# A New Model for Approximating RNA Folding Trajectories and Population Kinetics

Bonnie Kirkpatrick, Monir Hajiaghayi, and Anne Condon

February 8, 2013

**Abstract**

RNA participates both in functional aspects of the cell and in gene regulation. The interactions of these molecules are mediated by their secondary structure which can be viewed as a planar circle graph with arcs for all the chemical bonds between pairs of bases in the RNA sequence. The problem of predicting RNA secondary structure, specifically the chemically most probable structure, has many useful and efficient algorithms. This leaves RNA folding, the problem of predicting the dynamic behavior of RNA structure over time, as the main open problem. RNA folding is important for functional understanding, because some RNA molecules change secondary structure in response to interactions with the environment.

The full RNA folding model on at most $O(3^n)$ secondary structures is the gold standard. We present a new subset approximation model for the full model, give methods to analyze its accuracy, and discuss the relative merits of our model as compared with a pre-existing subset approximation. The main advantage of our model is that it generates Monte Carlo folding pathways with the same probabilities with which they are generated under the full model. The pre-existing subset approximation does not have this property.

**MSC Classification:** 62P10, 92B05, 60J27

**Keywords: RNA folding, continuous-time Markov chain, discrete state space, approximation**

# 1 Introduction

RNA molecules are involved in many cellular functions, including DNA transcription into mRNA, translation of RNA into proteins by tRNA [1], control of the plasmid copy number in *Escherichia coli* [2] and gene regulation via mechanisms that degrade mRNA [3]. RNA is also involved in gene expression regulation via splicing [4], and via competition [5]. The possible splice isoforms of the mRNA transcript are partly regulated by RNA structures [6].

When folding into their functional structures, i.e., three-dimensional shapes, RNA molecules dynamically move through a sequence of intermediate structures. In some cases these intermediate structures also contribute to the biological function of the molecule. For example, the function of the flavinmononucleotide riboswitch depends on this molecule's ability to change its structure [7]. In other examples, a molecule may be bistable, i.e., have two stable functional structures [8, 7].

While it would be ideal to have the full stereochemical description of RNA three-dimensional structure folding dynamics that describes changes in the bonds and bond angles over time, RNA *secondary structure* already provides a level of description that yields much insight into thermodynamics and folding kinetics of RNA [9]. Secondary structure is simply a set of C-G, A-U, and G-U base pairs - see Figure 1 for planar representations of two possible secondary structures for an RNA sequence of 23 nucleotides. A *folding pathway trajectory* is a sequence of secondary structures and *holding times*, where each successive secondary structure differs from the previous one by a single base pair and the holding time is the elapsed time in transitioning from one structure to the next. If we were to imagine a very large number of copies of the same RNA, say all initially in the unfolded state, and simulate a folding pathway for each copy, then we could count the fraction of copies occupying each secondary structure at each time point. The vector of these fractions over time is the *approximate population kinetics* of the RNA molecule (with respect to the initial unfolded state). For a given time point, the *exact* population kinetics are essentially the diffusion limit of the folding trajectory simulation at that time point, i.e., when the number of copies of the RNA is taken to infinity [10].

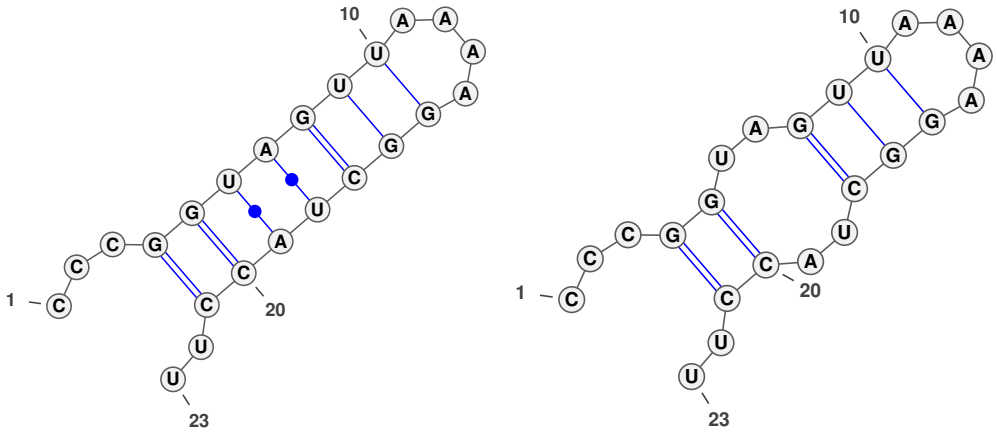



Figure 1: **Two RNA secondary structures** drawn using VARNA [11]. The structure on the left has energy −8.30 kcal/mol. These are two out of the many possible secondary structures.

RNA population kinetics were initially researched as a way to improve prediction of the functional secondary structure of RNA molecules [12, 13, 14]. Structure prediction could be improved by predicting structures that are very common in the population kinetics but that are not necessarily the thermodynamically most favourable, i.e., minimum free energy (MFE)

structure [15]. Just as the quality of MFE-based secondary structure prediction depends on the quality of the thermodynamic energy parameters used, it is also the case that the quality of population kinetics prediction depends on the kinetic rates used by the predictive model. Work that relies on or predicts these rates are grouped into two broad categories: 1) work at thermodynamic equilibrium without force acting on the system [16, 17, 18] and 2) work at a thermodynamic non-equilibrium where a force *is* acting on the system [19, 20].

In recent years, computational methods for RNA folding trajectory simulation have received increasing attention. Methods such as the Multistrand Simulator [21] and Kinfold [8] can be useful for understanding the structural mechanisms of RNA function in the above examples. For a given input RNA molecule and initial structure, these methods do a Monte-Carlo simulation of the folding trajectory, where the underlying state space is the set of all possible secondary structures for the input molecule and kinetic rates determine holding times and transition probabilities from one secondary structure to another. Since it seems quite difficult to obtain the necessary kinetic rates experimentally, in these simulators the kinetic transition rates are derived from nearest neighbor thermodynamic parameters [22]. Flamm et al. [8] describe an example of a 100-nucleotide long biological RNA, the Q$\beta$ variant S-11, whose MFE structure is a long hairpin interrupted by loops but whose functionally active secondary structure is very different with four distinct hairpins. They also demonstrate, through a folding pathway simulation, that the molecule is much more likely to fold into the functionally active structure and, once there, is unlikely to quickly fold further to reach the MFE structure because of a high energy barrier needed to do so. The Multistrand Simulator has also been used in the design of nano-scale machines that have potential health applications. For example, there is a designed RNA that can detect a cancer mutation and activate the cell death pathway [23]. In another example, designed RNAs were used to map simultaneous RNA expression patterns in intact biological samples [24].

Because the simulation model of Flamm et al. has all secondary structures in its state space, we refer to it as the *full model*. This is our gold standard. The full model can also be used to infer exact or approximate population kinetics (see our Background section 2). However, inferring the population kinetics from the full model is slow, in part because of the size of the underlying state space. For an RNA sequence of length $n$ there can be as many as $\binom{n}{2k}$ ways to create a secondary structure of $k$ base pairs, and thus $O(3^n)$ possible secondary structures (this follows from a bound on the Catalan numbers and on the sum of binomial coefficients). For this reason, many alternative approaches for approximate folding trajectory simulations have been described [25, 26, 27, 28, 29, 30, 31, 32, 33]. Some of this work, such as that of Munsky and Khammash [31], presents a useful approximate model that provides a lower bound on the probabilities in the full model. However, samples from their model are not valid folding trajectories, because one state of their model is not a secondary structure but rather represents a collection of secondary structures. In contrast, Tang et al [32] present a model that has no known accuracy bounds but whose state space is a subset of the set of all secondary structure for the input RNA molecule (i.e., the state space of the full model). Because of this, the Tang et al. simulator can produce valid folding trajectories, however the dynamics of the Tang et al. model are slow relative to the full model (detailed

discussion in Claim 2).

In our work we focus on *subset models* such as that of Tang et al.—models in which the state space is a subset of the set of all secondary structures for the input molecule. There are several challenges in designing such a model, even when the subset $S$ of secondary structure states has been chosen. How should kinetic rates between states be defined? How might different subset models be compared, in terms of providing the best approximation to the population kinetics?

Our new contributions in this paper address these challenges. First, we present a subset model that, for a given state space $S$ of structures which is a connected subset of the set of all possible secondary structures $U$ (the state space of the full model), has kinetic transition rates that preserve the holding times of the full model with respect to the subset $S$. This is important, because it means that folding trajectories are generated by our subset model with the same probabilities as by the full model when conditioned to produce only trajectories over $S$. Second, we present a measure of the accuracy of a folding trajectory simulation using a subset model, in terms of the quality of its estimate of population kinetics. This makes possible a rigorous empirical comparison of different subset models on small RNA's, which we present here, the results of which indicate that our subset model performs with greater accuracy than that of Tang et al. on short time scales. In the long term, we find that, on our examples, both our subset model and that of Tang et al. converge to stationary distributions whose accuracies are numerically indistinguishable.

The rest of the paper is organized as follows. In section 2 we provide some background on RNA secondary structures and folding trajectories. We review the full model of Flamm et al. and the subset model for RNA folding simulation of Tang et al. and describe how population kinetics can be computed, either approximately using a Monte Carlo simulation or exactly using the solution to a system of ordinary differential equations. Equipped with this background, in Section 3, we present our new approximative model of RNA folding whose holding times match those of the full model. We also introduce our measure for assessing the population kinetic inaccuracy of a particular approximative model relative to the full model. In Section 4, using our inaccuracy function, we empirically compare our new subset model with the subset model of Tang et al. [32] on five biological RNA's. Finally, we provide a simulation paradigm that can be used to do broader comparison of two subset models on synthetic data, and discuss the relative merits of the two approximative RNA folding models.

# 2 Background

Here and throughout, assume that an RNA molecule of length $n$ is given. We can formalize a *base pair* as tuple $(a, b)$ such that $a < b$ which specifies that the base at position $a$ in the sequence is bonded to the base at position $b$ to form an allowed base pair, one of C-G, A-U, and G-U. A secondary structure is a set of base pair tuples, $i = \{(a_1, b_1), (a_2, b_2), ..., (a_{|i|}, b_{|i|})\}$, that satisfies the following property: for every pair of tuples, $(a_j, b_j)$ and $(a_{j'}, b_{j'})$, we have that $a_j \neq b_j \neq a_{j'} \neq b_{j'}$. The secondary structure is *pseudoknot free*, if neither $a_j < a_{j'} <$

$b_j < b_{j'}$ nor $a_{j'} < a_j < b_{j'} < b_j$ can be true, for all tuples $(a_j, b_j)$ and $(a_{j'}, b_{j'})$. If a secondary structure is not pseudoknot free, it is said to have a *pseudoknot*.

We will restrict our attention to pseudoknot-free secondary structures which is a large and important class of structures. All claims and models in this paper can trivially be extended to the case of pseudoknots by simply including psudoknotted structures in the model and using free energies of pseudoknots to derive kinetic rates. Our empirical results depend non-trivially on the assumption of pseudoknot-free structures. However, we are doing calculations on small RNA which tend to be pseudoknot free, and the restriction to pseudoknot free structures enable us to build on the simulator of Flamm et al., which is for pseudoknot free structures.

A *folding pathway* of $m$ steps consists of a sequence of secondary structures $\sigma = i_1, i_2, ..., i_m$ where each successive secondary structure differs from the previous one by a single base pair, i.e. the Hamming distance between $i_{j-1}$ and $i_j$ is exactly 1 for all $2 \leq j \leq m$. A *folding pathway trajectory* of $m$ steps is a sequence of tuples $(s_0, t_0), (s_1, t_1), ..., (s_m, t_m)$ where $s_i$ is the secondary structure at time $i$ and $t_i$ is the holding time for state $s_i$ that was drawn at step $i$. The collection of structures in a folding pathway or trajectory is not necessarily distinct.

The equilibrium probability that an RNA molecule will occupy a particular secondary structure is

$$k_i = \frac{e^{-E(i)/(\rho\tau)}}{Z_k}$$

where $E(i)$ is the free energy of $i$, $\rho$ is the gas constant, $\tau$ is the temperature, and the partition function $Z_k = \sum_{j \in U} e^{-E(j)/(\rho\tau)}$, where $U$ is the set of all possible (pseudoknot free) secondary structures for the molecule. Then the free energy, $E(i)$, can be computed in time $O(n)$ and the partition function can be computed using an $O(n^3)$ dynamic programming algorithm [34]. The vector $k$ is a distribution since the sum of the $k_i$'s is one, and is known as the Boltzmann distribution.

For example $\tau = 310.15$ K is roughly body temperature. The gas constant is $\rho = 1.985 \times 10^3$ kcal/(K mol), making $\rho\tau = 6.1565 \times 10^5$ kcal/mol. The gas constant is related to the Boltzmann constant as $\rho = N_A k_B$ where $k_B$ is the Boltzmann constant and $N_A$ is Avogadro's number. In Figure 1, the energy of the left structure is $-8.30$ kcal/mol. Since $\rho\tau$ and $E$ have the same units, this makes the Boltzmann distribution unit-less.

## 2.1 The RNA Folding Models

Flamm et al. [8] introduced the RNA folding model that was inspired by chemical reaction systems modeled using continuous-time Markov chains (CTMCs). A CTMC is defined via its infinitesimal generator matrix and an initial distribution over its states. Like discrete Markov chains, many CTMCs have a stationary distribution. In this paper, we will define three CTMCs used for RNA folding and give their stationary distributions. Two of these CTMCs appear in this background section as they were discovered by other authors. The novel CTMC that we present appears in the methods section along with the other new contributions of this paper.

Assume that an RNA molecule is given, and let $U$ be the space of pseudoknot free secondary structures for that molecule. Let $N(i) \subset U$ be the neighborhood of $i$ where $\{i\} \cap N(i) = \emptyset$ and symmetry is preserved (if $j \in N(i)$ then $i \in N(j)$ for all $i \neq j, i, j \in U$). For the full model defined directly below, the neighborhood of $i$ includes every structure that differs from $i$ by a single base pair.

We say that secondary structures $i$ and $j$ are connected if there is a path of structures from $i$ to $j$ where each subsequent structure is a neighbor of the previous one, i.e. $i = s_1, s_2, ..., s_{\ell-1}, s_\ell = j$ such that $s_\ell \in N(s_{\ell-1})$ for all $\ell$. Let $S \subset U$ be a set of connected nodes. The choice of $S$ is a complicated open problem and consideration of this is deferred to Section 5 and mentioned in the Appendix.

**The Full Model [8].** Kawasaki introduced an 'inverse-symmetric' transition rule, by this we mean that $K_{ij} = K_{ji}^{-1}$ for $i \neq j$. The Kawasaki transition rule gives the infinitesimal generator matrix for $i \neq j$ as

$$K_{ij} = \begin{cases} e^{(E(i)-E(j))/(2\rho\tau)} & \text{if } j \in N(i) \\ 0 & \text{if } j \notin N(i) \end{cases}$$

and

$$K_{ii} = -\sum_{j \neq i} K_{ij}.$$

The neighborhood of $i$ is every structure that differs from $i$ by exactly one base pair. Notice that although Flamm et al. also included as neighbors a pair of structures that differ by shifted base pairs, we do not. This means that every pair of secondary structures is connected. We use the 'full model' to refer to the CTMT with infinitesimal matrix $K$ and with the initial state, $x$, being that distribution over the structures that the RNA molecule begins in. Typically $x$ is the point-mass on either the open chain or the minimum free energy state. The concepts introduced in what follows apply to any other initial state distribution.

Informally, the *population kinetics* at a given time $t$ is a vector whose $j$th entry is the probability that the CTMT is in state (secondary structure) $j$ at time $t$. Formally, for a CTMT with initial distribution $x$ and infinitesimal generator matrix $Q$, the population kinetics is given by $xe^{tQ}$. In the case that $Q = K$, as $t \to \infty$ the population kinetics converges to the stationary distribution given by the Boltzmann distribution $k$ (this follows from the fact that $kK = 0$). We note that a CTMC with stationary distribution $\pi$ and generator matrix $Q$ is called *reversible* if and only if the following equation, namely the *detailed balance* condition, holds for every pair of states $i$ and $j$, $\pi_i Q_{ij} = \pi_j Q_{ji}$. Process $K$ with stationary distribution $k$ is reversible and satisfies detailed balance.

**The Subset Model [35].** The main challenge with obtaining the matrix $K$ is that the size of the state space $U$ is exponential in the number of bases in an RNA molecule. For this reason, some approximations take a subset of connected structures $S \subseteq U$ [31], such that the initial (unfolded) state is in $S$. Some models design a CTMC on the subset of nodes $S$

that converges to the Boltzmann distribution normalized on the subset $S$ of structures [35]. Next, we describe such a model.

More specifically, if we again use the Kawasaki rule as in Tang et al. [35], we obtain the infinitesimal rate matrix

$$L_{ij} = \begin{cases} K_{ij} & \text{if } i \in S \text{ and } j \in N(i) \cap S \\ 0 & \text{otherwise} \end{cases}$$

and

$$L_{ii} = -\sum_{j \neq i} L_{ij} \geq K_{ii}.$$

Notice that since $L_{ii} \geq K_{ii}$ the process $L$ will have slower holding times on average than does the process $K$. More discussion of the holding times will follow.

The process with infinitesimal rate matrix $L$ converges to the distribution on $S$ which is proportional to the Boltzmann distribution

$$\ell_i = \frac{e^{-E(i)/(\rho\tau)}}{Z_\ell}$$

where $Z_\ell = \sum_{j \in S} e^{-E(j)/(\rho\tau)}$.

We will refer to an infinitesimal generator matrix and its associated process by the same name, i.e. the process having infinitesimal generator matrix $L$ is process $L$. We will also call this model the 'subset' model due to its treatment of the subset $S$ as if they were a full set of possible secondary structures.

## 2.2 Desired Calculations

Recall that the overall goal of of this paper is to introduce and evaluate a new approach for approximating population kinetics of an RNA molecule. In this section, we describe two ways that the population kinetics of an RNA molecule can be calculated: either from the transition probabilities of the corresponding CTMT or from a Monte Carlo simulation.

For an arbitrary finite CTMC with infinitesimal generator matrix $Q$, the *transition probability* $P^Q(t)_{ij}$ is defined to be the probability that the CTMT transitions to state $j$ at time $t$, given that it is in state $i$. The corresponding matrix of transition probabilities $P^Q(t)$ satisfies the ODE

$$\frac{dP^Q(t)}{dt} = QP^Q(t), \quad t \geq 0.$$

This ODE can be solved using an ODE solver or matrix exponentiation (see Norris [10] 2.1.1 for a proof):

$$P^Q(t) = e^{tQ} := \sum_{\nu=1}^{\infty} \frac{(Qt)^\nu}{\nu!}. \tag{1}$$

From the matrix exponential, we can immediately obtain the exact population kinetics which are the vectors $xe^{tQ}$, where $x$ is the initial distribution of the CTMC.

Equation 1 can be computed through matrix exponentiation methods such as spectral decomposition or the eigenvalues using the Lanczos algorithm [36]. However, these matrix exponentiation algorithms have exponential running time, due to $|U|$, and thus the size of the matrix, being exponential. Munsky and Khammash provide an alternative method for approximating the matrix exponential [31].

A second way to compute the population kinetics is the Monte Carlo simulation of folding pathway trajectories from process $Q$. A single sample of such a simulation tracks the progress of a single RNA molecule as it moves across the secondary structure state space. Monte Carlo simulation (or Doob-Gillespie algorithm [37]) is accomplished using the exponential holding-time distribution and the embedded Markov chain [38]:

**(1)** The holding time $T_i$ is exponentially distributed with parameter $-Q_{ii}$.

**(2)** The *embedded (discrete) Markov chain* has transition matrix $H_{ij} = -Q_{ij}/Q_{ii}$ when $Q_{ii} < 0$ whereas $H_{ii} = 0\ \forall i$. The self-loop transition probability $H_{ii}$ is zero to represent that self-transitions are invisible in continuous time.

Initially $s_i$ is distributed according to the initial distribution $x$. Steps (1)-(2) yield a single tuple, $(s_i, t_i)$, and a trajectory is made of $m$ tuples, $\theta = (s_0, t_0), (s_1, t_1), ..., (s_m, t_m)$. Let $\Theta$ be the set of trajectories sampled. Given these $|\Theta|$ samples, approximate population kinetics at a given time $t$ can be computed by calculating the fraction of samples in which the RNA molecule is in a given secondary structure at time $t$. Recall that the population kinetics is a vector and it's $i$'th component is approximated as

$$(\tilde{p}(t))_i = (1/|\Theta|) \cdot \left| \left\{ s_k = i : t \in \left[ \sum_{j<k} t_j, \sum_{j<k+1} t_j \right] \right\} \right|$$

The approximation $\tilde{p}(t)$ converges to the exact population kinetics, $xe^{tQ}$, as the number of samples grows [10]. This Monte Carlo simulation is the one performed by the RNA folding software Kinfold [8] with process $K$ and with both Metropolis-Hastings and Kawasaki versions of process $K$. The Metropolis-Hastings rule for our method will be discussed in detail in the Appendix.

# 3 Methods

We introduce the model $\Gamma$ which we will show has holding times equal to the holding times of $K$. Therefore Monte Carlo simulation of $\Gamma$ is equivalent to simulating trajectories from the full model and rejecting those that do not contain nodes in $S$. This equality of the holding time of $K$ and $\Gamma$ is obtained by sacrificing proportionality of the stationary distribution to the Boltzmann distribution.

**The Trajectory Subset Model.** Define $\Gamma$ as follows

$$\Gamma_{ij} = \begin{cases} \frac{\sum_{k \in U_{-i}} K_{ik}}{\sum_{k \in S_{-i}} K_{ik}} K_{ij} & \text{if } i \in S \text{ and } j \in N(i) \cap S \\ 0 & \text{if } i \notin S \text{ or } j \notin N(i) \cap S. \end{cases}$$

where we use the short notation $U_{-i} = U \setminus \{i\}$ and where

$$\Gamma_{ii} = -\sum_{j \neq i} \Gamma_{ij}.$$

The stationary distribution $\gamma$ of this model $\Gamma$ is related to the Boltzmann distribution as follows

$$\gamma_i = \frac{e^{E(i)/(\rho\tau)} \sum_{k \in S_{-i}} e^{-E(k)/(2\rho\tau)}}{Z_\gamma \sum_{k \in U_{-i}} e^{-E(k)/(2\rho\tau)}} \text{ where } Z_\gamma = \sum_{i \in S} \frac{e^{E(i)/(\rho\tau)} \sum_{k \in S_{-i}} e^{-E(k)/(2\rho\tau)}}{\sum_{k \in U_{-i}} e^{-E(k)/(2\rho\tau)}}.$$

The proof appears in the Appendix. A trajectory subset model can also be defined for the full process using the Metropolis-Hastings rule. This is also discussed in the Appendix.

We will argue that the trajectory subset model better approximates population kinetics of the full model at short time scales than the subset model. This appears to be due to the Claims 1 and 2 which show that on one-step time-scales, the trajectory subset model is more accurate.

Recall that a folding pathway trajectory of $m$ steps of the Monte Carlo simulation is a sequence of tuples $(s_0, t_0), (s_1, t_1), ..., (s_m, t_m)$ where $s_i$ is the secondary structure at time $i$ and $t_i$ is the holding-time for state $s_i$ that was drawn at step $i$. Now we can introduce the main claim.

**Claim 1.** *Let $\theta = (s_0, t_0), (s_1, t_1), ..., (s_m, t_m)$ be a trajectory of the full model $K$ having probability $p_K(\theta)$. Let $p_K(\theta|S)$ represent the probability of a trajectory when conditioning on it containing only structures from set $S$. $\theta$ is a trajectory of the $\Gamma$ model such that $p_K(\theta|S) = p_\Gamma(\theta)$ if and only if all the states of $\theta$ are in the subset, i.e. $s_i \in S$ for all $i$.*

*Proof.* A trajectory for model $Q$ has probability which is the scalar product of the exponentially-distributed holding-time probabilities and transition probabilities for every step.

$$\begin{aligned} p_Q(\theta) &= \prod_{i=0}^{m} -Q_{s_i,s_i} e^{Q_{s_i,s_i} t_i} \cdot \left( \frac{Q_{s_i,s_{i+1}}}{-Q_{s_i,s_i}} \right) \\ &= \prod_{i=0}^{m} e^{Q_{s_i,s_i} t_i} \cdot Q_{s_i,s_{i+1}} \end{aligned}$$

For process $K$, we can condition on trajectories being in set $S$ by simply checking whether all the structures $s_i$ for $1 \leq i \leq m$ are in $S$. Then, we can write $p_K(\theta|S) = \prod_{i=0}^{m} e^{K_{s_i,s_i} t_i} \cdot K_{s_i,s_{i+1}} / (\sum_{s_j \in S_{-i}} K_{s_i,s_j})$ where the transition probability is re-normalized to consider only transitions to states in $S$.

For process $\Gamma$ we will write the trajectory probability after two short calculations. From the properties of the CTMC, we have

$$\Gamma_{ii} = -\sum_{k \in U_{-i}} K_{ik} = K_{ii},$$

$$H_{ij} = \frac{-\Gamma_{ij}}{\Gamma_{ii}} = \frac{K_{ij}}{\sum_{j \in S_{-i}} K_{ij}},$$

and $H_{ii} = 0 \ \forall i$. Now we can write the trajectory probability as

$$\begin{aligned} p_\Gamma(\theta) &= \prod_{i=0}^{m} e^{K_{s_i,s_i} t_i} \cdot \left( \frac{K_{s_i,s_{i+1}}}{\sum_{s_j \in S_{-i}} K_{s_i,s_j}} \right) \\ &= p_K(\theta|S) \end{aligned}$$

□

This claim tells us that that simulating under $\Gamma$ is *equivalent* to simulating trajectories from the full model and rejecting those with states not in $S$. Therefore, every trajectory of $\Gamma$ is a trajectory of $K$.

The same cannot be said for the subset model $L$. Trajectories under $L$ cannot be trajectories for the full model, because the holding times are not drawn from the same distribution. Instead the trajectories under $L$ behave as if the set $S \subseteq U$ were the entire universe of secondary structures rather a proper subset of $U$. We formalize this in the following claim.

**Claim 2.** *Trajectories under $L$ have average holding times that are at least as long as those under $\Gamma$ and $K$. The average holding times under $U$, $\Gamma$, and $K$ are equal if and only if $S = U$.*

*Proof.* Recall that the parameter for the holding time of model $\Gamma$ is

$$\Gamma_{ii} = -\sum_{k \in U_{-i}} K_{ik} = K_{ii}.$$

We can also derive the parameter for the holding time of model $L$ as follows

$$L_{ii} = -\sum_{j \neq i} L_{ij} = -\sum_{k \in S_{-i}} K_{ik}.$$

The holding-time parameters $|L_{ii}|$ and $|K_{ii}|$ for the exponential distribution are related by

$$|L_{ii}| \leq |K_{ii}| = |\Gamma_{ii}|. \tag{2}$$

Since the mean holding times are $1/|L_{ii}|$ and $1/|K_{ii}|$, respectively, this means that the mean holding time for $L$ is at least as long as that for $K$ and $\Gamma$. Since Inequality 2 is equal if and only if $S = U$, this is the only condition under which the average holding times of all three models are equal. □

## 3.1 Accuracy: Comparing Population Kinetics

Here, we introduce a function that can be used to compare the quality of the population kinetics approximations inferred from the subset model and trajectory subset model. The function can also be applied more generally. First, recall that the population kinetics of the full model $K$ at a given time $t$ has distribution $xe^{Kt}$, which converges to the Boltzmann distribution $k$ as $t \to \infty$. As an example, two components (i.e., vector entries) of the vector-valued population kinetics are illustrated in Figure 2; for one state, the population density is initially 0 and then quickly increases, stabilizing at approximately 0.6, while the other initially has density 1 but quickly approaches 0. Now, suppose that $Q$ is a different process, such as the subset model or our trajectory subset model, and let $q$ be the stationary distribution of $Q$; that is, $xe^{Qt} \to q$ as $t \to \infty$. We wish to measure how well $Q$ approximates $K$. We define the inaccuracy of $Q$ at a given point $t$ to be

$$\begin{aligned} A(Q,t) &:= \left\| xe^{Kt} - xe^{Qt} \right\|_1 \\ &= \left\| xP_K(t) - xP_Q(t) \right\|_1 \\ &\to \|k - q\| \text{ as } t \to \infty \end{aligned} \tag{3}$$

where $x$ is the initial distribution on secondary structures and where the $L_1$-norm for discrete distributions $y$ and $y'$ is $\|y - y'\| = \sum_i |y_i - y'_i|$. The function $A(Q,t)$ takes real values in the interval $[0,2]$ since $xe^{Kt}$ and $xe^{Qt}$ are distributions on which the $L_1$-norm is taken. Figure 2 illustrates the inaccuracy as well; for a particular secondary structure, the contribution to the inaccuracy is the absolute value of the difference between one of the subset models and the full model population kinetics.

In a strict sense the above equation is ill-defined, since $K$ and $Q$ might have support over a different state space, as in the case of $K$ and $L$ or $K$ and $\Gamma$. There are two choices for how to compute $A(Q,t)$.

**(Support $U$)** Sum over all the secondary structures in $U$ and define $P_Q(t)_{ij} = 0$ for all $i$ or $j \in U \setminus S$. If this choice is made, the global dynamics dominate the behavior of the inaccuracy function. Specifically, the rate at which probability leaves set $S$ largely determines the inaccuracy function

**(Support $S$)** Sum the $L_1$-norm over only the common support of the two distributions, i.e. the secondary structures $S$. In this case, we choose to project the process $K$ onto states $S$ such that we obtain a (re-normalized) distribution over $S$. Under this comparison, it seems that local behaviors and the value of the stationary distribution dominate the behavior of the inaccuracy.

To compare the inaccuracy of model $L$ with model $\Gamma$, we will compute and compare $A(L,t)$ and $A(\Gamma,t)$ in Equation 3. We favor support $U$ over support $S$ due to the misleading aspect that the renormalization makes the stationary distributions look surprisingly different. This is contrary to empirical observations that the stationary distributions do not differ significantly. In the body of the paper we will present the inaccuracy for support $U$ where we compare $A(L,t)$ and $A(\Gamma,t)$, while the Appendix will contain results with both support $U$ and $S$.

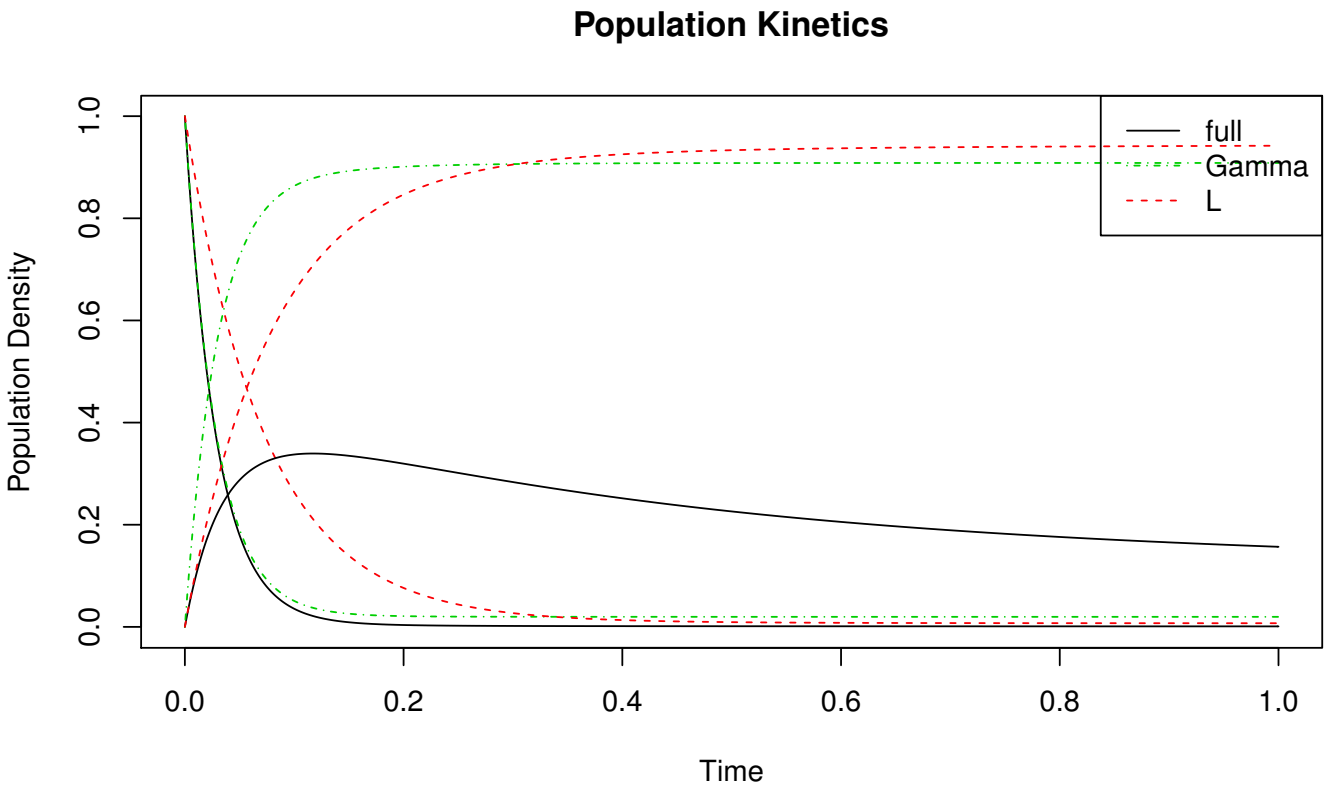



Figure 2: **Population kinetics of the full model compared with the subset models.** This figure shows the significant components of the population kinetic vector that appear in all the models, i.e. the larger components of $xe^{Kt}$ where $x$ was the point-mass at the open chain. There is one type of line for each model, solid black for the full model, green dot-dashed line for the $\Gamma$ model, and red dashed line for the $L$ model. Each line of one type is for a different secondary structure. The sum of the absolute difference between pairs of lines approximates the inaccuracy.

# 4 Results and Discussion

We consider both biological RNA sequences and synthetic data. Biological sequences are useful for generating realistic CTMC models. On the other hand, the synthetic data lets us investigate a *family* of RNA-like models. This allows us to investigate the average inaccuracy of the two subset models over multiple models. Since the biological sequences have various total numbers of secondary structures, this averaging would be difficult to do with real sequences. In both settings, we use the inaccuracy method above to test whether the $L$ or $\Gamma$ model is more accurate at many time points.

**Simulation Time.** Time, in this context, is a dimensionless quantity, meaning that it can be scaled. See the discussion about dimensions for the Boltzmann distribution in Section 2. The Boltzmann distribution and the transition probabilities for CTMCs, $K$, $L$, and $\Gamma$, are dimensionless. The best approach would be to compare the simulation to data and pick the scaling of time that best matches the data. In the absence of data, we pick a scale convenient for plotting and use the same scale for all the models. This is fine since we are examining relative inaccuracy and are not concerned with absolute inaccuracy.

When we discuss comparative timescales using the words 'short' and 'long', we mean short relative to the magnitudes of the eigenvalues. A look at ODEs and the interpretation of eigenvalues and eigenvectors provides more illumination. When sorted by increasing magnitude, the first eigenvalue is zero. Since the eigenvalues are non-positive, the population kinetics stabilize over time. The zero eigenvalue contributes to the stationary distribution. The other eigenvalues decay to zero and contribute to faster effects. The larger the magnitude of the eigenvalue the faster the effect.

## 4.1 Biological Sequences

Using biological sequences and thermodynamic parameters, we can compute energies that inform the kinetic rates of the full model. This gives us a test-bed for comparing the accuracy of the two subset models. We consider both exact inaccuracy calculations and approximate ones.

**Exact Accuracy Results.** Given a biological sequence, we can enumerate the secondary structures, $U$, and compute the energy of each structure. Next, we enumerate the one-base-pair neighbors of each secondary structure, $i$, which is the set $N(i)$. From this, we compute the rate matrix for $K$ and use the matrix exponential to compute the population kinetics. The rate matrix for $K$ is now fixed, and we average the inaccuracy over multiple choices of subset $S$ for each model $L$ and $\Gamma$.

The choice of $S$ can be randomized such that $|S| = (1/10)|U|$. This is done by initializing the set $S^1 = \{s_{mfe}\}$ where $s_{mfe}$ is the minimum free energy state. $S^{i+1}$ is obtained from $S^i$ in a randomized iterative process where a state in $s_i \in S^i$ is chosen uniformly at random, a neighbor $s_{i+1} \in N(s_i)$ is drawn uniformly at random, and the neighbor is added to the set to obtain $S^{i+1} = S^i \cap \{s_{i+1}\}$. Recall that a set contains only distinct elements. So, this procedure continues until we have $|S^i| = |S|$, and we let $S := S^i$. From this, we compute the rate matrix for each of the subset models: $L$, and $\Gamma$. Again the matrix exponential, together with the start distribution being the point-mass at the minimum free energy structure, gives us the population kinetics, and we compute the inaccuracy.

We modeled sequences varying in length from 12 - 15 nucleotides. These are the biological RNA sequences (PDB IDs: 1AFX, 1HJI:A, 1C0O, 1XV6, 1VOP) shown in Table 1. We will show results only for the first three in Figure 3, the last one in the next section, and the remaining RNA results are in the Appendix. For all these RNA, we see a modest improvement in the inaccuracy of $\Gamma$ relative to $L$ for short timescales. At long timescales we

| Sequence | Length | $\lvert U\rvert$ | $\lvert S\rvert$ |
|---|---|---|---|
| 1AFX | 12 | 70 | 7 |
| 1HJI:A | 15 | 195 | 20 |
| 1C0O | 14 | 410 | 41 |
| 1XV6 | 12 | 48 | 5 |
| 1VOP | 13 | 110 | 11 |
| RNA1 | 18 | 876 | 16 and 21 |

Table 1: Biological RNA Sequences

did not see any differences large enough to be detected (i.e. larger than some numerical error threshold $\epsilon = 0.0000001$). By this we mean that taking the known stationary distributions and using them to compute the stationary inaccuracy, the two models' inaccuracies were indistinguishable.

**Approximate Accuracy Results.** As described in Section 2.2, the alternative to exact computation of the population kinetics is the Monte Carlo simulation. This approach generates trajectories of single RNA molecules as they begin in the open chain and move through the available secondary structures. Therefore, this simulation needed to enumerate the one-base pair neighbors of each secondary structure in the set $S$ (for the models $L$ and $\Gamma$) or the set $U$ (for the model $K$). The population kinetics will then be calculated from the sampled trajectories using the Doob-Gillespie algorithm [37].

In order to be more robust in approximating the population kinetics, we simulate 1000 folding pathway trajectories for each model on a specific set. We next compute the population kinetics of a particular secondary structure at time $t$ by averaging its appearance over all trajectories at that time.

Figure 4 shows the approximate inaccuracy results of the models $L$ and $\Gamma$ compared to $K$ for the biological sequence RNA1 on two different subsets with size 16 and 21, respectively (the size of $U$ is 876). This molecule is an 18-nucleotides RNA (CGCGCUACUCCUAGAGCU) that contains a hairpin conformation in the "native" state [39]. The choice of $S$ for this sequence is done in the same way explained in the "exact accuracy results" section except that the size of $S$ was chosen to match results in the literature [39]. The Monte Carlo simulation results, on RNA1, still show a modest improvement for the model $\Gamma$ over $L$ on short timescales. They also show that the relative performance of the models $\Gamma$ and $L$ is somehow dependent on the choice of $S$ and can vary with different choices.

## 4.2 Synthetic Data

The simulation in this section relies on an energy distribution, because using such a distribution generates a *family* of RNA-like CTMCs that can be studied using statistics. This allows generalizations that are not possible from looking at a single sequence. From our simulation results of 1000 replicates of 100 secondary structures in Figure 6, we conclude that the $\Gamma$

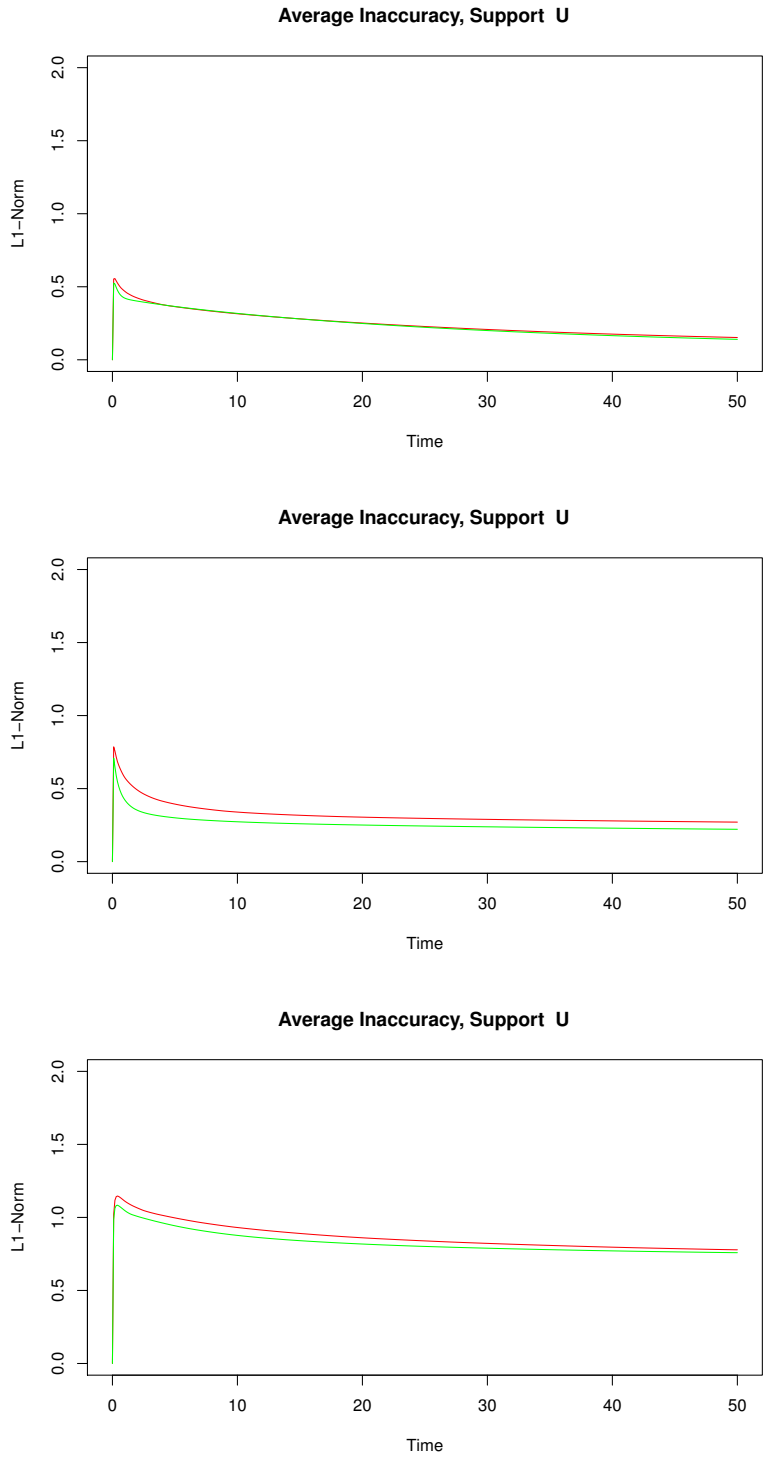
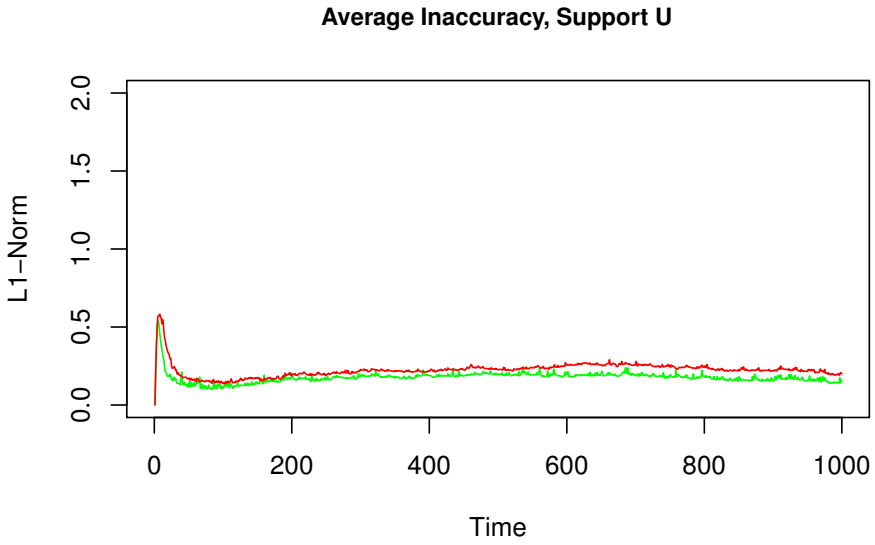



Figure 3: **Average inaccuracy for 100 replicates** with each drawing a set $S$. These are the results for RNA sequences 1AFX, 1HJI:A, and 1C0O top to bottom. The red line is for the $L$ model and the green line for $\Gamma$. The inaccuracies are averaged for 100 choices of subset $S$.



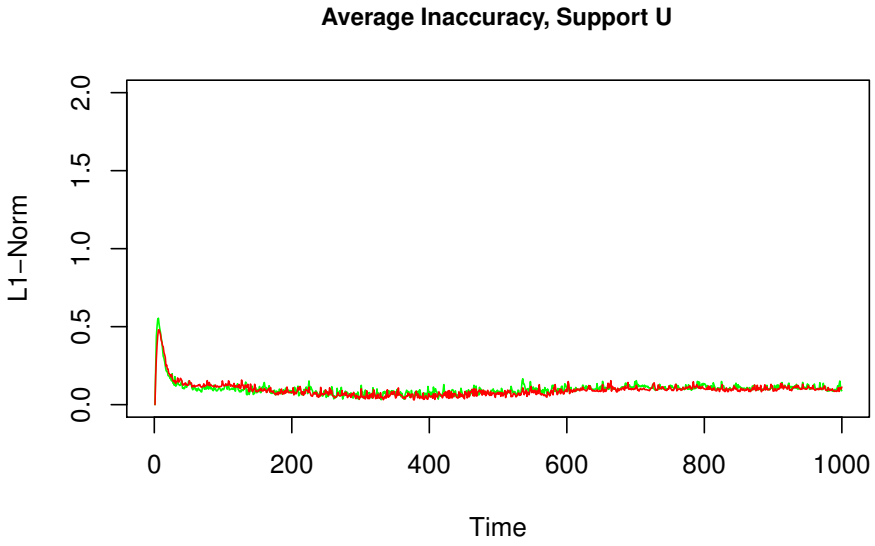




Figure 4: **Average inaccuracy for 1000 replicates** on two different subsets of the full space of RNA1 with sizes 16 and 21 respectively. These are the results for the biological RNA sequence RNA1 with 18 nucleotides. The red line is for the $L$ model and the green line for $\Gamma$. The inaccuracies are averaged over 1000 sampled RNA folding trajectories on each subset for all three models. The start structure was the open chain. Lower curves are better.

model performs at least as well as the $L$ model. This is a broad setting, so it confirms that $\Gamma$ is reasonable.

The simulation was performed with the statistical software `R`. There were three parameters for the simulation: 1) $|U|$: the number of secondary structures to simulate in the full model, 2) $|S|$ where $|S| \leq |U|$: the number of secondary structures in the subset models, and 3) $f$: the general fraction of the edges to select for the neighborhood. For each structure in the full model, a free energy was simulated from a bimodal Gaussian. The parameters of this distribution were selected relative to the value $\rho\tau$ such that the population kinetics of most simulations changed over an observable time frame. We have very little knowledge of what the actual distribution of free energies is where the distribution is taken over the free energies of all the possible secondary structures of an RNA. Due to this lack of knowledge, there currently is considerable freedom in selecting the free energy distribution. Once the free energies are chosen, the remaining choice is the neighborhood, $N(i)$, of each secondary structure $i$. Most of the neighborhood is chosen randomly with a fraction of $f = 0.2$ of the possible edges being selected for a neighborhood. Edges joining the modes of the bimodal Gaussian, where half of the free energies have been generated under one Gaussian and the other half under another Gaussian, are less dense with a fraction of $f/4$ edges being selected. This process is illustrated in Figure 5. Since we are interested in connected sets $S$, we check for connectedness and reject sets $S$ that are not connected.

The above simulation allows us to obtain a matrix $K$ with values given by the Kawasaki rule for the simulated free energies and neighborhoods. In order to observe the population kinetics for this model, the matrix exponential was computed using `R`'s function `eigen()`. This function provides a spectral decomposition for the matrix $K$ by finding its eigenvectors and eigenvalues. Using these, we can directly compute Equation 1. The output of the full simulation is a plot showing the components of the vector $xP^K(t)$ obtained by multiplying the initial population density vector $x$ by the transition probability matrix $P^K(t)$. As an example, the solid lines in Figure 2 shows the population kinetics of a simulation with $|U| = 100$ secondary structures. Rather than plotting all $|U|$ components of the vector, we plot only components for secondary structures with relatively high probability under the Boltzmann distribution that appear also in subset $S$.

The free energies were chosen according to a bimodal Gaussian and the choice of neighborhood was selected such that the two modes of the Gaussian were less densely connected to each other than within each mode. This effectively creates two basins in the energy landscape, one of which will have the minimum free energy (MFE) node as the local minimal free energy structure. The other basin will have a low energy node as the local minima. RNA molecules will be able to circulate more freely within a basin than between basins due to the choice of neighborhoods.

To obtain a simulation of the subset models, we selected $|S| = |U|/10$ connected secondary structures out of the $|U|$ from the full model. We did this in the way described in Section 4.1. Again, the start distribution was the point-mass on the minimum free energy state. We then created the models $L$ and $\Gamma$ as detailed above. Once again we can obtain the eigenvectors and eigenvalues for $L$ and $\Gamma$, allowing us to compute $A(L, t)$ and $A(\Gamma, t)$.

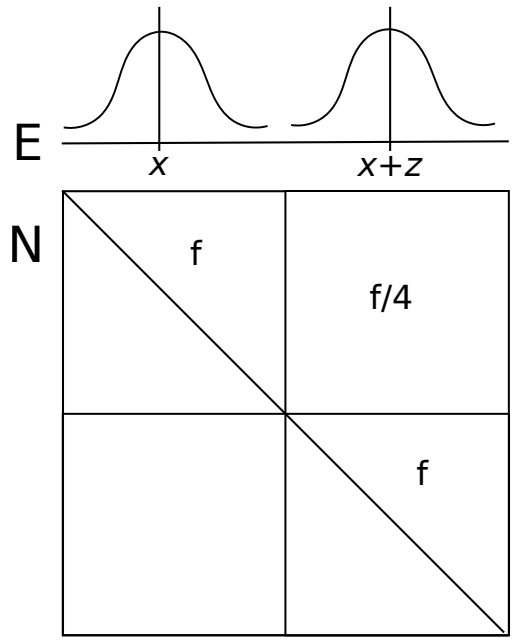



Figure 5: **Cartoon of $E$ and $N$.** The structures are split into two groups, $i = 1, ..., |U|/2$ and $|U|/2 + 1, ..., |U|$. For the first group the free energies are simulated from a Gaussian with mean $x$, and the second group from a Gaussian with mean $x+z$ where $z$ is small relative to the variance. Let the neighborhood matrix $N$ be a symmetric matrix having $N_{i,j} = N_{j,i}$ when $j \in N(i)$ and $i \in N(j)$. The neighborhoods $N$ are simulated differently depending on which quadrant of the matrix the entry $i, j$ where $i < j$ falls in. If $i, j$ is in a quadrant labeled $f$, then a fraction $f$ of the entries are set to 1. For the quadrant marked $f/4$ that fraction of entries are chosen. Since the matrix $N$ is symmetric, we only choose values for the upper-triangle.

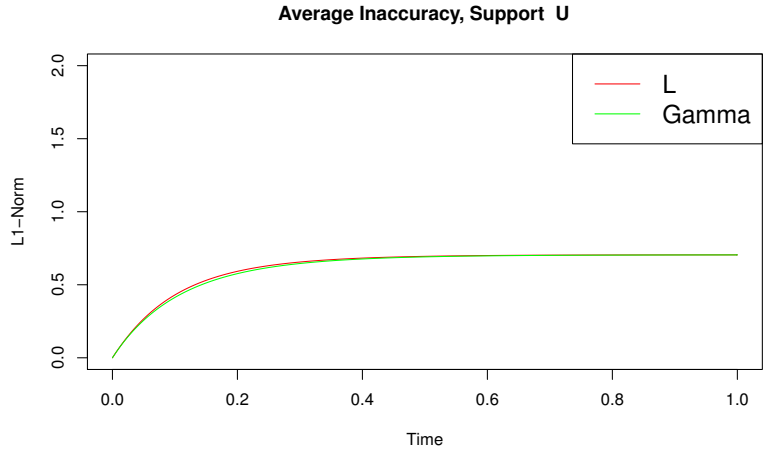



Figure 6: **Accuracy with support** $U$ on a subset model with 10 secondary structures with respect to the full model with 100 secondary structures. Here, the set $S$ contains the MFE structure. The red line, and upper line, is for $A(L,t)$ and the green line for $A(\Gamma,t)$ There is little difference between $\Gamma$ and $L$ except for the region around $t = 0.175$ when there is slightly worse accuracy for $L$. Here the start state was the minimum free energy structure.

Figure 6 shows the results of the inaccuracy computed for support $U$, as defined in Section 3.1. Results for the inaccuracy with support $S$ appear in the supplement. For all inaccuracy functions, values are in the interval $[0, 2]$ and are obtained by taking the $L_1$-norm between two distributions. A lower $L_1$ value is better.

## 4.3 Running Times

The running-time of the two subset models is the same (for both the exact calculation of the matrix exponential and with the Monte Carlo simulation). The difference in running-time between the full model and the subset models depends on the size of set $S$.

We do exact calculations on the full model for sequences of around 15 nucleotides. We chose this size of simulation, because we wanted to compute the exact matrix exponential for a number of replicates. The matrix exponential algorithm is limited to about $500 \times 500$ matrices which is roughly a 14-15 nucleotide RNA sequence. For example, computing 1000 simulation replicates of synthetic data took several hours. Now, consider the size of a subset model on which the exact calculation can be performed. If the matrix is limited to 500 structures and $|S| = 0.1 \times |U|$, then the full set of structures $|U| = 5000$ which indicates a sequence of length roughly 20 nucleotides.

Notice that simulating from the Doob-Gillespie algorithm under $\Gamma$ and $L$ is easier than simulation under $K$. This is because when sampling trajectories under $K$ the embedded Markov chain is used to choose randomly between the $N(i)$ neighbors of $i$. This takes time proportional to $N(i)$ and to the time required for the free energy calculation. However, in $\Gamma$ and $L$ there are fewer neighbors, because the process is selecting randomly from $N(i) \cap S$.

Thus the running-time of the Doob-Gillespie algorithm under $\Gamma$ and $L$ is at least as fast as under $K$ and depends on $|S|$.

Monte Carlo simulations using the Doob-Gillespie algorithm for the full model can go to about 40-50 nucleotides and for the subset models to 120-140 ones, depending on the size of $S$ and the number of trajectories sampled. In other words, the running time of the Monte Carlo simulation on a subset model at time $t$ would be $O(|S|\alpha)$, where $\alpha$ is the number of sampled trajectories. This holds since there will be at most $|S|$ neighbors for each secondary structure for a subset model on the set $S$. There is also a preparation time, which is the time to generate the transition probabilities on the set $S$ for each subset model, which costs at most $O(|S|^2)$.

# 5 Conclusions

This paper introduced a new RNA folding model, $\Gamma$, on a subset, $S$, of the space of secondary structures $U$. This model is as compared to the 'full' RNA folding model, $K$, over the entire space $U$ of secondary structures [8]. We introduced two rigorous methods for comparing the $\Gamma$ and $K$. As a second comparison point, we consider the pre-existing model, $L$, used by Tang et al. [35], and also compare it to $K$.

The model that we introduce is similar to Tang et al's model in that both models do not have known accuracy bounds. Furthermore our model and Tang et al.'s are guaranteed to produce valid trajectories. For these reasons the discussion focused on a comparison of our method to Tang et al.'s method. To address the lack of accuracy bounds, this paper also introduced an empirical method of assessing the accuracy of subset models which assumes that the Flamm et al model [8] is the gold standard. This is important, because it is difficult to get data for detailed population kinetics.

Our synthetic data simulations, introduced in this work, produce RNA-like folding landscapes by simulating from a bimodal Gaussian and creating a neighborhood graph that has two basins in the energy landscape. While these simulations approximate real RNA folding processes, they provide us a much closer look at the process than we can get using real RNA sequences or experimental data. This is because simulation allows us to compute statistics for a variety of models $K$ with similar features. The simulation results suggest that the $\Gamma$ model is at least as good as the $L$ model on average.

Not surprisingly, on small RNA sequences, the analysis of several biological sequences shows that the $\Gamma$ model has a modest improvement over the $L$ model on short timescales. We attribute this to the $\Gamma$ model having holding times that match the full model whereas the $L$ model is has slower holding times. At longer timescales, we found that the inaccuracy of the stationary distributions of the two models was indistinguishable. Because we need to also calculate population dynamics of the full model in order to do a rigorous comparison between the two subset models, it is not possible at this time to compare the two subset models on long RNA's. In future work we will investigate ways to assess which of two methods has better accuracy even when it is infeasible to calculate population dynamics of the full model.

This paper focused on ways to rigorously compare subset models to the full RNA folding

model. Choosing the connected set $S$ to optimize the accuracy of a given subset model is also difficult and an open problem. An initial exploration of the choice of $S$ is discussed in the Appendix. Tang et al. [35] provided some heuristic choices, but no rigorous justification in terms of accuracy. Ideally, we would choose the set $S$ to optimize accuracy and then choose the subset model to optimize accuracy in the time periods that we are most interesting in studying. While this paper provides a method for the latter, along with an accuracy measure, there is no rigorous method for choosing a set $S$. This is an interesting direction for future work that can leverage the contributions of this paper.

Another challenging problem is to determine highly reliable kinetic rates to use in the simulation of the full and subset models. The rates used in our work and earlier work on folding pathway simulation are computed from models of the free energy of RNA molecules. However, in the ideal world, the rates would be obtained from experiments performed at thermodynamic equilibrium. he challenge is to measure or derive a rate for every pair of structures that differ by one base pair. Some useful data on kinetic rates is available, such as that obtained by Zhang et al [18] for multi-stranded DNA complexes, RNA rates that have been measured under force [19, 20, 40, 41], rates measured for aggregate states [42], or observations of population kinetics without rate measurements [43, 44]. However, there is still paucity of experimental rate data relative to what is required by the simulators. In any case, our simulator and subset model can easily be adapted for use with improved rate parameters.

# Acknowledgments

We thank the anonymous reviewers and Ian Mitchell for helpful comments.

# References

[1] Andrei Korostelev and Harry F. Noller. The ribosome in focus: new structures bring new insights. *Trends Biochem. Sci.*, 32:434–441, 2007.

[2] Kenn Gerdes and E Gerhart H Wagner. RNA antitoxins. *Current Opinion in Microbiology*, 10(2):117–124, April 2007.

[3] David P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116:281–297, 2004.

[4] Dmitri D. Pervouchine, Ekaterina E. Khrameeva, Marina Yu. Pichugina, Oleksii V. Nikolaienko, Mikhail S. Gelfand, Petr M. Rubtsov, and Andrei A. Mironov. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA*, 2011.

[5] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier P. Pandolfi. A ceRNA hypothesis: The Rosetta stone of a hidden RNA language? *Cell*, 146(3):353–358, August 2011.

[6] Saba Valadkhan. The spliceosome: caught in a web of shifting interactions. *Curr. Opin. Struct. Biol.*, 17:310–315, 2007.

[7] J. Kenneth Wickiser, Wade C. Winkler, Ronald R. Breaker, and Donald M. Crothers. The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol. Cell*, 18:49–60, 2005.

[8] Christoph Flamm, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[9] D Thirumalai, Namkyung Lee, Sarah A Woodson, and DK Klimov. Early events in RNA folding. *Annual Review of Physical Chemistry*, 52(1):751–762, 2001.

[10] J.R. Norris. *Markov Chains*. Cambridge Series in Statistical and Proba- bilistic Mathematics. Cambridge U. Press, 1997.

[11] Kevin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–1975, 2009.

[12] Jan Pieter Abrashams, Mirjam Van Den Berg, Eke Van Batenburg, and Cornelis Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Research*, 18(10):3035–3035, May 1990.

[13] Hugo M. Martinez. An RNA folding rule. *Nucleic Acids Research*, 12(1 Pt 1):323–334, January 1984.

[14] Manfred Tacker, Walter Fontana, Peter F. Stadler, and Peter Schuster. Statistics of RNA melting kinetics. *European Biophysics Journal*, 23(1):29–38, April 1994.

[15] Gutell R.R., Lee J.C., and Cannone J.J. The accuracy of ribosomal rna comparative structure models. *Current Opinion in Structural Biology*, 12(3):301–310, 2002.

[16] H. Isambert. The jerky and knotty dynamics of RNA. *Methods*, 49(2):189–96, 2009.

[17] Shi-Jie Chen. Rna folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annual Review of Biophysics*, 37(1):197–214, 2008. PMID: 18573079.

[18] David Yu Zhang and Erik Winfree. Control of DNA strand displacement kinetics using toehold exchange. *J. Am. Chem. Soc.*, 131(47):17303–17314, 2009.

[19] I Tinoco, D Collin, and P T X Li. The effect of force on thermodynamics and kinetics: unfolding single RNA molecules. *Biochem Soc Trans*, 32(Pt 5):757–60, 2004.

[20] Paul Maragakis, Felix Ritort, Carlos Bustamante, Martin Karplus, and Gavin E. Crooks. Bayesian estimates of free energies from nonequilibrium work data in the presence of instrument noise. *J Chem Phys*, 129(2), 2008.

[21] Joseph Schaeffer. The Multistrand Simulator: Stochastic simulation of the kinetics of multiple interacting DNA strands. *California Institute of Technology*, page 123, 2012.

[22] DAVID H. MATHEWS. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):11781190, 2004.

[23] Suvir Venkataraman, Robert M. Dirks, Christine T. Ueda, and Niles A. Pierce. Selective cell death mediated by small conditional RNAs. *Proceedings of the National Academy of Sciences*, 107(39):16777–16782, 2010.

[24] Harry M T Choi, Joann Y Chang, Le A Trinh, Jennifer E Padilla, Scott E Fraser, and Niles A Pierce. Programmable in situ amplification for multiplexed imaging of mRNA expression. *Nat Biotech*, 28:1208– 1212, 2010.

[25] Ivan Dotu, William A. Lorenz, Pascal Van Hentenryck, and Peter Clote. Computing folding pathways between rna secondary structures. *Nucleic Acids Research*, 38(5):1711–1722, 2010.

[26] Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift fr Physikalische Chemie*, 216(2):155–173, 2002.

[27] Christoph Flamm and Ivo L Hofacker. Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatshefte fr Chemie Chemical Monthly*, 139(4):447–457, 2008.

[28] Michael Geis, Christoph Flamm, Michael T. Wolfinger, Andrea Tanzer, Ivo L. Hofacker, Martin Middendorf, Christian Mandl, Peter F. Stadler, and Caroline Thurner. Folding kinetics of large rnas. *Journal of Molecular Biology*, 379(1):160 – 173, 2008.

[29] Ivo L. Hofacker, C. Flamm, Christian Heine, M.T. Wolfinger, Gerik Scheuermann, and Peter F Stadler. Barmap: Rna folding on dynamic energy landscapes. *RNA*, 16(7):1308–1316, 2010.

[30] Yuan Li and Shaojie Zhang. Predicting folding pathways between rna conformational structures guided by rna stacks. *BMC Bioinformatics*, 13(Suppl 3):S5, 2012.

[31] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys*, 124(4):044104–1 – 044104–13, 2006.

[32] Xinyu Tang, Bonnie Kirkpatrick, Shawna Thomas, Guang Song, and Nancy M. Amato. Using motion planning to study RNA folding kinetics. *Journal of Computational Biology*, 12(6):862–881, 2005.

[33] Peinan Zhao, Wen-Bing Zhang, and Shi-Jie Chen. Predicting secondary structural folding kinetics for nucleic acids. *Biophys J*, 98(8):1617–25, 2010.

[34] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, May 1990.

[35] Xinyu Tang, Shawna Thomas, Lydia Tapia, and Nancy M. Amato. Tools for simulating and analyzing RNA folding kinetics. In *Proceedings of the 11th annual international conference on Research in computational molecular biology*, RECOMB'07, pages 268–282, Berlin, Heidelberg, 2007. Springer-Verlag.

[36] Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*. Halsted Press, New York, NY, 1992.

[37] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[38] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Process*. Oxford University Press, New York, 2nd edition, 1992.

[39] Xinyu Tang, Shawna Thomas, Lydia Tapia, David P Giedroc, and Nancy M Amato. Simulating RNA folding kinetics on approximated energy landscapes. *Journal of molecular biology*, 381:1055–1067, September 2008.

[40] Hans A. Heusa, Elias M. Puchnerb, Aafke J. van Vugt-Jonkera, Julia L. Zimmermannb, and Hermann E. Gaubb. Atomic force microscope-based single-molecule force spectroscopy of RNA unfolding. *Analytical Biochemistry*, 414:1–6, 2011.

[41] R. Bundschuh and U. Gerland. Dynamics of intramolecular recognition: Base-pairing in DNA/RNA near and far from equilibrium. *European Physical Journal E — Soft Matter*, 19(3):319–329, 2006.

[42] X. Huang, Y. Yao, G. R. Bowman, J. Sun, L. Guibas, G. Carlsson, and V. S. Pande. Constructing multi-resolution markov state models (msms) to elucidate rna hairpin folding mechanisms. *Pacific Symposium on Biocomputing*, 15:228–239, 2010.

[43] P.G Higgs. Overlaps between rna secondary structures. *Phys. Rev. Lett.*, 76:704–707, 1996.

[44] Nathan J. Baird, Steven J. Ludtke, Htet Khant, Wah Chiu, Tao Pan, and Tobin R. Sosnick. Discrete structure of an RNA folding intermediate revealed by cryo-electron microscopy. *Journal of the American Chemical Society*, 132(46):16352–16353, 2010.

# 6 Appendix: Additional Results

**Biological Sequences.** The exact inaccuracy calculation was run on models formed from sequences of lengths between 12 -14 nucleotides. These are the biological RNA sequences (PDB IDs: 1AFX, 1HJI:A, 1C0O, 1XV6, 1VOP) shown in Table 1. The results are the inaccuracy values averaged over various choices of subset $S$. The first three sequences have their results shown in the body of the paper. Here, we give the results for the last two in Figure 7 for accuracy support $U$ and in Figure 8 the accuracy support $S$. For all these RNA, we see a modest improvement in the accuracy of $\Gamma$ relative to $L$ for short timescales. On the longer timescales, we saw that the inaccuracies of the stationary distributions of $\Gamma$ and $L$ were indistinguishable. This means that all the inaccuracy curves converge to the same inaccuracy as $t \to \infty$.

Consider the choice of $S$ that would minimize the inaccuracy of a particular subset model. This is a difficult and open problem. One approach would be to compute the area under the inaccuracy curve and to choose the subset $S$ that minimizes this area. As an illustration using sequence 1AFX, in Figure 9 each panel represents the inaccuracy under a different choice of subset $S$ with there being two such choices. The challenge is to pick the better $S$ for a particular subset model, $L$ or $\Gamma$. The choice of $S$ in the top panel clearly favors model $L$, since the area under the inaccuracy curve is much less for $L$ in the top panel than the bottom panel. On the other hand, it is a bit less obvious as to whether the $\Gamma$ model has a smaller area under the curve in the top or the bottom panel. One difficult question is: in integrating to obtain the area under the curve, at what time point should the integration stop? So it is not clear what rule to use in choosing set $S$ or how to apply a rule efficiently.

# 7 Appendix: Subset Models for the Metropolis-Hastings and Kawasaki Rules

For a restriction of a full model to the states in set S, there seems to be a trade-off between having a stationary distribution proportional to the full model and having a correct local distribution, such as the holding time.

For example the $L$ process defined above satisfies detailed balance and converges to the Boltzmann distribution on the states of $S$. Notice that this Boltzmann distribution, while being proportional to the stationary distribution for the full model $K$ on states $U$, is not the same as the Boltzmann distribution on $K$. Since the process $L$ is not converging to the correct distribution, perhaps giving up proportionality of the stationary distribution will not make the stationary distribution much more incorrect.

Furthermore, the process $L$ on states $S$, as defined, does not generate trajectories that are correct under the full model $K$. There is a process on the states $S$ which generates trajectories with the same probability as they would be generated under the full process $Q$ on states $U$. Obtaining this process requires giving up proportionality of the stationary distribution. Despite having given this up, we see that this process converges to a distribution that is a function of the Boltzmann distribution. So, by accepting only slightly less accuracy on the

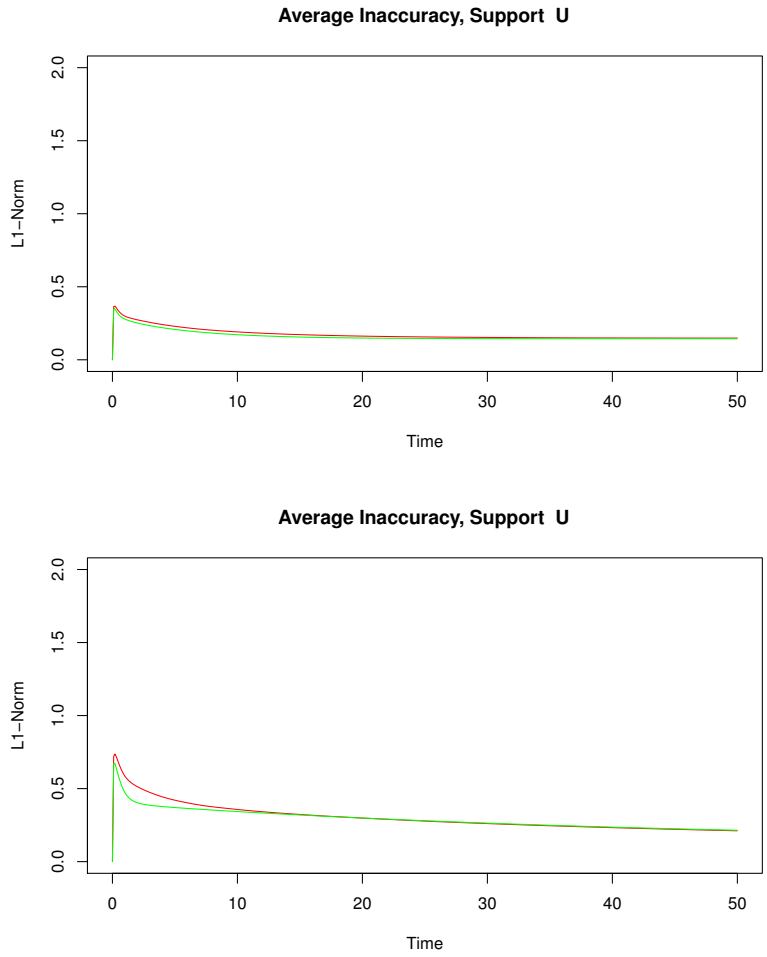



Figure 7: **Average inaccuracy for support $U$ for 100 replicates** with each drawing a set $S$. These are the results for RNA sequences 1XV6 and 1VOP, top to bottom. The red line is for the $L$ model and the green line for $\Gamma$.

stationary distribution of $\Gamma$, which we cannot hope to get correct, we can recover the correct probabilities for the trajectories, so that every trajectory occurs with the same probability in $\Gamma$ as it does in process $K$. We can make the same statement for the Metropolis-Hastings model where there is a process $\Gamma$ on states $S$ that generates trajectories with exactly the same probability as under the full process $K$ on states $U$.

## 7.1 Subset Models

Let $Q$ be an arbitrary reversible CTMC on state-space $U$. Let $\hat{Q}$ and $\tilde{Q}$ be the subset and trajectory subset models for $Q$, defined next. Reversibility of $Q$ is used to prove a result about the stationary distributions of $\hat{Q}$ and $\tilde{Q}$. Let the stationary distributions of these models be $\pi$, $\hat{\pi}$, and $\tilde{\pi}$, respectively for $Q$, $\hat{Q}$, and $\tilde{Q}$.

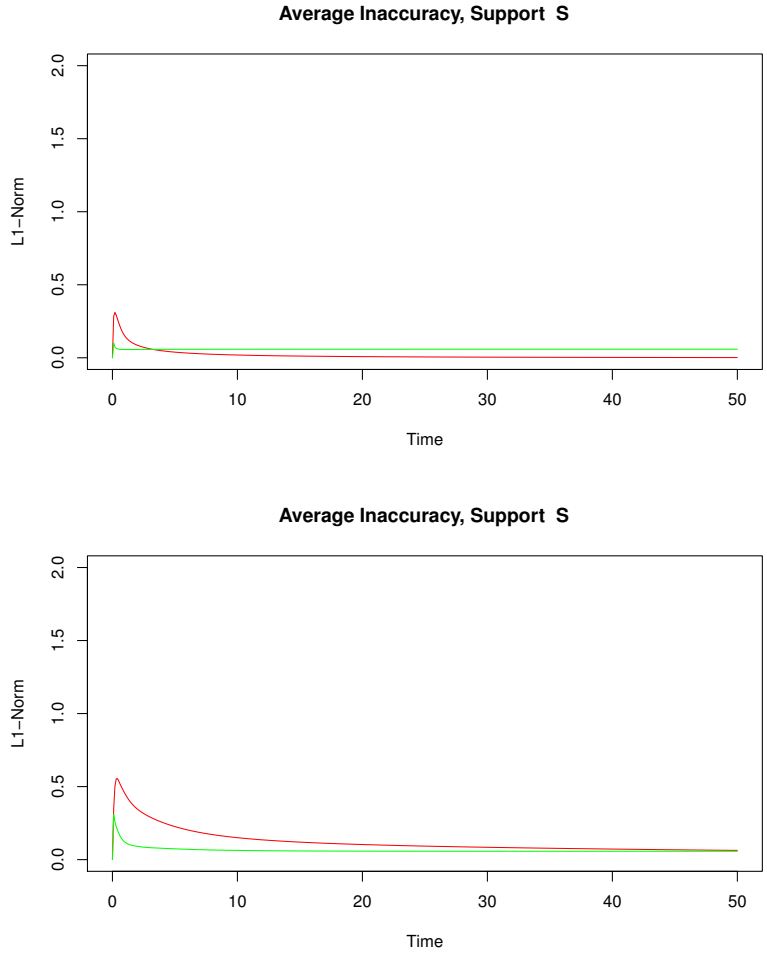



Figure 8: **Average inaccuracy for support $S$ for 100 replicates** with each drawing a set $S$. These are the results for RNA sequences 1XV6 and 1VOP, top to bottom. The red line is for the $L$ model and the green line for $\Gamma$.

For $i$, define its neighbors $N(i)$ as the set of states $j$ for which $Q_{ij} > 0$.
The subset model is defined as

$$\hat{Q}_{ij} = \begin{cases} Q_{ij} & \text{if } i \in S \text{ and } j \in N(i) \cap S \\ 0 & \text{if } i \notin S \text{ or } j \notin N(i) \cap S. \end{cases}$$

and

$$\hat{Q}_{ii} = -\sum_{j \neq i} \hat{Q}_{ij}.$$

The trajectory subset model $\tilde{Q}$ which has holding times equal to the holding times of $Q$

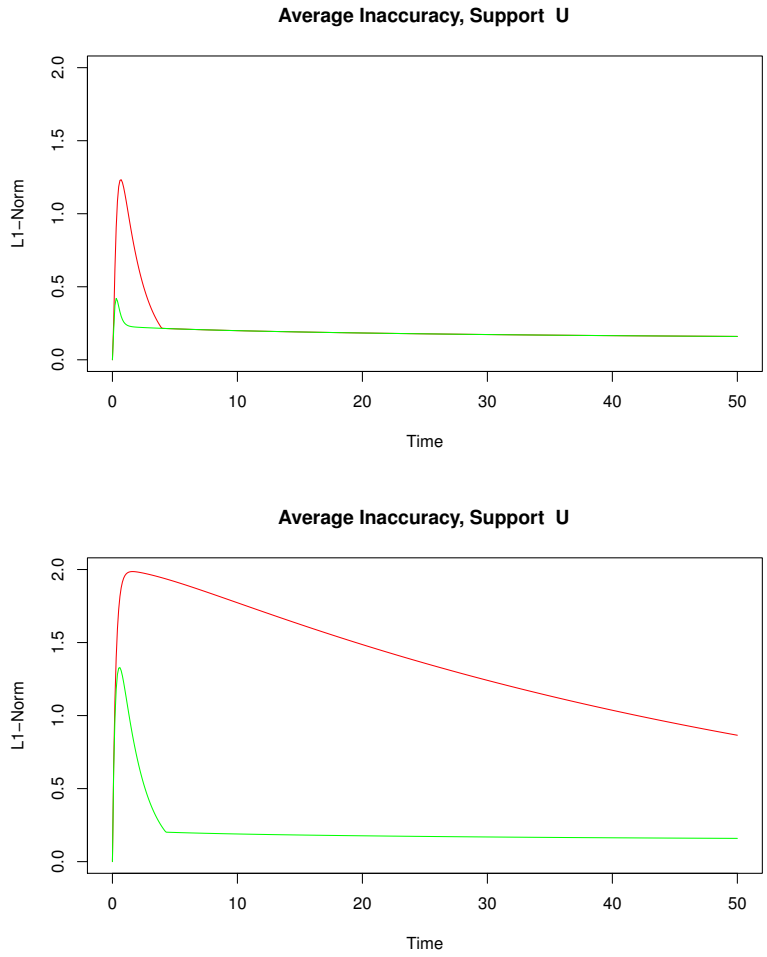



Figure 9: **Inaccuracy for several choices of $S$.** These are the results for RNA sequence 1AFX under two choices of subset $S$, one in the top panel and the other in the bottom panel. The red line is for the $L$ model and the green line for $\Gamma$. For a given model, we want to choose the set $S$ and the respective panel that has the smallest area under the inaccuracy curve. As before, in a single panel, we know that the inaccuracies of the stationary distributions of these two models are nearly identical. For model $L$ the top panel is the clear winner. However, for $\Gamma$ it is a little less clear whether the better choice is the top or bottom panel. This is largely because $\Gamma$ in the top panel seems to converge to the stationary inaccuracy value slower than in the bottom panel, and this difference could make up for the difference in the peak height at the short timescales.

is

$$\tilde{Q} = \begin{cases} \frac{\sum_{k \in U_{-i}} Q_{ik}}{\sum_{k \in S_{-i}} Q_{ik}} Q_{ij} & \text{if } i \in S \text{ and } j \in N(i) \cap S \\ 0 & \text{if } i \notin S \text{ or } j \notin N(i) \cap S. \end{cases}$$

where the notation $U_{-i} = U \setminus \{i\}$ and where

$$\tilde{Q}_{ii} = -\sum_{j \neq i} \tilde{Q}_{ij}.$$

**The Subset Model.** The stationary distribution $\hat{\pi}$ for process $\hat{Q}$ is the Boltzmann distribution

$$\hat{\pi}_i = \frac{\pi_i}{Z_\ell}$$

where $Z_\ell = \sum_{j \in S} \pi_i$. We leave it as an exercise to verify that if detailed balance holds for $Q$ then it holds for $\hat{Q}$ and that the process $\hat{Q}$ is reversible. The holding times of this process are exponential with parameter

$$|\hat{Q}_{ii}| = \sum_{j \neq i} \hat{Q}_{ij} = \sum_{k \in S_{-i}} Q_{ik}.$$

Let us take a quick look at the embedded Markov chain that corresponds to $\hat{Q}$. The transition probabilities are

$$\frac{\hat{Q}_{ij}}{-\hat{Q}_{ii}} = \frac{Q_{ij}}{\sum_{j \in S_{-i}} Q_{ij}}$$

for $i \neq j$.

**The Trajectory Subset Model.** We now investigate the properties of $\tilde{Q}$, including the properties of the holding time.

**Claim 3.**

$$\tilde{Q}_{ii} = -\sum_{k \in U_{-i}} Q_{ik}.$$

*Proof.* This is verified by a short calculation

$$\begin{aligned} \tilde{Q}_{ii} &= -\sum_{j \neq i} \tilde{Q}_{ij} \\ &= -\frac{\sum_{k \in U_{-i}} Q_{ik}}{\sum_{k \in S_{-i}} Q_{ik}} \sum_{j \in S_{-i}} Q_{ij} \\ &= -\sum_{k \in U_{-i}} Q_{ik}. \end{aligned}$$

□

Thus the holding times of $\tilde{Q}$ are equal to the holding times of the larger process $Q$.

Now we need to consider the stationary distribution, $\tilde{\pi}$ of $\tilde{Q}$. It turns out that $\tilde{\pi}$ is a function of the Boltzmann probabilities.

**Claim 4.** *The stationary distribution for $\tilde{Q}$ is*

$$\tilde{\pi}_i = \frac{\pi_i \sum_{k \in S_{-i}} Q_{ik}}{Z_{\tilde{\pi}} \sum_{k \in U_{-i}} Q_{ik}} \text{ where } Z_{\tilde{\pi}} = \sum_{i \in S_{-i}} \frac{\pi_i \sum_{k \in S_{-i}} Q_{ik}}{\sum_{k \in U_{-i}} Q_{ik}}$$

*Proof.* To prove that $\tilde{\pi}$ is the stationary distribution, we need to prove that $\tilde{\pi}\tilde{Q} = \vec{0}$.

$$\begin{aligned}
(\tilde{\pi}\tilde{Q})_i &= \sum_{j \in S} \tilde{\pi}_j \tilde{Q}_{ji} + \tilde{\pi}_i \tilde{Q}_{ii} \\
Z_{\tilde{\pi}}(\tilde{\pi}\tilde{Q})_i &= \sum_{j \in S} \pi_j \frac{\sum_{k \in S_{-j}} Q_{jk}}{\sum_{k \in U_{-j}} Q_{jk}} \frac{\sum_{k \in U_{-j}} Q_{jk}}{\sum_{k \in S_{-j}} Q_{jk}} Q_{ji} \\
&\quad -\pi_i \frac{\sum_{k \in S_{-i}} Q_{ik}}{\sum_{k \in U_{-i}} Q_{ik}} \sum_{k \in S} \frac{\sum_{k \in U_{-i}} Q_{ik}}{\sum_{k \in S_{-i}} Q_{ik}} Q_{ik} \\
&= \sum_{j \in S} \pi_j Q_{ji} - \pi_i \sum_{k \in S} Q_{ik} \\
&= 0
\end{aligned}$$

The last line is due to $\pi$ being the stationary distribution of $Q$ and to $Q$ being reversible. This completes the proof that $\tilde{Q}$ has stationary distribution $\tilde{\pi}$. □

**Claim 5.** *$\tilde{Q}$ is reversible*

*Proof.* In order to show the reversibility of $\tilde{Q}$, we need to prove that the detailed balance condition is satisfied for every pair of states $i$ and $j$ in process $\tilde{Q}$. First, we can write:

$$\begin{aligned}
\tilde{\pi}_i \tilde{Q}_{ij} &= \frac{\pi_i \sum_{k \in S_{-i}} Q_{ik}}{Z_{\tilde{\pi}} \sum_{k \in U_{-i}} Q_{ik}} \frac{\sum_{k \in U_{-i}} Q_{ik}}{\sum_{k \in S_{-i}} Q_{ik}} Q_{ij} \\
&= \frac{\pi_i Q_{ij}}{Z_{\tilde{\pi}}}.
\end{aligned}$$

Similarly we can have $\tilde{\pi}_j \tilde{Q}_{ji} = \frac{\pi_j Q_{ji}}{Z_{\tilde{\pi}}}$. Therefore, we can conclude $\tilde{\pi}_i \tilde{Q}_{ij} = \tilde{\pi}_j \tilde{Q}_{ji}$, since $Q$ is reversible and $Z_{\tilde{\pi}}$ is also a constant. □

Now, let us consider the embedded Markov chain transition probabilities for $\tilde{Q}$. These are

$$\frac{\tilde{Q}_{ij}}{-\tilde{Q}_{ii}} = \frac{(Q_{ij}) \sum_{k \in U_{-i}} Q_{ik}}{(-\tilde{Q}_{ii}) \sum_{k \in S_{-i}} Q_{ik}} = \frac{(Q_{ij}) \sum_{k \in U_{-i}} Q_{ik}}{\sum_{k \in U_{-i}} Q_{ik} \sum_{k \in S_{-i}} Q_{ik}} = \frac{Q_{ij}}{\sum_{k \in S_{-i}} Q_{ik}}$$

for $i \neq j$. This is proportional to the embedded Markov chain transition probabilities for $Q$.

So the process $\tilde{Q}$ has holding times equal to the holding times of process $Q$ while having transition probabilities that are proportional to those of $Q$. The process $\tilde{Q}$ is able to do this because it does not have a stationary distribution proportional to the Boltzmann distribution, and instead has a stationary distribution that is a function of the Boltzmann distribution.

## 7.2 Comparison of Kawasaki and Metropolis-Hastings

**Kawasaki Model.** For the Kawasaki rule the full model is $K$. The subset model $L = \hat{K}$ and $\Gamma = \tilde{K}$. These three models were defined explicitly in the text.

**Metropolis-Hastings Model.** For the Metropolis-Hastings rule, the full model is $M$. The subset model is $\hat{M}$ and the trajectory subset model is $\tilde{M}$.

We define the full Metropolis-Hastings infinitesimal rate matrix as

$$M_{ij} = \begin{cases} e^{(E(i)-E(j))/(\rho\tau)} & \text{if } E(j) - E(i) \geq 0 \text{ and } j \in N(i) \\ 1 & \text{if } E(j) - E(i) < 0 \text{ and } j \in N(i) \\ 0 & \text{if } j \notin N(i) \end{cases}$$

and $N(i)$ is the neighborhood of $i$ and $\{i\} \cap N(i) = \emptyset$. Flamm's definition of the neighborhood $N(i)$ is all the structures $j$ which differ by one base pair from $i$. Let

$$M_{ii} = -\sum_{j \neq i} M_{ij}.$$

The two subset models are $\hat{M}$ and $\tilde{M}$.

**Tables of Comparison.** For the Metropolis-Hastings rule, $Q = M$. For the Kawasaki rule, $Q = K$.

| Process | $Q = M$ | $Q = K$ |
|---|---|---|
| $Q$ | $\frac{e^{-E(i)/(\rho\tau)}}{Z}$ | $\frac{e^{-E(i)/(\rho\tau)}}{Z}$ |
| $\hat{Q}$ | $\frac{e^{-E(i)/(\rho\tau)}}{\hat{Z}}$ | $\frac{e^{-E(i)/(\rho\tau)}}{\hat{Z}}$ |
| $\tilde{Q}$ | $\frac{e^{-E(i)/(\rho\tau)} \sum_{k \in S_{-i}} M_{ik}}{Z^M_{\tilde{\pi}} \sum_{k \in U_{-i}} M_{ik}}$ | $\frac{e^{-E(i)/(\rho\tau)} \sum_{k \in S_{-i}} e^{-E(k)/(2\rho\tau)}}{Z^K_{\tilde{\pi}} \sum_{k \in U_{-i}} e^{-E(k)/(2\rho\tau)}}$ |

Table 2: **Stationary Distributions**

To define quantities referenced in the table for stationary distributions in Table 2, let

$$\begin{aligned} Z &= \sum_{j\in U} e^{-E(j)/(\rho\tau)}, \\ \hat{Z} &= \sum_{j\in S} e^{-E(j)/(\rho\tau)}, \\ Z_{\tilde{\pi}}^{M} &= \sum_{i\in S} \frac{e^{-E(i)/(\rho\tau)} \sum_{k\in S_{-i}} M_{ik}}{\sum_{k\in U_{-i}} M_{ik}}, \text{ and} \\ Z_{\tilde{\pi}}^{K} &= \sum_{i\in S} \frac{e^{-E(i)/(\rho\tau)} \sum_{k\in S_{-i}} e^{-E(k)/(2\rho\tau)}}{\sum_{k\in U_{-i}} e^{-E(k)/(2\rho\tau)}}. \end{aligned}$$

Table 3 summarizes the derived embedded Markov chain transition probabilities and holding times.

| Process | Holding Time | Transition Probabilities |
|---|---|---|
| $Q$ | $-\sum_{k\in U_{-i}} Q_{ik}$ | $\frac{Q_{ij}}{\sum_{k\in U_{-i}} Q_{ik}}$ |
| $\hat{Q}$ | $-\sum_{k\in S_{-i}} Q_{ik}$ | $\frac{Q_{ij}}{\sum_{k\in S_{-i}} Q_{ik}}$ |
| $\tilde{Q}$ | $-\sum_{k\in U_{-i}} Q_{ik}$ | $\frac{Q_{ij}}{\sum_{k\in S_{-i}} Q_{ik}}$ |

Table 3: **Monte Carlo (Doob-Gillespie Algorithm) Properties**