A Universal Method for Analysing Copolymer Growth

Benjamin Qureshi,[1] Jordan Juritz,[1] Jenny M. Poulton,[2] Adrian Beersing-Vasquez,[3] and Thomas E. Ouldridge[1]
[1)]*Department of Bioengineering and Centre for Synthetic Biology, Imperial College London, London SW7 2AZ, United Kingdom*
[2)]*Foundation for Fundamental Research on Matter (FOM) Institute for Atomic and Molecular Physics (AMOLF), 1098 XE Amsterdam, The Netherlands*
[3)]*Faculty of Science, Amsterdam Science Park 904, 1098 XH Amsterdam*

(*Electronic mail: t.ouldridge@imperial.ac.uk)

(Dated: 17 February 2023)

Polymers consisting of more than one type of monomer, known as copolymers, are vital to both living and synthetic systems. Copolymerisation has been studied theoretically in a number of contexts, often by considering a Markov process in which monomers are added or removed from the growing tip of a long copolymer. To date, the analysis of the most general models of this class has necessitated simulation. We present a general method for analysing such processes without resorting to simulation. Our method can be applied to models with an arbitrary network of substeps prior to addition or removal of a monomer, including non-equilibrium kinetic proofreading cycles. Moreover, the approach allows for a dependency of addition and removal reactions on the neighbouring site in the copolymer, and thermodynamically self-consistent models in which all steps are assumed to be microscopically reversible. Using our approach, thermodynamic quantities such as chemical work; kinetic quantities such as time taken to grow; and statistical quantities such as the distribution of monomer types in the growing copolymer can be derived either analytically or numerically directly from the model definition.

I. INTRODUCTION

Copolymers are polymers consisting of more than one type of monomeric unit; the order of these monomers in the chain defines the copolymer *sequence*. Broadly, copolymerisation mechanisms can be classified into two main categories: free copolymerisation that does not rely on a template[1], as shown in figure 1(a); and templated copolymerisation, in which a template (usually another copolymer) is used to bias the distribution of sequences produced, as shown in figure 1(b) and figure 1(c). Polymers produced via both types of mechanism are of relevance to both biological and industrial systems. In living systems, O-glycans are sequences of monosaccharides that grow by free copolymerisation from serine or threonine amino acids[2]. They play a key role as a physical protective barrier for cells from pathogens, as well as participating in other cellular processes[2,3]. Free copolymerisation is also a common method for producing plastics and rubbers in commercial and industrial systems[4,5]. Additionally, there have been recent experimental designs for free copolymerisation systems to produce specific products utilising DNA-nanotechnology-based reaction schemes[6,7].

Templated copolymerisation is the mechanism by which DNA, RNA and polypeptides are produced in DNA replication, RNA transcription, and protein translation, respectively. These processes are at the heart of the central dogma of molecular biology[8] and are the basis of the informational and biochemical complexity of life. In DNA replication, DNA templates the production of copies of itself; in transcription, DNA templates the production of RNA; and, in translation, mRNA is the template for the production of a polypeptide[9]. Inspired by these biological templated copolymerisation mechanisms, there has been recent interest in designing synthetic systems that can produce other sequence-controlled molecules via templated copolymerisation[10–16].

Free polymerisation can be modelled as a Markovian growth process under which monomers bind to the end of a growing polymer at a certain rate. Early free copolymerisation models[17–19] built on this framework to allow for copolymerisation via the incorporation of multiple types of monomeric unit, as shown in figure 1(a), albeit with irreversible polymerisation reactions. In particular, Mayo and Lewis[19] emphasised that in polymerisation models, if the monomer binding events are irreversible and their rates are conditional on the terminal monomer type, then intra-sequence correlations are generated within the copolymer.

Although the use of models with irreversible transitions is reasonable in many contexts, thermodynamically self-consistent models require all transitions to be microscopically reversible.[20] Specifically, if a transition from state A to state B is possible, then transitions from B to A must also be possible. Models with fully microscopically reversible polymerisation reactions, as in figure 1(a), are more challenging to analyse but can be interpreted in a thermodynamic sense.[1,21,22]

Templates can affect the rate at which monomers are added or removed from a growing copolymer, and hence templated copolymerisation models can be more complex than free copolymerisation models. When the template consists of just one type of templating monomer (homopolymer), a templated copolymerisation process can be mapped onto a free copolymerisation model. Further, if one assumes some symmetries regarding interactions between monomers in the growing copolymer and those in the template (such as all complementary bonds have equal strength and all non-complementary bonds have equal strength), models of sequence-bearing templates may be mapped onto models with homopolymeric templates, and hence to models of free copolymerisation[23–27].

Templated copolymerisation models can be further divided into two main categories: templated self-assembly (figure 1(b)[1,27–41] and autonomously-separating mechanisms (fig-

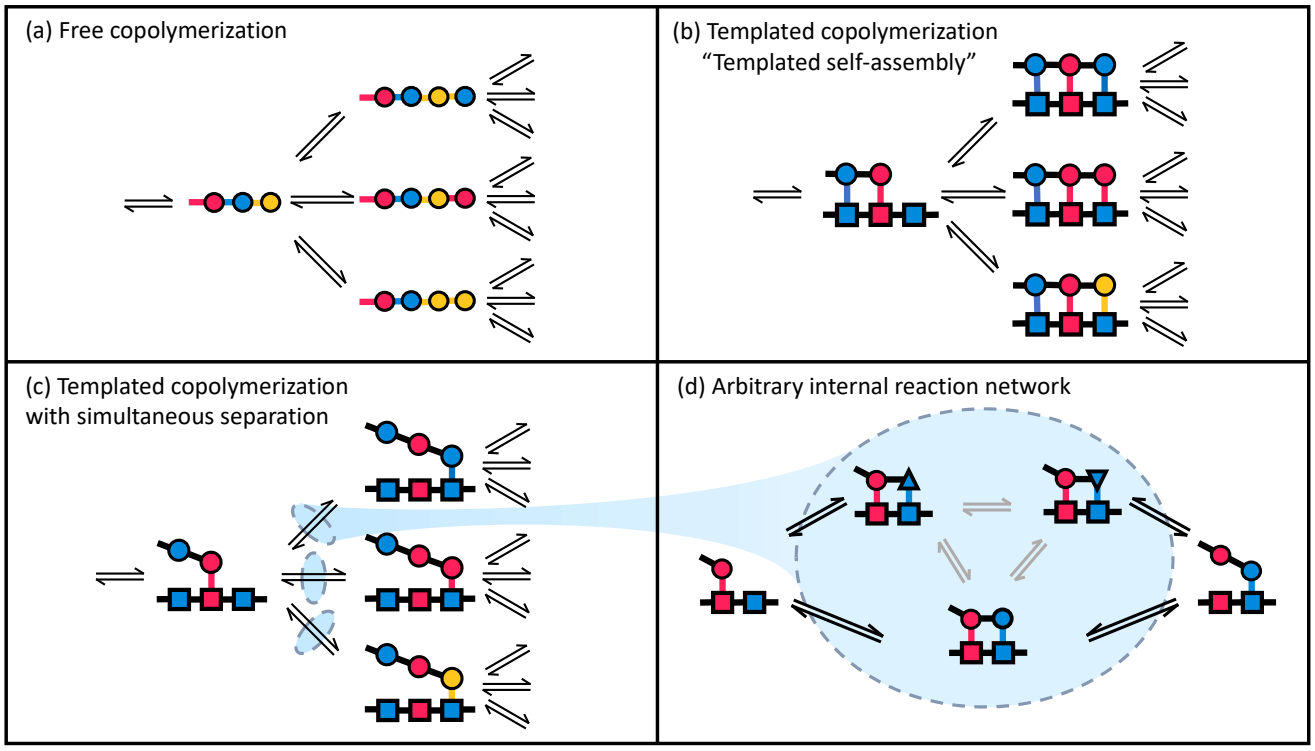


FIG. 1: Comparison of the different types of copolymerisation mechanism with three types of monomer (blue, red and yellow). (a) shows free copolymerisation, (b) templated self assembly, and (c) templated copolymerisation with autonomous separation. In (b) and (c) the template is shown with squares and the growing polymer with circles. (d) An example of a more detailed reaction scheme used to select the next monomer. In each of these sub-figures, different colours represent different monomer types, with bonds coloured accordingly when their strength might depend on the monomer. In (d), different activation states of the monomer undergoing incorporation are represented by different shapes. The dashed bubble indicates how the arbitrary set of reactions in (d) may replace the simple reaction surrounded by a dashed bubble in (c) for a more complex model.

ure 1(c)[23,24,42]. Templated self-assembly models are those in which all the monomers in the growing copolymer remain bound to the template. In autonomously-separating models, the growing copolymer detaches as it extends[23–25]. There has been recent interest in explicitly modelling autonomous separation in templated growth in an attempt to understand models that give a better description of transcription or translation[23–25]. In autonomously-separating models, the simultaneous growth and separation of the copolymer and template mean that the copy-template interactions are not permanent, and therefore free energy released from such interactions cannot be part of the driving force of polymerisation. Additionally, since these copy-template bonds are temporary, they cannot stabilise the accurate copy directly in the long time limit. Further, an ensemble of accurate polymers is a lower entropy state than an ensemble of random polymers. These conditions mean that non-equilibrium driving is required to generate accurate copies of the template if the copies are to spontaneously detach[42]. Moreover, the separation of the lagging tail from the template as the copolymer grows naturally causes intra-sequence correlations within the product.[23]

The models described above are maximally coarse-grained, in that they treat the binding of monomers to the growing tip of the copolymer as a simple, usually single-step, process. However, more generally, one may wish to study models in which polymerisation occurs via a more detailed series of steps, as in figure 1(d). For instance, in order to explain the high accuracy observed in biological polymer copying systems, Hopfield[43] and Ninio[44] independently introduced the concept of kinetic proofreading: a reaction motif in which a monomer undergoes a free energy consuming activation reaction before it is polymerised into the copolymer. The introduction of kinetic proofreading reaction motifs presaged the investigation

of more complex copolymerisation mechanisms[35,45].

In summary, models that allow for multiple monomer types, intra-sequence correlations, reversible reactions, and general, multi-step monomer inclusion reactions represent a wide class of copolymerisation processes. Previous techniques[17–19,22,23,27–35,39–41,46–51] have not allowed analysis of thermodynamically self-consistent models of generalised free copolymerisation processes in which monomer addition is given by an arbitrarily complex network of reversible reactions with rates that may depend on the terminal monomer type, and templated copolymerisation models with high symmetry that can be mapped to these free processes. Investigating the most general type of model in this class would require simulation.

In this paper we present a universal method for studying this large class of copolymerisation models. Drawing on the work of Gaspard and Andrieux[52] for analysing linear copolymerisation processes, and Hill[53,54] for analysing absorbing Markov processes, we present analytical methods for extracting: explicit expressions for the probability of inclusion of a given monomer; the growth rate of a copolymerisation process; and the chemical work done by the process. Our method removes the need to extract the same features by simulation and often produces simple, analytic results.

In section II A, we review and refine methods relating to absorbing Markov chains that are crucial to understanding our approach. In section II B 1 we present our method. In section III, we apply the method to a few example processes to demonstrate its use and power when considering models with certain features. First, we apply the method to models for which the rate of adding new monomers only depends on the monomer type being added. Next we apply the method to templated copolymerisation systems with autonomous-separation that do not have non-equilibrium kinetic proofreading cycles. Finally, we solve a generalised version of Hopfield's kinetic proofreading model applied to a templated copolymerisation system with an autonomously separating product.

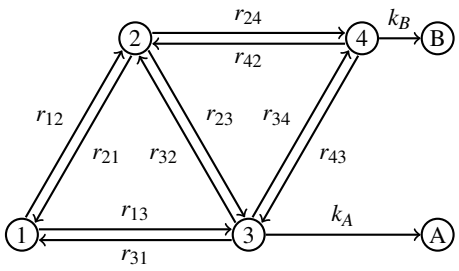
II. METHODS

A. Absorbing Markov Chains

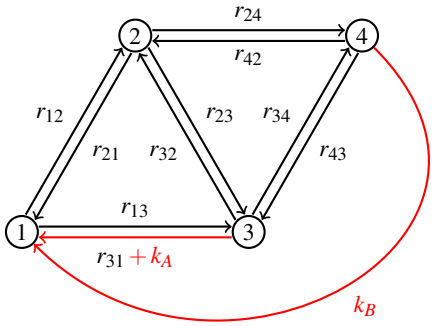
We begin by reviewing and adapting some diagrammatic techniques introduced by Hill to analyse absorbing Markov chains[53,54]. An absorbing Markov chain is a Markov chain for which any trajectory through its state space with arbitrary initial conditions will reach an absorbing state in finite time almost surely[55]. We can decompose the state space of an absorbing Markov chain into absorbing states, $\mathscr{A}$, and transient states, $\mathscr{X}$, such that the state space is $V = \mathscr{A} \cup \mathscr{X}$. Let us denote the rate function that describes the chain as $K : V \times V \to \mathbb{R}^+$, such that $K(x,y)$ is the rate of the transition from state x to state y. Then we denote a Markov process as the tuple, (V,K).

Throughout this section, we shall refer to the absorbing Markov chain given in figure 2(a), which possesses two absorbing states and non-trivial cycles, for illustrative purposes.

(a) Example absorbing Markov process



(b) Closed process



(c) Cycle process for cycle $1 \to 2 \to 3 \to 1$

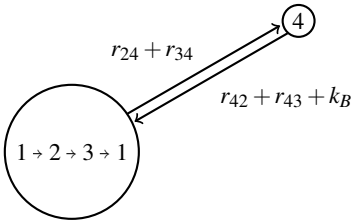


FIG. 2: Graphical representations of an absorbing Markov process to illustrate the methodology outlined in section II A. (a) Example absorbing Markov process $(\mathscr{X} \cup \mathscr{A}, K)$, with two absorbing states, $\mathscr{A} = \{A,B\}$, and four transient states $\mathscr{X} = \{1,2,3,4\}$. (b) The closed process starting at state 1, $(\mathscr{X}, K_1)$. (c) The cycle process $(\{c\} \cup \mathscr{X}/\{1,2,3\} = \{c,4\}, K_{1,C})$ for the cycle $C = 1 \to 2 \to 3 \to 1$ or $C' = 1 \to 3 \to 2 \to 1$.

1. *Expectations of an absorbing process are steady-state averages of a "closed process"*

We will derive expressions for four main quantities: the probabilities of reaching certain absorbing states, the expected time taken to absorption, the expected net number of times traversing a given edge before absorption and the expected number of times that a trajectory goes round a cycle before absorption. These quantities depend on the starting (transient) state $\sigma \in \mathscr{X}$ and can be found in terms of the "closed" process[54]. The closed process is a modified version of an absorbing Markov chain in which transitions to the absorbing states are redirected to the starting state. Figure 2(b) shows

the closed process starting at state 1 of our example absorbing chain of figure 2(a). The closed process for a Markov process $(\mathscr{X}\cup\mathscr{A},K)$ starting at state σ is a new Markov process $(\mathscr{X},K_\sigma)$ with a rate function given by:

$$K_\sigma(x,\sigma)=K(x,\sigma)+\sum_{A\in\mathscr{A}}K(x,A) \tag{1}$$

for $x\in\mathscr{X}$ and agreeing with K on $\mathscr{X}\times\mathscr{X}/\{\sigma\}$.

The closed process has a unique stationary distribution for the following reasons. From the definition of an absorbing Markov chain, there exists a path from any state to an absorbing state, taking finite time. Thus, in the closed process, there is a path from any state to the starting state, taking finite time. The set of states including the starting state and all those that may be reached from the starting state is therefore positive recurrent and further, this set is the only recurrent set of states and will be reached from any other state. Since there is only one recurrent set of states, there is a unique stationary distribution[55].

Expected quantities of an absorbing Markov chain, such as the expected probability that a particular absorbing state is reached, can be found in terms of steady state quantities in the closed process. Whenever a trajectory of the original process reaches an absorbing state, in the closed process that same trajectory would have been reset back to the starting state. Hence, running the closed process for long times is equivalent to generating many independent trajectories to absorption for the original chain. Thus, averaging quantities in the steady state of the closed process is equivalent to taking expectations over independent trials of quantities in the absorbing chain. It is worth noting that the dependence of expected quantities on the starting state is encoded in the definition of the closed process. Finally, we can see that the definition of the closed process may permit self-transitions, $\sigma\to\sigma$, which, for continuous time Markov processes, have little meaning. However, for the purposes of calculating steady state probabilities of the closed process they may be ignored.

2. Steady State averages of the closed process are calculated using the Markov chain tree theorem

Given that we can turn the calculation of expectations of absorbing processes into steady-state averages over closed processes, we can make use of tools developed for analysing the steady state of Markov processes, such as the Markov chain tree theorem (MCTT)[56]. The MCTT states that the steady state distribution of a Markov chain with a unique stationary distribution may be found by summing over rooted spanning trees of the process, where the transition rates are taken as weights on the edges of the graph. Explicitly, let $\mathscr{G}$ be a directed weighted graph, with weight $K(e)$ for an edge e of $\mathscr{G}$. A spanning tree of $\mathscr{G}$, rooted at a vertex, v, is a subgraph of $\mathscr{G}$ with no cycles that connects all the vertices of $\mathscr{G}$ and for which the out degree of every vertex, except v, is one. The sets of spanning trees rooted at nodes 3 and 4 of the closed process, figure 2(b), are shown in figure 3. Denote by $\mathscr{T}(x)$ the set of all spanning trees rooted at x. The MCTT states that the steady state probability to be in state x, $\pi(x)$, is given by:

$$\pi(x)=\frac{\sum_{T\in\mathscr{T}(x)}\prod_{e\in T}K(e)}{\sum_{v\in\mathscr{X}}\sum_{T\in\mathscr{T}(v)}\prod_{e\in T}K(e)}, \tag{2}$$

with $e\in T$ representing the edges of the tree. The denominator here is simply a normalisation constant.

We can define steady state currents from the steady state distribution of the closed process that corresponds to expected currents of the absorbing chain. Let a subscript σ denote quantities in the closed process starting at state σ. Then π_σ is the steady state probability distribution and K_σ the rate function. The current along a given edge, $e=x\to y$, is given by the probability to be in state x, $\pi_\sigma(x)$, multiplied by the rate along said edge. We can, therefore, write the steady state current along all edges that originally led to absorbing states as:

$$J_{\mathrm{Tot}}(\sigma)=\sum_{A\in\mathscr{A}}\sum_{x\in\mathscr{X}}\pi_\sigma(x)K(x,A), \tag{3}$$

where, as before, $\mathscr{X}$ is the set of transient states and $\mathscr{A}$, the set of absorbing states. $J_{\mathrm{Tot}}(\sigma)$ is the expected total current to absorbing states from state σ, and, therefore, its reciprocal is the expected time to absorption.

For the example process shown in figure 2, we present the spanning trees of the corresponding closed process rooted at node 3 and 4 in figure 3. Given the spanning trees, we can directly write down the total current to absorbing states as:

$$\begin{aligned}J_{\mathrm{Tot}}(1)=\frac{1}{\mathscr{N}}\Big[&k_A\big[r_{12}r_{24}r_{43}+r_{12}r_{23}(r_{42}+r_{43}+k_B)\\&+r_{13}((r_{43}+k_B)(r_{21}+r_{23}+r_{24})+r_{42}(r_{21}+r_{23}))\big]\\&+k_B\big[r_{12}r_{23}r_{34}+r_{12}r_{24}(r_{31}+k_A+r_{32}+r_{34})\\&+r_{13}r_{34}(r_{21}+r_{23}+r_{24})+r_{13}r_{32}r_{24}\big]\Big],\end{aligned} \tag{4}$$

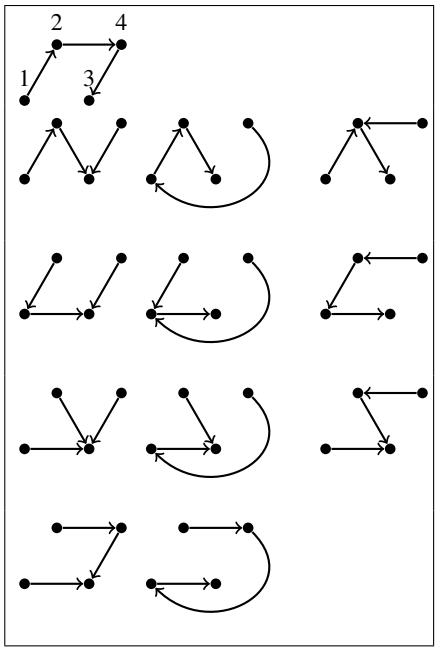
where $\mathscr{N}$ is the normalisation term, given in appendix B. The terms multiplied by k_A are the partial current to absorbing state A, i.e. the current along transition $3\to A$, coming from the trees rooted at node 3, and equivalently for k_B with state B, i.e. the current along transition $4\to B$, coming from trees rooted at node 4.

3. Absorbing probabilities

Given a Markov chain with multiple absorbing states, we can ask for the probability of absorption in each absorbing state in the long time limit. The probability that a trajectory eventually ends in a specific absorbing state can be calculated from the closed process, by dividing the expected current along transitions that originally led to the absorbing state in question by $J_{\mathrm{Tot}}(\sigma)$ (eqn. 3). Therefore the absorption probabilities can be written

$$\mathbb{P}[\sigma\to A]=\frac{\sum_{x\in\mathscr{X}}\pi_\sigma(x)K(x,A)}{\sum_{B\in\mathscr{A}}\sum_{x\in\mathscr{X}}\pi_\sigma(x)K(x,B)}, \tag{5}$$

(a) Spanning trees rooted at node 3.



(b) Spanning trees rooted at node 4.

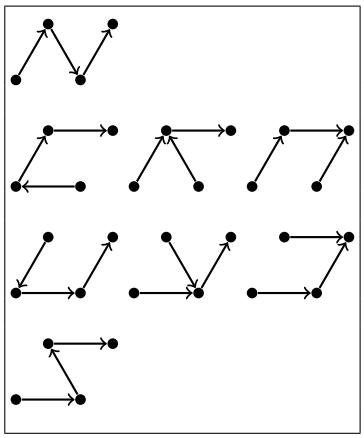


FIG. 3: Spanning trees of the closed processes rooted at (a) node 3 and (b) node 4 derived from figure 2 (b), with nodes labelled in the first spanning tree and all other trees following the same positioning. The spanning trees have been arranged in terms of the self-avoiding walk between nodes 1 and 3 for the trees rooted at node 3 and arranged in terms of the self-avoiding walks between nodes 1 and 4 for the trees rooted at node 4. More details on the relationship between self avoiding walks and spanning trees are given in appendix A.

using the notation $\mathbb{P}[\sigma \to A]$ to denote probability of being absorbed to A given that the trajectory started in state σ. It is worth noting here that given that this quantity is a ratio of currents, there is a factor of π_σ in both the denominator and the numerator of the expression. In practice, we see that the normalisation factor from the MCTT (eqn. 2) cancels out, which simplifies the quantities in the calculation.

For our example process shown in figure 2, we can use the partial currents to absorbing states A and B to write down the absorbing probabilities:

$$\begin{aligned}\mathbb{P}\left[1 \to A\right] &= \frac{1}{\mathscr{N} J_{\mathrm{Tot}}(1)} k_A \big[r_{12}r_{24}r_{43} + r_{12}r_{23}(r_{42}+r_{43}+k_B) \\ &+ r_{13}((r_{43}+k_B)(r_{21}+r_{23}+r_{24}) + r_{42}(r_{21}+r_{23}))\big], \\ \mathbb{P}\left[1 \to B\right] &= \frac{1}{\mathscr{N} J_{\mathrm{Tot}}(1)} k_B \big[r_{13}r_{34}(r_{21}+r_{23}+r_{24}) + r_{13}r_{32}r_{24} \\ &+ r_{12}r_{23}r_{34} + r_{12}r_{24}(r_{31}+k_A+r_{32}+r_{34})\big]\Big]. \end{aligned} \tag{6}$$

The normalisation factor, $\mathscr{N}$, propagated through from eqn. 2, conveniently cancels out with the $1/\mathscr{N}$ implicit in J_{tot}.

4. *Counting edge and cycle transitions*

In this subsection, we shall calculate the expected net number of times traversing a given edge of an absorbing Markov process before absorption. Additionally, we shall calculate the expected number of times a non-recurrent cycle of an absorbing Markov process is traversed before absorption. Both of these will be of use later in defining a notion of chemical work.

To calculate the net number of times crossing a given edge we find the expected current along the transition, $x \leftrightharpoons y$, between states x and y of an absorbing process, $(\mathscr{X} \cup \mathscr{A}, K)$, as in section II A. The expected current through this edge, denoted $J_{x \leftrightharpoons y}(\sigma)$, given starting in state $\sigma \in \mathscr{X}$, can be calculated from the closed process, $(\mathscr{X}, K_\sigma)$, as in eqn. 1, as the difference between the steady state probability to be in state x multiplied by the rate from $x \to y$ and the steady state probability to be in state y multiplied by the rate from $y \to x$,

$$J_{x \leftrightharpoons y}(\sigma) = \pi_\sigma(x)K(x,y) - \pi_\sigma(y)K(y,x). \tag{7}$$

The net number of times traversing this edge (number of observed transitions $x \to y$ - number of observed transitions $y \to x$) before absorption is then just the ratio between this current and total current to absorbing states:

$$N_{x \leftrightharpoons y}(\sigma) = \frac{J_{x \leftrightharpoons y}(\sigma)}{J_{\mathrm{Tot}}(\sigma)}. \tag{8}$$

The current, eqn. 7, is intimately linked to the notion of cycles as pointed out by Wachtel et al.[57] and detailed in appendix C. Thus, we also wish to find the expected number of times traversing a non-recurrent cycle. We define a non-recurrent cycle for a Markov chain to be a cycle of

states, where each state, aside from the originating state, does not appear more than once in the cycle. For example, the cycle A→B→C→D→A is non-recurrent, but A→B→C→D→B→A is recurrent. Note that the originating state is arbitrary, and so A→B→C→D→A is equivalent to B→C→D→A→B. For a stationary process, the expected frequency with which a cycle is completed can be calculated from the one-way cycle current[54,58], which is the probability current going around the cycle. For a chosen non-recurrent cycle, the one-way cycle current can be calculated diagrammatically from three terms. First, a cycle term given by the product of rates around the cycle in the chosen direction. Second, a spanning tree term that can be found by collapsing the nodes in the cycle into a single node (in figure 2c, the cycle $1 \to 2 \to 3 \to 1$ has been collapsed in this way) and finding the sum of spanning trees of this new graph rooted at the collapsed cycle node. Finally, there is a normalisation factor, which is the same normalisation factor as for the current, $\mathscr{N}$. Explicitly, consider an absorbing Markov chain, $\mathscr{G} = (\mathscr{X} \cup \mathscr{A}, K)$ and its closed process starting at $\sigma \in \mathscr{X}$, $\mathscr{G}_\sigma = (\mathscr{X}, K_\sigma)$. Let C denote both the set of edges and set of nodes of a cycle in the closed process. To calculate the spanning tree term for the one-way cycle current, construct a new Markov chain, the cycle process, $\mathscr{G}_{\sigma,C} = (\{c\} \cup (\mathscr{X}/C), K_{\sigma,C})$, where $\{c\} \cup (\mathscr{X}/C)$ is the set of transient states of the original process with the states in the cycle replaced by the single node, c, and $K_{\sigma,C}$ is given by:

$$\begin{aligned} K_{\sigma,C}(x,c) &= \sum_{i \in C} K_\sigma(x,i) \\ K_{\sigma,C}(c,x) &= \sum_{i \in C} K_\sigma(i,x) \\ K_{\sigma,C}(c,c) &= 0 \end{aligned} \tag{9}$$

for $x \in \mathscr{X}/C$ and agreeing with K_σ elsewhere. The cycle process for the cycle $C = 1231$ (or $C' = 1321$) of the example system in figure 2(a) is shown in figure 2(c). Let $\mathscr{T}_\sigma(x), \mathscr{T}_C(x)$ be the sets of spanning trees rooted at x of the closed process, $\mathscr{G}_\sigma$, and cycle process, $\mathscr{G}_{\sigma,C}$, respectively. Then, the cycle current is given by[58]:

$$J_{\mathrm{Cyc}}(\sigma,C) = \frac{\overbrace{\left(\prod_{e \in C} K(e)\right)}^{\text{Cycle}} \overbrace{\sum_{T \in \mathscr{T}_C} \prod_{e \in T} K_{\sigma,C}(e)}^{\text{Spanning Trees}}}{\underbrace{\sum_{x \in \mathscr{X}} \sum_{T \in \mathscr{T}_\sigma(x)} \prod_{e \in T} K_\sigma(e)}_{\text{Normalisation}}}. \tag{10}$$

Note for the cycle term, the edges are taken from the original process rather than the closed process. Given the cycle current for the closed process, the expected number of circulations of the cycle before absorption, $N_{\mathrm{Cyc}}(\sigma,C)$, is the ratio of the cycle current to the total current to absorbing states:

$$N_{\mathrm{Cyc}}(\sigma,C) = \frac{J_{\mathrm{Cyc}}(\sigma,C)}{J_{\mathrm{Tot}(\sigma)}}. \tag{11}$$

For our example process, the expected number of circulations of $C = 1231$ is

$$N_{\mathrm{Cyc}}(1,C) = \frac{(r_{12}r_{23}r_{31})(r_{42}+r_{43}+k_B)}{\mathscr{N} J_{\mathrm{Tot}}(1)}, \tag{12}$$

with the same implicit cancellation of normalisation as before, since $J_{\mathrm{Tot}} \propto \frac{1}{\mathscr{N}}$.

For an absorbing process starting at a given state, we may divide the cycles into internal and external cycles. External cycles are those which appear in the closed process and involve edges which were absorbing edges in the original process. The set of all cycles, sorted into internal and external, for the example process fig 2, is shown in appendix D. The external cycles correspond to the pathways from the starting state to an absorbing state. Therefore, the expected number of times traversing an external cycle before absorption will be at most one and corresponds to the probability of following a given path to absorption. Further, the sum of eqn. 11 over all external cycles will be one.

B. Copolymer Methods

1. *Philosophy of coarse-graining complex underlying copolymerisation reactions networks*

Armed with the techniques for solving absorbing Markov chains, here we set out the method for the analysis of copolymerisation processes. Gaspard and Andrieux[52] presented a method to analyse Markov polymerisation processes in which each monomer is added in a single step (i.e. if the internal reaction network shown in figure 1d were trivial), assuming long polymers. We shall present a method for mapping more complex models for the individual polymerisation step onto coarse-grained descriptions that can be analysed using this framework, and then subsequently show how to back out the behaviour of the full model from the results.

Consider a growing copolymer with M monomer types, which are assumed to be present in the environment at fixed concentrations. At a coarse grained level, we can define a a state space of finite length sequences $\{x_1x_2\cdots x_l \mid x_i \in \{1,2,\cdots M\}, l \geq 0\}$, where l is the length of the sequence. Let us refer to the coarse-grained states in this state space as *completed states*. On this coarse-grained level, a sequence of length l may increase in length by one unit by polymerising one of M units at the growing tip ($x_1x_2\cdots x_l \to x_1x_2\cdots x_lx_{l+1}$), or it may decrease in length by one unit ($x_1x_2\cdots x_l \to x_1x_2\cdots x_{l-1}$). Such a coarse grained model is depicted in figure 1(a,b,c) for free polymerisation, templated self-assembly and templated polymerisation with simultaneous separation.

In general, copolymerisation processes may be best described by models in which the underlying copolymerisation reaction networks are complex, featuring multiple substeps in arbitrarily complex networks connecting the completed states, as suggested in figure 1d. Hence, overall, we could consider a copolymerisation process as having a tree-like structure with networks of reactions connecting completed states, as in figure 4. Such a class of models is wide-

reaching, with many examples from the literature included in this class[17–19,22,23,27,28,30–35,39–41,45–51].

We will define a Markov process at the level of the coarse-grained completed states that, by construction, preserves probabilities of transitions between the completed states of the fine-grained process, and therefore preserves the statistics of the sequences produced. The coarse-grained Markov process does not preserve the distribution of transition times between completed states implied by the fine-grained model, which will in general be non-Markovian. Moreover, it does not provide fine-grained information on trajectories between the coarse-grained completed states. However, temporal details and information about the fine-grained dynamics can be added back in at a later stage, once statistics have been analysed at the coarse-grained level.

2. Identifying propensities in the coarse-grained model

We find the transition rates of the coarse-grained model (hereafter labelled propensities to avoid confusion with the underlying rates of the fine-grained process) by considering first passage problems between completed states. From a given completed state, there are $M+1$ completed states that may be reached, corresponding to the M possible additions of a monomer and the removal of the monomer currently at the tip of the copolymer. For a first passage problem, we can convert each of these reachable completed states into an absorbing state by removing the transitions out of said states, as in figure 4(a), in the same vein as Cady and Qian[59]. Let us refer to this absorbing Markov process as the *step-wise process* and define step to mean the addition/removal of a monomer.

We shall work with the assumption that the transition rates depend on the two monomers at the growing tip of the copolymer, following[1,17–19,21–23,36–38,52,60]. There will therefore be M^2 flavours of this process corresponding to the combinations of the two terminal monomers of the copolymer, the central state $\&xy$ (here & represents an arbitrary sequence). We wish to find the absorbing probabilities, $\mathbb{P}[\&xy \to \&xyz]$, $z \in \{1,\cdots M\}$, $\mathbb{P}[\&xy \to \&x]$ given an initial condition of the central state, $\&xy$. As outlined in the previous section, eqn. 5, we can find these probabilities by constructing the closed process and finding sums over spanning trees rooted at different states. The step-wise process has $M+1$ petal-like graphs each connected to the central state, but disconnected from each other. Due to this structure, any sums over spanning trees of the full process will factorise into a product of sums over spanning trees of the petals. Thus, we find that the absorbing probabilities take the following form:

$$\mathbb{P}[\&xy \to \&xyz] = \frac{1}{\mathscr{N}}\Lambda^{+}(z,y)\left[\prod_{z'\neq z} Q(z',y)\right] Q(y,x),$$
$$\mathbb{P}[\&xy \to \&x] = \frac{1}{\mathscr{N}}\Lambda^{-}(y,x)\prod_{z} Q(z,y). \tag{13}$$

Here $z \in \{1,\cdots M\}$, $\mathscr{N}$ is the normalisation factor from eqn. 2; $\Lambda^{+}(z,y)$ is the sum over spanning trees of the petal connecting states monomers $\&xy$ and $\&xyz$, rooted at the forward completed state, $\&xyz$; $\Lambda^{-}(y,x)$ is the sum over spanning trees of the petal connecting states monomers $\&x$ and $\&xy$, rooted at the backwards completed state, $\&x$, figure 4(b); and $Q(y,x)$ is the sum over spanning trees of the petal connecting states $\&x$ and $\&xy$, linked back to the central state and rooted at the central state, i.e. with edges redirected to the starting state as in the closed process, as in figure 4(c). Since Q is a sum over spanning trees rooted at the node to which edges have been redirected, the sum takes the same form for both the forwards and backwards petals, only depending on which two completed states it is connecting.

From these probabilities, we see that choosing propensities $\omega_{\pm yx}$ for the transitions $\&x \xrightarrow{\omega_{+yx}} \&xy$ and $\&xy \xrightarrow{\omega_{-yx}} \&x$ such that

$$\omega_{\pm yx} = \frac{\Lambda^{\pm}(y,x)}{Q(y,x)} \tag{14}$$

not only preserves the ratios of probabilities of transitions to completed states, but also ensures that $\omega_{\pm yx}$ only depends on monomers x and y. We note here that this coarse graining process is different from lumping[55,61], in which the state space is reduced while attempting to retain trajectory dynamics. In our approach, the coarse-grained process does not reproduce the dynamics of the fine-grained process, only the statistics of the completed states that are visited. However, dynamic quantities may be extracted exactly from the step-wise process, as we show in Sec. II B 4.

3. Solving the coarse-grained model

We now use the methods developed by Gaspard and Andrieux[52] to solve the coarse-grained Markov model over the completed states, with propensities, $\omega_{\pm yx}$. Gaspard and Andrieux's approach considers a frame of reference that is comoving with the tip of the growing polymer, and assumes that the state of the tip and nearby monomers reaches a stationary distribution, to derive quantities at this steady state, such as the set of tip incorporation velocities, v_x (the rates of adding monomers to a copolymer $\&x$), the tip probabilities, $\mu(x)$ (the probability at a given time that the growing polymer is in state $\&x$), and the pair tip probabilities, $\mu(x,y)$ (the probability of being in state $\&xy$). However, we note that the time-dependent information is not physical at this stage due to the coarse-graining process. The above quantities are found from solving the following equations[52]:

$$v_x = \sum_{y=1}^{M} \frac{\omega_{+yx} v_y}{\omega_{-yx} + v_y}, \tag{15}$$

$$\mu(x) = \sum_{y=1}^{M} \frac{\omega_{+xy}}{\omega_{-xy} + v_x}\mu(y), \tag{16}$$

$$\mu(x,y) = \frac{\omega_{+yx}}{\omega_{-yx} + v_y}\mu(x). \tag{17}$$

Using μ and v, we can calculate the statistics of the copolymer sequence far behind the growing tip[52]. We note that the

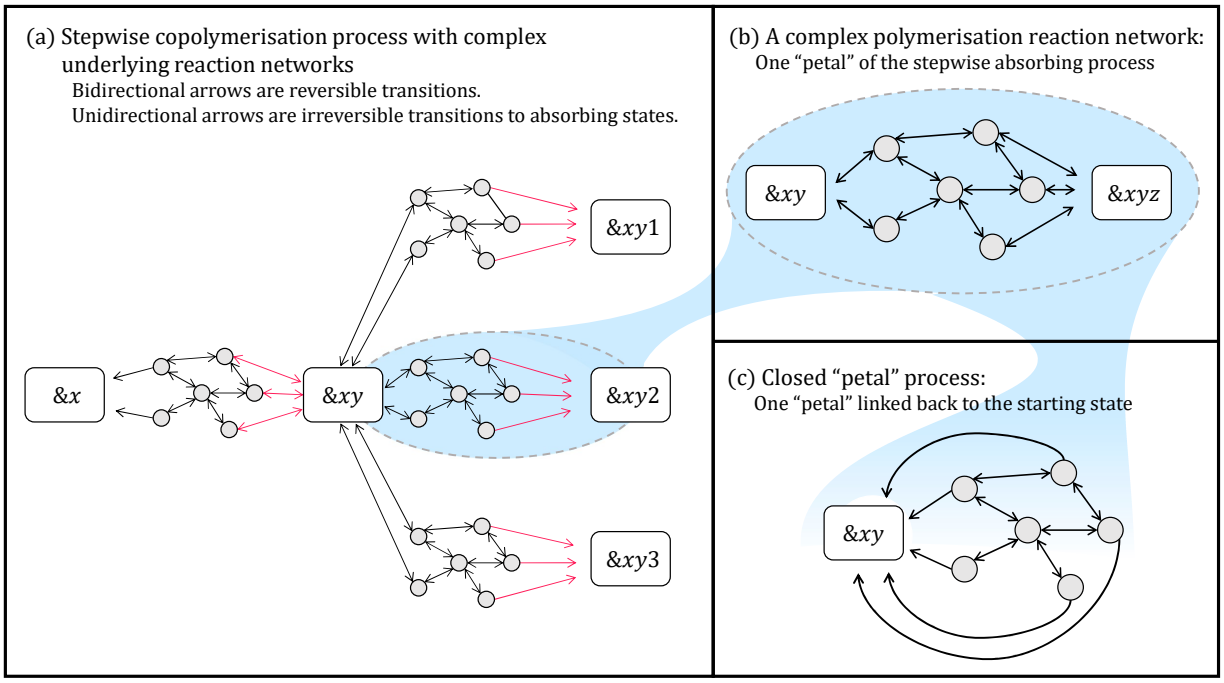


FIG. 4: a) The step-wise process for an arbitrary model with 3 monomer types. This step-wise process is for a copolymer &xy. The edges coloured red are the *completion* edges. The flower like structure of the step-wise process can be seen with 4 petals each connected at the starting state, &xy. b) One of the petals of the step-wise process, which we use to define $\Lambda^{\pm}(z,y)$. $\Lambda^{+}(z,y)$ is defined as the sum of spanning trees rooted at the rightmost state, &xyz and $\Lambda^{-}(z,y)$ the sum of trees rooted at the leftmost state, &xy. c) One of the petals (connecting &xy to &xyz, as in b)) which has been linked back to the starting state. This graph is used to define $Q(z,y)$ as the sum of spanning trees rooted at the leftmost state, &xy.

distribution of monomers at the tip, $\mu(x)$, is different from the distribution of monomers at sites behind the tip; we assume that this distribution reaches some limit far behind the growing tip, in the bulk of the copolymer. This limiting distribution describes the probability that a monomer in the bulk of the copolymer takes a value x. Using $\varepsilon(x)$ to denote the frequency of monomer x in the bulk of the copolymer,[52]

$$\varepsilon(x) = \frac{\mu(x)v_x}{\sum_y \mu(y)v_y}. \tag{18}$$

We may similarly define $\varepsilon(y|x)$ as the probability that in the bulk of the copolymer, a monomer y is observed given a monomer x behind it. $\varepsilon(x)$ and $\varepsilon(y|x)$ fully characterise the statistics of the bulk copolymer since under our assumptions - transitions only depend on the two monomers at the tip - the completed copolymer sequence is itself a Markov chain[23].

4. *Extracting properties of the fine-grained model from the solution of the coarse-grained model*

The easiest quantities to extract are the frequencies of monomers in the bulk of the copolymer. These quantities are identical in the coarse-grained and fine-grained models, since the coarse-graining preserves the statistical distribution of sequences produced. Therefore $\varepsilon(x)$ as defined in eqn. 18 and $\varepsilon(y|x)$ apply directly to the fine-grained process.

The tip probabilities, μ, above give the fraction of time spent in each tip state in the coarse-grained model. However, the coarse-grained model will not reproduce the time series of the fine-grained model, only the sequences of completed states visited. We therefore quotient out the lifetime of tip state (x,y), $\tau(x,y)$, to obtain the frequency with which the tip

states are visited in the coarse-grained model,

$$\xi(x,y) = \frac{1}{\sum_{x',y'=1}^{M} \frac{\mu(x',y')}{\tau(x',y')}} \frac{\mu(x,y)}{\tau(x,y)}, \tag{19}$$

$$\tau(x,y) = \frac{1}{\omega_{-yx} + \sum_{z=1}^{M} \omega_{+zy}}. \tag{20}$$

This frequency defines a new tip distribution, ξ. $\xi(x,y)$ is the frequency that a given pair of monomers x,y is observed at the tip of the growing copolymer in the sequence of transitions. This distribution, $\xi(x,y)$, applies to both the coarse-grained model and the sequence of completed states visited in the full fine-grained model. It can therefore be used to find averages of key dynamic properties.

For example, we can calculate the probability, P, that a growing copolymer increases in length at each step of the step-wise process. P is calculated by averaging the probability of adding a monomer over the possible states $\&xy$:

$$P = \sum_{x,y=1}^{M} \xi(x,y) \frac{\sum_{z=1}^{M} \omega_{+zy}}{\omega_{-yx} + \sum_{z=1}^{M} \omega_{+zy}}. \tag{21}$$

Upon averaging out the sequence information we may treat the growth of a polymer as a random walk with probability P of stepping forwards and $1-P$ of stepping back. We can find the expected number of monomer inclusion/removal steps per net forward step as $1/(2P-1)$ (for proof see appendix E). A number of quantities scale with the total number of steps rather than the net number of steps, making the number of steps per net forward step a necessary quantity. For example in order to find the expected time taken per net forward step, one can find the expected time to absorption for the step-wise process, figure 4(a), $T(x,y)$, for a copolymer in state $\&xy$ by calculating $1/J_{\mathrm{Tot}}(\&xy)$ for the step-wise process using eqn. 3. The expected time per net forward step is then

$$\tau_{\mathrm{step}} = \frac{1}{2P-1} \sum_{x,y=1}^{M} \xi(x,y) T(x,y). \tag{22}$$

$1/\tau_{\mathrm{step}}$ is therefore the physical average growth rate of the copolymer in the fine-grained model.

We may also calculate the chemical work done by the system in producing the copolymer. In a purely chemical system, with no time-varying externally applied protocols, the entropy increase of the universe is given by the decrease in the generalised free energy of the chemical system, including any coupled reservoirs of fuel molecules.[20]. Since the total free energy must decrease, any increase in one contribution must be paid for by a decrease of at least the same magnitude in another contribution. It is common to describe the latter subsystem as doing work on the former.

For the polymerisation systems analysed here, the generalised free energy can be split into a term corresponding to the chemical free energy of the system, averaged over the uncertain state of the system, and a term related to the entropy arising due to the uncertainty of the state occupied.[62]

$$\mathscr{G} = \sum_a p(a) G_{\mathrm{chem}}(a) + \sum_a p(a) \ln p(a), \tag{23}$$

where we use natural units such that $k_B T = 1$. Here, a is a chemical state of the system as a whole, $G_{\mathrm{chem}}(a)$ is the chemical free energy of state a, and $p(a)$ is the probability that the system occupies the state a. $G_{\mathrm{chem}}(a) = -\ln Z_a$, where Z_a is the partition function of the system (explicitly including any large chemical buffers) restricted to the chemical state a, and represents the contribution of concentrations and bond strength to the favourability of a molecular state. The principle of detailed balance[20] states that the chemical free energy change associated with a transition from a to b is given by

$$G_{\mathrm{chem}}(b) - G_{\mathrm{chem}}(a) = -\ln\left(\frac{K(a,b)}{K(b,a)}\right). \tag{24}$$

The second term in eqn. 23 is information theoretic in character; it is equal to the negative of the Shannon entropy associated with the distribution over chemical states. For the systems studied here, in which we consider infinitely long copolymers that have reached steady state growth, the only relevant contribution to this term is the increase in Shannon entropy of the copolymer sequence produced as the polymer gets longer. Since the copolymer sequence is itself a discrete time Markov chain[23] the additional entropy per net forward step (the entropy rate) can be readily calculated[63]:

$$H = -\sum_{x,y=1}^{M} \varepsilon(x) \varepsilon(y|x) \ln \varepsilon(y|x), \tag{25}$$

with x,y representing the monomer types. Since the purpose of a copolymerisation system is often to produce a low entropy (or "accurate") sequence, it is reasonable to think of the chemical free-energy decrease per net forward step as the chemical work done to reduce the information entropy of eqn. 25 below that of a uniform, random polymer. Extending the definition provided by Poulton et al.[23], we may define the efficiency of copolymerisation as:

$$\eta = \frac{\ln M - H}{\ln M + \mathscr{W}_{\mathrm{chem}}} \leq 1, \tag{26}$$

$\ln M$ is the entropy per monomer (or entropy rate) of a uniform, random copolymer with M monomer types, and $\mathscr{W}_{\mathrm{chem}}$ is the average decrease in chemical free energy per net forward step. This efficiency is then ratio between the entropy drop due to the accuracy of the copolymer compared to a random one ($\ln M - H$) and the chemical work used to drive the system ($\mathscr{W}_{\mathrm{chem}}$) above that required to make a random copolymer in equilibrium (-$\ln M$)[64].

The expected work done during a transition adding or removing a monomer given starting in completed state $\&xy$ can be calculated by summing the contribution from eqn. 24 multiplied by the expected net current along the edge $a \leftrightharpoons b$ prior

to absorption over all edges in the step-wise process:

$$w_{\text{chem}}(x,y) = \tag{27}$$

$$-\Delta G_{\text{chem}}(x,y) = \sum_{b>a} \ln\left(\frac{K(a,b)}{K(b,a)}\right) N_{a\leftrightharpoons b}(\&xy), \tag{28}$$

where $N_{a\leftrightharpoons b}(\&xy)$ is the expected net number of times traversing edge $a \leftrightharpoons b$ before absorption giving starting in the central state of the step-wise process, $\&xy$, as in eqn. 8. This sum will also require contributions from edges which lead to absorbing states. For such edges, the rate for the reverse transition in the logarithm of eqn. 28 is the rate from the full process.

Equivalently, however, as outlined in appendix C, we may find this chemical work by considering the non-recurrent cycles of the process[57]. For a given internal cycle, C, we may define the affinity[20],

$$A(C) = \ln \frac{\prod_{e\in C} K(e)}{\prod_{e\in C'} K(e)}, \tag{29}$$

where the sum is over the edges, e, composing the cycle and C' is the cycle with edges in revered direction. For external cycles, we may define the affinity in the same way, inferring the rate for the reversed edge of the transition to absorbing states from the full process. The expected work done before absorption of the cycle, C, given starting in the state $\&xy$ is

$$w_{\text{chem}}(x,y) = \tag{30}$$

$$-\Delta G_{\text{chem}}(x,y) = \sum_C A(C) \frac{J_{\text{Cyc}}(\sigma,C) - J_{\text{Cyc}}(\sigma,C')}{J_{\text{Tot}}(\&xy)}. \tag{31}$$

Averaging $w_{\text{chem}}(x,y)$ with ξ and multiplying by the expected number of steps per net forward step gives the expected chemical work done per net forward step,

$$\mathscr{W}_{\text{chem}} = \frac{1}{2P-1} \sum_{x,y=1}^{M} \xi(x,y) w_{\text{chem}}(x,y). \tag{32}$$

Further, the forms of eqns. 22 and 32 may be applied to an arbitrary quantity for which one can find the expected value in the step-wise process starting in state $\&xy$. Let this arbitrary quantity be $A(x,y)$. One can then average this using the distribution, ξ, to obtain the expected value of the quantity per step. Then, if appropriate, multiplying by $1/(2P-1)$, gives the expected value of the quantity per net forward step. In practice, as shall be seen in section III C, since the quantities we wish to calculate may be written in terms of sums over spanning trees, the quantities for the step-wise process may be written as a sum over the terms per petal, with the quantity for a given petal factorising into some quantity which depends on the petal multiplied by Q's for the other petals.

5. Stalled growth

Explicit simulation of copolymer growth is particularly challenging in regimes where $P \gtrsim 0.5$, since many backward and forwards steps are taken per net forwards step. At $P = 0.5$, then the process will not reliably produce copolymers; for $P < 0.5$ polymers will tend to shrink. In general, for $P = 0.5$, we can say the model has stalled. Our approach is particularly beneficial in this case; indeed, it is possible to check whether a model is at the stall point by considering an $M \times M$ dimensional matrix of the ratios of forward to backwards propensities[52], $Z_{yx} = \left(\frac{\omega_{+yx}}{\omega_{-yx}}\right) = \frac{\Lambda^+(y,x)}{\Lambda^-(y,x)}$. The model is at the stall point if and only if:

$$\det(Z - \mathbb{1}_M) = 0, \tag{33}$$

where $\mathbb{1}_M$ is the $M \times M$ identity matrix, and shrinking if negative. Since Z gives the ratios of adding a monomer to removing one, this condition essentially says that models will stall if the total rate of adding a monomer is equal to the total rate of removing one.

In a typical model, there exists at least one parameter that controls the driving. Often this parameter is related to the backbone strength of the polymer produced: *e.g.* the free energy drop associated with the formation of a generic backbone bond, ΔG_{pol}. This parameter will be present in the rates of each external cycle so that by tuning it, the model can be moved all the way from stalling to irreversible driving, whereby monomers cannot be removed once polymerised. If such a parameter exists, we may rephrase the stall condition, eqn. 33, in terms of this parameter. For example, for the case of the parameter being ΔG_{pol}, we may find some threshold value Γ such that the model will stall for $\Delta G_{\text{pol}} = \Gamma$.

6. Limiting behaviour

We shall note two limits for which we may give analytic expressions for the frequency of monomer types in the copolymer bulk for all models. First, consider the case that the system is at the stall point (eqn. 33). In general, entropy production can still occur within cycles in the step-wise process; therefore, these frequencies cannot be determined from equilibrium arguments and are non-trivial. Nonetheless, at the stall point, we may express the monomer frequencies in the bulk relatively simply. The frequency of monomer x, $\varepsilon_{\text{stall}}(x)$, is proportional (up to normalisation) to the cofactor of the diagonal element (corresponding to monomer x) of the matrix $(\mathbb{1}_M - Z)$, as proven in appendix F. For example, for $M = 2$,

$$\begin{aligned} \varepsilon_{\text{stall}}(1) &\propto 1 - \frac{\omega_{+22}}{\omega_{-22}}, \\ \varepsilon_{\text{stall}}(2) &\propto 1 - \frac{\omega_{+11}}{\omega_{-11}}, \end{aligned} \tag{34}$$

and for $M = 3$, we have

$$\begin{aligned} \varepsilon_{\text{stall}}(1) &\propto \left(1 - \frac{\omega_{+22}}{\omega_{-22}}\right)\left(1 - \frac{\omega_{+33}}{\omega_{-33}}\right) - \frac{\omega_{+23}}{\omega_{-23}}\frac{\omega_{+32}}{\omega_{-32}}, \\ \varepsilon_{\text{stall}}(2) &\propto \left(1 - \frac{\omega_{+11}}{\omega_{-11}}\right)\left(1 - \frac{\omega_{+33}}{\omega_{-33}}\right) - \frac{\omega_{+13}}{\omega_{-13}}\frac{\omega_{+31}}{\omega_{-31}}, \\ \varepsilon_{\text{stall}}(3) &\propto \left(1 - \frac{\omega_{+11}}{\omega_{-11}}\right)\left(1 - \frac{\omega_{+22}}{\omega_{-22}}\right) - \frac{\omega_{+12}}{\omega_{-12}}\frac{\omega_{+21}}{\omega_{-21}}. \end{aligned} \tag{35}$$

On the other end of the spectrum, we can also solve for monomer bulk frequencies in the irreversible limit, where $\omega_{-yx} = 0$ for all x, y. Intuitively, we could consider the Markov process on the state space $\{1, \cdots, M\}$ representing copolymers with a given monomer at its tip, and transitions between those states with rates, $K_{\text{irrev}}(x \to y) = \omega_{+yx}$. The steady state of this process will give the time dependent frequencies of having a given monomer at the tip of the copolymer. Therefore, dividing by the time spent in each state will give the bulk frequencies. A nice way to write out these frequencies in the style of the methods described thus far is as a sum over the spanning trees on the complete graph on M vertices with rate functions $K_{\text{irrev}}(x,y) = \omega_{+yx}$. Explicitly, we may write these frequencies (up to normalisation) as:

$$\varepsilon_{\text{irrev}}(x) \propto \left(\sum_{T \in \mathscr{T}(x)} \prod_{e \in T} K_{\text{irrev}}(e) \right) \sum_{y=1}^{M} \omega_{+yx}, \tag{36}$$

where $\mathscr{T}(x)$ is the set of spanning trees of the complete graph on M vertices. This expression is derived formally in appendix G. For example, with $M = 2$,

$$\begin{aligned} \varepsilon_{\text{irrev}}(1) &= \frac{\omega_{+12}(\omega_{+11} + \omega_{+21})}{\omega_{+12}(\omega_{+11} + \omega_{+21}) + \omega_{+21}(\omega_{+12} + \omega_{+22})}, \\ \varepsilon_{\text{irrev}}(2) &= \frac{\omega_{+21}(\omega_{+12} + \omega_{+22})}{\omega_{+12}(\omega_{+11} + \omega_{+21}) + \omega_{+21}(\omega_{+12} + \omega_{+22})}. \end{aligned} \tag{37}$$

7. *Simplification for factorisable propensities*

The presented method applies to arbitrary complex copolymerisation models obeying the structure of figure 4. However, if we make some further common assumptions, much of the analysis simplifies. For example, consider the case in which the ratios of propensities may be factored:

$$\frac{\omega_{+yx}}{\omega_{-yx}} = \frac{\Lambda^{+}(y,x)}{\Lambda^{-}(y,x)} = Y(y)X(x), \tag{38}$$

where Y is a function of monomer y only and X is a function of monomer x only. Intuitively, such a condition holds in the cases where there is no direct, type-dependent interactions between monomers in the growing polymer, such as when monomers only interact with a template[1,23,24,27–38]. Under such an assumption, multiple calculations simplify, see appendix H. For example, the stall condition becomes simply that the model will stall at

$$\sum_{x} X(x)Y(x) = 1, \tag{39}$$

Bulk frequencies at stall are just:

$$\varepsilon_{\text{stall}}(x) = X(x)Y(x). \tag{40}$$

III. EXAMPLE APPLICATIONS

We shall now consider some exemplar classes of models to: provide examples of how to utilise the methods; validate their accuracy; and to show the types of quantities and information that may be extracted.

A useful initial classification of models is into those which we shall call balanced. We shall refer to a model as being balanced if its petals (see figure 4(b)) are detailed balanced. Such models are useful baseline checks as their cycles all have zero affinity, meaning no chemical work is done internally and hence the only contributions to chemical work are from external cycles. Further, these models exhibit a proper equilibrium at the stall point, and as such allow for equilibrium arguments to validate the method at this point. It is worth noting that although related to the notion of detailed balanced, the full model with its infinite state space is not detailed balanced.

A. Stalling behaviour in a polymerisation model with no neighbour-neighbour interactions

We shall start with the simplest case, where the propensities in the coarse-grained model only depend on the monomer type being added/removed: $\omega_{\pm yx} = \omega_{\pm y}$, such as in a simple model for templated self assembly, figure 1(b). Assume there exists a backbone free energy, ΔG_{pol} controlling the driving as in section II B 5. Any spanning tree in $\Lambda^{\pm}$ must involve at least one incidence of ΔG_{pol}, since it appears in every external cycle. Therefore, we can split the ratio of propensities as follows:

$$\frac{\omega_{+y}}{\omega_{-y}} = e^{\Delta G_y} e^{\Delta G_{\text{pol}}}, \tag{41}$$

where ΔG_y encompasses the rest of the details about the models. We note in general, ΔG_y may be a function of ΔG_{pol}, however in many cases, it is not. These cases include when there is only one completion reaction (highlighted in red in figure 4(a)) that contains the dependence on ΔG_{pol} or if the model is balanced. We may then interpret $-\Delta G_y$ as an effective binding free energy of monomer y. If we think of ΔG_{pol} as the free energy drive of the model away from stall, we look for a threshold value $\Delta G_{\text{pol}} = \Gamma$ above which the model will not stall. Using eqn. 39, we see that

$$\Gamma = -\ln\left(\sum_{y} e^{\Delta G_y} \right) = -\ln \mathscr{Z}, \tag{42}$$

where $\mathscr{Z}$ is the partition function for a system with one state for each monomer type, each state labelled by y and with free energy $-\Delta G_y$. Furthermore, using eqn. 40, the bulk frequencies at the stall point may be written:

$$\varepsilon_{\text{stall}}(y) = \frac{e^{\Delta G_y}}{\sum_x e^{\Delta G_x}} = \frac{1}{\mathscr{Z}} e^{\Delta G_y}, \tag{43}$$

which is the probability of selecting a state y, with free energy, $-\Delta G_y$ as predicted by equilibrium statistical mechanics. In these results, $-\Delta G_y$ looks like the equilibrium contribution to free energy, and the results follow fairly directly in equilibrium. However, these results hold even if the process involves fuel-consuming cycles: entropy may still be being produced at stall. In such cases, the effect of breaking equilibrium will be to change the effective free energies of selecting a given monomer type.

B. Balanced models of templated polymerisation with autonomous separation

Next we shall consider a class of models where the ratio of propensities may be written:

$$\frac{\omega_{+yx}}{\omega_{-yx}} = e^{\Delta G_y} e^{-\Delta G_x} e^{\Delta G_{\rm pol}}. \tag{44}$$

As before, $\Delta G_{\rm pol}$, coming from the polymerisation reactions represents the driving of this process. Such a class of models includes, most notably, balanced models of templated polymerisation with autonomous separation,[23]. In these cases the breaking of the previous copy-template bond every time a new bond is formed enforces the structure in eqn. 44. We shall assume, as in Ref. 23, that ΔG_y is independent of $\Delta G_{\rm pol}$.

Using eqn. 39 and eqn. 40, we find the stall point to be $\Delta G_{\rm pol} = \Gamma = -\ln M$ and bulk frequencies at stall, $\varepsilon_{\rm stall}(y) = \frac{1}{M}$, where M the number of monomer types. Physically, we can understand these results by considering balanced models of templated polymerisation with autonomous separation. For such models, by definition there is no entropy production in internal cycles and therefore, the stall point must be thermodynamic equilibrium. In such models, the only driving comes from the polymerisation, $\Delta G_{\rm pol}$, and the entropic effect having M monomers to choose. These two effect balance at equilibrium.[64]

Next let us consider the limit that the completion reactions highlighted in red in figure 4 (a) are much slower than the other reactions. Explicitly, let k be some rate constant at the same order of magnitude of the rates of the process that are not the rates for the completion transitions indicated in red in figure 4 (a). Write the completion rates as $k_{\rm com} R^+_{\rm com}(y,x)$, where $k_{\rm com} \ll k$ is a rate constant controlling the overall speed of the completion reactions and $R^+_{\rm com}(y,x)$ provides any sequence dependence. Similarly, the reverse transitions along the completion edges have the rate $k_{\rm com} R^-_{\rm com}(y,x)$. Further, let there be $n_{\rm com}$ such completion reactions in a given petal of the step-wise process (we shall assume this number is the same for all pairs of monomers, x, y)

Assume for simplicity that all completion reactions, $R^\pm_{\rm com}(y,x)$, take the same form in a given petal. Then, we can write the sum over spanning trees, $Q(y,x)$ as

$$Q(y,x) = \frac{1}{n_{\rm com}} \frac{\Lambda^-(y,x)}{k_{\rm com} R^-_{\rm com}(y,x)} + \mathscr{O}\left(\frac{k_{\rm com}}{k}\right), \tag{45}$$

since $\Lambda^-(y,x)$ has first order terms in $k_{\rm com}/k$. This fact can be seen from noting that the leading order terms in $Q(y,x)$ are the trees with no completion reactions and the leading order terms in $\Lambda^-(y,x)$ are those same leading order trees of $Q(y,x)$, except with one completion reaction added in. There are $n_{\rm com}$ such completion reactions and each adds the same leading order term to $\Lambda^-(y,x)$. With $Q(y,x)$ taking this form, and remembering eqn. 44, the propensities take the following form:

$$\omega_{+yx} = n_{\rm com} k_{\rm com} R^-_{\rm com}(y,x) e^{\Delta G_y - \Delta G_x + \Delta G_{\rm pol}} + \mathscr{O}\left(\frac{k_{\rm com}}{k}\right)^2,$$
$$\omega_{-yx} = n_{\rm com} k_{\rm com} R^-_{\rm com}(y,x) + \mathscr{O}\left(\frac{k_{\rm com}}{k}\right)^2. \tag{46}$$

The $n_{\rm com} k_{\rm com}$ term cancels in ratios of $\omega_{\pm yx}$ variables, and therefore does not affect the sequence statistics. Thus, in the slow completion limit, such models are only affected by the binding free energy differences ($\Delta G_y - \Delta G_x$), the driving ($\Delta G_{\rm pol}$), and the nature of the final completion step ($R^-_{\rm com}$). Therefore, the fine details do not affect the statistics of the polymers.

Assuming that all completion edges are associated with the same free energy change $-\Delta G_{\rm pol}$, so that $R^-_{\rm com}(y,x) = e^{-\Delta G_{\rm pol}}$, we may solve for the statistics explicitly. For the case of two monomer types, $M = 2$, we find the bulk frequency to be (appendix I):

$$\varepsilon(1) = \left(1 - \frac{1}{2}(e^{-\Delta G_{\rm pol}} - 1)(e^{-DG} - 1) + \frac{1}{2}\sqrt{(e^{-\Delta G_{\rm pol}} - 1)^2 (e^{-DG} - 1)^2 + 4e^{-DG}}\right)^{-1}, \tag{47}$$

where $DG = \Delta G_1 - \Delta G_2$. This expression is plotted in figure 5 for $DG = 4$. From this expression, we can confirm explicitly by substituting in the stall driving, $\Delta G_{\rm pol} = -\ln 2$, that the bulk frequency indeed becomes $\varepsilon(1) = \frac{1}{2}$. Further, taking the irreversible limit, $\Delta G_{\rm pol} \to \infty$, we find the bulk frequency becomes:

$$\varepsilon(1) = \frac{e^{\Delta G_1}}{e^{\Delta G_1} + e^{\Delta G_2}}, \tag{48}$$

the equilibrium statistical mechanics probability of choosing state 1 with free energy $-\Delta G_1$, given state 2 has free energy $-\Delta G_2$. Since the completion reactions are slow and irreversible, in this limit, the process selecting the monomers is allowed to equilibriate. Therefore, copolymerisation is simply sampling from the equilibrium distribution of this process, and hence tends to the result predicted by equilibrium statistical mechanics.

Eqn.46 shows that in the slow completion limit, the fine details of the reaction network leading to selection of a specific monomer become unimportant and the models collapse onto a single accuracy curve determined by DG, $\Delta G_{\rm pol}$ and $R^-_{\rm com}$. Conversely, if we fix all parameters except $k_{\rm com}$, we seem to see that the bulk frequencies will tend monotonically to their limits as $k_{\rm com}/k \to 0$, either from above or below.

We can use this fact to compare bulk frequencies for certain types of model. For example, we may compare on-rate discrimination,[26] where incorrect monomers bind more slowly, to off-rate discrimination,[26] where incorrect monomers unbind more quickly. An example model comparing on-rate and off-rate discrimination is plotted in figure 5 for a model defined in appendix J. Consider the bulk frequency of an incorrect monomer. On-rate discrimination benefits from fast polymerisation and therefore tends to its slow polymerisation limit from below, whereas off-rate discrimination benefits

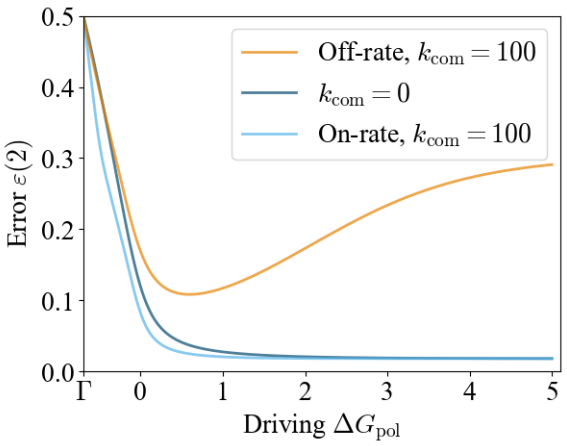


FIG. 5: Plots of the frequency of the less stably-bound (incorrect) monomer with smallest binding free energy, labelled 2 for on- and off-rate discrimination balanced models with $k_{\text{com}} = 100$ and $k_{\text{com}} \to 0$. The binding free-energy difference for these models is $DG = \Delta G_1 - \Delta G_2 = 4$. The models are topologically the Hopfield model as in figure 6a, with $\Delta G_{\text{act}} = 0$ and $M_{\text{in}} = M_{\text{act}} = 1$. However, for the on-rate discrimination, the free-energy terms are in the binding reactions instead of the unbinding ones. The specific models are given in appendix J.

from allowing the process selecting monomers to equilibriate and hence tends to its slow copolymerisation limit from above. This fact sets up a hierarchy for a given set of parameters, and moderate or strong driving, for the bulk frequency of incorrect monomers, off-rate discrimination > slow copolymerisation > on-rate discrimination. This observation is consistent with the results of Sartori and Pigolotti[26] and Poulton et al.[23] for kinetic (on-rate) and energetic (off-rate) discrimination.

C. Hopfield's Kinetic Proofreading in a model of templated copying with autonomous separation

For our final example, we shall consider an explicit model of copolymerisation, with Hopfield's kinetic proofreading mechanism incorporated into a templated copolymerisation system with autonomously separating product in a thermodynamically valid way. From this setup, we can provide a fully worked example of an explicit model, as well as demonstrating the power of the method for analysing sequences of models with recursive structures as we look at a generalised version of Hopfield's proofreading incorporated into a model of templated polymerisation with autonomous separation.

Explicitly, we first consider the one-loop model of kinetic proofreading shown in figure 6 (a). There are two monomer types, the right ones $x = r$ and wrong ones $x = w$. Note that we have already transformed the model so that the sequence of the copy is defined relative to that of the template[35]. These monomer types exist in inactive and active states with concentrations M_{in} and M_{act}, respectively, relative to some reference concentration, with each monomer type having the same concentration. As previously, we shall assume the environment is sufficiently large such that these concentrations remain constant.

The monomers may bind to the template either in an active or inactive state with binding free energies $-\Delta G_x$ for monomer type x. Inactive monomers may be activated on the template with a free-energy change of ΔG_{act}. Finally, active monomers may be polymerised into the copolymer chain, with free-energy change $-\Delta G_{\text{pol}}$. Subsequently, the penultimate monomer of the copolymer unbinds from the template. Each of these reactions is assigned a forwards and reverse reaction rate consistent with the thermodynamic model; the full model is illustrated in figure 6 (a). Conceptually, the proofreading motif functions by providing two opportunities to reject the unwanted monomer w: first, when the un-activated monomer binds, and second, after it has been activated. To be effective, a non-zero affinity is required to drive the system around the cycle of states in the correct order: unbound template site $\to$ unactivated monomer bound $\to$ activated monomer bound.[20,43] We emphasise that this model differs from Hopfield's original description in two important ways: firstly, we consider a full, microscopically reversible polymerisation process, rather than a single incorporation step with irreversible polymerisation; and secondly, we embed the proofreading motif into a non-trivial polymerisation process involving autonomous detachment from the template.

Given the model as described in figure 6 (a), we first identify the propensities ω_{xy} connecting completed states. Due to the petal-like structure, we can follow eqn. 13 and simply consider spanning trees of the petal sub-processes illustrated in figure 6; $\Lambda^-(y,x)$ rooted at $\&x$, $\Lambda^+(y,x)$ rooted at $\&xy$, and $Q(y,x)$, for a petal connecting $\&x$ and $\&xy$. Explicitly writing out the sums of spanning trees, we obtain:

$$\Lambda_1^+(y,x) = \left[k_1 k_{\text{act}} M_{\text{in}} + k_{KP} M_{\text{act}} (k_1 e^{-\Delta G_y} + k_{\text{act}})\right] k_{\text{pol}} e^{-\Delta G_x}, \tag{49}$$

$$\Lambda_1^-(y,x) = \left[k_1 k_{\text{act}} e^{\Delta G_{\text{act}} - \Delta G_y} + k_{KP} e^{-\Delta G_y} (k_1 e^{-\Delta G_y} + k_{\text{act}})\right] k_{\text{pol}} e^{-\Delta G_{\text{pol}}}, \tag{50}$$

$$Q_1(y,x) = \left[k_1 k_{\text{act}} e^{\Delta G_{\text{act}} - \Delta G_y} + (k_{KP} e^{-\Delta G_y} + k_{\text{pol}} e^{-\Delta G_x})(k_1 e^{-\Delta G_y} + k_{\text{act}})\right]. \tag{51}$$

Here, we add a subscript 1 to denote these as for the simple, "1-loop", Hopfield model, which we shall extend to allow more loops later. We note that the ratio, $\Lambda^+(y,x)/\Lambda^-(y,x)$ factorises as eqn. 38 and so we can easily write down the stall

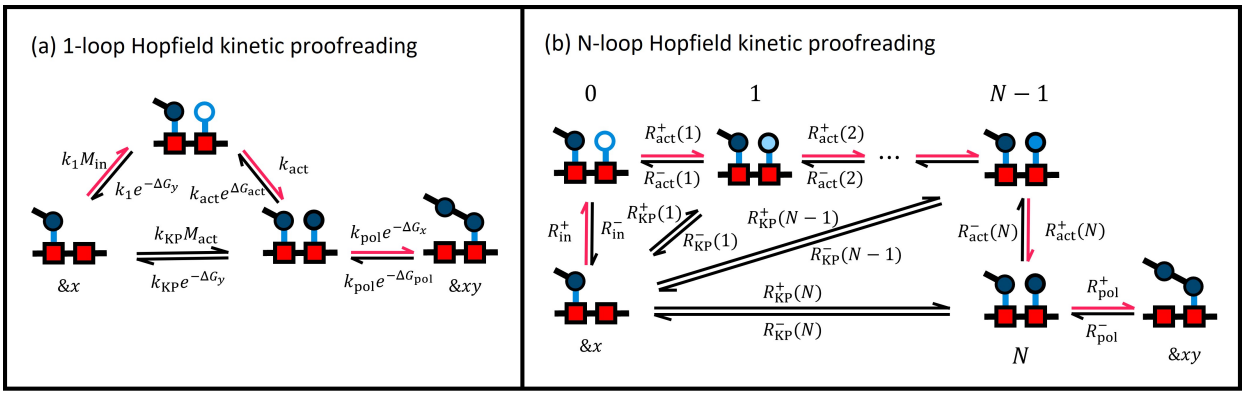


FIG. 6: Reaction rates of the (a) 1-loop and (b) N-loop Hopfield kinetic proofreading models implemented in a templated polymerisation model with autonomous separation system. Each of these subfigures represents a single petal of the step-wise process as in figure 4, going from completed state $\&x \to \&xy$. In both cases, the template is represented by red squares. In (a) the inactive monomer is represented by a white circle and the activated monomers by a dark blue circle. In (b), different levels of activation are represented by increasingly dark shades of blue circles. Further, in (b) the numbers by the states represent the activation level of the monomer. In each case, the desired pathway is highlighted with red arrows.

condition as $\Delta G_{\rm pol} = \Gamma$ with

$$\Gamma = -\ln\left(\frac{k_1 k_{\rm KP} M_{\rm act} e^{-\Delta G_r} + k_{\rm act}(k_{\rm KP}M_{\rm act} + k_1 M_{\rm in})}{k_1 k_{\rm KP} e^{-\Delta G_r} + k_{\rm act}(k_1 e^{\Delta G_{\rm act}} + k_{\rm KP})} + \frac{k_1 k_{\rm KP} M_{\rm act} e^{-\Delta G_w} + k_{\rm act}(k_{\rm KP}M_{\rm act} + k_1 M_{\rm in})}{k_1 k_{\rm KP} e^{-\Delta G_w} + k_{\rm act}(k_1 e^{\Delta G_{\rm act}} + k_{\rm KP})}\right). \quad (52)$$

Note that setting $M_{\rm in} = M_{\rm act} = 1, \Delta G_{\rm act} = 0$, in eqn. 52, Γ collapses to $-\ln 2$ as these conditions reduce the system to a balanced one with a stall point at equilibrium, as in Section III B.

The frequency of right and wrong monomers, $\varepsilon(x = r, w)$ may be calculated from eqn. 18 (the calculation is implemented in the supporting information). We plot copying error, as represented by $\varepsilon(w)$, in figure 7 (a), and demonstrate that it agrees well with the results found from a Gillespie simulation[65] of the same model. We also compare to a "0-loop" version of the model, in which the inactivated monomers and the inactivated monomer bound state are omitted. As can be seen, the proofreading motif generally improves accuracy when driven above its stall point $\Delta G_{\rm pol} = \Gamma$. Indeed, we may write down expressions for the bulk frequency in the irreversible limit ($\Delta G_{\rm pol} \to \infty$) using eqn. 36. In this irreversible limit, we recover Hopfield's classic argument by taking some further limits consistent with his analysis. Namely, letting $M_{\rm act}, k_{\rm act}, k_{\rm pol} \to 0$, we find $\varepsilon(w)/\varepsilon(r) = e^{2(\Delta G_w - \Delta G_r)}$. In this limit, the ratio of incorrect monomers to correct ones involves the square of the binding free energy difference, reflecting the fact that two steps of discrimination have occurred.

We may also write down expressions for the expected chemical work done per net step of the process. This quantity will involve the total current to absorbing states of the step-wise process for starting with a copolymer $\&xy$, which we may write as:

$$J_{\rm Tot}(y,x) = \frac{1}{\mathcal{N}(y,x)}(\Lambda_1^+(r,y)Q(w,y)Q(y,x) + \Lambda_1^+(w,y)Q(r,y)Q(y,x) + \Lambda_1^-(y,x)Q(r,y)Q(w,y)), \quad (53)$$

where $\mathcal{N}$ is a normalisation factor that will cancel out of calculations. In order to track each of the terms here, we shall break down the contributions to the chemical work done into three parts, one for each of the petals present in the step-wise process. These three petals correspond to adding a monomer type r, adding a monomer type w or removing a monomer type y. Let us label each of these contributions to the chemical work with a subscript, $\mathcal{G}_r(y,x)$ for the transition $\&xy \to \&xyr$, $\mathcal{G}_w(y,x)$ for the transition $\&xy \to \&xyw$, and $\mathcal{G}_q(y,x)$ for the transition $\&xy \to \&x$. From the r petal, we have:

$$\mathcal{G}_r(y,x) = \Bigg[\left(-\Delta G_{\rm act} + \ln\frac{M_{in}}{M_{\rm act}}\right) k_1 k_{\rm act} k_{KP} e^{-\Delta G_r}(M_{in} + M_{\rm act}e^{\Delta G_{\rm act}}) + (\Delta G_{\rm pol} + \Delta G_r - \Delta G_y + \ln M_{in} - \Delta G_a)(k_1 k_{\rm act} k_{\rm pol} M_{in} e^{-\Delta G_y}) + (\Delta G_{\rm pol} + \Delta G_r - \Delta G_y + \ln M_{\rm act}) k_{KP} k_{\rm pol} M_{\rm act} e^{-\Delta G_y}(k_1 e^{-\Delta G_r} + k_{\rm act})\Bigg] \times \frac{Q(w,y)Q(y,x)}{\mathcal{N}(y,x)J_{Tot}(y,x)}. \quad (54)$$

The first line of eqn. 54 corresponds to the chemical work associated with the internal cycle (inactive monomer binds,

gets activated, and activated monomer unbinds). The second line corresponds to an external cycle: an inactive monomer binds to the template, is activated and is polymerised into the chain with the previous monomer, y, detaching from the template. The third line corresponds to the alternative external cycle: an active monomer binds to the template and is polymerised with monomer y unbinding from the template. We may similarly write down $\mathscr{G}_w(y,x)$ as eqn. 54, except swapping r and w. Finally, the contribution to the chemical work from the petal for removing monomer y may be written:

$$\begin{aligned}\mathscr{G}_q(y,x) = \Big[&-(\Delta G_{\rm pol}+\Delta G_y-\Delta G_x+\ln M_{in}-\Delta G_a)(k_1 k_{\rm act} k_{\rm pol} e^{\Delta G_{\rm act}-\Delta G_y-\Delta G_{\rm pol}}) \\ &-(\Delta G_{\rm pol}+\Delta G_y-\Delta G_x+\ln M_{\rm act}) k_{KP} k_{\rm pol} e^{-\Delta G_{\rm pol}-\Delta G_y}(k_1 e^{-\Delta G_r}+k_{\rm act})\Big] \\ &\times \frac{Q(r,y)Q(w,y)}{\mathscr{N}(y,x)J_{Tot}(y,x)}.\end{aligned} \tag{55}$$

Here, only external cycles are possible. The first line corresponds to monomer x rebinding to the template, monomer y being depolymerised, this monomer being deactivated and an inactive monomer y unbinding from the template; and the second line to x rebinding, y being depolymerised and active monomer y unbinding from the template. The distribution, $\xi(y,x)$ may be calculated from eqn. 20 and P from eqn. 21 (both demonstrated in the supporting information), letting the chemical work done per net step of the the 1-loop model be written:

$$\Delta\mathscr{G} = \frac{1}{2P-1}\sum_{x,y\in\{r,w\}} \xi(y,x)\left(\mathscr{G}_r(y,x)+\mathscr{G}_w(y,x)+\mathscr{G}_q(y,x)\right). \tag{56}$$

This chemical work done is plotted for a certain set of parameters in figure 7 (b) and is also compared both to the results of direct simulation and the simpler "0-loop" model which has chemical work, $\Delta G_{\rm pol}$. The free-energy cost of the proofreading mechanism diverges as $\Delta G_{\rm pol}\to\Gamma$ since there will be a finite chemical work done per monomer addition/removal step due to the proofreading internal cycle, and the number of addition/removal steps per net step diverges. Further, for large $\Delta G_{\rm pol}$, the work tends to be dominated by $\Delta G_{\rm pol}$, albeit very slowly, as shown by the orange line gradually approaching $\Delta G_{\rm pol}$ (the blue line) in figure 7(b).

Additionally, we can find an expression for the time taken per net step forwards, eqn. 22. For this quantity, we need the explicit expression for the normalisation, $\mathscr{N}$. Similarly to the chemical work, we can split this term into contributions from the petal adding an r, $\mathscr{N}_r(y,x)$; from the petal adding a w, $\mathscr{N}_w(y,x)$; from the petal removing monomer y, $\mathscr{N}_q(y,x)$ and a contribution from the central node. These normalisation terms come from the sums of spanning trees directed to the individual nodes in the closed step-wise process. We see that

$$\begin{aligned}\mathscr{N}_r(y,x) \;=\; &\Big[k_1 M_{\rm in}(k_{\rm act}e^{\Delta G_{\rm act}}+k_{KP}e^{-\Delta G_r}+k_{\rm pol}e^{-\Delta G_y}) \\ &+k_{KP}k_{\rm act}M_{\rm act}e^{\Delta G_{\rm act}}+k_1 k_{\rm act}M_{\rm in}+k_{KP}k_{\rm act}M_{\rm act}+k_1 k_{KP}e^{-\Delta G_r}\Big] \\ &\times\; Q(y,x)Q(w,y),\end{aligned} \tag{57}$$

with a similar result for $\mathscr{N}_w(y,x)$ except swapping r and w. Finally, for the monomer removal petal, we have:

$$\begin{aligned}\mathscr{N}_q(y,x) = k_{\rm pol}e^{-\Delta G_{\rm pol}}(&k_1 e^{-\Delta G_y}+k_{\rm act} \\ &+k_{\rm act}e^{\Delta G_{\rm act}})Q(r,y)Q(w,y).\end{aligned} \tag{58}$$

The total normalisation is then:

$$\begin{aligned}\mathscr{N}(y,x) = &\mathscr{N}_r(y,x)+\mathscr{N}_w(y,x)+\mathscr{N}_q(y,x) \\ &+Q(y,x)Q(r,y)Q(w,y),\end{aligned} \tag{59}$$

with the last term being the contribution from the starting, central node. This normalisation can be used in eqn. 53 to give the current to absorbing states, which can be used in eqn. 22 to find the expected time per net step. This time is plotted in figure 7 (c), alongside a simulation of the same model and the simplified 0-loop model for comparison. Like the chemical work in figure 7 (b), the time per net step diverges as $\Delta G_{\rm pol}\to\Gamma$, since each monomer addition/removal step will take finite time, but the number of such steps required for a net forwards step diverges. Unsurprisingly, the time taken for a given driving for the Hopfield model is longer than that of the simple model, due to the proofreading cycle.

Hopfield's model for proofreading may be naturally extended to include N activation stages instead of just one.[48,50] We shall call these extensions the N-loop Hopfield models. These models can be solved recursively to write down expressions for the sums over spanning trees, $\Lambda_N^{\pm}(y,x)$, $Q_N(y,x)$, as a function of the number of loops, N. We shall consider the model as in figure 6 (b). A detailed derivation of the sums over spanning trees is given in appendix K. From these sums over spanning trees, we calculate the bulk frequencies, the time taken per net step and the chemical work done per net step using recursive relations (see appendix K).

For simplicity, we shall discuss the case where the monomer binding free energy is only dependent on monomer type, not on activation stage; each activation stage is associated with a free energy change of $\Delta G_{\rm act}$; each active monomer is present in the environment at a concentration $M_{\rm act}$ except the inactive monomers at concentration $M_{\rm in}$; and the overall rate constants are k_1 for binding of inactive monomer, $k_{\rm KP}$ for binding of active monomers, $k_{\rm act}$ for activation of monomers. Under these assumptions, the corresponding rates are given in appendix K.

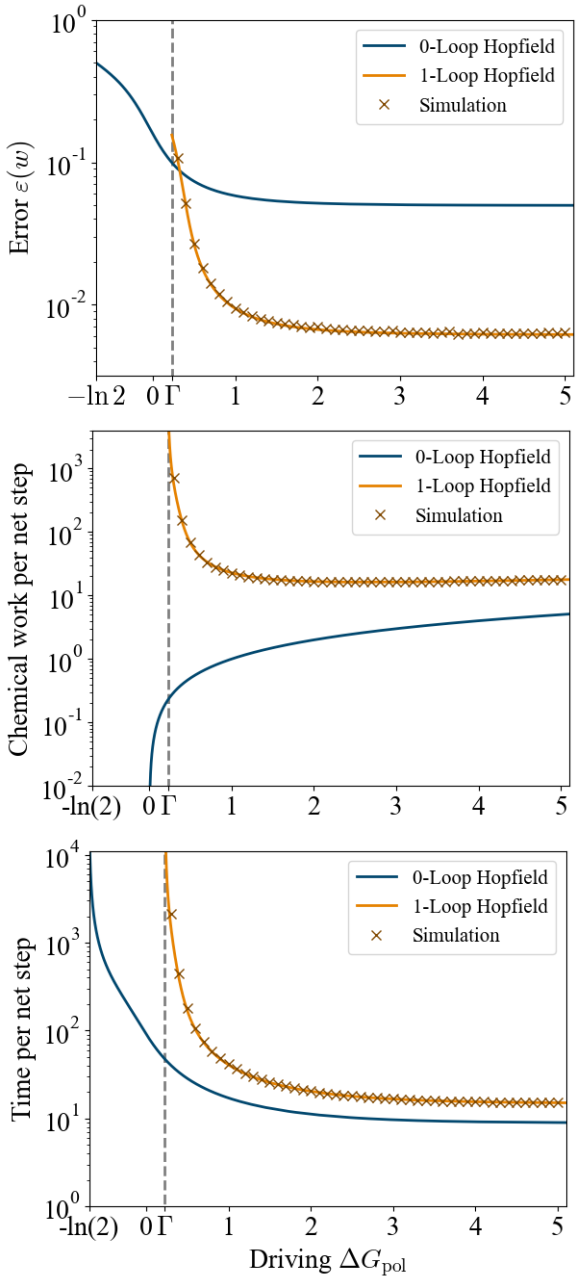


FIG. 7: The analytical method applied to a 1-loop proofreading model (figure 6 (a)), compared to Gillespie simulation of the same model and a simpler 0-loop model. For these data, the following parameters were used: $\Delta G_r = 2$, $\Delta G_w = -2$, $\Delta G_{\text{act}} = -1$ $M_{\text{in}} = 1$, $M_{\text{act}} = 0.01$, $k_1 = k_{\text{act}} = k_{\text{KP}} = 1$. The stall point, Γ, is marked on each of the plots. The Gillespie simulations used a template of length 2000 and were run till completion with the first monomer being chosen as either r or w with probability 0.5. The statistics were averaged over 2000 copolymers per data point. The chemical work was calculated from the simulation as $(\text{Inactive monomers}) * (\Delta G_{\text{act}} - \ln(M_{\text{act}}/M_{\text{in}})) + L * (\Delta G_{\text{pol}} + \ln M_{\text{act}})$ where "Inactive monomers" is the number of inactive monomers taken out of the environment and L is the length of the template.

To reduce the frequency of incorrect monomers in the product, we wish to have a low concentration M_{act} of active monomers in solution to force the system into utilising the proofreading cycles. Indeed, the bulk error probability in the irreversible limit (calculated using eqn. 36 and plotted in figure 8 (a) shows a strong improvement with loop number for low M_{act}, but larger values of M_{act} lead to much worse performance and limited (or negative) returns to increasing the number of loops.

However, for finite driving strength ΔG_{pol}, we cannot allow this concentration to be arbitrarily small. To see why, consider the stall point, $\Gamma(N)$, derived in appendix K and plotted for a certain set of parameters in figure 8 (b). It is observed that the stall point driving increases monotonically with N, and that this increase is faster and tends to a higher limit for smaller M_{act}. We find that the limiting Γ scales approximately linearly with $-\ln(M_{\text{act}})$. Intuitively, introducing more monomer states at low concentration in the environment destabilises the polymer. For small M_{act} and driving ΔG_{pol}, the depolymerisation of the polymer into these activated states competes with its tendency to grow by binding to and activating the inactive monomers.

One drawback of proofreading with a large number of loops is therefore that the tendency to disassemble the growing polymer increases. A second effect is a tendency to introduce errors by alternate pathways if M_{act} is non-zero. Specifically, for $M_{\text{act}} \neq 0$, we observe in figure 8 (a) a minimum in $\varepsilon_{\text{irrev}}(w)$ for a relatively small value of N. This minimum can be explained by splitting the pathways by which a monomer can go from solution to being incorporated into the polymer into two, either starting from a fully inactive monomer or from a partially activated one. The pathway starting with an inactive monomer will have the highest discrimination between right and wrong monomers and will improve exponentially with more loops, as demonstrated by the exponential decrease in error for $M_{\text{act}} = 0$. However, the probability that a monomer, taking this path, will reach polymerisation falls exponentially with loop number at the same time. On the other hand, the pathway from partially active monomers will give an error that reaches some non-zero limit as the number of loops, N, increases. Further, the rate with which activated monomers bind to an available template site and subsequently get incorporated into the polymer will also tend to a constant. As such, the error will initially decrease exponentially with N, but for non-zero M_{act} will eventually become dominated by the less discriminating, partially active monomer pathways through which monomers are more likely to be incorporated into the polymer.

Having calculated the error probability $\varepsilon(w)$ at finite driving, plotted in figure 9 (a); used $\varepsilon(x,y)$ to calculate the entropy rate; and calculated $\Delta\mathscr{G}$; we can evaluate the efficiency η, as in eqn. 26 (see supporting information for demonstrations). This efficiency is plotted in figure 9 for $N = 0, 1, 5, 10$ and a certain set of parameters. Although accuracy is generally increased above the stall point, we see that in this particular model kinetic proofreading requires much more work than the minimum required to generate information and as such are inefficient. Additionally, the gradient of the efficiency at min-

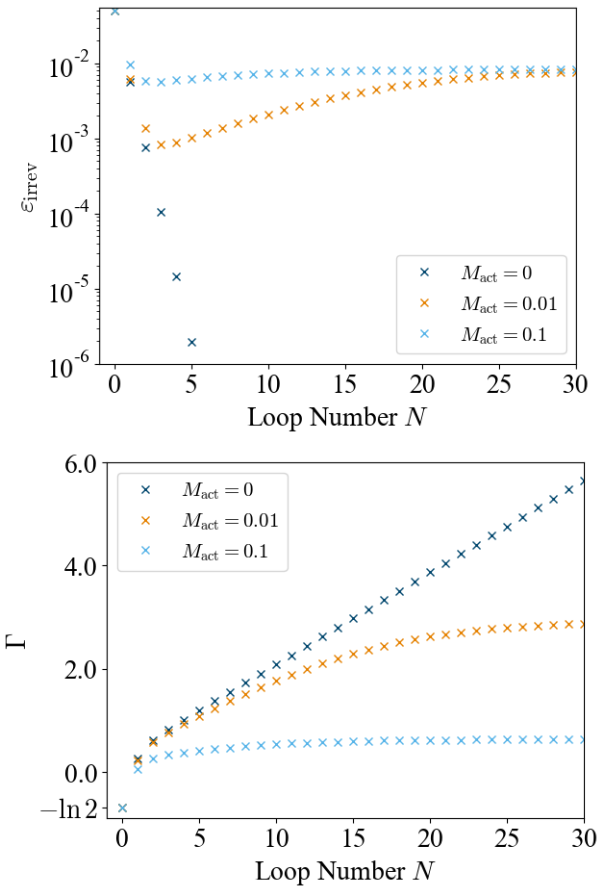


FIG. 8: Many-loop models have limited efficacy for finite M_{act} for the proofreading model introduced in figure 6. Plots of the error in the irreversible limit, $\varepsilon_{\text{irrev}}(w)$ and the stall point driving, Γ, for different values of the active monomer concentrations, M_{act}, and other parameters: $\Delta G_r = 2$, $\Delta G_w = -2$, $M_{\text{in}} = 1$, $\Delta G_{\text{act}} = -1$, $ks = 1$. The $N = 0$, error is not shown for clarity, but is 0.5 for all M_{act}.

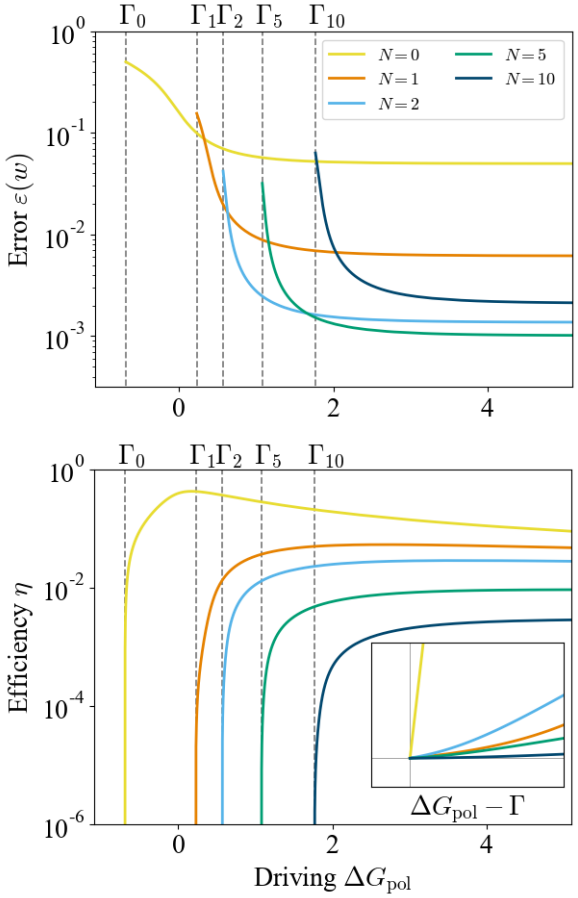


FIG. 9: Plots of (a) the error and (b) efficiency of the N-loop proofreading model (figure 6) for a range of N, with the same parameters as in the one loop Hopfield case, figure 7, as a function of driving ΔG_{pol}. Proofreading is observed to generally increase accuracy above its stall point, but in a thermodynamically inefficient way. The enhanced plot in the second graph shows the efficiencies near the stall point for each of the loop numbers on a non-logarithmic scale to emphasise the decreasing gradient at stall.

imum driving, Γ_N, is zero for $N > 0$, reflecting how at minimum driving, the number of monomer addition/removal steps diverges, but the chemical work done per such step remains finite.

IV. CONCLUSION

We have presented a method for analysing copolymerisation models with complex networks of reactions leading to the incorporation or removal of monomers. By coarse graining, a model may be transformed into a simpler model which may be solved and then afterwards, information from the fine-grained process may be put back into the model to extract thermodynamic or kinetic quantities such as chemical work done, molecule exchange or time taken. The approach allows for complex incorporation motifs to be considered alongside nearest neighbour interactions in a thermodynamically well-defined model of polymerisation with microscopic reversibility. We note that all of these features were present in the kinetic proofreading example in Section III C. Moreover, phenomena such as the shift in stall point with loop number and

the non-monotonicity of error rate with loop number rely on these features being present in the model.

In general, this method provides a way to extract model predictions numerically quickly and without the need for simulations. Doing so is particularly useful when simulating polymer growth is slow, either due to the details of the incorporation process or because the polymer is near its stall point. Additionally, the approach makes screening of a large parameter space for a given model topology feasible.

In addition to the numerical performance, the approach allows for analytic results in simpler models or those with helpful symmetries, as well as in certain limits for more complex models. The process of summing over spanning trees is particularly well suited to identifying the structure of the process and providing simplified results.

Moving forwards, it is an open question as to whether components of the techniques developed here can be applied outside of the context of infinitely long polymers whose tips have reached steady state. An obvious goal would be a simplified way to analyse finite-length "oligomers".[24]. More generally, we believe the key equation of this paper, eqn. 14, may be applied more generally for the coarse graining of Markov processes. Specifically, that if a set of states are enclosed between two boundary states, in the sense that any path from one of the trapped states to outside must pass through one of the boundary states, then this set of states may be replaced by a pair of edges analogously to eqn. 14 which shall preserve steady state properties of the Markov process.

This framework could be applied to explore models of copolymerisation processes such as those presented in[17–19,22,23,27,28,30–35,39–41,45–51] more straightforwardly or more thoroughly. Alternatively, the method would allow for more complex reaction steps to be included in such models. The framework presented here is particularly useful when backwards steps are relevant, either when the system is weakly driven and thus operating near stall, or when thermodynamics is of importance or interest. We also predict that it will be useful to guide design principles for synthetic copolymerisation systems, which are often particularly well-described by the class of models studied here.

SUPPLEMENTARY MATERIAL

The supplementary material contains a C++ script implementing the Gillespie algorithm that reproduces the data for the 1-loop Hopfield kinetic proofreading model presented in figure 7, and a MATLAB script for numerically calculating quantities of the 1-Loop and N-Loop Hopfield kinetic proofreading models presented in section III C and shown in the solid lines of figure 7, the points of figure 8 and figure 9.

ACKNOWLEDGEMENTS

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 851910). T.E.O. is supported by a Royal Society University Fellowship. J.J. is supported by a Royal Society PhD studentship.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

All authors conceived of the project. B.Q. produced the methodology and analysis and wrote the initial draft. All authors interpreted results and reviewed and edited this paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are reproducible from the code openly available in Zenodo at `https://doi.org/10.5281/zenodo.7271702`.

[1] P. Gaspard, Philosophical Transactions of the Royal Society A **374**, 20160147 (2016).
[2] A. P. Corfield, Biochimica et Biophysica Acta (BBA) - General Subjects **1850**, 236 (2015).
[3] S. Pinzón Martín, P. H. Seeberger, and D. Varón Silva, Frontiers in Chemistry **7**, 710 (2019).
[4] M. Chanda, *Introduction to polymer science and chemistry: a problem-solving approach* (CRC Press, Boca Raton, FL, 2013).
[5] C. G. Overberger, Journal of Polymer Science: Polymer Symposia **72**, 67 (1985).
[6] W. Meng, R. A. Muscat, M. L. McKee, P. J. Milnes, A. H. El-Sagheer, J. Bath, B. G. Davis, T. Brown, R. K. O'Reilly, and A. J. Turberfield, Nature Chemistry **8**, 542 (2016).
[7] H. Zhang, Y. Wang, H. Zhang, X. Liu, A. Lee, Q. Huang, F. Wang, J. Chao, H. Liu, J. Li, *et al.*, Nature Communications **10**, 1 (2019).
[8] F. Crick, Nature **227**, 561 (1970).
[9] B. Alberts, A. Johnson, J. Lewis, D. Morgan, and M. Raff, *Molecular Biology of the Cell* (Garland Science, Taylor & Francis Group, New York, NY, 2014).
[10] J. Niu, R. Hili, and D. R. Liu, Nature Chemistry **5**, 282 (2013).
[11] D. Kong, W. Yeung, and R. Hili, ACS Combinatorial Science **18**, 355 (2016).
[12] A. E. Stross, G. Iadevaia, D. Núñez-Villanueva, and C. A. Hunter, Journal of the American Chemical Society **139**, 12655 (2017).
[13] J.-F. Lutz, *Sequence-controlled polymers* (Wiley-VCH, Weinheim, Germany, 2018).
[14] D. Núñez-Villanueva, M. Ciaccia, G. Iadevaia, E. Sanna, and C. A. Hunter, Chemical Science **10**, 5258 (2019).
[15] D. Núñez-Villanueva and C. A. Hunter, Accounts of Chemical Research **54**, 1298 (2021).
[16] J. Cabello-Garcia, W. Bae, G.-B. V. Stan, and T. E. Ouldridge, ACS Nano **15**, 3272 (2021).
[17] F. T. Wall, Journal of the American Chemical Society **63**, 1862 (1941).
[18] F. T. Wall, Journal of the American Chemical Society **66**, 2050 (1944).
[19] F. R. Mayo and F. M. Lewis, Journal of the American Chemical Society **66**, 1594 (1944).
[20] T. E. Ouldridge, Natural Computing **17**, 3 (2018).
[21] S. Whitelam, R. Schulman, and L. Hedges, Physical Review Letters **109**, 265506 (2012).

[22]M. Nguyen and S. Vaikuntanathan, Proceedings of the National Academy of Sciences of the United States of America **113**, 14231 (2016).
[23]J. M. Poulton, P. R. ten Wolde, and T. E. Ouldridge, Proceedings of the National Academy of Sciences of the United States of America **116**, 1946 (2019).
[24]J. M. Poulton and T. E. Ouldridge, New Journal of Physics **23**, 063061 (2021).
[25]J. Juritz, J. M. Poulton, and T. E. Ouldridge, The Journal of Chemical Physics **156**, 074103 (2022).
[26]P. Sartori and S. Pigolotti, Physical Review Letters **110**, 188101 (2013).
[27]P. Sartori and S. Pigolotti, Physical Review X **5**, 041039 (2015).
[28]M. Sahoo, N. Arsha, P. R. Baral, and S. Klumpp, Physical Review E **104**, 034417 (2021).
[29]Y.-S. Song, Y.-G. Shu, X. Zhou, Z.-C. Ou-Yang, and M. Li, Journal of Physics: Condensed Matter **29**, 025101 (2016).
[30]Y. Song and C. Hyeon, The Journal of Physical Chemistry Letters **11**, 3136 (2020).
[31]Q.-S. Li, P.-D. Zheng, Y.-G. Shu, Z.-C. Ou-Yang, and M. Li, Physical Review E **100**, 012131 (2019).
[32]S. Pigolotti and P. Sartori, Journal of Statistical Physics **162**, 1167 (2016).
[33]F. Wong, A. Amir, and J. Gunawardena, Physical Review E **98**, 012420 (2018).
[34]W. D. Piñeros and T. Tlusty, Physical Review E **101**, 022415 (2020).
[35]C. H. Bennett, BioSystems **11**, 85 (1979).
[36]P. Gaspard, Physical Review E **93**, 042420 (2016).
[37]P. Gaspard, Physical Review E **93**, 042419 (2016).
[38]D. Andrieux and P. Gaspard, The Journal of Chemical Physics **130**, 014901 (2009).
[39]R. Rao and L. Peliti, Journal of Statistical Mechanics: Theory and Experiment **2015**, P06001 (2015).
[40]K. Banerjee, A. B. Kolomeisky, and O. A. Igoshin, Proceedings of the National Academy of Sciences of the United States of America **114**, 5183 (2017).
[41]D. Chiuchiù, Y. Tu, and S. Pigolotti, Physical Review Letters **123**, 038101 (2019).
[42]T. E. Ouldridge and P. R. ten Wolde, Physical Review Letters **118**, 158103 (2017).
[43]J. J. Hopfield, Proceedings of the National Academy of Sciences of the United States of America **71**, 4135 (1974).
[44]J. Ninio, Biochimie **57**, 587 (1975).
[45]J. D. Mallory, O. A. Igoshin, and A. B. Kolomeisky, The Journal of Physical Chemistry B **124**, 9289 (2020).
[46]A. Murugan, D. A. Huse, and S. Leibler, Physical Review X **4**, 021016 (2014).
[47]A. Murugan, D. A. Huse, and S. Leibler, Proceedings of the National Academy of Sciences of the United States of America **109**, 12034 (2012).
[48]Q. Yu, A. B. Kolomeisky, and O. A. Igoshin, Journal of the Royal Society Interface **19**, 20210883 (2022).
[49]M. Sahoo and S. Klumpp, Journal of Physics: Condensed Matter **25**, 374104 (2013).
[50]M. Ehrenberg and C. Blomberg, Biophysical Journal **31**, 333 (1980).
[51]V. Galstyan and R. Phillips, The Journal of Physical Chemistry B **123**, 10990 (2019).
[52]P. Gaspard and D. Andrieux, The Journal of Chemical Physics **141**, 044908 (2014).
[53]T. L. Hill, Journal of Theoretical Biology **10**, 442 (1966).
[54]T. L. Hill, Proceedings of the National Academy of Sciences of the United States of America **85**, 2879 (1988).
[55]J. G. Kemeny and J. L. Snell, *Finite Markov chains* (Springer, New York, NY, 1983).
[56]V. Anantharam and P. Tsoucas, Statistics & Probability Letters **8**, 189 (1989).
[57]A. Wachtel, R. Rao, and M. Esposito, New Journal of Physics **20**, 042002 (2018).
[58]H.-H. Kohler and E. Vollmerhaus, Journal of Mathematical Biology **9**, 275 (1980).
[59]F. Cady and H. Qian, Physical Biology **6**, 036011 (2009).
[60]P. Gaspard, Physical Review Letters **117**, 238101 (2016).
[61]M. Esposito, Physical Review E **85**, 041125 (2012).
[62]T. E. Ouldridge, R. Brittain, and P. R. ten Wolde, in *The Energetics of Computing in Life & Machines*, edited by C. Kempes, D. H. Wolpert, P. F. Stadler, and J. A. Grochow (SFI Press, Santa Fe, NM, 2019) pp. 307–351.
[63]T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, Inc., Hoboken, NJ, 2006).
[64]M. Esposito, K. Lindenberg, and C. Van den Broeck, Journal of Statistical Mechanics: Theory and Experiment **2010**, P01008 (2010).
[65]D. T. Gillespie, Journal of Computational Physics **22**, 403 (1976).
[66]D. Chiuchiù, J. Ferrare, and S. Pigolotti, Physical Review E **100**, 062502 (2019).
[67]M. Das and H. Kantz, Physical Review E **103**, 032110 (2021).
[68]V. Galstyan, K. Husain, F. Xiao, A. Murugan, and R. Phillips, eLife **9**, e60415 (2020).
[69]P. Gaspard, Physical Review E **96**, 042403 (2017).
[70]D. A. Harville, *Matrix Algebra From a Statistician's Perspective* (Springer, New York, NY, 1998).

Appendix A: Factorising sums of spanning trees

We note here that sums of spanning trees can be factorised in terms of Self-Avoiding Walks (SAWs), a result which is both useful for generating sets of spanning trees and allows us to make statements about ratios of propensities of balanced models. For a given process, $\mathscr{G} = (\mathscr{X}, K)$, for which we wish to find the sum of spanning trees rooted at $x_1 \in \mathscr{X}$, we may factorise this sum in terms of self-avoiding walks (SAWs) between two vertices in the graph. Select some other arbitrary vertex $x_2 \in \mathscr{X}/\{x_1\}$ and let $\mathscr{S}(x_2, x_1)$ be the set of SAWs from x_2 to x_1. For each $S \in \mathscr{S}(x_2, x_1)$, we can construct $\mathscr{G}_S = (\{s\} \cup (\mathscr{X}/S), K_S)$ analogously to eqn. 9, whereby we collapse the nodes in the SAW, S, into the single node s. The sum over spanning trees rooted at x_1 may then be written:

$$\sum_{T \in \mathscr{T}(x_1)} \prod_{e \in T} K(e) = \sum_{S \in \mathscr{S}(x_2, x_1)} \underbrace{\left[\prod_{e \in S} K(e)\right]}_{\text{SAW term}} \underbrace{\left[\sum_{T \in \mathscr{T}_S(s)} \prod_{e \in T} K_S(e)\right]}_{\text{Spanning tree term}}, \tag{A1}$$

where $\mathscr{T}(x), \mathscr{T}_S(x)$ are the sets of spanning trees directed to x for the original process, $\mathscr{G}$, and the new process, $\mathscr{G}_S$. For example, in figure 3(a), the spanning trees are arranged in terms of SAWs from node 1 to node 3, with the first row for SAW: $1 \to 2 \to 4 \to 3$; the second row for $1 \to 2 \to 3$; and the last three rows for $1 \to 3$. Similarly for figure 3(b), the trees are arranged in terms of SAWs from node 1 to node 4 with row one for $1 \to 2 \to 3 \to 4$; row two for $1 \to 2 \to 4$; row three for $1 \to 3 \to 4$; and row four for $1 \to 3 \to 2 \to 4$.

Appendix B: Normalisation constant for example absorbing Markov process

The normalisation constant for the closed example process, figure 2(b), can be found by considering the spanning trees rooted at each of the nodes. Factorising these in terms of SAWs, we write:

$$\begin{aligned}
\mathscr{N} = &\big[r_{34}r_{42}r_{21} + r_{32}r_{24}k_B + r_{32}r_{21}(r_{43} + r_{42} + k_B) + r_{34}k_B(r_{24} + r_{23} + r_{21}) \\
&+ (r_{31} + k_A)(r_{21}r_{43} + r_{21}k_B + r_{42}r_{21} + r_{23}r_{43} + r_{23}k_B + r_{42}r_{23} + r_{24}r_{43} + r_{24k_B})\big] \\
&+ \big[r_{13}r_{34}r_{42} + r_{13}r_{32}(r_{43} + r_{42} + k_B) \\
&+ r_{12}(r_{34}r_{42} + r_{43}r_{32} + r_{32}k_B + r_{32}r_{42} + r_{43}(r_{31} + k_A) + k_B(r_{31} + k_A) + r_{42}(r_{31} + k_A) + r_{34}k_B)\big] \\
&+ \big[r_{12}r_{24}r_{43} + r_{12}r_{23}(r_{42} + r_{43} + k_B) + r_{13}((r_{43} + k_B)(r_{21} + r_{23} + r_{24}) + r_{42}(r_{21} + r_{23}))\big] \\
&+ \big[r_{13}r_{34}(r_{21} + r_{23} + r_{24}) + r_{13}r_{32}r_{24} + r_{12}r_{23}r_{34} + r_{12}r_{24}(r_{31} + k_A + r_{32} + r_{34})\big].
\end{aligned} \tag{B1}$$

The first square bracket corresponds to the trees rooted at node 1, organised by SAWs from node 3; the second to trees rooted at 2 organised by SAWs from 1; the third to trees rooted at 3 organised by SAWs from 1 and the fourth to trees rooted at 4 organised by SAWs from 1.

Appendix C: Equivalence between chemical work calculated from Edges and cycles.

Here, we shall show the equivalence of chemical work for a process calculated by summing over edges versus summing over cycles. For this, consider a process $(\mathscr{X}, K)$, without any absorbing states (for simplicity) and such that every edge is microscopically reversible and let $\pi(x)$ be the steady state probability to be in state x. For an edge $x \leftrightharpoons y$, as described in section II A 4, the net current through this edge is:

$$J_{x \leftrightharpoons y} = \pi(x)K(x,y) - \pi(y)K(y,x). \tag{C1}$$

We can write $\pi(x)$ in terms of spanning tress by MCTT, and by appendix A, we may expand the sum over spanning trees by SAWs from y to x. For $\pi(y)$, we may expand by SAWs from x to y such the spanning tree terms of both expansions are the same and only the direction of edges in the SAW terms is flipped. The net current may then be written:

$$J_{x \leftrightharpoons y} = \frac{1}{\mathscr{N}} \sum_{S \in \mathscr{S}(y,x)} \left[K(x,y)\prod_{e \in S} K(e) - K(y,x)\prod_{e' \in S} K(e')\right] \left[\sum_{T \in \mathscr{T}_S(s)} \prod_{e \in T} K_S(e)\right] \tag{C2}$$

where $\mathscr{S}(x,y)$ is the set of SAWs from node x to node y; $\mathscr{N}$ is the normalisation as in eqn. 2, and e' is the edge in the opposite direction, i.e. if $e = x \to y$, $e' = y \to x$; and the last bracketed term is the spanning tree part for SAW, S, as in eqn. A1. One of

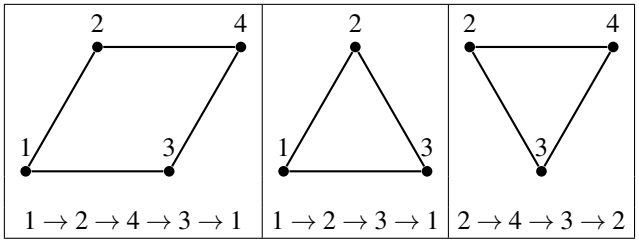
the SAWs from y to x will simply be the single transition $x \to y$, however, this term will cancel out from the sum leaving just the non-trivial SAWs. Taking a non-trivial SAW from y to x and multiplying by the rate $K(x,y)$ gives a cycle containing the edge $x \to y$. Therefore, the current may be written as a sum over cycle currents, as in section II A 4, of cycles which contain the edge $x \to y$ minus those which contain $y \to x$. Each of the edges contains a contribution to chemical work $\ln\left(\frac{K(x,y)}{K(y,x)}\right)$. The total chemical work before absorption is the sum over all edges of these contributions:

$$\mathscr{W}_{\text{chem}} = \sum_{x \rightleftharpoons y} \ln\left(\frac{K(x,y)}{K(y,x)}\right) \frac{J_{x \rightleftharpoons y}}{J_{\text{Tot}}}. \tag{C3}$$

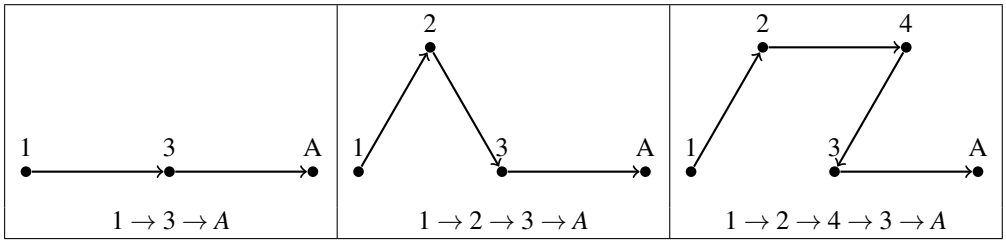
Since, in this sum the $J_{x \rightleftharpoons y}$ may be split up as a sum over cycles, we may collect the parts of this corresponding to given cycles and convert the sum over edges into a sum over cycles. Doing so we find the contribution to the chemical work from cycle, C, to be $ln\left(\frac{A(C)}{A(C')}\right)$, i.e. the affinities as we might expect. Hence, the sum over cycles is equivalent to the sum over edges.

Appendix D: Cycles of the example absorbing process

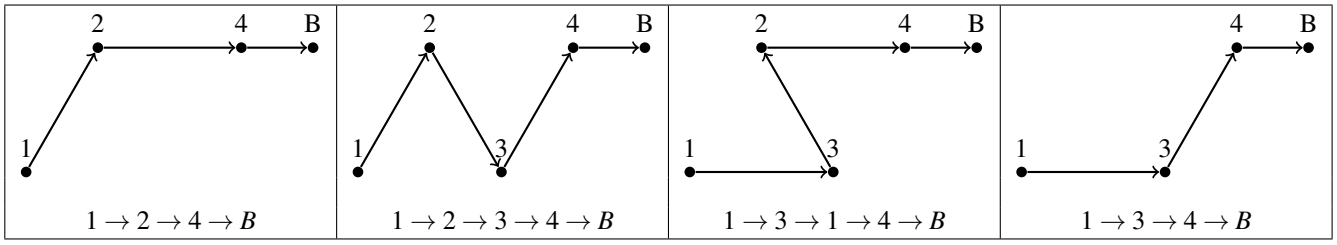
We make divide the cycles of the example process, figure 2(a), into internal cycles, external cycles to absorbing state A and external cycles to absorbing state B. Firstly, the internal cycles are:



where the cycle is written out below in the clockwise direction. Similarly, we find the external cycles to state A:



Finally, the external cycles to absorbing state B are:



Appendix E: Number of steps per net forward step of a random walk

Here we shall derive the number of steps per net forward step of a random walk. Let us set up a random walk as follows. Let the state space be the nodes $\{0,1,\cdots L\}$ where L is the length of the walk (polymer). Let the transition $0 \to 1$ have probability 1, $i \to i+1$ for $i = 1,\cdots L-1$ have probability p, $i \to i-1$ for $i = 1,\cdots L-1$ have probability $q = 1-p$ and let state L be an absorbing state as in figure 10.

We then wish to find the expected number of steps to absorption, given starting in state 0, for which we can utilise the spanning tree methods with eqn. 22. Since the total rate out of any state sums to one, the expected number of steps equals the expected

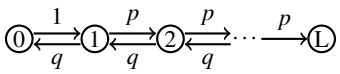


FIG. 10: Graphical representation of the random walk process considered.

time to absorption. Thus, we can form the closed process starting at 0. Let $f(n)$ be the sum over spanning trees rooted at node n for the closed process. $f(n)$ is given by:

$$f(n) = \begin{cases} \sum_{i=0}^{L-1} p^i q^{L-1-i} & \text{for } n = 0 \\ p^{n-1} \sum_{i=0}^{L-1-n} p^i q^{L-1-n-i} & \text{for } n = 1, \cdots, L-1 \end{cases}. \tag{E1}$$

From this, the expected number of steps before absorption is:

$$\mathbb{E}[\text{steps}] = \frac{\sum_{n=0}^{L-1} f(n)}{p f(L-1)}. \tag{E2}$$

By utilising the formulae for finite geometric series, we can find the expected number of steps to be:

$$\mathbb{E}[\text{steps}] = \frac{1}{2p-1}\left(L - 1 - \frac{q}{p^L}\left(\frac{p^L - q^L}{p-q}\right) + \frac{q^L}{p^L} + \frac{p^L - q^L}{p^{L-1}}\right). \tag{E3}$$

Most of this expression is sub-linear in L, and as such:

$$\lim_{L\to\infty} \frac{\mathbb{E}[\text{steps}]}{L} = \frac{1}{2p-1}, \tag{E4}$$

which is the net number of steps per net forward step.

Appendix F: The frequency at stall is given by the diagonal cofactors of a matrix

We wish to show that, at stall, the frequency with which a monomer appears in the bulk of the copolymer is proportional to the cofactor of the corresponding diagonal element of a matrix:

$$\varepsilon(x) \propto A_{xx}, \tag{F1}$$

where A_{ij} is the cofactor of element i, j of the matrix $\mathbb{1} - Z$. To show this relation we will rely on the relationship between cofactors and vectors of the nullspace of a matrix. Let M be an arbitrary matrix with a one dimensional nullspace, and let A be its matrix of cofactors. Recall that

$$MA^T = \det(M)\mathbb{1} = 0. \tag{F2}$$

Thus, any column of A^T is in the nullspace of M. In anticipation, let $\overrightarrow{\mu}$ be a vector in the nullspace of M and $\overrightarrow{v}$ be a vector in the nullspace of M^T. Since M has a one dimensional nullspace, then

$$\frac{\mu_x}{\mu_y} = \frac{A_{ix}}{A_{iy}}, \tag{F3}$$

for some arbitrary i. Similarly,

$$\frac{v_x}{v_y} = \frac{A_{xj}}{A_{yj}}, \tag{F4}$$

for arbitrary j.

Looking at eqns. 15, 16, noting that near the stall point, $v_z << \omega_{\pm y,x}$, we see that, the tip probabilities, $\mu(x)$, form a vector in the nullspace of $\mathbb{1}_M - Z$ and the tip velocities, v_x, form a vector in the nullspace of $\mathbb{1}_M - Z^T$. Hence, we have that

$$\frac{\mu(y)v_y}{\mu(x)v_x} = \frac{A_{jy}A_{yi}}{A_{jx}A_{xi}}, \tag{F5}$$

for arbitrary i, j. Thus, we may choose $j = y$ and $i = x$ leading to cancellation such that

$$\frac{\mu(x)v_x}{\mu(y)v_y} = \frac{A_{xx}}{A_{yy}}. \tag{F6}$$

Since

$$\varepsilon(x) = \frac{\mu(x)v_x}{\sum_y \mu(y)v_y} = \frac{A_{xx}}{\sum_y A_{yy}}, \tag{F7}$$

we get the required result.

Appendix G: The frequencies in the irreversible limit are given by the steady state of a process of the complete graph

We wish to find an expression for the frequency with which monomer x appears in the bulk of the copolymer in the irreversible limit. This limit is such that the backwards propensities, $\omega_{-yx} = 0$. With this assumption, from eqn. 15, we have

$$v_x = \sum_y \omega_{+yx}. \tag{G1}$$

With this form for the velocities, we may manipulate eqn. 16:

$$\begin{aligned} \mu(x) &= \sum_y \frac{\omega_{+yx}}{\sum_z \omega_{+zx}} \mu(y), \\ \sum_{z\neq x} \omega_{+zx}\mu(x) + \omega_{+xx}\mu(x) &= \sum_{y\neq x} \omega_{+xy}\mu(y) + \omega_{+xx}\mu(x). \end{aligned} \tag{G2}$$

This last line is the equation for steady state of a Markov process with probability $\mu(x)$ to be in state x and rate ω_{+yx} of transition from state x to state y. Thus, set $\mu(x)$ to be the steady state probability distribution of the Markov process on M states with transition rates from state x to y given by ω_{+yx}, and $v_x = \sum_y \omega_{+yx}$. Then, calculating

$$\varepsilon(x) \propto \mu(x)v_x, \tag{G3}$$

gives the required result. Finding the distribution, $\mu(x)$, in terms of spanning trees of the complete graph on M elements gives eqn. 36.

Appendix H: Simplification of results for factorisable ratios of propensities

We shall show that, if the ratio of propensities factorises as in eqn. 38, then we may simplify the stall condition and frequency of monomers at stall. Thinking of the functions X and Y as column vectors, since they have a discrete domain, the matrix Z may be written,

$$Z = \overrightarrow{X}\,\overrightarrow{Y}^T. \tag{H1}$$

By a well known result[70],

$$\det(\mathbb{1}_M - \overrightarrow{X}\,\overrightarrow{Y}^T) = 1 - \overrightarrow{Y}^T\overrightarrow{X} = 1 - \sum_x X(x)Y(x). \tag{H2}$$

rearranging gives eqn. 39. At stall, this bound is saturated. As shown, the frequency of monomer x in the bulk of the copolymer is given by the cofactor of the diagonal elements of $\mathbb{1}_M - \overrightarrow{X}\,\overrightarrow{Y}^T$. The cofactor, A_{xx}, may be written:

$$A_{xx} = \det(\mathbb{1}_{M-1} - \overrightarrow{X}_{[x]}\,\overrightarrow{Y}^T_{[x]}) = 1 - \sum_{y\neq x} X(y)Y(y) = X(x)Y(x), \tag{H3}$$

using the stall condition, and where $\overrightarrow{X}_{[x]}$ is the vector $\overrightarrow{X}$, missing element $X(x)$, i.e. $\overrightarrow{X}_{[x]} = (X(1), \cdots X(x-1), X(x+1), \cdots X(M))^T$. Additionally, because of the stall condition $\sum_x X(x)Y(x) = 1$,

$$\varepsilon_{\text{stall}}(x) = X(x)Y(x) \tag{H4}$$

is already normalised.

Appendix I: Frequency for a balanced model with two monomer types in the slow polymerisation limit

We shall derive the frequency of monomer x in the bulk of the copolymer with propensities given by eqn. 46, cancelling $n_{\rm com}k_{\rm com}$, with $R^-_{\rm com} = e^{-\Delta G_{\rm pol}}$, and with $M = 2$. With these propensities eqn. 15 becomes

$$v_1 = \frac{v_1}{e^{-\Delta G_{\rm pol}} + v_1} + \frac{e^{-DG} v_2}{e^{-\Delta G_{\rm pol}} + v_2} \tag{I1}$$

$$v_2 = \frac{e^{DG} v_1}{e^{-\Delta G_{\rm pol}} + v_1} + \frac{v_2}{e^{-\Delta G_{\rm pol}} + v_2}, \tag{I2}$$

where $DG = \Delta G_1 - \Delta G_2$. These equations may be solved by the following form the velocities:

$$v_x = e^{\Delta G_y - \Delta G_x} v_y. \tag{I3}$$

Doing so, reduces eqn. 15 to a quadratic equation,

$$0 = e^{DG} v_1^2 + (1 + e^{DG})(e^{-\Delta G_{\rm pol}} - 1) v_1 + e^{-\Delta G_{\rm pol}}(e^{-\Delta G_{\rm pol}} - 2) \tag{I4}$$

with one positive root,

$$v_1 = \frac{1}{2}\left((1 - e^{-\Delta G_{\rm pol}})(e^{-DG} + 1) + \sqrt{(e^{-\Delta G_{\rm pol}} - 1)^2 (e^{-DG} - 1)^2 + 4e^{-DG}} \right), \tag{I5}$$

when the system is not stalling. v_2 can be found from in terms of v_1 as $v_2 = e^{DG} v_1$. Further, a quick check confirms $v_1 = 0$ if $\Delta G_{\rm pol} = -\ln 2$. Further, with v_y known, eqn. 16 is a simple linear equation, μ can be found as the eigenvector of the matrix

$$\begin{pmatrix} \frac{1}{e^{-\Delta G_{\rm pol}} + v_1} & \frac{e^{DG}}{e^{-\Delta G_{\rm pol}} + v_1} \\ \frac{e^{-DG}}{e^{-\Delta G_{\rm pol}} + v_2} & \frac{1}{e^{-\Delta G_{\rm pol}} + v_2} \end{pmatrix}, \tag{I6}$$

with eigenvalue 1 and normalised to sum to 1. Combining the solutions for μ and v, using eqn. 18, gives eqn. 47.

Appendix J: Model used for balanced on-rate vs off-rate discrimination comparisons

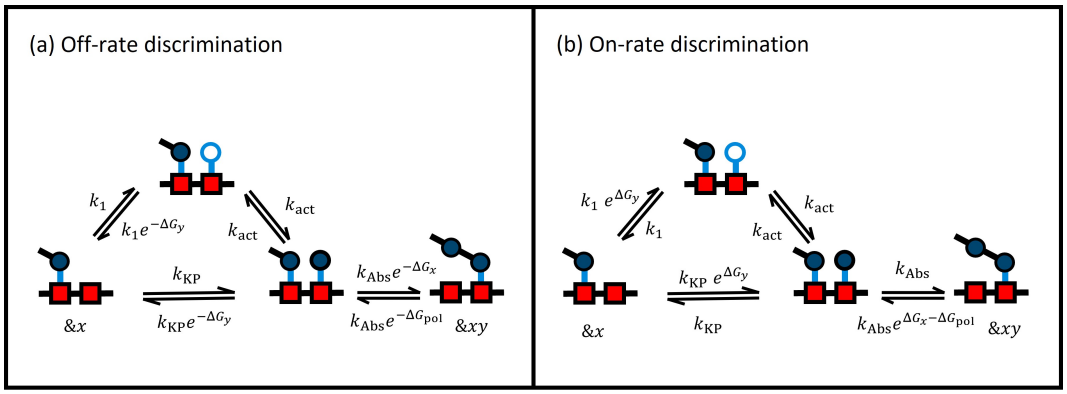


FIG. 11: Reaction rates of the (a) off-rate and (b) on-rate discrimination models used to produce the results of figure 5. These reactions represent a single petal of the step-wise process (figure 4) between completed states &x and &xy. For the results in figure 5 for the off-rate and on-rate curves, the following parameters were takes, $\Delta G_1 = 2$, $\Delta G_2 = -2$, $k_1 = k_{\rm KP} = k_{\rm act} = 1$, $k_{\rm com} = 100$.

Appendix K: Equations for N-loop Hopfield model

The sum over spanning trees of the N-loop model can be written in terms of sums over spanning trees of the lower loop number models. We label the reaction rates for the N-loop process as shown in figure 6. The N-loop model has one more node

and two more edges than the $N-1$-loop model. Let a subscript, N, denote the sums over spanning trees for the N-loop models. Tracking the spanning trees, we see,

$$\Lambda_N^+ = R_{\rm act}^+(N)\Lambda_{N-1}^+ + R_{KP}^+(N)\frac{R_{\rm pol}^+}{R_{\rm pol}^-}\sum_{i=0}^{N}\left[\prod_{j=0}^{i-1}R_{\rm act}^+(N-j)\right]\Lambda_{N-1-i}^-, \tag{K1}$$

$$\Lambda_N^- = R_{\rm act}^-(N)\Lambda_{N-1}^- + R_{KP}^-(N)\sum_{i=0}^{N}\left[\prod_{j=0}^{i-1}R_{\rm act}^+(N-j)\right]\Lambda_{N-1-i}^-, \tag{K2}$$

$$Q_N = \frac{1}{R_{\rm pol}^-}\left(\Lambda_N^- + R_{\rm pol}^+\sum_{i=0}^{N}\left[\prod_{j=0}^{i-1}R_{\rm act}^+(N-j)\right]\Lambda_{N-1-i}^-\right), \tag{K3}$$

with initial conditions

$$\begin{aligned}\Lambda_{-1}^{\pm} &= R_{\rm pol}^{\pm},\\ \Lambda_0^{\pm} &= R_{\rm pol}^{\pm}R_{\rm in}^{\pm},\\ Q_0 &= R_{\rm pol}^+ + R_{\rm in}^-.\end{aligned} \tag{K4}$$

The sum-product can be eliminated by subtracting terms proportional to $\Lambda_{N-1}^{\pm}, Q_{N-1}$, leaving just:

$$\Lambda_N^+ = \left(R_{\rm act}^+(N) + \frac{R_{KP}^+R_{\rm act}^+(N)}{R_{KP}^+(N-1)}\right)\Lambda_{N-1}^+ - \frac{R_{KP}^+(N)R_{\rm act}^+(N)R_{\rm act}^+(N-1)}{R_{KP}^+(N-1)}\Lambda_{N-2}^+ + R_{KP}^+(N)\frac{R_{\rm pol}^+}{R_{\rm pol}^-}\Lambda_{N-1}^-, \tag{K5}$$

$$\Lambda_N^- = \left(R_{\rm act}^-(N) + R_{KP}^-(N) + \frac{R_{KP}^-(N)R_{\rm act}^+(N)}{R_{KP}^-(N)}\right)\Lambda_{N-1}^- - \frac{R_{KP}^-(N)R_{\rm act}^+(N)R_{\rm act}^-(N-1)}{R_{KP}^-(N-1)}\Lambda_{N-2}^-, \tag{K6}$$

$$Q_N = R_{\rm act}^+(N)Q_{N-1} + \frac{\Lambda_N^-}{R_{\rm pol}^-} + (R_{\rm pol}^+ - R_{\rm act}^+(N))\frac{\Lambda_{N-1}^-}{R_{\rm pol}^-}, \tag{K7}$$

with the same initial conditions as above. This system of recursion relations may be used to generate the terms of the spanning tree sums quickly.

In certain simple cases eqns. K5, K6, K7 can be solved as a function of N. For example, when the reaction rates are not a function of N, such as:

$$\begin{aligned}R_{\rm in}^+ &= k_1 M_{\rm in},\\ R_{\rm in}^- &= k_1 e^{-\Delta G_y},\\ R_{\rm act}^+(n) &= k_{\rm act},\\ R_{\rm act}^-(n) &= k_{\rm act}e^{\Delta G_{\rm act}},\\ R_{KP}^+(n) &= k_{KP}M_{\rm act},\\ R_{KP}^-(n) &= k_{KP}e^{-\Delta G_y},\end{aligned} \tag{K8}$$

for $n \in \{1,\cdots N\}$, for the step-wise process with monomer tip $\&xy$, where k_i are some overall rates, $M_{\rm in}$, $M_{\rm act}$ represent the concentrations of inactive or active monomers, $\Delta G_{\rm act}$ represents the chemical work upon moving a monomer up one activation stage. The rates in eqn. K8 are used for the numeric results in figures 8 and 9. In this case, the sums over spanning trees are:

$$\begin{aligned}\Lambda_N^+(y,x) = k_{\rm pol}e^{-\Delta G_x}\Bigg(&k_{\rm act}^N\Big(k_1M_{\rm in} - k_1M_{\rm act} + k_{\rm act}e^{\Delta G_y}M_{\rm act}(e^{\Delta G_{\rm act}}-1)\Big)\\ &+ \frac{k_{KP}M_{\rm act}e^{-\Delta G_y}}{\Delta}\left[\frac{(k_1\lambda_+ + k_{\rm act}(k_{KP}-k_1))}{(\lambda_+ - k_{\rm act})^2}\lambda_+^{N+1} - \frac{(k_1\lambda_- + k_{\rm act}(k_{KP}-k_1))}{(\lambda_- - k_{\rm act})^2}\lambda_-^{N+1}\right]\Bigg),\end{aligned} \tag{K9}$$

$$\Lambda_N^-(y,x) = \frac{k_{\rm pol}e^{-\Delta G_{\rm pol}}e^{-\Delta G_y}}{\Delta}\Big[\left(k_1\lambda_+ + k_{\rm act}(k_{KP}-k_1)\right)\lambda_+^N - \left(k_1\lambda_- + k_{\rm act}(k_{KP}-k_1)\right)\lambda_-^N\Big], \tag{K10}$$

$$\begin{aligned}Q_N(y,x) = &\frac{k_{\rm pol}e^{-\Delta G_x}}{\Delta}\Big[(k_1e^{-\Delta G_y} + k_{\rm act} - \lambda_-)\lambda_+^N - (k_1e^{-\Delta G_y} + k_{\rm act} - \lambda_+)\lambda_-^N\Big]\\ &+ \frac{e^{-\Delta G_y}}{\Delta}\Big[\left(k_1\lambda_+ + k_{\rm act}(k_{KP}-k_1)\right)\lambda_+^N - \left(k_1\lambda_- + k_{\rm act}(k_{KP}-k_1)\right)\lambda_-^N\Big],\end{aligned} \tag{K11}$$

where

$$\lambda_{\pm} = \frac{1}{2}\left(k_{\text{act}} + k_{KP}e^{-\Delta G_y} + k_{\text{act}}e^{\Delta G_{\text{act}}} \pm \Delta\right), \tag{K12}$$

$$\Delta = \sqrt{\left(k_{\text{act}} + k_{KP}e^{-\Delta G_y} + k_{\text{act}}e^{\Delta G_{\text{act}}}\right)^2 - 4(k_{\text{act}})^2 e^{\Delta G_{\text{act}}}}. \tag{K13}$$

From eqns. K9, K10 we may write the stall condition. Noting that, $\Lambda^+(y,x)$ is independent of ΔG_{pol} and $\Lambda^-(y,x)$ is proportional to $e^{-\Delta G_{\text{pol}}}$, we may write the stall point as

$$\Gamma(N) = -\ln\left(\frac{\Lambda_N^+(r,r)}{e^{\Delta G_{\text{pol}}}\Lambda_N^-(r,r)} + \frac{\Lambda_N^+(w,w)}{e^{\Delta G_{\text{pol}}}\Lambda_N^-(w,w)}\right), \tag{K14}$$

such that the dependence on ΔG_{pol} in the logarithm is cancelled out.