*,

$$\frac{K(x, x' \mid \beta, \theta)}{K(x', x \mid \beta, \theta)} = \frac{\pi(x' \mid \beta)}{\pi(x \mid \beta)} = e^{-\beta\cdot\left(\Delta G(x') - \Delta G(x)\right)} \quad \text{for all} \quad x, x' \in \mathcal{X}\,, \tag{5}$$

which is a sufficient condition for recovering the Gibbs-Boltzmann distribution (1) in the steady-state limit.

2.1.3 Context-dependent Arrhenius rates

Kinetic models of Arrhenius type factorise the rate coefficient using a *pre-exponential factor* A and an *activation energy* E that couples to the temperature. They can be parameterised in ways that make the kinetic behavior of an elementary step dependent on local context features surrounding the affected base pair [82, 25], e.g., by assuming multiplicative (log-additive) effects selected by $C : \mathcal{X}^2 \to 2^{\mathcal{C}}$ from a set of transition context features $\mathcal{C}$,

$$\ln \bar{K}(x, x'|\theta, \beta) := \sum_{c \in C(x,x')} \ln A_{\theta,c} - \beta \sum_{c \in C(x,x')} E_{\theta,c}\,, \qquad \theta := (\ln \alpha_\theta, (\ln A_{\theta,c})_{c\in\mathcal{C}}, (E_{\theta,c})_{c\in\mathcal{C}})\,. \tag{6}$$

In the Arrhenius model we consider in this work, $\mathcal{C}$ comprises topological *half contexts*, which refer to base pairing structures on a single side of the affected base pair in an elementary transition [82]. Symmetry of $\ln \bar{K}$ is ensured by applying the same features to the forward and reverse directions of a transition. This model differentiates between seven half contexts[2] which categorise transitions into 28 equivalence classes of *local contexts*[3]. It therefore contains 15 free kinetic parameters, whereas the Metropolis kinetic model originally used in *Multistrand* [59, 60] has only two, $(k_{\text{uni}}, k_{\text{bi}}) \equiv (A_\theta, \alpha_\theta \cdot A_\theta)$.

3 Bayesian model for kinetic parameters

3.1 Prior over kinetic parameters

Our new prior imposes an independent, weak log-normal distribution on the multiplicative rate contribution from each kinetic parameter dimension,

$$\ln \alpha_\theta \sim \mathcal{N}\left(\mu = -2.3,\ \sigma^2 = 800\right) [\ln(M^{-1})]\,, \tag{7}$$
$$E_{\theta,c} \sim \mathcal{N}\left(\mu = 0.0,\ \sigma^2 = 25\right) [\text{kcal/mol}]\,, \quad \ln A_{\theta,c} \sim \mathcal{N}\left(\mu = 7.5,\ \sigma^2 = 110\right) [\ln s^{-1/2}]\,.$$

Assuming fixed thermodynamic parameters, this prior effectively describes a temperature-dependent Gaussian law over the elementary log-rates, and provides support for all values that may be physically possible. It leads to $\ln \bar{K} \sim \mathcal{N}(15.0, 364.8)$ at 25°C and $\ln \bar{K} \sim \mathcal{N}(15.0, 342.3)$ at 50°C, which is compatible with experimental measurements of elementary rates [47, 15] as well as values from past calibration of the Metropolis model [59, 60, 68, 82, 83].

Notably, the pre-exponential factors A_c have non-negative support, and the particularly weak prior for $\ln \alpha$ is centred at $-\ln(10)$, which corresponds to the assumption that the numerical value of the bimolecular elementary rate coefficient is approximately one order of magnitude smaller than the numerical value of the unimolecular rate coefficient, when measured in the standard units of $M^{-1}s^{-1}$ and s^{-1}, respectively [59, 60, 68, 82, 83]. Our weak

[2] $\mathcal{C}$:= {stack, loop, end, stack+loop, stack+end, loop+end, stack+stack}.

[3] The number of unordered pairs of half contexts is the multiset coefficient $\left(\!\binom{7}{2}\!\right) = \binom{8}{2} = 28$.

assumptions for the parameters E_c are that each elementary scale Arrhenius activation energy should be less, in magnitude, than experimentally estimated macroscale activation energies [9, 3, 50, 52], and that the elementary scale and macroscale activation energies will be closer in value for simple reactions such as hairpin closing/opening, helix association/dissociation and bubble closing, especially when the strands are short. Furthermore, centering E_c at zero amounts to a regularisation towards the functional form of the original Metropolis model.

3.2 Likelihood of observations

Our observation model for macroscopic rates follows an approach that was previously used for posterior inference [82] and for maximum likelihood estimation [82, 83] of elementary step rates. Given a kinetic parameterisation θ and a state space $\mathcal{X}$, a *deterministic* rate prediction for a reaction is simulated as the inverse of the expected MFPT τ_F from the initial state distribution π_0, representing reactants, to the non-empty final region F, representing products. More precisely, the rate coefficient prediction $\hat{k}$ depends on the MFPT as:

$$\hat{k} = \begin{cases} \dfrac{1}{\mathbb{E}_{\pi_0}[\tau_F \mid \mathcal{X}, \beta, \theta]} & \text{for reactions of the form } A \to B \text{ or } A \to B + C \\ \dfrac{1}{u\ \mathbb{E}_{\pi_0}[\tau_F \mid \mathcal{X}, \beta, \theta]} & \text{for reactions of the form } A + B \to C \text{ or } A + B \to C + D. \end{cases} \tag{8}$$

The former case describes unimolecular reactions, for which $\hat{k}$ has the meaning of a first-order reaction rate coefficient. Its estimate is based on the assumption that there are no intermediate association steps along the reaction pathway, which holds in our dataset. The latter case describes bimolecular reactions, for which $\hat{k}$ has the meaning of a second-order reaction rate coefficient[4]. By dimensional analysis, it requires a concentration quantity to relate to a time quantity. In general, second-order rates cannot be determined directly from the CTMC model of the *Multistrand* simulator, but the simple estimate above, which uses the initial concentration u of reactants, is justified in the limit of low concentrations, where the initial association step of second order is rate-limiting for the full reaction[5].

For each reaction r, our noise model then centers a log-normal distribution at the predicted rate coefficient to obtain a probabilistic synthetic observation for the log-rate coefficient, i.e., $\log_{10} k_r \sim \mathcal{N}(\log_{10} \hat{k}_r, \sigma_r^2)$, where the variance σ_r^2 is taken as the experimental variance of the log-rate coefficients among the reactions of the same *group*. A group of reactions is defined by its experimental publication and reaction type, and corresponds to a row in Appendix Table 3. In the case where $\hat{k}$ is a non-physical prediction, which occurs when the sparse solver for the MFPT fails or produces a negative or infinite prediction, we instead apply a constant likelihood close to zero, $\mathcal{N}(5\sigma_r^2 \mid 0, \sigma_r^2)$. It should be noted that this noise model is an *ad hoc* choice for constructing a likelihood kernel and, in its current form, cannot be understood as a physically motivated generative model.

[4] Note that the experimental rate coefficients in our dataset were not necessarily computed under the same assumptions, and, for instance, we have several strand displacement reactions in our dataset whose rate coefficients were estimated with first order fits. This warrants reconsideration in future work, particularly if new reactions at higher concentrations are included in the dataset.

[5] See [59, Ch. 7,8] for a discussion about estimating second-order rates from *Multistrand* simulation statistics, and [82, Sec. 5.2] for the modelling assumptions in (8).

3.3 Approximations of the intractable likelihood

In order to *truncate* the intractable state space underlying the *Multistrand* simulator, we employ the deterministic method in [82], which we call *Assumed Pathway* (AP), and the more recent stochastic method *Pathway Elaboration* (PE) [83]. A distinct truncated state space $\mathcal{X}$ is precomputed once for each reaction and is treated as fixed during inference.

The states included in the AP approximation are the non-pseudoknotted secondary structures whose base pairs occur in either the initial or final secondary structure, and which are reachable by elementary step transitions where the affected base pair is always at the boundary of a hybridized domain. Thus the AP method only considers a small subset of states and transitions which are assumed to cover the dominant pathway. For instance, in a helix association reaction, any hairpin formations prior to or during helix association would not be modelled, although they could significantly affect reaction rates [27]. To avoid these sorts of limitations, the PE method constructs truncated CTMCs stochastically, using the states found through a succession of distance-biased and unbiased trajectory samples [83].

The criteria for the initial and final regions in our simulations are reported in Appendix Table 1. In many cases, these criteria define regions that are much broader than those considered in previous work [83]. The initial states are treated as Boltzmann distributed according to the thermodynamic model, for which we use Nupack 3.2.2 [76]. The number of states in the PE approximations that satisfy the criteria for our endpoint regions is stochastic: Our simulations led to initial regions with up to 3689 states and final regions with up to 1080 states. In contrast, the AP state spaces include exactly one initial and one final state.

The computational cost of the PE method depends strongly on the choice of hyperparameters and on the size and energy landscape of the true state space. We aim to construct truncated CTMCs using a set of hyperparameters suggested in the original reference ($n_b = 128$, $n_\kappa = 256$, $b = 0.4$, $\kappa = 16ns$) [83]. Each CTMC construction attempt for all reactions in our dataset is given 7 days with at least 10GB of RAM on a single CPU core. However, these resources proved insufficient for our initial choice of hyperparameters for many reactions. We therefore attempted eight different hyperparameter settings, which we report and name in Appendix Table 2. We give preference to the hyperparameter settings in order of the expected resulting state space size. We did not use the δ-pruning step of the PE method, because, in the existing implementation, its computational cost incurred during the CTMC construction did not warrant the speedup that could have been achieved in inference. All PE state spaces were generated with the Metropolis kinetic parameters

$$k_{uni} = 3.61 \times 10^6 \ [s^{-1}], \quad k_{bi} = 1.12 \times 10^5 \ [M^{-1}\, s^{-1}], \tag{9}$$

which are equivalent to the Arrhenius parameters

$$\ln \alpha = -3.47 \ [\ln M^{-1}], \quad \forall c \in \mathcal{C}.\ \ln A_c = 7.55 \ [\ln s^{-1/2}], \quad E_c = 0 \ [\text{kcal/mol}] \tag{10}$$

in the sense of Section 2.1.3. These parameters were the result of previous parameter tuning on PE state spaces via gradient-free maximum a posteriori (MAP) optimisation [83].

4 Bayesian inference for kinetic parameters

4.1 Bayesian inference methods

Following the approach taken in [82], we used Markov chain Monte Carlo (MCMC) [18, 53, 42] for approximate Bayesian inference, but with a different implementation choice. In particular, our Bayesian model was expressed in the probabilistic programming framework PyMC [56], using custom operations to construct and solve the sparse linear equations for the MFPT, and the *random walk Metropolis* (RWM) algorithm [48] was used for inference. In addition to the inference algorithm implementation, PyMC provides a number of standard tools [7, 55] for diagnosing the behaviour of MCMC samplers, e.g., trace plots and the *effective sample size* (ESS) estimated using Geyer's initial monotone sequence criterion [29, 30].

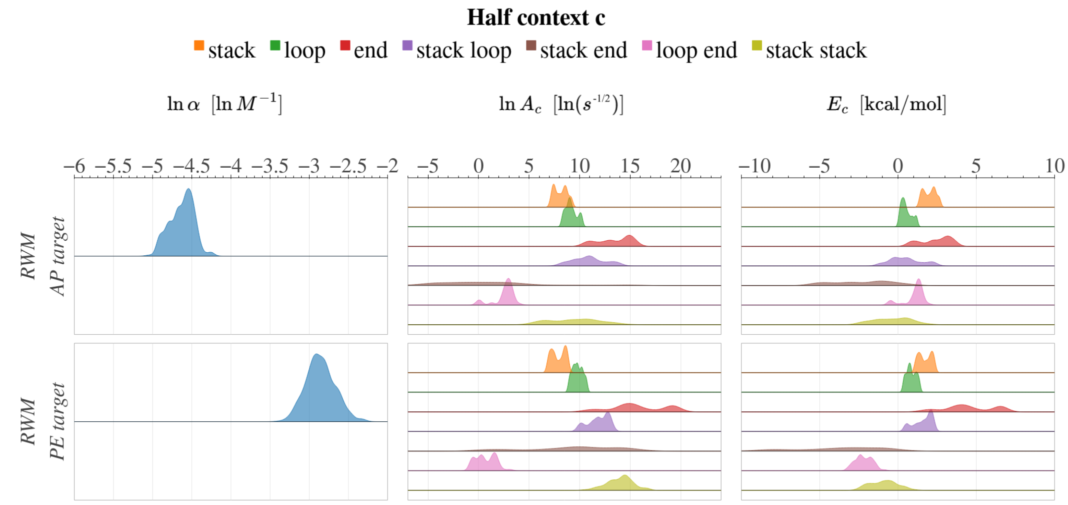


Figure 1 Posterior densities over the Arrhenius kinetic parameters in Section 2.1.3. Approximations are obtained from 600 samples using RWM, on AP and PE inference targets as defined in Section 4.2. The corresponding trace plots are shown in Appendix Figure 7.

4.2 Dataset and likelihood approximation results

The dataset used for parameter inference, summarised in Appendix Table 3, consists of 1105 DNA reactions and includes hairpin opening/closing [9, 8, 37], bubble closing [3], helix association/dissociation [49, 52], and three-way strand displacement [52, 77, 44][6]. We define the following two posterior inference targets, which use different combinations of data subsets and state space approximations.

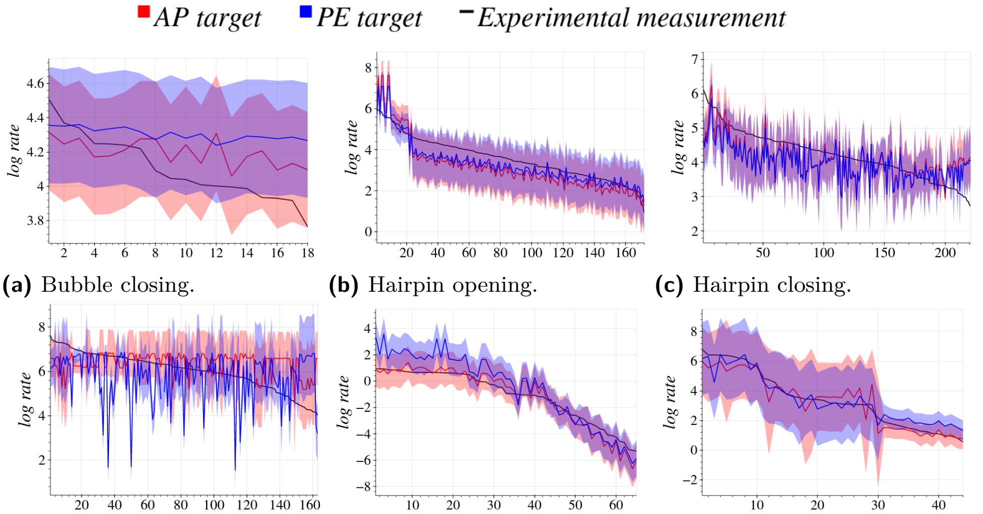
1. *AP target*: All 1105 reactions, using state spaces constructed by the AP method.
2. *PE target*: Only 683 reactions, using all the valid CTMCs that could be constructed by the existing PE implementation, with a preference over hyperparameter values as described in Section 3.3.

These inference targets allow us to indirectly compare the state space approximation methods through their behaviour during inference. For each inference target, the dataset and state spaces are summarised in Appendix Table 4. The *stack* and *loop* half contexts together account for more than 80% of the half context occurrences in each group of truncated CTMCs.

4.3 Posterior approximation results

We ran RWM for 800 total steps, 200 of which were discarded as burn-in. These choices, which proved sufficient for an analysis of the current model and its most significant bottlenecks, were motivated by computational resource constraints and past posterior inference attempts [83].

[6] The dataset of reaction rate coefficients was collected by Sedigheh Zolaktaf. A small set of 14 four-way strand exchange reactions were also collected [15], which we exclude due to computational limitations of the PE implementation. Small subsets of the collected data have been used for parameter tuning or Bayesian inference in previous work [82, 84, 83].



(a) Bubble closing. **(b)** Hairpin opening. **(c)** Hairpin closing.

(d) Helix association. **(e)** Helix dissociation. **(f)** Strand displacement.

Figure 2 Posterior predictive distributions for each reaction type, using the AP and PE likelihood approximations and the corresponding posteriors in Figure 1. The horizontal axis is the reaction index, ordered by decreasing experimental reaction rate. Red and blue solid lines show the mean of the posterior predicted log-rates, and the shaded regions are the 4-96 percentile ranges. We only display the 683 reactions that appear in both inference targets. Both predictive distributions were approximated by taking 100 samples from the likelihood kernel for each posterior sample.

The hyperparameter settings for the RWM sampler were chosen such that approximately the same number of sparse matrix solves are performed in each posterior inference attempt. We designated the kinetic parameters in Equation (10), which are in the 2-dimensional subspace of the Metropolis model, as the seed for all MCMC experiments, expecting that this initial point will mostly yield CTMCs that are physically possible and numerically stable.

The resulting posterior densities are shown in Figure 1. While distinctly multimodal in many dimensions, the shapes of the marginal posterior approximations for each $\ln A_c$ roughly mirror the shapes of the corresponding E_c, indicating a strong correlation between the kinetic parameters associated with the same half context. When comparing posterior results from different chains, the bimolecular scaling parameter $\ln \alpha$ appears to be multimodal, and seems to correlate with the *loop-end* and *stack-stack* half contexts, which are of low frequency in our CTMCs. There has only been one other reported attempt at Bayesian inference on the Arrhenius parameters [82], which used a smaller dataset of 376 reactions and AP state spaces. Despite our different prior, likelihood width, dataset, and inference method, most of our high density intervals for the $\ln A_c$ and E_c dimensions overlap with those reported in [82]. Furthermore, their posterior correlation matrix reflected a correlation structure between corresponding $\ln A_c$ and E_c dimensions that is qualitatively similar to ours.

In general, the approximation quality of Bayesian inference is determined by a complex interaction of the inference method and its hyperparameters with the forward model and the dataset, and each component should be assessed as a potential cause for poor behaviour in inference [28]. As a first step of sampling diagnostics, the RWM trace plots in Appendix Figure 7 display varied behaviour across different half contexts. For example, the *stack-stack* parameters are consistently explored much more broadly than the *stack* and *loop* parameters. In contrast, predictive checks are useful for understanding a prior or a posterior in terms of the distribution of predictions it generates [10]. In Figure 2, we compare the posterior predictive distributions from RWM using the AP and PE targets. Even though the PE state spaces often cover more of the energy landscape, their posterior predictive log-rates are not consistently more accurate than those resulting from the AP state spaces. For further quantitative analyses of the inference results we refer to [43].

4.4 Discussion

4.4.1 Next steps for the kinetic model

The kinetic parameter dimensions that are least explored in our RWM chains (Appendix Figure 7, rows 1-3) roughly correspond to the half contexts that occur most frequently. This might suggest that some of the half contexts are underspecified, and that the multimodalities in the posterior approximations (Figure 1, rows 1-3) arise from significant differences in kinetic behaviour based on features that are not made explicit by the current kinetic model. It would therefore be worth considering kinetic models that partition the transitions into more fine-grained equivalence classes. For instance, the half contexts could be defined in a way that accounts for stack and loop sizes, or for base identities. These refinements could also improve posterior rate predictions, particularly for hairpin closing reactions, whose rate coefficients are not well-captured in our current model (see Figure 2c), and whose experimental rate measurements suggest sequence-dependent behavior [32].

4.4.2 Next steps for MCMC inference

Because RWM consistently recovers an expected correlation structure that is not incorporated into the prior or the proposal, significant sampling effort is spent discovering these correlations. The sample efficiency could be improved by using *Gibbs sampling*, in which a single sample from the target is constructed by iteratively drawing from the conditional distributions of parameter dimensions, while treating all other parameters as observed. Sampling from such intractable conditional distributions is often performed via nested Metropolis-Hastings steps, and the procedure is known as *Metropolis within Gibbs*. It might also be beneficial to partition the parameters into *Gibbs blocks*, containing subsets of mutually highly correlated parameters which are proposed jointly. In our case, the RWM posteriors suggest grouping $(\ln A_{\theta,c}, E_{\theta,c})$ for each $c \in \mathcal{C}$.

4.4.3 Next steps for the likelihood formulation

Within the high density intervals of our posterior approximations, the posterior rate predictions are more strongly influenced by the state space approximation than by the kinetic parameters. This suggests that our current likelihood model cannot further distinguish between different kinetic parameters on the state spaces considered. Moreover, while the PE state spaces cover a higher proportion of the full energy landscapes than the AP state spaces, they do not consistently yield more accurate posterior rate predictions, as indicated by Figure 2. We attribute this finding primarily to the high proportion of sparse linear solver failures or non-physical solutions during inference, which arise from ill-conditioned MFPT equations and for which we apply a constant likelihood close to zero (see Section 3.2). Hence, while the MFPT equations evaluated at our posterior samples yield valid solutions, we expect that the truncation-dependent and solver-dependent likelihood term biases the information extracted during inference from the experimental data. A more detailed quantification of the numerical issues around PE, MFPT equations and posterior inference can be found in [43].

This bottleneck cannot easily be resolved by expanding the dataset or by increasing the approximation quality of the truncated state spaces. The numerical stability of the MFPT computation could in principle be improved by using an iterative solver with suitable preconditioning techniques that might render more of the systems solvable. However, regardless of the preconditioner, the current way of estimating the rate coefficients via the MFPT is a simple, scalar observation model which conceals some transient kinetic effects and which ignores the variety of regression techniques used to estimate reaction rates from experimental observations. Hence, in addition to improved kinetic model features and state space approximations, a reformulation of the observation model to better incorporate transient observations appears prudent.

4.4.4 Software

Multistrand is an efficient and general forward simulator, but its implementation is not amenable to the inverse problem of parameter inference and does not support flexible parallelism. Any task that would require online updates to the state probabilities or transition rates cannot be achieved without considerable overhead in software development and resource usage. This includes forward simulation via replica exchange, Bayesian model averaging, and kinetic parameter inference methods that sample trajectories in an inner loop.

Python libraries used for the work in Section 4 include PyMC [56], Aesara [17, 70], ArviZ [38], Xarray [35], Joblib [71], SciPy, and UMFPACK [72, 16]. At the time of publication of this manuscript, a new official minor release of *Multistrand* will port it to Python 3 and will provide an Apptainer/Singularity container [39], making it simple to run on Linux host systems including HPC clusters. This will be accompanied by a software release specific to this work[7], consisting of the posterior approximations in Section 4.3 and the post-processing scripts used for the *NUPACK* and *Multistrand* simulations in Section 5.

5 Case study

5.1 Motivation

Among the reaction types in the dataset, the most pronounced difference between the predictions made by the AP and PE state spaces is on the helix association reactions in Figure 2d, suggesting that the additional secondary structures in the latter significantly affect kinetic behavior. Many studies have assessed secondary structure effects on the reaction rate of hybridization [27, 61, 33], although they typically extrapolate kinetic behaviour from thermodynamic properties.

This case study focuses on a recent set of 47 hybridization reactions by Hata et al. [33], which we will assess using PE state spaces and the mode of our new posterior distribution over kinetic parameters. The examined sequences have equal length and similar melting temperatures, and were designed to avoid very stable secondary structures as well as stable misnucleation and mishybridization. Nevertheless, at a fixed temperature of 25° C and single-strand concentration of 50 nM, the empirically estimated rate constants varied by more than two orders of magnitude. The authors suggest that decreases in hybridization rates can be explained by decreases in nucleation rates, caused by intra-molecular base pairs that, although not necessarily thermodynamically stable, render nucleation sites inaccessible.

[7] To be found at: `https://github.com/UBC-Mol-Prog/hybridization-profiling`

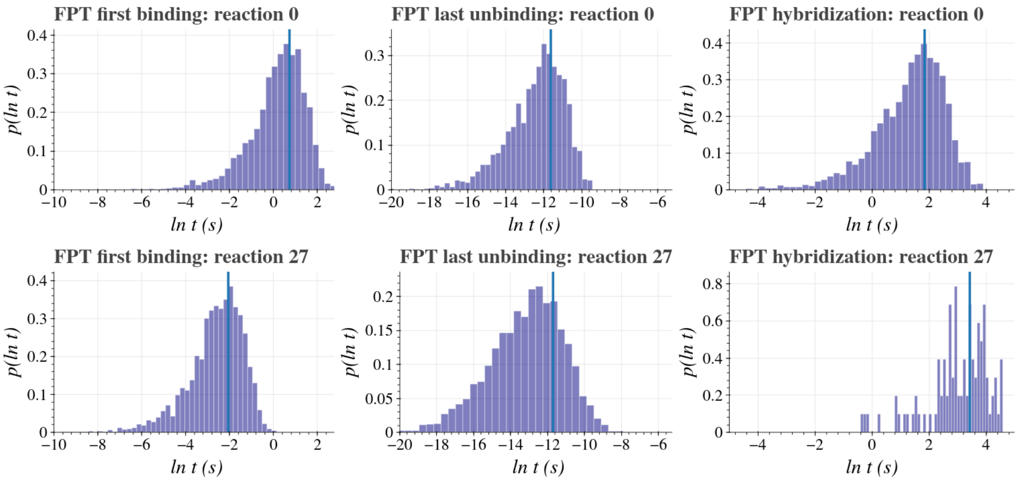


Figure 3 Samples of the first passage times for the first binding, last unbinding, and hybridization times for reactions no. 0 and no. 27 from [33]. The x-axis is the simulated ln-time in seconds, and the blue vertical lines indicate the simulated ln-MFPT. For each event, we attempted to gather 3000 samples using KPS within 24 hours, with 30GB of RAM and 20 CPUs. However, it was only possible to generate 102 samples for the hybridization of reaction no. 27.

The authors also estimated that mishybridized secondary structures (wherein an unwanted stack of successive base pairs forms between strands) were unstable and infrequent, and therefore that their effect on the overall kinetics was negligible.

To assess these hypotheses, we first employed the kinetic path sampling (KPS) implementation in DISCOTRESS [66, 67] to sample times from the first passage distributions for forming the first inter-strand bond, for breaking the only inter-strand bond, and for overall hybridization in the PE CTMCs. KPS is an enhanced sampling technique for rare event simulation, and requires a partition of states into communities, which we specified manually[8]. Results of the simulations on reactions no. 0 and no. 27 are given in Figure 3. Although the simulated first binding rate is much faster in reaction no. 27, its simulated total hybridization rate is much slower, which is compatible with experimental estimates. Furthermore, the rate of dissociating from a state with a single base pair is comparable between the two reactions. Therefore, although these rates are important in general, they do not consistently account for experimental differences in hybridization rates. We therefore expect significant sequence-dependent kinetic behavior to occur between the moments of first binding and stable nucleation.

5.2 Boltzmann statistics of initial states

Starting from the classical nucleation-zippering model [1, Ch. 8.2], Hata et al. emphasise the importance of single-strand secondary structures with positive Gibbs free energy of formation. In particular, they argue that the nucleation phase is rate-limiting, and propose an empirical model in which the hybridization rate is proportional to the expected effective nucleation site density of each strand. As a consistency check for this model and similar metrics over the Boltzmann distribution of single-stranded structures, we show in Figure 4 the average number of free bases per strand, the product over the average relative number of

[8] Secondary structures were binned according to the nearest (edit-distance) structure in the AP model.

nucleation sites in each strand (similar to Hata et al. [33]), and the average number of free and compatible (mis-)nucleation sites per strand pair. All three metrics demonstrate some positive correlation with the experimental rate, although the correlations appear too weak to validate a kinetic model for hybridization based solely on Boltzmann statistics over the unbound states.

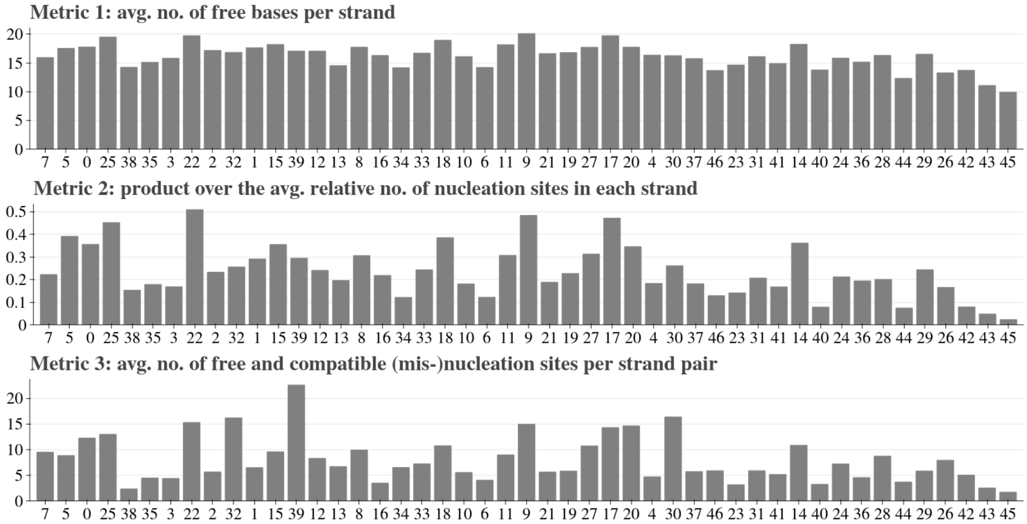


Figure 4 Secondary structure metrics[9] over the Boltzmann distribution of single-stranded structures, estimated from 10k samples of each strand in *NUPACK*. Reactions are ordered by decreasing experimental reaction rate and labeled using the same indices as in the reference [33].

5.3 Trajectory statistics after initial binding

Therefore, to assess potential secondary structure effects beyond the first moment at which the two complementary strands bind, we employ *Multistrand*'s "first step mode" (FSM), stopping the trajectory when the strands either fully dissociate or fully hybridize. Simulations are performed at the same strand concentration and temperature as the physical experiments. To analyze the trajectories, we define 8 different state types, which partition the secondary structures based on occurrences of hairpins[10], correctly hybridized stacks[11], and/or mishybridized stacks[12]. We label each of these types by three characters, which indicate whether there is at least one correctly hybridized stack (**S**) or not (**0**), at least one mishybridized stack (**M**) or not (**0**), and at least one hairpin (**H**) or not (**0**). For instance, **SM0** represents the type of states with at least three consecutive correctly hybridized base pairs, at least three consecutive mishybridized base pairs, and at most two consecutive intra-strand base pairs. Using these definitions, results of our kinetic simulations are provided in Figure 5. The proportion of first step trajectories ending in hybridization varies over

[9] **1.** $\frac{1}{2}\left(\mathbb{E}_{\pi_0^A}\left[n_b^A\right] + \mathbb{E}_{\pi_0^B}\left[n_b^B\right]\right)$, **2.** $\mathbb{E}_{\pi_0^A}\left[\frac{n_s^A}{n_s^{A,\max}}\right] \cdot \mathbb{E}_{\pi_0^B}\left[\frac{n_s^B}{n_s^{B,\max}}\right]$, **3.** $\mathbb{E}_{\pi_0^A \times \pi_0^B}\left[n_p^{A,B}\right]$, where for a strand A with complement B, π_0^A is the Boltzmann distribution over secondary structures, n_b^A is the no. of free bases, n_s^A is the no. of free nucleation sites, $n_s^{A,\max}$ is the max of n_s^A (and analogously for B), and $n_p^{A,B}$ is the no. of free and compatible (mis-)nucleation sites for the pair (A, B).

[10] 3+ consecutive base pairs occurring within a strand

[11] 3+ consecutive base pairs occurring between the two strands at the desired site

[12] 3+ consecutive base pairs occurring between the two strands in an undesired site

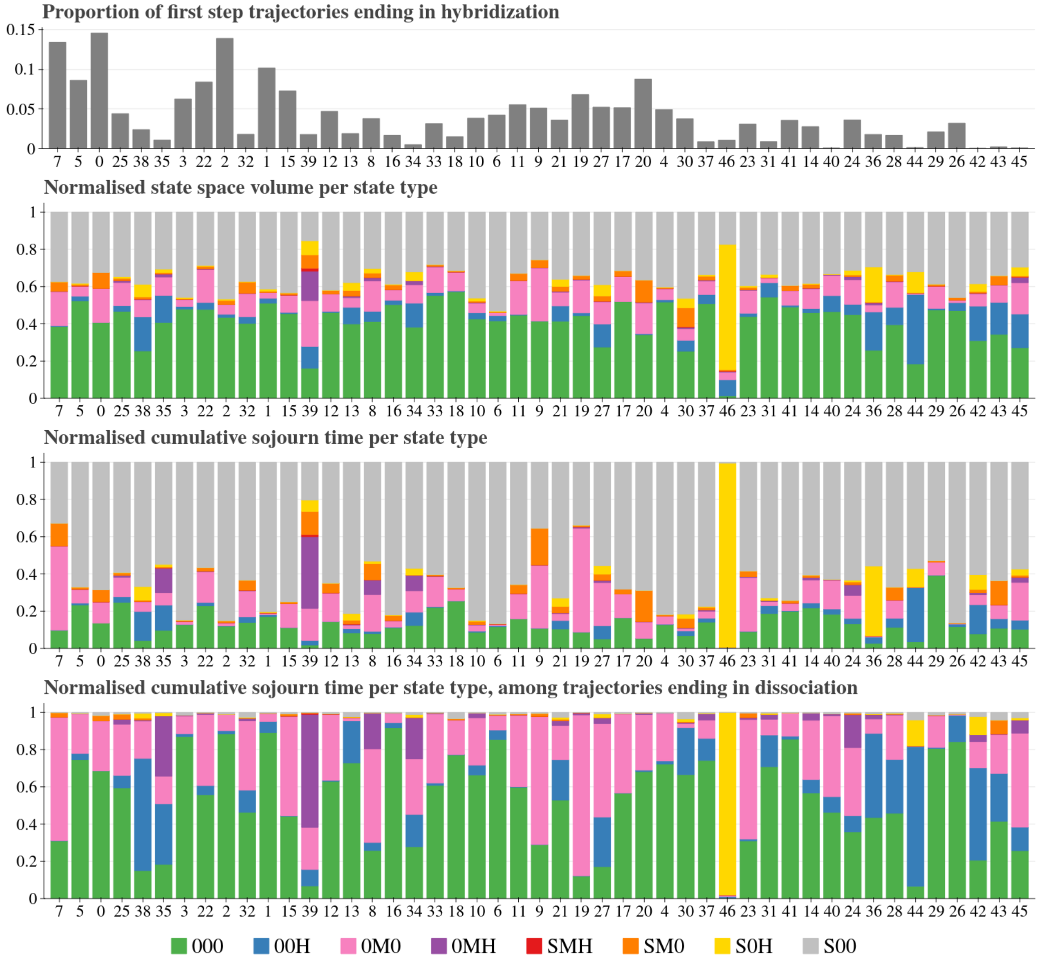


Figure 5 Results of FSM simulations in *Multistrand*. Reactions are ordered by decreasing experimental reaction rate. For each reaction, we accumulate reactive and non-reactive trajectory samples until at least 5 million elementary steps (maximum 155 million) and at least 500 hybridization trajectories (maximum 2305) have been sampled[13].

two orders of magnitude, although high hybridization proportions do not always correspond to fast reaction rates, such as in reaction no. 35. The proportion of different state types appears relatively consistent across different reactions, with some notable exceptions, such as reactions no. 39 and no. 46. In most reactions, we observe a high proportion of states with mishybridized stacks, which is contrary to one of the hypotheses in the experimental source [33]. Furthermore, the total time spent in conformations with misnucleation and/or hairpins is often significant, indicating that the mishybridized states, although potentially thermodynamically unfavourable, behave as kinetic traps. A small number of reactions, such as no. 1 and no. 3, appear to have folding pathways dominated by desired stacks, which is the underlying assumption in the AP model, while other reactions, such as no. 42 and no. 44, appear to have a high proportion of dissociating trajectories caused by simple hairpin formation after first binding, a phenomenon explored in other studies [27, 61].

[13] With the exception of reaction no. 46, whose hybridization trajectories are much longer. Our estimates therefore only include the 26 successful trajectories that could be stored using 30G of RAM, and contained 4.5 million secondary structures on average.

In Figure 6, the time spent in each state type is shown for 20 successful hybridization trajectories in six different reactions. These trajectories illustrate a high degree of sequence dependence, as well as a high variance of reactive pathways within reactions. Overall, these findings suggest that short sequence-level features, such as the nucleation capability proposed by Hata et al. [33], can be influential both before and after the first binding, and that in order to reach higher accuracy, simplified kinetic models of hybridization should in general consider misnucleation and hairpins in their choice of transition states. More generally, while the state predicates above are useful for assessing the reaction pathways from stochastic simulation, the correlation between the considered state predicates and the experimental rates appears too weak, and the relative strength of the considered kinetic effects too varied across reactions, to motivate or validate simplified mechanistic models for hybridization that are derived solely from state-based features. This reinforces the need for elementary step kinetic simulation methods that directly enable the parameterisation and estimation of local and global transient behaviour.

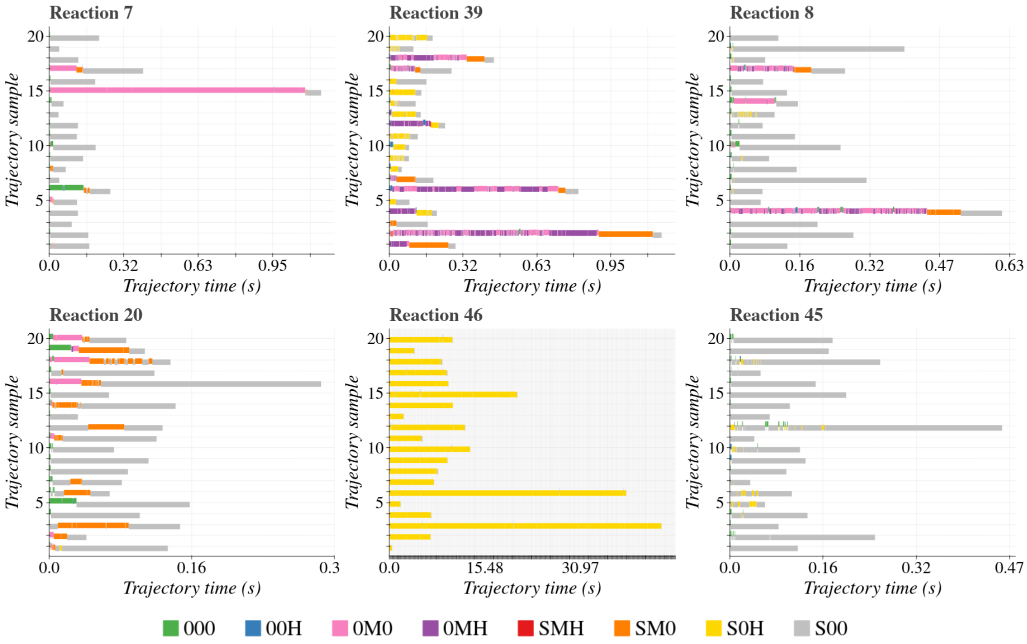


Figure 6 Examples of successful hybridization trajectories sampled using *Multistrand*'s FSM.

6 Conclusion and outlook

In this work, we extended a previous Bayesian inference approach to the calibration problem of an Arrhenius-type model of elementary step DNA kinetics. We introduced a new prior distribution over the desired kinetic parameters and expanded the training dataset to over 1000 reactions, using truncated state spaces generated with the *Assumed Pathway* (AP) method for the full dataset, and the *Pathway Elaboration* (PE) approach when computationally feasible. Posterior inference was performed with the *random walk Metropolis* (RWM) algorithm.

Our posterior distributions, though only preliminary in terms of hyperparameter tuning and convergence analysis, display expected strong correlations between physically tightly coupled kinetic parameters, and thus establish compatibility with past work. The behavior of our MCMC chains varied significantly across the dimensions of our kinetic parameters, and correlated with the half context frequency in our truncated state spaces, suggesting that some half contexts are underspecified. Hence, it would be worth extending the dataset with new reaction types contributing to the transition classes of currently low frequency, as well as developing kinetic models that partition the transitions into more fine-grained equivalence classes. Parameters for transition contexts which are prevalent in the energy barrier regions, e.g., as suggested by the trajectory analyses in Section 5 for hybridization, can be expected to be particularly influential for the overall kinetics. Another important finding is that the posteriors obtained for the AP and PE targets are visibly different in Figure 1, but still produce remarkably similar predictive distributions over reaction rates in Figure 2. This suggests that the posterior approximations are concentrated towards a parameter region in which the rate predictions are more strongly influenced by the state space approximation than by the kinetic parameters. It would therefore be worth improving the likelihood approximation method, such that the forward model dynamically regenerates the state space with different parameters, rather than using a constant state space truncation.

Our results also reveal severe, previously undocumented numerical instability in the current likelihood model, which predicts the MFPT by solving an often ill-conditioned linear system. Effectively, this is a truncation-dependent and solver-dependent likelihood term which penalises parts of the parameter space, influencing in a nontrivial way both the exact posterior and its numerical approximation. In principle, this issue could be mitigated by a suitably preconditioned iterative solver. However, this ill-conditioning is intrinsic for metastable systems, and problem-specific preconditioners have not yet been developed, beyond rescaling by the equilibrium distribution. Furthermore, the MFPT observation model might in general be too low-dimensional and even misspecified with respect to various regression methods used to extract reaction rates from experimental measurements.

Despite these challenges, we were able to use our new Arrhenius parameters to suggest why rates of helix association reactions vary by two orders of magnitude, even when the interacting strands have little or no stable intra-strand secondary structure. This shows the potential of elementary step models for gaining insight into the kinetic behavior of nucleic acid reactions, once properly calibrated and augmented with effective tools for truncation and coarse-graining. In future work, it would be valuable to implement a simulator that allows mathematically separate model components (e.g., experimental conditions, state spaces, initial distributions, final regions, thermodynamic and kinetic models) to be defined and parameterised independently, that supports flexible parallelism, and that provides enhanced path sampling. With such capabilities, the elementary step model could become a standard tool for designing sequences in a way that accounts for transient behavior, marking a significant improvement over thermodynamic sequence design techniques.

References

1 Victor a Bloomfield, Donald M. Crothers, and Ignacio Tinoco. *Nucleic Acids: Structures, Properties and Functions.* University Science Books, Sausalito, Calif, 2000.

2 Daniel P. Aalberts, John M. Parman, and Noel L. Goddard. Single-strand stacking free energy from DNA beacon kinetics. *Biophysical Journal*, 84(5):3212–3217, 2003. doi:10.1016/S0006-3495(03)70045-9.

3 Grégoire Altan-Bonnet, Albert J. Libchaber, and Oleg Krichevsky. Bubble dynamics in double-stranded DNA. *Physical Review Letters*, 90(13):138101, 2003. doi:10.1103/PhysRevLett.90.138101.

4 Yaniv Amir, Eldad Ben-Ishay, Daniel Levner, Shmulik Ittah, Almogit Abu-Horowitz, and Ido Bachelet. Universal computing by DNA origami robots in a living animal. *Nature Nanotechnology*, 9(5):353–357, 2014. doi:10.1038/nnano.2014.58.

5 Jonathan Bath, Simon J. Green, and Andrew J. Turberfield. A free-running DNA motor powered by a nicking enzyme. *Angewandte Chemie International Edition*, 44(28):4358–4361, 2005. doi:10.1002/anie.200501262.

6 Jonathan Bath and Andrew J. Turberfield. DNA nanomachines. *Nature Nanotechnology*, 2(5):275–284, 2007. doi:10.1038/nnano.2007.104.

7 Michael Betancourt. A short review of ergodicity and convergence of Markov chain Monte Carlo estimators. *arXiv e-prints*, 2021. doi:10.48550/arXiv.2110.07032.

8 Grégoire Bonnet. *Dynamics of DNA Breathing and Folding for Molecular Recognition and Computation*. PhD Thesis, Rockefeller University, 2000.

9 Grégoire Bonnet, Oleg Krichevsky, and Albert Libchaber. Kinetics of conformational fluctuations in DNA hairpin-loops. *Proceedings of the National Academy of Sciences*, 95(15):8602–8606, 1998. doi:10.1073/pnas.95.15.8602.

10 George E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):383–404, 1980. doi:10.2307/2982063.

11 Ronald R. Breaker and Gerald F. Joyce. The expanding view of RNA and DNA function. *Chemistry & Biology*, 21(9):1059–1065, 2014. doi:10.1016/j.chembiol.2014.07.008.

12 Thomas R. Cech and Joan A. Steitz. The noncoding RNA revolution: Trashing old rules to forge new ones. *Cell*, 157(1):77–94, 2014. doi:10.1016/j.cell.2014.03.008.

13 Arkadiusz Chworos, Isil Severcan, Alexey Y. Koyfman, Patrick Weinkam, Emin Oroudjev, Helen G. Hansma, and Luc Jaeger. Building programmable jigsaw puzzles with RNA. *Science*, 306(5704):2068–2072, 2004. doi:10.1126/science.1104686.

14 Ibrahim I. Cisse, Hajin Kim, and Taekjip Ha. A rule of seven in Watson-Crick base-pairing of mismatched sequences. *Nature Structural & Moleuclar Biology*, 19(6):623, 2012. doi:10.1038/nsmb.2294.

15 Nadine L. Dabby. *Synthetic Molecular Machines for Active Self-Assembly: Prototype Algorithms, Designs, and Experimental Study*. PhD Thesis, California Institute of Technology, 2013. doi:10.7907/T0ZG-PA07.

16 Timothy A. Davis. Algorithm 832: UMFPACK v4.3: An unsymmetric-pattern multifrontal method. *ACM Transactions on Mathematical Software*, 30(2):196–199, 2004. doi:10.1145/992200.992206.

17 Aesara Developers. Aesara. https://aesara.readthedocs.io/en/latest/. Accessed: 2023-04-10.

18 P. Diaconis and L. Saloff-Coste. What do we know about the Metropolis algorithm? *Journal of Computer and System Sciences*, 57(1):20–36, 1998. doi:10.1006/jcss.1998.1576.

19 Hendrik Dietz, Shawn M. Douglas, and William M. Shih. Folding DNA into twisted and curved nanoscale shapes. *Science*, 325(5941):725–730, 2009. doi:10.1126/science.1174251.

20 Shawn M. Douglas, Ido Bachelet, and George M. Church. A logic-gated nanorobot for targeted transport of molecular payloads. *Science*, 335(6070):831–834, 2012. doi:10.1126/science.1214081.

21 Shawn M. Douglas, Hendrik Dietz, Tim Liedl, Bjorn Hogberg, Franziska Graf, and William M. Shih. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature*, 459(7245):414–418, 2009. doi:10.1038/nature08016.

22 Eric C Dykeman. An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update. *Nucleic Acids Research*, 43(12):5708–5715, 2015. doi:10.1093/nar/gkv480.

23 Christoph Flamm, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6(03):325–338, 2000. doi:10.1017/S1355838200992161.

24 Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306, 2013. doi:10.1086/670067.

25 Mark E. Fornace. *Computational Methods for Simulating and Parameterizing Nucleic Acid Secondary Structure Thermodynamics and Kinetics*. PhD Thesis, California Institute of Technology, 2022. doi:10.7907/ayeg-at42.

26 Jinglin Fu, Yuhe Renee Yang, Alexander Johnson-Buck, Minghui Liu, Yan Liu, Nils G. Walter, Neal W. Woodbury, and Hao Yan. Multi-enzyme complexes on DNA scaffolds capable of substrate channelling with an artificial swinging arm. *Nature Nanotechnology*, 9(7):531–536, 2014. doi:10.1038/nnano.2014.100.

27 Yang Gao, Lauren K. Wolf, and Rosina M. Georgiadis. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Research*, 34(11):3370–3377, 2006. doi:10.1093/nar/gkl422.

28 Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv e-prints*, 2020. doi:10.48550/arXiv.2011.01808.

29 Charles Geyer. Introduction to Markov chain Monte Carlo. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, volume 20116022. Chapman and Hall/CRC, 2011. doi:10.1201/b10905.

30 Charles J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992. doi:10.1214/ss/1177011137.

31 Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977. doi:10.1021/j100540a008.

32 Noel L. Goddard, Grégoire Bonnet, Oleg Krichevsky, and Albert Libchaber. Sequence dependent rigidity of single stranded DNA. *Physical Review Letters*, 85(11):2400, 2000. doi:10.1103/PhysRevLett.85.2400.

33 Hiroaki Hata, Tetsuro Kitajima, and Akira Suyama. Influence of thermodynamically unfavorable secondary structures on DNA hybridization kinetics. *Nucleic Acids Research*, 46(2):782–791, 2018. doi:10.1093/nar/gkx1171.

34 Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003. doi:10.1093/nar/gkg599.

35 Stephan Hoyer and Joe Hamman. Xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, 5(1):10, 2017. doi:10.5334/jors.148.

36 Yonggang Ke, Luvena L. Ong, William M. Shih, and Peng Yin. Three-dimensional structures self-assembled from DNA bricks. *Science*, 338(6111):1177–1183, 2012. doi:10.1126/science.1227268.

37 Jiho Kim, Sören Doose, Hannes Neuweiler, and Markus Sauer. The initial step of DNA hairpin folding: a kinetic analysis using fluorescence correlation spectroscopy. *Nucleic Acids Research*, 34(9):2516–2527, 2006. doi:10.1093/nar/gkl221.

38 Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33):1143, 2019. doi:10.21105/joss.01143.

39 Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459, 2017. doi:10.1371/journal.pone.0177459.

40 Martin Langecker, Vera Arnaut, Thomas G. Martin, Jonathan List, Stephan Renner, Michael Mayer, Hendrik Dietz, and Friedrich C. Simmel. Synthetic lipid membrane channels formed by designed DNA nanostructures. *Science*, 338(6109):932–936, 2012. doi:10.1126/science.1225624.

41 Tim Liedl, Björn Högberg, Jessica Tytell, Donald E. Ingber, and William M. Shih. Self-assembly of three-dimensional prestressed tensegrity structures from DNA. *Nature Nanotechnology*, 5(7):520–524, 2010. doi:10.1038/nnano.2010.107.

42 Pavel Loskot, Komlan Atitey, and Lyudmila Mihaylova. Comprehensive review of models and methods for inferences in bio-chemical reaction networks. *Frontiers in Genetics*, 10, 2019. doi:10.3389/fgene.2019.00549.

43 Jordan Lovrod. Bayesian modelling of DNA secondary structure kinetics: revisiting path space approximations and posterior inference in exponentially large state spaces. MSc Thesis, University of British Columbia, 2023. doi:10.14288/1.0431312.

44 Robert R. F. Machinek, Thomas E. Ouldridge, Natalie E. C. Haley, Jonathan Bath, and Andrew J. Turberfield. Programmable energy landscapes for kinetic control of DNA strand displacement. *Nature Communications*, 5(1):5324, 2014. doi:10.1038/ncomms6324.

45 Chengde Mao, Weiqiong Sun, Zhiyong Shen, and Nadrian C. Seeman. A nanomechanical device based on the B–Z transition of DNA. *Nature*, 397(6715):144–146, 1999. doi:10.1038/16437.

46 David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999. doi:10.1006/jmbi.1999.2700.

47 Sean A. McKinney, Alasdair D. J. Freeman, David M. J. Lilley, and Taekjip Ha. Observing spontaneous branch migration of Holliday junctions one step at a time. *Proceedings of the National Academy of Sciences*, 102(16):5715–5720, 2005. doi:10.1073/pnas.0409328102.

48 Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi:10.1063/1.1699114.

49 Larry E. Morrison and Lucy M. Stols. Sensitive fluorescence-based thermodynamic and kinetic measurements of DNA hybridization in solution. *Biochemistry*, 32(12):3095–3104, 1993. doi:10.1021/bi00063a022.

50 G. Eric Plum, Kenneth J. Breslauer, and Richard W. Roberts. 7.02 - Thermodynamics and kinetics of nucleic acid association/dissociation and folding processes. In *Comprehensive Natural Products Chemistry*, pages 15–53. Pergamon, Oxford, 1999. doi:10.1016/B978-0-08-091283-7.00056-4.

51 Brittany Rauzan, Elizabeth McMichael, Rachel Cave, Lesley R. Sevcik, Kara Ostrosky, Elisabeth Whitman, Rachel Stegemann, Audra L. Sinclair, Martin J. Serra, and Alice A. Deckert. Kinetics and thermodynamics of DNA, RNA, and hybrid duplex formation. *Biochemistry*, 52(5):765–772, 2013. doi:10.1021/bi3013005.

52 Luis P. Reynaldo, Alexander V. Vologodskii, Bruce P. Neri, and Victor I. Lyamichev. The kinetics of oligonucleotide replacements. *Journal of Moleuclar Biology*, 297(2):511–520, 2000. doi:10.1006/jmbi.2000.3573.

53 Christian P. Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999. doi:10.1007/978-1-4757-4145-2.

54 Paul W. K. Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006. doi:10.1038/nature04586.

55 Vivekananda Roy. Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7:387–412, 2020. doi:10.1146/annurev-statistics-031219-041300.

56 John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016. doi:10.7717/peerj-cs.55.

57 John SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4):1460–1465, 1998. doi:10.1073/pnas.95.4.1460.

58 John SantaLucia Jr. and Donald Hicks. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004. doi:10.1146/annurev.biophys.32.110601.141800.

59 Joseph M. Schaeffer. *Stochastic Simulation of the Kinetics of Multiple Interacting Nucleic Acid Strands*. PhD Thesis, California Institute of Technology, 2013. doi:10.7907/JEBY-6X69.

60 Joseph M. Schaeffer, Chris Thachuk, and Erik Winfree. Stochastic simulation of the kinetics of multiple interacting nucleic acid strands. In *DNA Computing and Molecular Programming*, volume 9211 of *Lecture Notes in Computer Science*, pages 194–211, 2015. doi:10.1007/978-3-319-21999-8_13.

61 John S. Schreck, Thomas E. Ouldridge, Flavio Romano, Petr Šulc, Liam P. Shaw, Ard A. Louis, and Jonathan P.K. Doye. DNA hairpins destabilize duplexes primarily by promoting melting rather than by inhibiting hybridization. *Nucleic Acids Research*, 43(13):6181–6190, 2015. doi:10.1093/nar/gkv582.

62 Verena J. Schüller, Simon Heidegger, Nadja Sandholzer, Philipp C. Nickels, Nina A. Suhartha, Stefan Endres, Carole Bourquin, and Tim Liedl. Cellular immunostimulation by CpG-sequence-coated DNA origami structures. *ACS Nano*, 5(12):9696–9702, 2011. doi:10.1021/nn203161y.

63 Georg Seelig, David Soloveichik, David Yu Zhang, and Erik Winfree. Enzyme-free nucleic acid logic circuits. *Science*, 314(5805):1585–1588, 2006. doi:10.1126/science.1132493.

64 Nadrian C. Seeman. Nucleic acid junctions and lattices. *Journal of Theoretical Biology*, 99(2):237–247, 1982. doi:10.1016/0022-5193(82)90002-9.

65 Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152(1-2):17–24, 2013. doi:10.1016/j.cell.2012.12.024.

66 Daniel J. Sharpe and David J. Wales. Efficient and exact sampling of transition path ensembles on Markovian networks. *The Journal of Chemical Physics*, 153(2):024121, 2020. doi:10.1063/5.0012128.

67 Daniel J. Sharpe and David J. Wales. Nearly reducible finite Markov chains: Theory and algorithms. *The Journal of Chemical Physics*, 155(14):140901, 2021. doi:10.1063/5.0060978.

68 Niranjan Srinivas, Thomas E. Ouldridge, Petr Šulc, Joseph M. Schaeffer, Bernard Yurke, Ard A. Louis, Jonathan P. K. Doye, and Erik Winfree. On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Research*, 41(22):10641–10658, 2013. doi:10.1093/nar/gkt801.

69 Yuri Suhov and Mark Kelbert. *Markov Chains: A Primer in Random Processes and Their Applications*, volume 2. Cambridge University Press, 2008.

70 Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, 2016. doi:10.48550/arXiv.1605.02688.

71 Gael Varoquaux and Olivier Grisel. Joblib: Running Python functions as pipeline jobs. https://joblib.readthedocs.io/en/latest/, 2009. Accessed: 2023-04-15.

72 Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi:10.1038/s41592-019-0686-2.

73 Mark I. Wallace, Liming Ying, Shankar Balasubramanian, and David Klenerman. Non-Arrhenius kinetics for the loop closure of a DNA hairpin. *Proceedings of the National Academy of Sciences*, 98(10):5584–5589, 2001. doi:10.1073/pnas.101523498.

74 Anthony S. Walsh, HaiFang Yin, Christoph M. Erben, Matthew J. A. Wood, and Andrew J. Turberfield. DNA cage delivery to mammalian cells. *ACS Nano*, 5(7):5427–5432, 2011. doi:10.1021/nn2005574.

75 Erik Winfree. *Algorithmic Self-Assembly of DNA*. PhD Thesis, California Institute of Technology, 1998. doi:10.7907/HBBV-PF79.

76 Joseph N. Zadeh, Conrad D. Steenberg, Justin S. Bois, Brian R. Wolfe, Marshall B. Pierce, Asif R. Khan, Robert M. Dirks, and Niles A. Pierce. NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, 32(1):170–173, 2011. doi:10.1002/jcc.21596.

77 David Yu Zhang and Erik Winfree. Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47):17303–17314, 2009. doi:10.1021/ja906987s.

78 Jinny X. Zhang, John Z. Fang, Wei Duan, Lucia R. Wu, Angela W. Zhang, Neil Dalchau, Boyan Yordanov, Rasmus Petersen, Andrew Phillips, and David Yu Zhang. Predicting DNA hybridization kinetics from sequence. *Nature Chemistry*, 10(1):91–98, 2018. doi:10.1038/nchem.2877.

79 Jinny X. Zhang, Boyan Yordanov, Alexander Gaunt, Michael X. Wang, Peng Dai, Yuan-Jyue Chen, Kerou Zhang, John Z. Fang, Neil Dalchau, Jiaming Li, Andrew Phillips, and David Yu Zhang. A deep learning model for predicting next-generation sequencing depth from DNA sequence. *Nature Communications*, 12(1):4387, 2021. doi:10.1038/s41467-021-24497-8.

80 Yong-Xing Zhao, Alan Shaw, Xianghui Zeng, Erik Benson, Andreas M. Nyström, and Björn Högberg. DNA origami delivery system for cancer therapy with tunable release properties. *ACS Nano*, 6(10):8684–8691, 2012. doi:10.1021/nn3022662.

81 Sedigheh Zolaktaf. *Efficiently estimating kinetics of interacting nucleic acid strands modeled as continuous-time Markov chains.* PhD Thesis, University of British Columbia, 2020. doi:10.14288/1.0395346.

82 Sedigheh Zolaktaf, Frits Dannenberg, Xander Rudelis, Anne Condon, Joseph M. Schaeffer, Mark Schmidt, Chris Thachuk, and Erik Winfree. Inferring parameters for an elementary step model of DNA structure kinetics with locally context-dependent Arrhenius rates. In *DNA Computing and Molecular Programming*, volume 10467 of *Lecture Notes in Computer Science*, pages 172–187, 2017. doi:10.1007/978-3-319-66799-7_12.

83 Sedigheh Zolaktaf, Frits Dannenberg, Mark Schmidt, Anne Condon, and Erik Winfree. Predicting DNA kinetics with a truncated continuous-time Markov chain method. *Computational Biology and Chemistry*, 104:107837, 2023. doi:10.1016/j.compbiolchem.2023.107837.

84 Sedigheh Zolaktaf, Frits Dannenberg, Erik Winfree, Alexandre Bouchard-Côté, Mark Schmidt, and Anne Condon. Efficient parameter estimation for DNA kinetics modeled as continuous-time Markov chains. In *DNA Computing and Molecular Programming*, volume 11648 of *Lecture Notes in Computer Science*, pages 80–99, 2019. doi:10.1007/978-3-030-26807-7_5.

85 Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003. doi:10.1093/nar/gkg595.

A Appendix

A.1 Criteria for initial and final regions of a reaction

A.2 PE method

The PE algorithm has five hyperparameters, b, κ, n_b, n_κ, and δ (referred to as β, κ, N, K, and δ in the original reference [83]). For a CTMC with transition rate matrix K, probability matrix P and final region F, let $\mathbb{1}_{\text{dist}} : \mathcal{X}^2 \to \{0,1\}$ be the decreasing-distance indicator that maps adjacent states x and x' to 1 if the minimum distance (in steps) from x' to any absorbing state $x_f \in F$ is less than the distance from x to x_f, and to 0 otherwise. This indicator is used to define P_{bias}, which alters P by only allowing for transitions that decrease the distance to the final region, and P_b for any $b \in [0,1]$ is taken to be the convex combination of P_{bias} and P:

$$P_{\text{bias}}(x,x') = \frac{K(x,x')\mathbb{1}_{\text{dist}}(x,x')}{\sum_{x''\in\mathcal{X}} K(x,x'')\mathbb{1}_{\text{dist}}(x,x'')}\,, \tag{11}$$

$$P_b = bP + (1-b)P_{\text{bias}}\,. \tag{12}$$

For a given P, we refer to samples from P_b as b-biased. The PE algorithm can be summarised in four steps:

Table 1 Criteria for the initial and final regions. Note that, although some of these definitions allow for endpoint regions containing several secondary structure states, the state spaces constructed by the AP approximation only include one state satisfying each criterion.

Bubble closing	*Initial:* The microstate where all bases in the scope of the specified bubble are unpaired, and all other bases are paired as in the final state *Final:* The microstate with the fully formed hairpin
Hairpin opening	*Initial:* The microstate with the fully formed hairpin *Final:* The microstate with no base pairs
Hairpin closing	*Reverse of hairpin opening*
Helix association	*Initial:* Any microstate with no base pairs between strands *Final:* The microstate with the fully formed helix
Helix dissociation	*Reverse of helix association*
Strand displacement	*Initial:* Any microstate where the incumbent and substrate form a complex without the invader *Final:* Any microstate where the invader and substrate form a complex without the incumbent

1. **Pathway construction.** Sample n_b distance-biased trajectories from initial region I to final region F, such that the holding times are sampled according to the diagonal entries of K, while transitions are sampled according to P_b.
2. **State elaboration.** For each state x discovered during the construction step, sample n_κ unbiaed paths according to K, starting at x with time limit κ.
3. **Transition construction.** Construct a new rate matrix $\hat{K}$ such that for any states x and x' discovered during simulations, $\hat{K}(x, x') := K(x, x')$.
4. **δ-pruning.** Group all states that are within a fixed MFPT δ of any $x_f \in F$ into a single absorbing state, and update the rate matrix $\hat{K}$ accordingly.

The hyperparameters κ and b directly improve the approximation quality of the truncated CTMC, while n_κ and n_b only increase the approximation quality if κ and b are suitable for the CTMC in question. In brief, increases in b decrease the degree of bias in the pathway construction step, increases in κ increase the expected length of each unbiased trajectory sample, and increases in n_b/n_κ increase the number of biased/unbiased trajectory samples, respectively. In particular, when $b = 1$ and $n_\kappa = 0$, SSA is recovered as a special case. In practice, the current implementation of PE is inefficient in time and memory usage, which strongly limits the practically achievable hyperparameters and approximation quality.

A.3 PE hyperparameter value sets

Table 2 Sets of PE hyperparameter values attempted on each reaction. The ordering of the sets of values corresponds to decreasing expected state space size. This order was also used to prioritise which CTMCs to include in our inference targets (Appendix Table 4), for reactions where more than one set of hyperparameters produced a valid CTMC.

	P1B1	**P1B2**	**P2B1**	**P2B2**	**P3B1**	**P3B2**	**P4B1**	**P4B2**
n_b	128	128	64	64	32	32	16	16
n_κ	256	256	128	128	64	64	32	32
b	0.4	0.1	0.4	0.1	0.4	0.1	0.4	0.1
κ	16ns	16ns	16ns	16ns	16ns	16ns	16ns	16ns

A.4 Dataset

Table 3 Dataset of experimentally measured rate coefficients. The sign * next to the number of reactions indicates that the dataset includes mismatch experiments.

Reaction type	**no. reactions**	**no. bases**	**°C**	$\log_{10}\left(k\,[M^{-1}s^{-1}]\right)$	**ref.**
Bubble closing	18	62	21.9 – 48.7	3.8 – 4.5	[3]
Hairpin opening	63	22 – 40	10.3 – 48.8	1.4 – 4.6	[9]
	79	20 – 40	17.8 – 48.7	2.2 – 4.7	[8, 32]
	8	40	9.5 – 45.6	0.9 – 3.1	[73]
	22	8	9.9 – 60.5	4.6 – 6.0	[37]
Hairpin closing	62	22 – 40	10.0 – 49.4	3.4 – 4.8	[9]
	102	18 – 40	14.3 – 48.8	2.8 – 5.3	[8]
	8	40	9.8 – 45.6	3.2 – 3.4	[73]
	22	8	9.9 – 60.6	4.6 – 6.1	[37]
	27	31	14.1 – 41.1	2.7 – 4.0	[2]
Helix association	15	20 – 40	3.4 – 49.3	5.9 – 7.6	[49]
	47	46	25.0	4.0 – 6.7	[33]
	210	72	28.0 – 55.0	4.2 – 7.4	[78]
	39*	18 – 20	23.0 – 37.0	4.2 – 7.4	[14]
	9	50	23.0	4.9 – 6.2	[27]
	18	16	6.6 – 33.6	6.5 – 7.3	[51]
Helix dissociation	12	20 – 40	24.7 – 68.0	-2.7 – -1.0	[49]
	14	42 – 46	30.0 – 55.0	-5.3 – -2.9	[52]
	39*	18 – 20	23.0 – 37.0	-1.2 – 0.9	[14]
Strand displacement	30	78 – 96	25.0	0.9 – 7.0	[77]
	14	54 – 62	30.0 – 55.0	0.6 – 1.9	[52]
	36*	83 – 87	23.0	2.7 – 6.8	[44]
	211	89 – 102	28.0 – 55.0	1.3 – 8.2	[79]
Overall	1105	8 – 102	3.4 – 68.0	-5.3 – 8.2	

A.5 Inference targets

Table 4 State spaces used in each inference target.

	no. reactions	**avg. no. states**
AP target		
Bubble closing	18	758
Hairpin opening	221	46
Hairpin closing	172	44
Helix association	338	1843
Helix dissociation	65	275
Strand displacement	291	9626
Total	1105	3143
PE target		
Bubble closing	18	2048
Hairpin opening	221	3120
Hairpin closing	172	1268
Helix association	163	26113
Helix dissociation	65	23223
Strand displacement	44	27045
Total	683	11567

A.6 MCMC trace plots

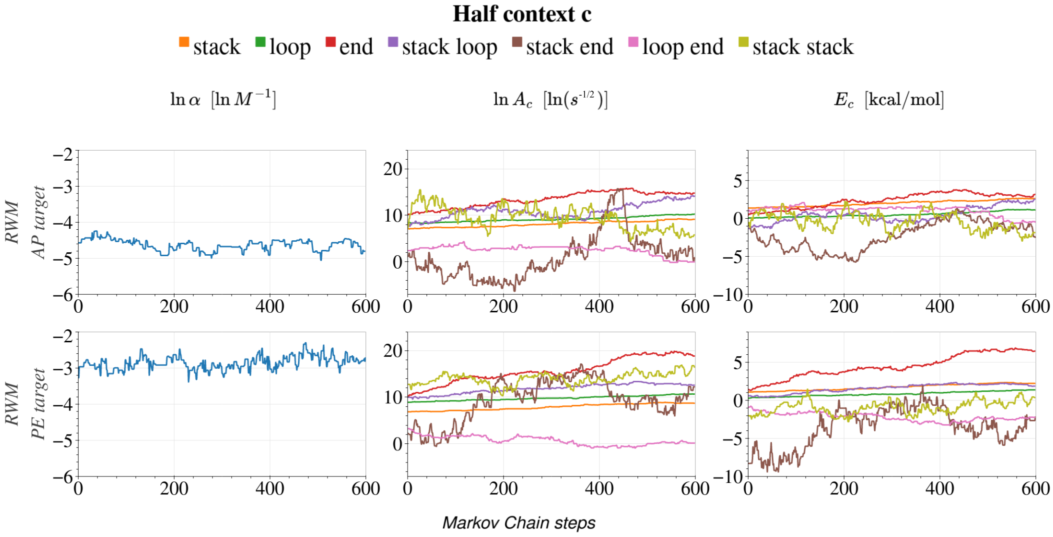


Figure 7 Trace plots from the RWM sampler for AP and PE inference targets, corresponding to Figure 1. Burn-in samples are not shown.