Title: What is that molecular machine really doing? Automated exploration of alternative transporter mechanisms

Authors: August George[1], Paola Bisignano[2], John Rosenberg[3], Michael Grabe[2], Daniel M Zuckerman[1*]

[*] zuckermd@ohsu.edu

Affiliations: [1]OHSU Center for Spatial Systems, Dept. of Biomedical Engineering, OHSU. [2]Cardiovascular Research Inst., UCSF. [3]Dept. of Biological Sciences, Univ. Pittsburgh.

Keywords: computational biology, chemical physics, molecular machines, kinetic models, biochemical networks, Monte Carlo, cotransporters, symporters, antiporters, computational methods, Markov models

Abstract: Do complex molecular machines tend to exhibit a single mechanism, as suggested in literature and textbook cartoons, or is a multiplicity of pathways possible or even prevalent? Motivated by growing evidence for pathway heterogeneity, we develop theoretical and computational approaches for investigating this question, which could have significant potential impact in both understanding molecular-scale cell biology and in the design of synthetic machines. Focusing on transport in this initial study, we pursue Metropolis Monte Carlo exploration of "model space." Trial moves adjust transition rates among an automatically generated set of conformational and binding states while maintaining fidelity to thermodynamic principles and user-supplied fitness/functionality goal. The simulations yield both single-mechanism and heterogeneous multi-mechanism models for cotransport in a simple environment with a single substrate present. In a "competitive" environment including an additional decoy substrate, several qualitatively distinct models are found which are capable of extremely high discrimination coupled to a leak of the driving ion. The array of functional models would be difficult to find by intuition alone in the complex state-spaces of interest.

Introduction:

The proteins and protein complexes known as molecular machines perform essential functions in the cell, such as transport, locomotion, energy production, and gene expression [1]. Of special interest in this initial study are the transport proteins that move ions and small molecules across a membrane [1] using an electrochemical gradient [1], such as the SGLT family of transporters. SGLT transporters are of biomedical interest due to the vital role that SGLT1 and SGLT2 play in the uptake of glucose in the kidney [2]. In part due to the continued development of SGLT2 inhibitors, there has been renewed interest in the mechanisms, kinetics, and functions of sodium glucose transport proteins [3] [4] [5] [6].

The biological mechanisms of transporters as well as other molecular machines can be modeled using biochemical networks, typically along with mass action kinetics [7]. In a biochemical network, the system is decomposed into discrete states connected by transition rates between states [7] forming a network of interconnected reactions in the state-space. These networks can be modeled using the chemical master equation: a set of differential equations describing the state probabilities and connected transition rates for each state [8]. Biochemical networks are generally Markovian [9], have a number of different control patterns [10], and adhere to specific design principles [11].

One of the most remarkable properties of some molecular machines is the ability to perform "proofreading" or error-correction [12] [13] [14]. Specifically, certain network topologies promote enhanced selectivity (i.e. reduced error) in systems with a competing substrate [12] [13] [14], a phenomenon known as proofreading; this enhancement in selectivity incurs a free energy cost, typically paid via hydrolysis of a phosphate bond. While some aspects of proofreading networks have been examined, such as the speed, accuracy, and dissipation tradeoffs [15][16], and its non-equilibrium applications [17], there has been little mention of transporters [18] [19], exhibiting a similar behavior in the literature.

Despite the complex state-spaces accessible to molecular machines such as transporters, their mechanisms are often described using a single-pathway, highly machine-like cartoon-like model [1] [20] [21]. While such models are helpful for a qualitative understanding of complex protein behavior and chemical networks, they may also build in unwarranted assumptions about the system. There is growing evidence that molecular machines may exhibit complexity beyond that embodied in typical 'textbook' cartoons models [22] [23]. Recent experimental studies have shown that certain traditional model assumptions such as fixed stoichiometry [24] [25], homogenous pathways [26], and constrained binding sites [27] may be incorrect.

Due to the challenges of modelling complex biochemical networks of interest, such as combinatorically large state-spaces, new approaches have been developed to computationally model these systems. Genetic-algorithm sampling methods have been used [28] [29] to evolve complex biochemical networks such as metabolic pathways [30]. In addition, a computationally efficient matrix method has been developed for use on ion channel network models [31]. While these methods have shown promising results, they do not impose thermodynamic consistency (cycle constraints based on detailed balance) which is a necessary condition for physical systems [7].

Motivated by the possibility of discovering new potential mechanisms and the limitations of current approaches, we have developed a new approach that systematically explores the biochemical network model space for a given molecular machine. Our approach generates diverse models that are both thermodynamically consistent and testable. Due to their biomedical significance, we first examined transporters motivated by SGLT type proteins. Our results suggest there is a diverse set of possible mechanisms for these cotransporters, including Hopfield-like kinetic proofreading.

Methods:

We have developed custom software, ModelExplorer, to model molecular machine behavior. Although this study is focused primarily on membrane co-transporters, the software is designed to be general enough for the exploration of other molecular systems. ModelExplorer automatically generates combinatorial state spaces and then uses a modified Monte Carlo Metropolis algorithm to sample the (Markov) model space with a user-defined "energy" or fitness function. This fitness function could embody experimental measurements via a weighted sum of residuals (loss function), but here we use functionality-motivated fitness functions. The software can also impose constraints motivated by structural or biochemical knowledge, such as prohibited states or a known order of binding events.

We create a system model consisting of states and connecting rate constants. Systems states are created from all the allowed combinations of user-specified conformational and chemical substates, and

placed into physically equivalent groupings (see SI). The Monte Carlo sampling will generate a "fitness trajectory" in model space that allows the selection of the fittest models, with a tempering procedure used to avoid trapping. Each model is assessed by its steady-state behavior, in the current implementation, although transient information could be employed. The generated models may then be analyzed for kinetic pathways as well as flow stoichiometry over a range of chemical potential conditions, resulting in experimentally testable models.

The behavior of each model is determined by the rate constants governing transitions among the states. To ensure thermodynamic consistency among all rate constants [7] we use an energy-based formulation where each state and transition state are assigned a free energy. To simplify equations, we use reduced units, with all energies expressed in units of k_BT. For a conformational transition from state $i \rightarrow j$, the Arrhenius-like first-order rate constant is expressed as:

$$k_{ij} = k_0 e^{-(E_{ij}^{\mathrm{bar}} - E_i)}. \tag{1}$$

The user-defined prefactor k_0 is set arbitrarily to $10^{-3}/s$, while E_{ij}^{bar} corresponds to the transition state (free) energy and $E_i < E_{ij}^{\mathrm{bar}}$ is the free energy of state *i*. Because k_0 is the pre-factor for all rate constants, its value does not affect the mechanisms discovered.

For a process $i \rightarrow j$ involving binding or unbinding, the ion or substrate concentration is built into an effective first-order rate constant, α_{ij} [7] given by

$$\alpha_{ij} = e^{\Delta\mu_{ij}} k_0 e^{-(E_{ij}^{\mathrm{bar}} - E_i)}, \tag{2}$$

with the same parameters as above except $\Delta\mu_{ij}$ is the non-equilibrium difference in the chemical potential for the $i \rightarrow j$ state transition:

$$\Delta\mu_{ij} = \begin{cases} -\frac{\Delta\mu_x}{2}, & \text{if the i} \rightarrow \text{j transition is a binding event for species x} \\ +\frac{\Delta\mu_x}{2}, & \text{if the i} \rightarrow \text{j transition is an unbinding event for species x} \end{cases} \tag{3}$$

Here, $\Delta\mu_x$ is the difference in the chemical potential across the membrane for species *x* (e.g. ion or substrate) [7]:

$$\Delta\mu_x = \mu_x^{\mathrm{in}} - \mu_x^{\mathrm{out}} = \ln \frac{[x_{\mathrm{inside}}]}{[x_{\mathrm{outside}}]} \tag{4}$$

where $[x_{\mathrm{inside}}]$ and $[x_{\mathrm{outside}}]$ are the intracellular and extracellular concentration of species. Note that by construction only one species can have a binding/unbinding event during an $i \rightarrow j$ state transition. Here the $\frac{1}{2}$ factor (eq. 3) is an arbitrary choice for dividing the driving force and was found empirically to have a minimal effect on the models found.

We use strings of characters to construct the state space for a molecular machine. This is best illustrated with an example for a hypothetical alternating-access transporter of substrate (S) driven by a sodium ion (N): the state "OF-Nb-Si" represents the outward-facing (OF) conformation with a sodium bound (Nb) and substrate inside the pertinent cell or organelle. The string OF-Nb-Si fully defines the system state in a sufficient way for our kinetic scheme explained below. The state string consists of a set

of 'base state' strings, each of which describes a different characteristic of the machine. For a transporter, this includes the conformational state (inward or outward facing; IF or OF), sodium-ion state (bound to protein, intracellular/"inside", or extracellular/"outside"; Nb, Ni, or No), and a substrate binding state (Sb, Si, or So), resulting in a 3D state space. To examine transport mechanisms in heterogeneous environments we add an additional binding base state for the competing substrate (W) following the same conventions (Wb, Wi, or Wo), resulting in a 4D state space. The set of states result from combinatorial joining of base-state strings, filtering out any user-defined exclusions such as mutually exclusive substrate binding, to create the final high-dimensional state space. This state-space is then mapped to a lower-dimensional space of conformationally unique (equivalent) states, allowing for more efficient state population calculations.

We use the Monte Carlo (MC) Metropolis algorithm [32][33] with tempering to sample our model space, where a trial move consists of randomly adjusting a state or transition energy, and its equivalent "tied" members. Here, "tied" members are considered any state (or transition) which is physically equivalent given a steady state where the outside and inside concentrations are fixed. For example, OF-Nb-Si and OF-Nb-So are physically equivalent states because concentrations of the **S**ubstrate "**i**nside" and "**o**utside" of the cell are constant, and the same for both of these states. Adjusting all the "tied" states during a trial move prevents physically equivalent states in the model from having non-equivalent energies, maintaining thermodynamic consistency. We then construct a rate matrix, $\boldsymbol{K}$, for the new model using the equilibrium and non-equilibrium (binding) energies for each state/transition along with the modified Arrhenius equation (shown above).

The rate matrix, $\boldsymbol{K}$, can be written with the state probabilities column vector, P, to form the chemical master equation (CME) [8][34]:

$$\frac{dP}{dt} = -\boldsymbol{K}P \tag{5}$$

where $K_{ij} = -k_{ji}$ for each state i≠j, and K_{ii} is the sum over incoming rate constants, $\sum_{i \neq j} k_{ij}$. The CME can be solved under steady state conditions using the definition of steady-state, $\frac{dP_i^{ss}}{dt} = 0$, and the additional constraint that $\sum_i P_i = 1$. This yields the steady state probabilities for each state, from which we can calculate the steady flows between states when combined with the rates [35]:

$$j_{ij} = P_i^{ss} k_{ij} \tag{6}$$

where j_{ij} is the steady-state flow from state *i* to state *j*, P_i^{ss} is the steady-state probability of being in state *i*, and k_{ij} is the transition rate between state *i* and *j*. The overall flux, J, of a species (e.g. substrate or ion) is then determined from the sum of *net* flows of user defined transitions, such as the substrate binding/unbinding transitions in the 'inward' facing conformation (see SI).

Each model (set of rate constants between states) will be have a Monte Carlo 'energy' E_{MC} assigned (i.e. fitness score) based on a user-defined general function (e.g., substrate flow into the cell). By convention, lower fitness scores are more fit. The current model 'energy' is compared to the previous model's energy and accepted/rejected based on the Metropolis criterion [32][33]:

$$\text{If } \Delta E_{MC} \leq 0\text{: Accept new model}$$

Else: Accept new model with transition probability, $p = e^{-\beta \Delta E_{MC}}$

where ΔE_{MC} corresponds to the change in the Monte Carlo 'energies' of the two models, and β is the effective inverse 'temperature' parameter: $(k_B T)^{-1}$. We emphasize that our MC procedure does not generate a true thermal ensemble, and the MC β parameter is used only to aid sampling.

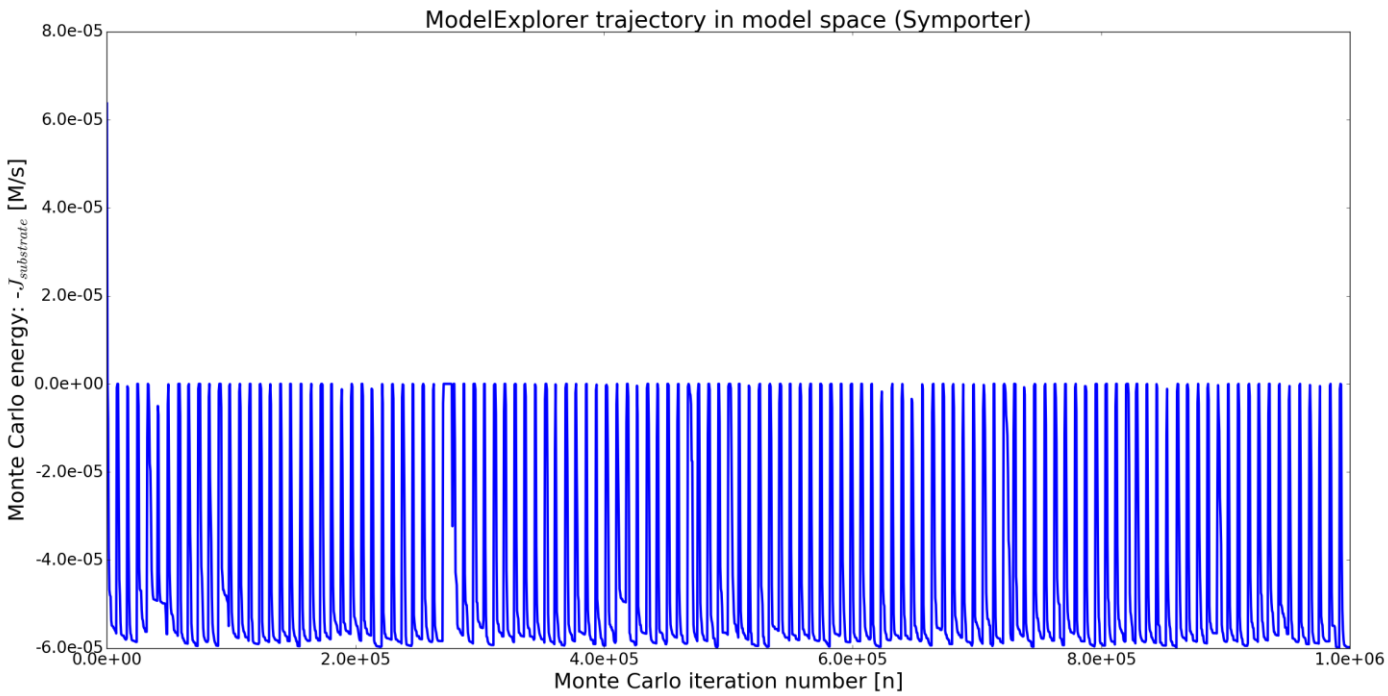

Figure 1: Exploration of model space using Markov-chain Monte Carlo. The plot shows a ModelExplorer trajectory based on symporter energy function (no decoy) during a 1e6 MC step simulation. Note that *each point represents a different kinetic model*. Energy minima correspond to more fit models using $E_{MC} = -J_{Substrate}$ as a fitness function, where $J_{substrate}$ is the flux of substrate. Models are initially at a high energy but quickly find local minima. A tempering schedule of alternating temperature increases and decreases prevents the simulation from being trapped in local minima.

Since each Monte Carlo step results in a model, each simulation yields trajectory in model space **[Figure 1]** – i.e., a sequence of models. In order to avoid trapping in deep "energy" (high fitness) basins we employ a tempering [36] procedure. This procedure cyclically raises and lowers the inverse-temperature β in the Metropolis acceptance criterion, facilitating exploration different areas in the model space because of the increased likelihood of acceptance. ModelExplorer allows for both automated and manual tempering. Automated tempering tracks the change in fitness of the previous models, decreasing β (heating) if the fitness has not changed over several models, and then increasing β (cooling) once a user-defined threshold is met (see SI). Manual tempering sets a fixed heating and cooling schedule for the duration of the simulation. The tempering procedure is fully customizable and increases the diversity of models found in a simulation. **[Figure 1]** is an example of the Monte Carlo "energy" trajectory in model space produced in ModelExplorer for a simple symporter system. Note that the initial models which are not easily visible at the very left of the graph exhibit poor fitness (positive MC energy), but negative-energy models are quickly found.

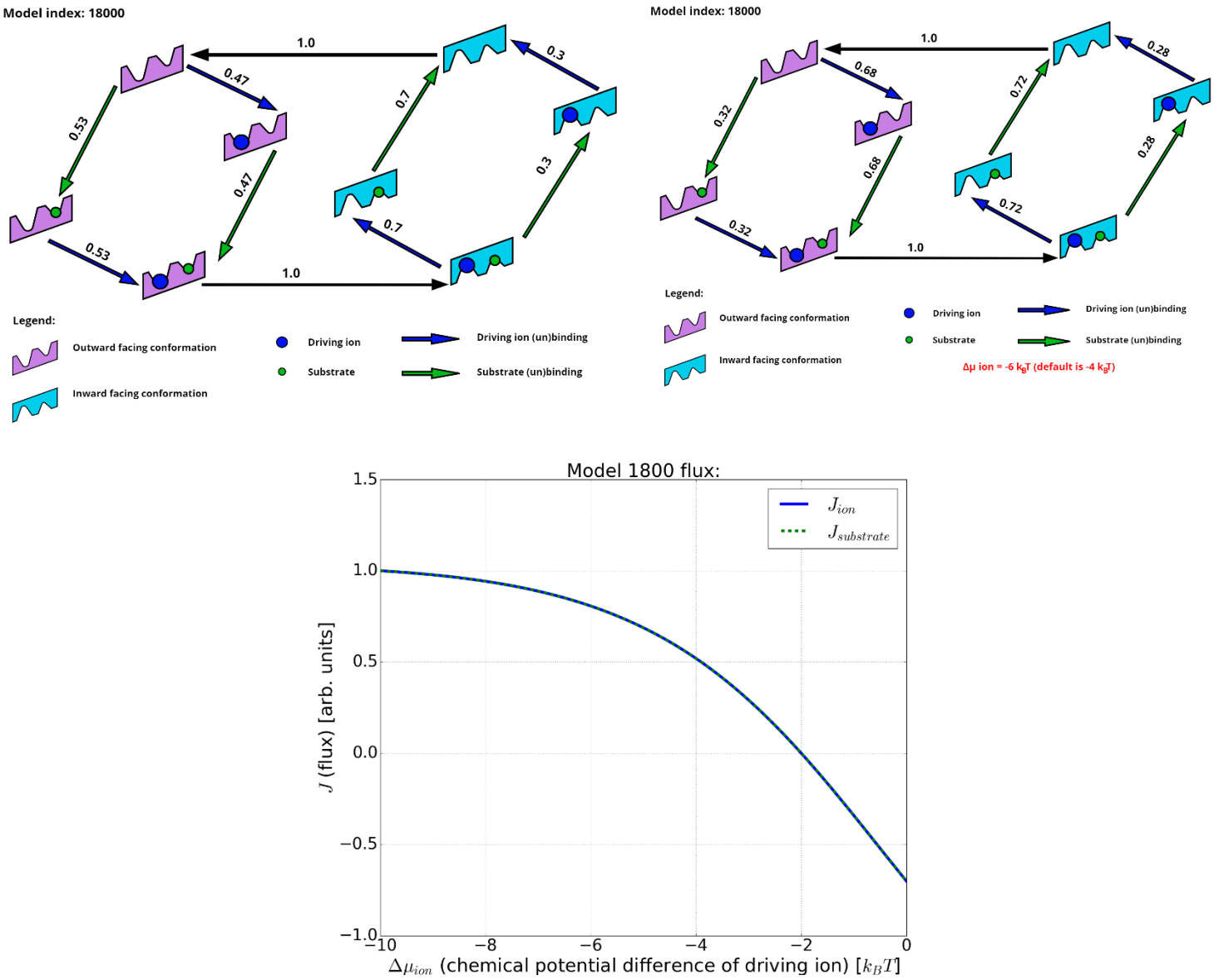

Figure 2. Example symporter model. **Figure 2a (top left):** Relative flow pathways, normalized by the maximum flow, are shown for a particular symporter model (index 18000) found during the 1e6 MC step symporter simulation. This model employs a combination of two simple pathways, thus differing from traditional symporter cartoons, while still maintaining a 1:1 ratio of ion to substrate flux. **Figure 2b (top right):** The same model can be analyzed at different fixed ion chemical potential differences (essentially concentrations). Here, the ion chemical potential difference is -6 k_BT instead of the default -4 k_BT (shown in figure 2a). Note that the network topology is the same but has different relative flows. **Figure 2c (bottom):** The flux (J) of the driving ion and substrate of the same model (index 18000) are shown over a range of ion chemical potential differences. This allows the analysis of stoichiometry and flux information for the given model. Note the 1:1 stoichiometry of the substrate to ion flux, indicating an ideal symporter with no leaks. Also note that the absolute flux values are not physically relevant due to an arbitrary overall constant prefactor (i.e. all flows can be scaled by an arbitrary constant and remain consonant with the governing equations). The ion and substrate flux have been scaled by the maximum flux for visual clarity.

Selected models can be analyzed in several ways using ModelExplorer: characterizing overall stoichiometries, kinetic pathway analysis, and manual adjustment of selected transition rates. The ion and substrate flux of a model can be calculated over a range of chemical potential differences by incrementing the desired chemical potential difference and updating the non-equilibrium energy terms for each state. The rates, steady state probabilities, flows, and then fluxes are then recalculated for each chemical potential increment, allowing for the examination of stoichiometry. Since ModelExplorer calculates the net flows between states, a kinetic pathway diagram for a specific model and chemical potential difference can be made. Furthermore, individual models may be perturbed by modifying specific state/transition energies, leading to insights on pathway characteristics (e.g. pathway with or without leaks). Combining the kinetic pathway diagram with the flux analysis provides an insightful method to investigate the dynamic behavior of molecular machines and provide testable mechanistic models **[figure 2]**. Note that the flux values for each model are not physically relevant due to an arbitrary overall constant prefactor (i.e. all flows and resulting fluxes can be scaled by an arbitrary constant and remain consonant with the governing equations). The ion and substrate flux have been scaled by the maximum flux (i.e. ion flux at largest chemical potential difference), and the net flows between states are normalized by the maximum net flow.

In addition to single model analysis, we have developed tools for the meta-analysis of the data – i.e., a systems analysis. A simple data pipeline allows for the analysis of sampling parameters, run-to-run model distributions, and model clustering, based on the differences in model flows. Sampling efficiency can be evaluated based on the average time to find a sufficiently different model (i.e. cluster) during the simulation as well as the run-to-run comparison. The run-to-run comparison tool computes the minimum distance between models found in simulation 'A' compared to simulation 'B' and vice versa, generating a model distribution between the runs. This provides insight into possibly unique models found under different simulation conditions. Models may also be clustered hierarchically [37], allowing for the discovery of different model classes. The model differences are calculated using the Euclidean distance between the vectors of normalized net flows between states for a given model.

ModelExplorer has a number of adjustable physical parameters (change in chemical potential), simulation parameters (seed, run length), sampling parameters (tempering schedule, amount to perturb a state/barrier energy), as well as constraints (illegal transitions and states) for a simple cotransport system. In order to investigate Hopfield-like enhanced discrimination [12], the cotransport system with a decoy substrate has additional parameters and constraints. The decoy-bound state energies differ from their equivalent substrate-bound states by a fixed amount ($\Delta\Delta G$ parameter), and have equal barrier energies. This constraint forces the enhanced selectivity to result from the difference in binding affinities alone, and not "internal proofreading", which is consistent with Hopfield's kinetic proofreading network (see discussion) [12]. Furthermore, for simplicity, occluded states and mutual binding states of the substrate and decoy have been disabled.

Results:

We present results for a simple symporter (cotransporter) system and for a transporter 'challenged' by a decoy substrate. For a simple cotransporter system without a decoy, the results include validation of ideal symporter/antiporter mechanisms, selected network topologies, and substrate/ion flux for a range of ion chemical potential differences. For cotransporters with an additional decoy substrate, we analyze

several features of four simulations: selected network topologies, substrate/decoy/ion flux for varying ion chemical potential differences, network heterogeneity via clustering, and possible alternative mechanisms.

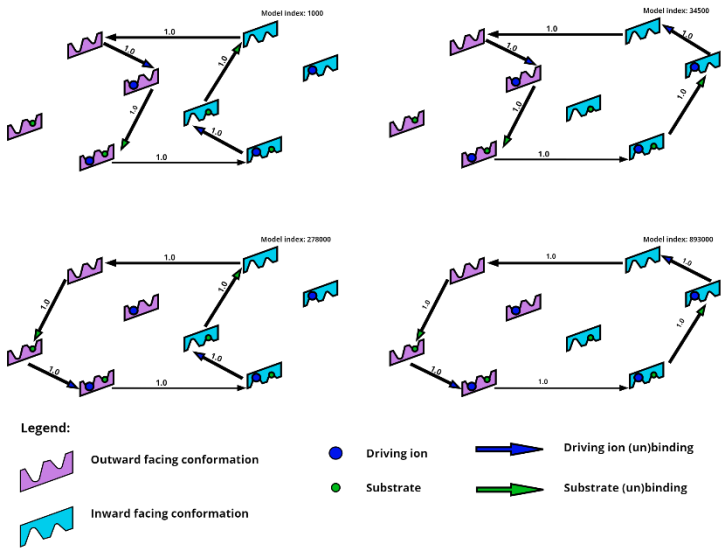

Figure 3: Finding multiple pathways. All four ideal symporter pathways are generated during a 1e6 MC step ModelExplorer simulation using a fitness function (Monte Carlo energy) $E_{MC,Symport} = - J_{Substrate}$. These ideal pathways correspond to the lowest number of connected states that still achieve the functional goal of symporting. The net flows normalized by the maximum flow.

We first studied a transporter system with a single substrate driven by an ion gradient. To explore the symporter behavior of this model, the substrate chemical potential difference ($\Delta\mu_{substrate}$) was set to 2 k_BT and the ion chemical potential difference ($\Delta\mu_{ion}$) was set to -4 k_BT. The fitness goal (Monte Carlo energy) was set to:

$$E_{\mathrm{MC,symport}} = -J_{\mathrm{substrate}} \tag{7}$$

where $J_{\mathrm{substrate}}$ is the flux of the substrate into the cell, (see SI). This promotes intracellular substrate flow along the ion gradient. Note that negative Monte Carlo energies are more fit by convention. The simulation of 1e6 MC steps yielded the four idealized symporter cycles, as well as combinations of the idealized symporter cycles. All of these models consist of a 1:1 ratio of substrate and ion flux into the cell - consistent with the theoretical predications of an optimized symporter **[figure 3]**.

Antiporter behavior of the same state space was tested by modifying the substrate chemical potential difference ($\Delta\mu_{substrate}$) to -2 k_BT and setting the fitness goal to:

$$E_{\mathrm{MC,antiport}} = J_{\mathrm{substrate}} \tag{8}$$

where $J_{\text{substrate}}$ is the flux of the substrate, (see SI). This promotes extracellular flow of the substrate opposite to the ion gradient and flow**.** The simulation of MC n = 1e6 steps yielded antiporter models consisting of a 1:1 ratio of substrate flux out of the cell and ion flux into the cell - consistent with the theoretical predications of an optimized antiporter **[figure 4].**

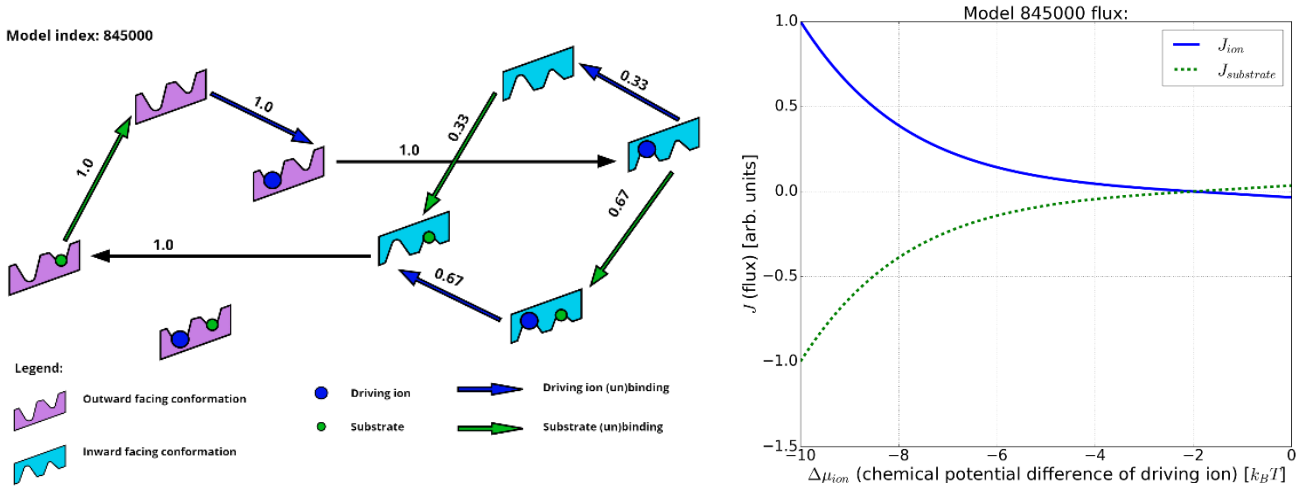

Figure 4a (left): Selected antiporter. This model (index 845000) shows a combination of pathways, which differ from the traditional antiporter representation while still maintaining a 1:1 ratio of ion to substrate flux. The net flows normalized by the maximum flow. **Figure 4b (right):** The flux of the driving ion and substrate of the same antiporter model (index 845000) are shown. The pathway in figure 4a (left) corresponds to an ion chemical potential difference of -4 k_BT. Note the 1:1 stoichiometry of the substrate efflux to ion influx, indicating an ideal antiporter with no leaks. The ion and substrate fluxes have been scaled by the maximum flux for visual clarity.

To investigate whether symporter models are capable of enhancing selectivity beyond that generated by their differing affinities ($\Delta\Delta G$ = 1 k_BT), a competing decoy substrate was added to a transport system with a substrate driven by an ion gradient. The substrate and decoy chemical potential differences ($\Delta\mu_{substrate}$, $\Delta\mu_{decoy}$) were set to 2 k_BT, and the ion chemical potential difference ($\Delta\mu_{ion}$) was set to -4 k_BT. The fitness goal (Monte Carlo energy) was set to:

$$E_{\text{MC}} = -J_{\text{substrate}} \left(\frac{|J_{\text{substrate}} + \varepsilon|}{|J_{\text{decoy}} + \varepsilon|} \right)^{\alpha} \tag{9}$$

where α allows for weighting of the substrate to decoy flow ratio, ε = 1e-15 improves numerical stability, and $J_{\text{substrate}}$ and $J_{\text{substrate}}$ are the flux of the substrate and decoy respectively (see SI). To reduce runtime and memory requirements, models were stored every 500 MC steps. Four separate simulations of 1e6 MC steps with α = 1 were conducted at different sampling conditions with each simulation outputting 2000 models. The models from each run were combined and then filtered based on the sodium ion to substrate ratio being greater than 0.1, and the substrate to decoy ratio being $10e^{\Delta\Delta G}$ (ten times the expected discrimination ratio at equilibrium) [34], where $\Delta\Delta G$ = 1 is the difference in binding affinities between the substrate and decoy substrate. This resulted in 1783 (22%) of the aggregate 8000 models exhibiting enhanced selectivity. These models were then clustered based on the normalized flows between states using a Euclidean distance metric with complete-linkage [37]. Based on a distance threshold of 0.65, found empirically to produce qualitatively different networks, there were four different mechanistic model classes with differing kinetic pathways **[figures 5-6]**.

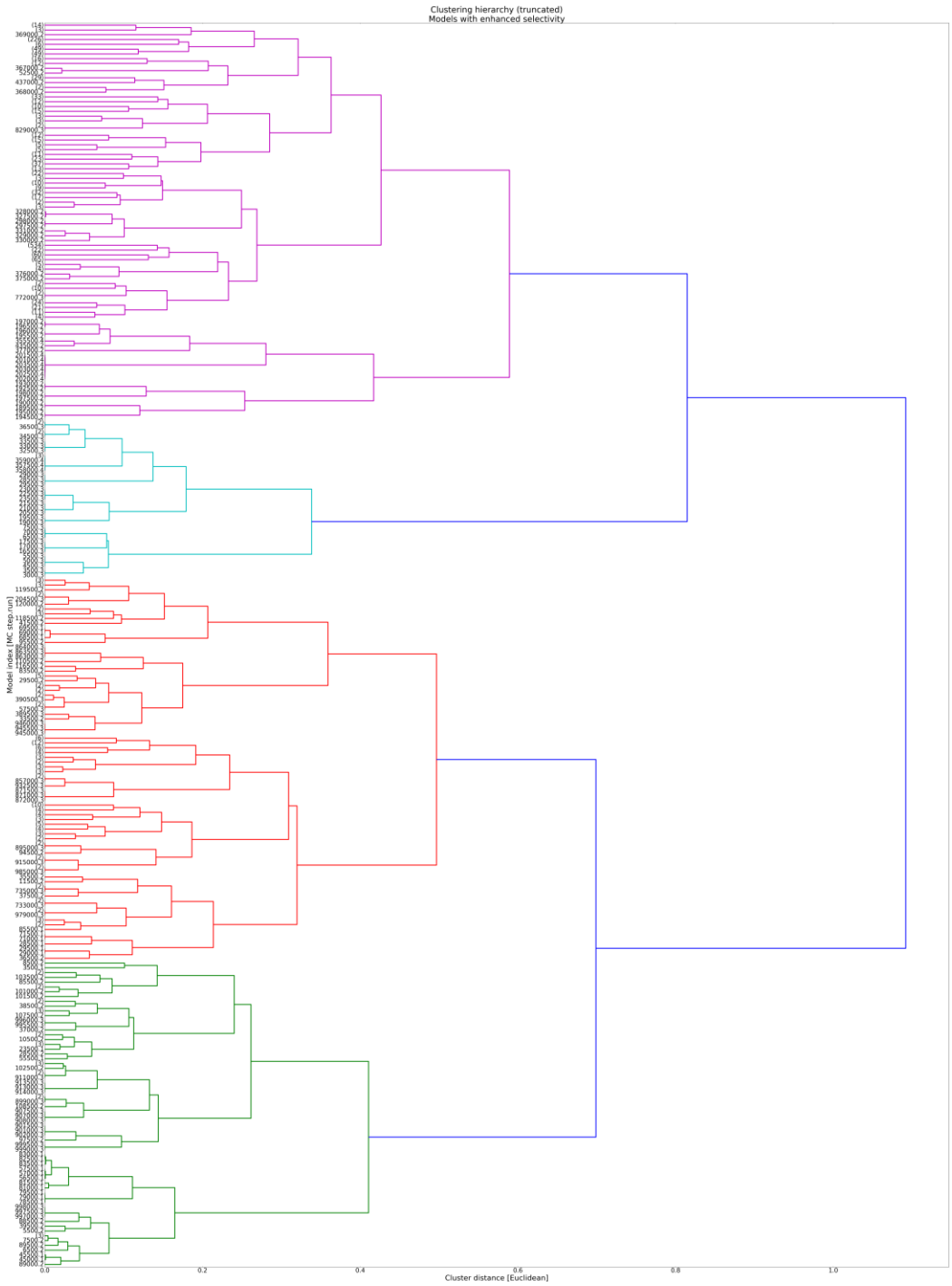

Figure 5: Discovery of highly diverse models. The plot shows hierarchical clustering (truncated) of cotransport models with enhanced selectivity of a substrate over a decoy substrate. 8000 models were aggregated from four separate runs with different sampling parameters. The models were filtered for enhanced selectivity, defined as models where the sodium ion to substrate flux ratio is greater than 0.1, and the substrate to decoy flux ratio being greater than $10e^1$ ($\Delta\Delta G = 1\ k_BT$).The models were clustered based on a Euclidean distance metric of the flows of each model using the complete-linkage strategy. The distance threshold of 0.65 is empirically found to lead to distinct mechanism classes with qualitatively different pathways. The dendrogram was truncated to eight levels for visual clarity. The number of nodes below the truncated node are shown in parenthesis.

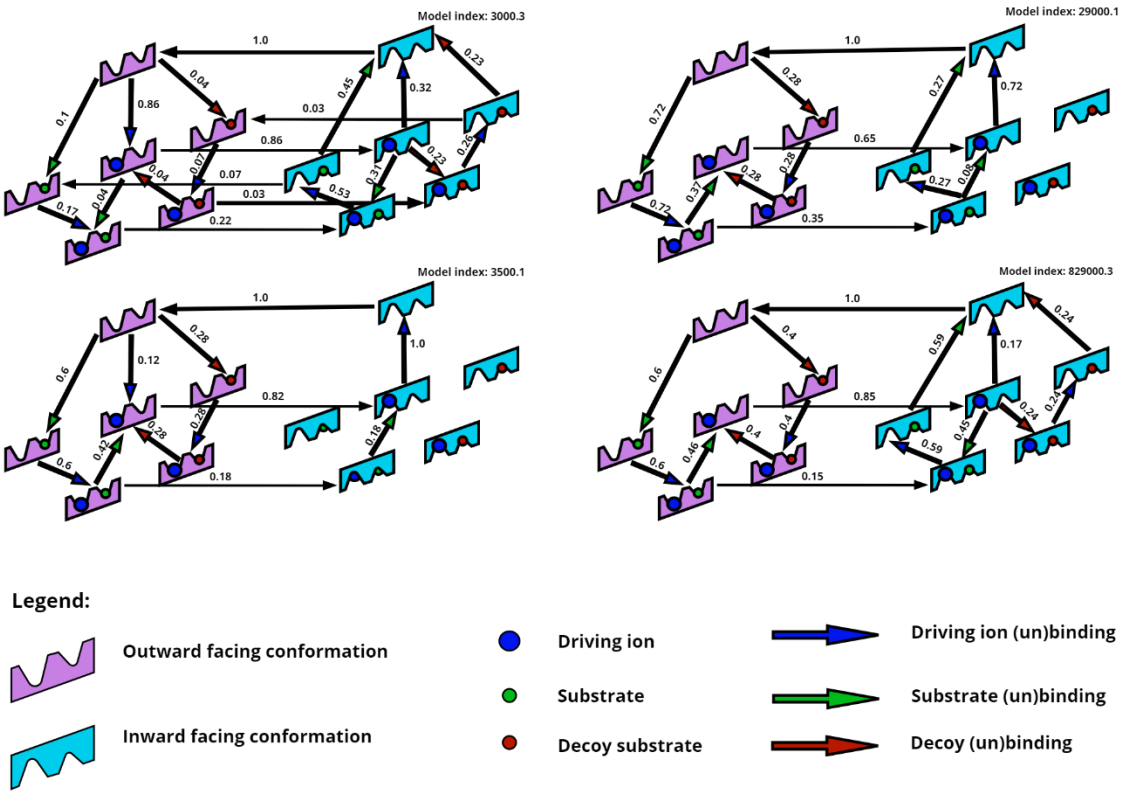

Figure 6: Collection of enhanced selectivity model classes from four combined simulations. The hierarchal clustering with complete-linkage found four clusters at a Euclidean distance threshold of 0.65. Each model represents a cluster, which exhibit a high level of selectivity and a moderate amount of model diversity. The net flows normalized by the maximum flow.

All model classes contained an ion leak coupled with an extra unbinding step for both the substrate and the decoy in the outward facing conformation, as well as symmetry in the flow pathways of the outward facing conformation. The most populated class is shown in **[figure 7a,b]** and corresponds to the purple cluster in **[figure 5].** To investigate the mechanism of enhanced selectivity, the ion leak was removed by increasing the energy barrier between the ion-bound outward-facing and inward-facing conformations by 100 k_BT. The resulting model exhibited a decrease in selectivity to the expected equilibrium value ($e^{\Delta\Delta G}$ with $\Delta\Delta G = 1$) **[figure 7c,d]**. This suggests that the ion leak is a potential driving mechanism for enhanced selectivity in Hopfield-like reaction networks. The model classes also contain non-intuitive pathways such as shown in **[figure 8]**. This model is fully connected (i.e. each legal state transition has a non-zero flow) and corresponds to the cyan cluster in **[figure 5].** These alternative mechanisms of enhanced selectivity highlight the moderate model diversity found during simulations, based on the preliminary data.

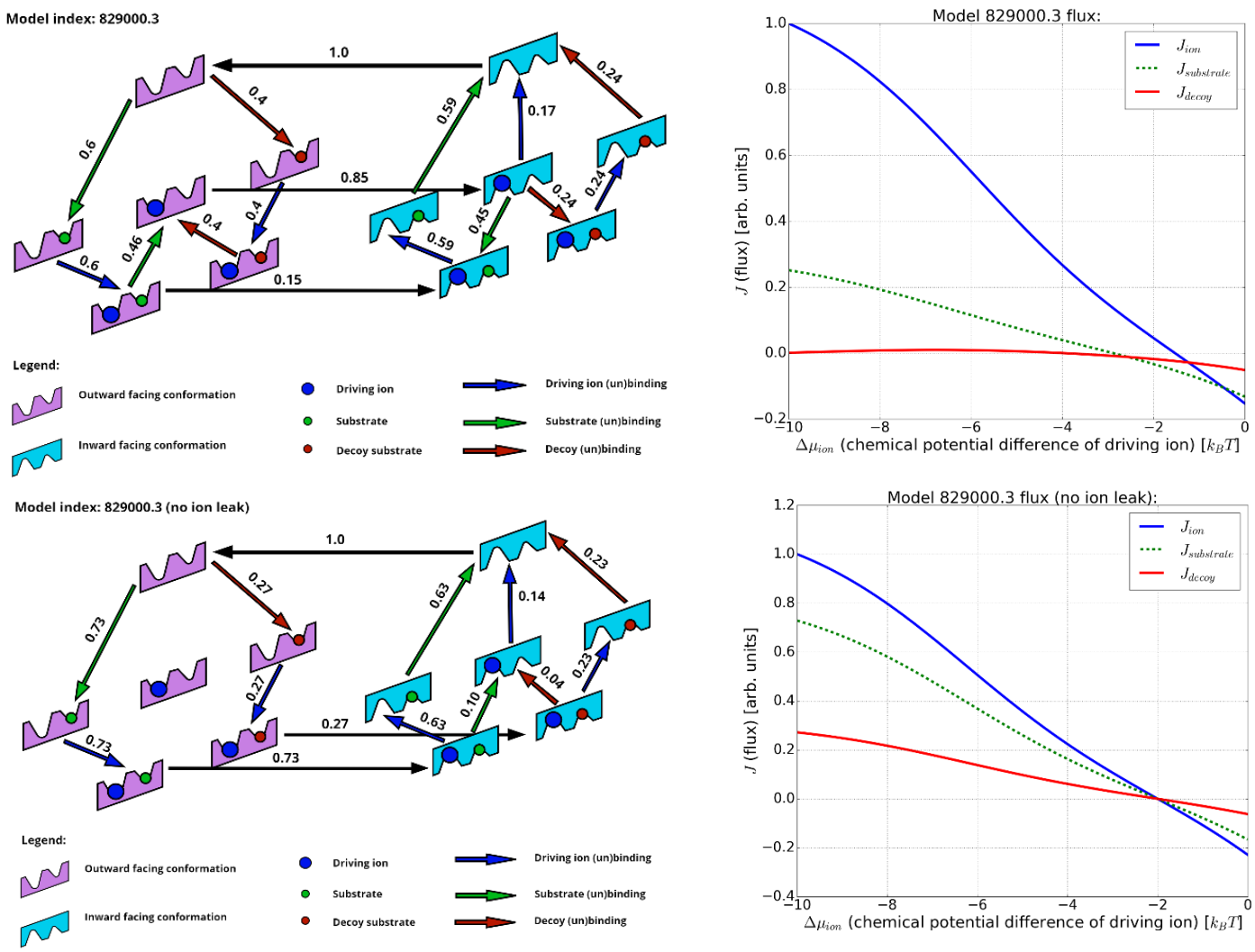

Figure 7. A Hopfield-like "proofreading" model. **Figure 7a (top left):** Model (index 829000.3) from a 1e6 MC step simulation. The majority of the models with enhanced selectivity (as defined earlier) have a network that resembles this model (purple cluster in dendrogram). The model has an extra unbinding step for both the substrate and decoy substrate. Due to the differences in binding affinities (ΔΔG = 1 k_BT) the substrate is favored to re-enter the reaction network compared to the decoy substrate. The net flows normalized by the maximum flow. **Figure 7b (top right):** Flux diagram of the same "proofreading" model. Note that while the model is optimized at ion chemical potential difference = -4 k_BT, there is also enhanced discrimination at larger ion chemical potential difference (concentration) values. Also, note that there is a regime with low ion chemical potential difference where the substrate and the decoy have negative flux (i.e. leak outside of the cell). Over a narrow range of concentrations, this model has an unbounded discrimination (substrate to decoy) ratio as the decoy flux approaches zero. The ion, substrate, and decoy fluxes have been scaled by the maximum flux for visual clarity. **Figure 7c (bottom left**): Pathway of the same model with the ion leak removed. The ion leak was removed by increasing the energy barrier between the ion-bound outward-facing and inward-facing conformations by 100 k_BT. Without an ion leak, the substrate and decoy unbind to the ion in the inward-facing state, increasing the ratio of substrate to decoy flux **[SI 1]**. **Figure 7d (bottom right):** Flux diagram of the same model with the ion leak removed. Without the ion leak, the substrate to decoy flux ratio is much lower (~2.7), such that the model does not discriminate beyond $e^{\Delta\Delta G}$, where ΔΔG = 1, the expected discrimination ratio at equilibrium. This suggests that the ion leak is a potential driving mechanism for enhanced proofreading in a Hopfield-like reaction network.

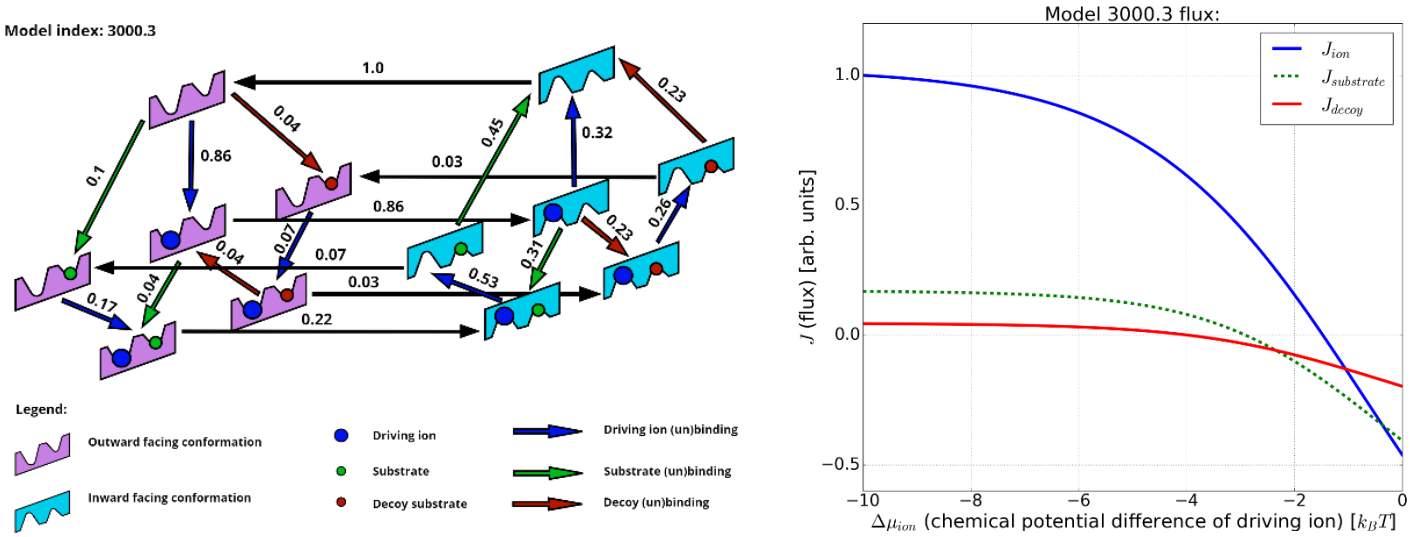

Figure 8. A non-intuitive proofreading model. **Figure 8a (left):** Kinetic pathway of another model (index 3000.3) with enhanced selectivity. Note that the states in this model are full connected, forming a complex network structure that is highly non-intuitive. Over a narrow range of concentrations, this model has an unbounded discrimination (substrate to decoy) ratio as the decoy flux approaches zero. The net flows normalized by the maximum flow. **Figure 8b (right):** Flux diagram of the same model (index 3000.3) with enhanced selectivity. Note that the model is optimized at ion chemical potential difference = -4 k_BT . Also, note that there is a regime with low ion chemical potential difference where the substrate and the decoy have negative flux (i.e. leak outside of the cell). Over a narrow range of concentrations, this model has an unbounded discrimination (substrate to decoy) ratio as the decoy flux approaches zero. The ion, substrate, and decoy fluxes have been scaled by the maximum flux for visual clarity.

Discussion:

The results indicate that Monte Carlo sampling of model space can find diverse, testable models that provide insight into complicated cotransport systems. The model-sampling approach systematically found kinetic pathways for both simple symporters and antiporters, including mixed ideal pathways, validating our approach as applied to simple transport systems. The preliminary data also indicates the possibility of cotransporters that exhibit enhanced selectivity similar to Hopfield's kinetic proofreading models. In these models, there is an ion leak which when removed prevents the enhanced discrimination. This suggests an apparently unreported mechanism for secondary active transporters, which can enhance selectivity to a remarkable degree in a limited range of conditions. In our systems analysis of all models sampled, clustering show a fairly diverse group of model classes that exhibit enhanced selectivity using different kinetic pathways.

To our knowledge, this is the first systems-based approach to modeling molecular machines that also satisfies the physically required detailed-balance constraints. Every kinetic model that is generated is also thermodynamically consistent. Since all biochemical reactions are governed by the laws of thermodynamics, this consistency is an important part of accurately modeling the mechanisms of molecular machines. The energy-based formulation combined with a name-based full state-space construction distinguishes our method from similar approaches.

All the models sampled in this study, by construction, were of the 'Hopfield type' [12] where substrate discrimination results from the interplay of an external driving force (due to the ion gradient) and a difference in binding affinity between the two substrates. No additional differences between the substrate and decoy were permitted: specifically, all barrier energies were constrained to be identical for both species. This type of proofreading differs from what might be called "internal proofreading" where additional discrimination results from differing barrier heights [14]. Biological proofreading can be expected to mix both mechanisms to some degree [15], but we chose to focus on Hopfield proofreading because an arbitrary degree of discrimination can result from differing barriers – for example, if the decoy species is a negligible on-rate for the transporter. Our approach differs somewhat from Hopfield's work in that *no irreversible steps enter our models.* The full reversibility appears to be a necessary ingredient in the ideal discrimination that occurs in some models for specific concentrations.

There are several areas of improvement for future work. Sampling models in the 'competitive' transporter state-space has proven to be challenging and will require further optimization. New functionality will need to be added to ensure sufficient sampling, such as initializing several separate simulations using a Latin hypercube [38]. Once the sampling issues have been resolved, we plan to provide tighter integration with experimental data. The Monte Carlo energy function can be modified into a loss function, which would minimize the difference between experimental and calculated values of interest, such as charge movement measured during a voltage-clamp experiment [39]. This would generate kinetic models, or potentially classes of models, that describe the experimental data. Finally, we plan to apply the approach to further study complex molecular machines such as the ATP synthase.

Conclusions:

In light of new evidence for the alternative behavior of molecular machines, we have developed a thermodynamically consistent approach to systematically explore a range of mechanisms, generating experimentally testable kinetic models based on a predetermined fitness function. In short, we have taken a 'systems biology' approach to analyzing mechanisms of molecular machines. The approach was first applied to a cotransporter system with and without a decoy substrate. Preliminary results have validated the method against 'textbook' symporter/antiporter models, and have generated a variety of mechanisms that enhance selectivity – including Hopfield-like proofreading networks, which could have important biological implications and biotechnology applications. The mechanisms discovered using an automated approach would be difficult to design using human intuition alone.

Acknowledgements: We thank Barmak Mostofian and Jeremy Copperman for helpful discussions. This research was supported by NSF Grant MCB-1715823.

References:

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, Molecular Biology of the Cell, New York: Garland Science, 2002.

[2] S. B. Poulsen, R. A. Fenton and T. Rieg, "Sodium-glucose cotransport," *Current Opinion in Nephrology and Hypertension,* vol. 24, no. 5, pp. 463-469, 2015.

[3] D. Oleg, Y. Tatiana, K. Dmitry and D. Oleg, "Analysis of the efficacy of SGLT2 inhibitors using semi-mechanistic model," *Frontiers in Pharmacology,* vol. 5, p. 218, 2014.

[4] D. Deng and N. Yan, "GLUT, SGLT, and SWEET: Structural and mechanistic investigations of the glucose transporters," *Protein Science,* vol. 25, pp. 546-558, 2016.

[5] P. Bisignano, C. Kalyanaraman, C. Ghezzi, E. M. Wright, J. Abramson, A. Paz, M. P. Jacobson, R. Friemann and M. Grabe, "Structural Insights into Sodium-Dependent Sugar Transporters and their Inhibition Mechanism," *Biophysical Journal,* vol. 112, no. 3, p. 128a, 2017.

[6] E. M. Wright, C. Ghezzi and D. D. F. Loo, "Novel and Unexpected Functions of SGLTs," *Physiology,* vol. 32, no. 6, pp. 435-443, 2017.

[7] T. L. Hill, Free Energy Transduction and Biochemical Cycle Kinetics, Springer, 1989.

[8] D. Schnoerr, G. Sanguinetti and R. Grima, "Approximation and inference methods for stochastic biochemical kinetics—a tutorial review," *Journal of Physics A: Mathematical and Theoretical,* vol. 50, no. 9, 2017.

[9] G. Jenkinson and J. Goutsias, "Markovian dynamics on complex reaction networks," *Physics Reports,* vol. 529, no. 2, pp. 199-264, 2013.

[10] V. Galstyan, L. Funk, T. Einav and R. Phillips, "Combinatorial Control through Allostery," *The Journal of Physical Chemistry B,* vol. 123, no. 13, pp. 2792-2800, 2019.

[11] W. Lim, C. Lee and Chao Tang, "Design Principles of Regulatory Networks: Searching for the Molecular Algorithms of the Cell," *Molecular Cell,* vol. 49, no. 2, pp. 202-212, 2013.

[12] J. J. Hopfield, "Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity," *PNAS,* vol. 71, no. 10, pp. 4135-4139, 1974.

[13] J. Ninio, "Kinetic amplification of enzyme discrimination," *Biochimie,* vol. 57, no. 5, pp. 587-595, 1975.

[14] A. Fersht, "Editing mechanisms in protein synthesis. Rejection of valine by the isoleucyl-tRNA synthetase," *Biochemistry,* 1977.

[15] K. Banerjee, A. B. Kolomeisky and O. A. Igoshin, "Elucidating interplay of speed and accuracy in biological error correction," *PNAS,* vol. 114, no. 20, pp. 5183-5188, 2017.

[16] J. Mallory, A. B. Kolomeisky, and O. A. Igoshin, “Trade-Offs between Error, Speed, Noise, and Energy Dissipation in Biological Processes with Proofreading,” *The Journal of Physical Chemistry B*, vol. 123, no. 22, pp. 4718-4725, 2019.

[17] A. Murugan, D. A. Huse and S. Leibler, "Discriminatory Proofreading Regimes in Nonequilibrium Systems," *Phys. Rev. X,* vol. 4, no. 2, p. 021016, 2014.

[18] W. D. Stein and B. Honig, "Models for the active transport of cations - The steady-state analysis," *Molecular and Cellular Biochemistry,* vol. 15, no. 1, pp. 27-44, 1977.

[19] T. Chou, K. Mallick and R. K. P. Zia, "Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport," *Reports on Progress in Physics,* vol. 74, no. 11, 2011.

[20] I. Sobczak and J. S. Lolkema, "The 2-Hydroxycarboxylate Transporter Family: Physiology, Structure, and Mechanism," *Microbiology and Molecular Biology Reviews,* vol. 69, no. 4, pp. 665-695, 2005.

[21] R. Schröder, S. F. Wulf, V. Ropars, S. Fujita-Becker, M. Oster, G. Hofhaus, L. G. Trabuco, O. Pylypenko, H. Lee Sweeney and A. Houdusse, "The Force Producing ADP State of Myosin Bound to Actin," *PNAS,* vol. 110, no. 3, p. 13a, 2016.

[22] K. E. Mangold, B. D. Brumback, P. Angsutararux, T. L. Voelker, W. Zhu, P. W. Kang, J. D. Moreno and J. R. Silva, "Mechanisms and models of cardiac sodium channel inactivation," *Channels,* vol. 11, no. 6, pp. 517-533, 2017.

[23] A. Bazzone, A. J. Zabadne, A. Salisowski, M. G. Madej and K. Fendler, "A Loose Relationship: Incomplete H+/Sugar Coupling in the MFS Sugar Transporter GlcP," *Biophysical Journal,* vol. 113, no. 12, pp. 2736-2749, 2017.

[24] R. K. Henderson, K. Fendler and B. Poolman, "Coupling efficiency of secondary active transporters," *Current Opinion in Biotechnology,* vol. 58, pp. 62-71, 2019.

[25] C. Kettner, A. Bertl, G. Obermeyer, C. Slayman and H. Bihler, "Electrophysiological Analysis of the Yeast V-Type Proton Pump: Variable Coupling Ratio and Proton Shunt," *Biophys J.,* vol. 85, no. 6, p. 3730–3738, 2003.

[26] A. E. Robinson, N. E. Thomas, E. A. Morrison, B. M. Balthazor and K. A. Henzler-Wildman, "New free-exchange model of EmrE transport," *PNAS,* vol. 117, no. 47, pp. E10083-E10091, 2017.

[27] L. Shi, M. Quick, Y. Zhao, H. Weinstein and Jonathan A. Javitch, "The Mechanism of a Neurotransmitter:Sodium Symporter—Inward Release of Na+ and Substrate Is Triggered by Substrate in a Second Binding Site," *Molecular Cell,* vol. 30, no. 6, pp. 667-677, 2008.

[28] A. Deckard and H. M. Sauro, "Preliminary Studies on the In Silico Evolution of Biochemical Networks," *ChemBioChem,* vol. 5, pp. 1423-1431, 2004.

[29] V. Menon, N. Spruston and W. L. Kath, "A state-mutating genetic algorithm to design ion-channel models," *PNAS,* vol. 106, no. 39, pp. 16829-16834, 2009.

[30] W. Gottstein, S. Müller, H. Herzel and R. Steuer, "Elucidating the adaptation and temporal coordination of metabolic pathways using in-silico evolution," *Biosystems,* vol. 117, pp. 68-76, 2014.

[31] Z. R. Teed and Jonathan R. Silva, "A computationally efficient algorithm for fitting ion channel parameters," *MethodsX,* vol. 3, pp. 577-588, 2016.

[32] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth and A. H. Teller, "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.,* vol. 21, p. 1087, 1953.

[33] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika,* vol. 57, no. 1, pp. 97-109, 1970.

[34] D. M. Zuckerman, "Physical Lens on the Cell," [Online]. Available: http://physicallensonthecell.org/cell-biology-phenomena/active-kinetic-proofreading. [Accessed 26 June 2019].

[35] R. Phillips, J. Kondev, J. Theriot, H. Garcia, R. Phillips, J. Kondev, J. Theriot and H. Garcia, Physical Biology of the Cell, Garland Science, 2012.

[36] S. Whitelam and and Phillip L. Geissler, "Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles," *J. Chem. Phys.,* vol. 127, p. 154101, 2007.

[37] L. Rokach and O. Maimon, "Clustering Methods," in *Data Mining and Knowledge Discovery Handbook*, Springer, 2005, pp. 321-352.

[38] M. D. McKay, R. J. Beckman and W. J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics,* vol. 21, no. 2, pp. 239-245, 1979.

[39] D. D. F. Loo, A. Hazama, S. Supplisson, E. Turk and E. M. Wright, "Relaxation kinetics of the Na+/glucose cotransporter," *Proc Natl Acad Sci ,* vol. 90, no. 12, p. 5767–5771, 1993.