IMPROVED MODELING OF BIOCHEMICAL REACTION NETWORKS:

APPLICATIONS TOWARDS TRANSPORTER MECHANISM DISCOVERY

by

August Daniel George

A DISSERTATION

Presented to the
Biomedical Engineering Department
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Biomedical Engineering

September 2023

Biomedical Engineering

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of
August Daniel George
has been approved.

Daniel M. Zuckerman
Advisor

Jeremy Goecks
Chair

Charles S. Springer, Jr.
Committee Member

James Faeder
Committee Member

Peter G. Jacobs
Committee Member

TABLE OF CONTENTS

ABSTRACT

Biochemical reaction networks, mathematical models that describe the dynamic interrelationships between biochemical species, facilitate studying complex biological processes such as cell signaling, metabolism, and enzyme kinetics. This work focuses on secondary active membrane transporter proteins which selectively move nutrients across the cell membrane using the energy stored in electrochemical gradients. The dysregulation of these transporters is implicated with diseases such as type 2 diabetes, anxiety, and depression. However, their exact mechanisms are often poorly understood due to challenges in experimentally observing the entire reaction pathway. Recent studies suggest that transporter proteins may exhibit more complexity than previously thought – such as alternative functionality and multiple reaction pathways. In response to these findings, we developed a novel computational method to engineer alternative functions for transporter proteins using biochemical reaction network models. We use a rules-based approach to systematically create the underlying network structure for a generic transporter, while also enforcing constraints to ensure physical consistency in the network models. This method was applied to study a theoretical cotransporter in a competitive environment with an ion, true substrate, and decoy substrate. We were able to engineer (in silico) transport mechanisms that exhibited a new functionality of enhanced selectivity of the true substrate, paving the way for future experimental investigation.

Alongside these advancements, solid-supported membrane electrophysiology, an emerging experimental technique, has improved the measurement of the transient behavior of transporter proteins under different perturbations. Despite the enhanced stability and signal-to-noise ratio offered by this method, it is unknown how much information is contained in these sparse and noisy time series datasets and how well transport mechanisms can be inferred from them. We developed a computational pipeline using reaction network models and Bayesian inference to generate robust descriptions of transporter kinetic parameters, experimental nuisance parameters, and their uncertainties. By applying this pipeline to synthetic datasets and models, we find a surprisingly rich amount of information in these datasets. Our results suggest that this pipeline can be used to select between competing mechanisms correctly and recommend assays that yield more informative data, suggesting its potential to enhance our understanding of transporter proteins significantly.

Dedication

For my lovely wife, Zhi Zeng, who inspired and supported me during this long journey.

And for my dear parents, Daniel and Leisa George, who have always encouraged my curious mind.

List of Figures

List of Tables

1 Introduction

1.1 Background and Motivation

This section first describes the primary research question that directs the dissertation, lists the specific objectives that the research aims to achieve, and states the underlying hypothesis of this research. The following sections examine the contextual background and motivation for this work, ongoing challenges, and the broader implications of this research.

1.1.1 Research objectives

This dissertation seeks to answer the following research question: “Can the development of robust computational methods provide a more accurate modeling of transporter proteins and their mechanistic complexities?"

To address this question, the dissertation aims to:

1. Develop robust computational tools to improve biochemical network modeling and data analysis techniques targeting transporter proteins
2. Utilize these computational tools to explore the mechanistic heterogeneity and complexity of transporter proteins *in silico*

This research is predicated on the hypothesis that robust computational methods will provide more detailed insights into transporter protein complexities than current methodologies.

1.1.2 The Essential Role of Transporter Proteins

Molecular Machinery in Biological Systems Biological systems are an intricate web of interactions between numerous components, all working towards sustaining life. These systems cover many levels of detail and complexity, from nanoscopic molecular interactions to planetary-scale ecosystems [6, 242]. Despite the enormous differences in scale, each system utilizes an elaborate network that functions toward its particular objective.

Working inside the cells of all living systems, from bacteria to archaea to eukaryotes, are “molecular machines" that operate at the molecular scale - thousands of times smaller than the width of a human hair [5, 19]. These “machines" are assembles of biomolecules, proteins and protein complexes, that use the available free energy to do mechanical work [5, 19]. Consider these proteins as sophisticated nanoscale constructions that perform a diverse range of biological tasks.

There are many types of molecular machines, which are traditionally thought to execute a designated function within the cell (Fig 1). Some are motor proteins such as kinesin and dynein [103, 5] that move cargo throughout the cell via transport along microtubule tracks, or large complexes such as flagella that propel cells forward using an ion gradient [18, 5]. Another molecular machine is the ATP synthase complex, which acts as a nanoscale turbine, harnessing the flow of protons into a mechanical rotation that catalyzes ATP production - the primary energy carrier of the cell [121, 5].

In addition, the ribosome represents another noteworthy molecular machine. This intricate assembly synthesizes proteins by mechanically “reading" an mRNA sequence and “writing" the amino acid to a polypeptide chain [195, 232, 5]. Remarkably, the ribosome also has a “proofreading" ability for error correction during synthesis [5, 178, 108],. This error correction functionality is not exclusive to the ribosome [47, 108, 178], and may occur in transporter proteins [23, 74], an essential topic of this dissertation

Then there are transporter proteins, primarily responsible for moving molecules across the cell membrane [5]. The subsequent sections will center on these proteins, examining their role within the context of cellular processes, and their complexity.

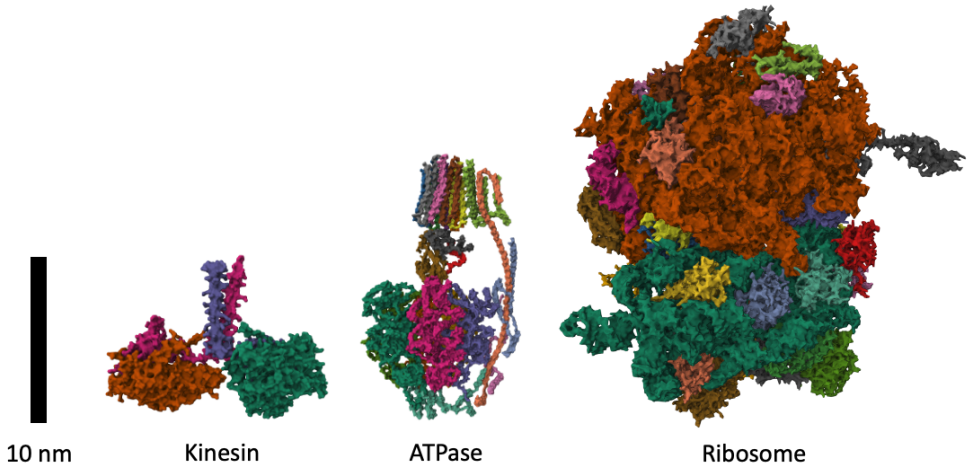


Figure 1: Selected Molecular Machine Structures Cartoon visualization of the 3D molecular structures for a kinesin subunit, ATPase complex, and ribosome complex, corresponding to IDs 3KIN, 5VOX, 4V4R from the protein data bank respectively [51].

An Introduction to Transporters Transporters are proteins that are embedded into the cell membrane, which acts as a semi-permeable barrier between the cell’s interior and environment [141, 5]. Small molecules, such as oxygen, can freely diffuse across the membrane, while other essential biochemical species, such as glucose, cannot. Transporters act as carriers that enable these substances to travel across the membrane. This is facilitated by their molecular structure, which consists of two primary components, the transmembrane domain and the binding sites [44, 5]. The transmembrane domain of the protein is anchored inside the cell membrane and typically contains multiple alpha helices that form a pathway for molecules to travel through. The binding sites of the transporter are the regions where the molecules attach themselves, inducing a change in the protein shape that enables transport through the transmembrane domain.

Transporter proteins are grouped into three categories based on how they utilize energy: passive, primary active, and secondary active [5, 19] (see Fig 2). Further, they may transport multiple different substances at a time, in either the same direction (symporter) or opposite directions (antiporter). Passive transporters do not need to expend energy but instead, leverage diffusion to facilitate the transfer of molecules or ions from high to low concentrations [5, 19]. They

essentially act as a gate, allowing specific materials to move down their concentration gradient. Notable passive transporters include glucose transporters (GLUTs), potassium channels, and aquaporins - channels that rapidly transport water.

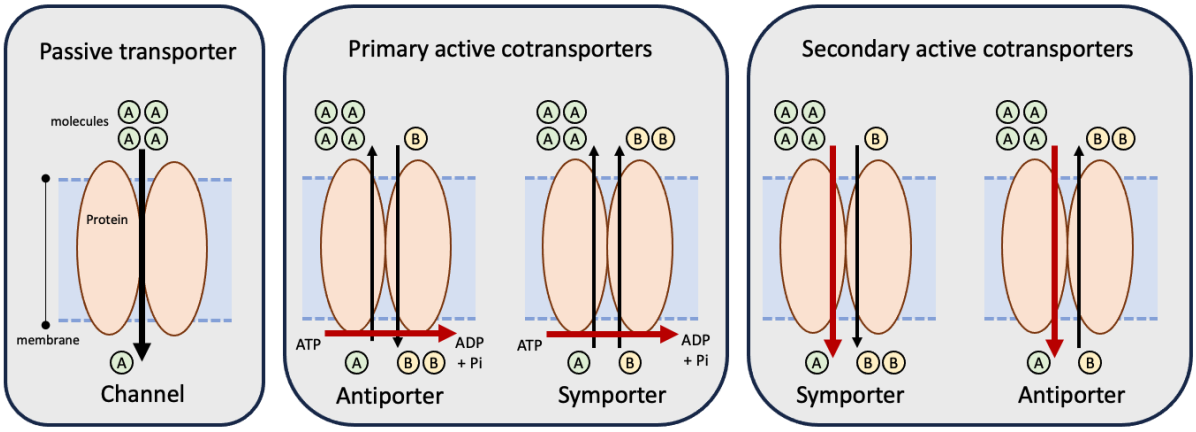


Figure 2: Transporter Proteins Schematic of passive, primary active, and secondary active transport processes.

In contrast to passive transporters, active transporters utilize energy to pump molecules against their concentration gradient. Primary active transporters generate energy from the breakdown of ATP (adenosine triphosphate) to ADP (adenosine diphosphate), which drives transport, acting as both a transporter and ATPase [121, 5]. Examples include the Na^+/K^+ ATPase, calcium ATPase, and proton pumps.

Secondary active transporters utilize energy, not by ATP hydrolysis, but through the energy stored in the concentration and voltage differences across the membrane - i.e., the electrochemical gradient maintained by primary active transporters [25, 5]. An important example of this interplay is the Na+/glucose transporter (SGLT). The Na^+/K^+ ATPase pumps sodium out of the cell and potassium into the cell, creating an electrochemical gradient that is used by the Na^+/glucose transporter to pump glucose into the cell, against its concentration gradient [25, 5] (see Fig 3). These secondary active transporters are the primary focus of this dissertation.

Transporters in a Cellular Context Living systems, particularly cells, regulate their internal conditions, such as temperature, pH, and substance concentrations, to ensure stability and adaptability in response to their environment [5, 40]. This homeostasis is largely facilitated by the activity of transporter proteins, which control the flow of materials between the cell and its environment[5, 40]. Transporters achieve this through their many different functions, including maintaining ion gradients, uptake of nutrients, signaling, and removal of waste [5, 40], which have considerable biomedical significance.

Transporters such as ion pumps and channels play a crucial role in the creation and maintenance of ion gradients - an essential part of cellular homeostasis. Ion gradients help maintain the electric potential across the membrane, regulate cell volume, and help drive secondary active transport processes [5, 40]. A notable example is the gastric H^+/K^+ ATPase which enables the acidification of the stomach by pumping protons into the stomach lumen, facilitating enzyme

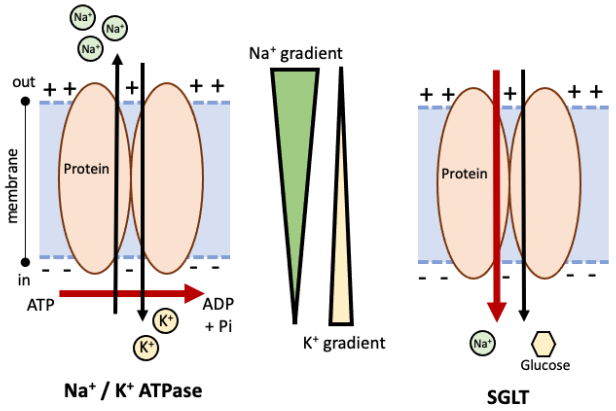


Figure 3: Coupled Transporters Diagram of the Na+/K+ ATPase, which maintains an ion gradient that is utilized by a sodium-glucose transporter to influx glucose.

activation and food digestion. These pumps are targets for inhibitors that reduce the acidity of the stomach in order to treat gastroesophageal reflux disease [205].

In addition, some transporters uptake essential nutrients such as amino acids to synthesize proteins or glucose which acts as a primary energy source for cellular processes [5, 40]. Glucose transporters (GLUTs), such as GLUT4 found in muscle and fat cells, transfer glucose from regions of high to low concentration via passive transport [109, 135, 28]. In response to the presence of insulin, GLUT4 is moved to the cell membrane where it facilitates the uptake of glucose, regulating blood glucose levels [109, 135, 28]. On the other hand, sodium-glucose transporters (SGLTs) [258, 49, 259] such as SGLT2 in the kidneys, actively carry glucose against its concentration gradient using the sodium ion gradient. This process reabsorbs glucose from filtrate in the kidneys which prevents it from being excreted. Both transporters help maintain glucose homeostasis, and, as such, are associated with diseases such as diabetes. A lowered response to insulin from GLUT4 results in hyperglycemia, while inhibitors targeting SGLT2 can reduce glucose re-absorption and effectively lower blood sugar levels [236, 120, 120].

Transporters also play an essential role in cell signaling that enables cells to sense and adapt to changes in their internal and external environment. This is achieved by moving signaling molecules such as hormones and neurotransmitters that allow the cell to send and receive signals and respond accordingly [5, 40]. For instance, the serotonin transporter (SERT) actively regulates the concentration of serotonin in the synaptic cleft by pumping serotonin back into the presynaptic neuron [157, 170]. This is essential to control the duration and intensity of the serotonin signal, which is associated with physiological processes such as mood regulation. As such, SERT is a target of selective serotonin reuptake inhibitors (SSRIs) used to increase the concentration of serotonin in the presynaptic neuron and thereby treat symptoms of depression [207, 157].

Finally, a critical function of transporters is removing waste products from the cell, which prevents a buildup of potentially toxic chemical byproducts such as urea. Urea transporters (UT) located in the kidney pump urea from the bloodstream that is eventually excreted in urine. Additionally, some transporters play a significant role in the export of

drugs from cells, such as ATP-binding cassette (ABC) transporters, multidrug resistance proteins (MRP), and small multidrug resistance proteins (SMR). These proteins, including P-glycoprotein (P-gp), MRP1, and the *Escherichia coli* multidrug resistance transporter (EmrE) are known to export a range of structurally different substances which helps the cell remove a variety of potentially harmful substances. However, this mechanism also leads to multidrug resistance, which is a challenge in cancer chemotherapy [199] as well as antibiotic and antimicrobial treatments [62, 115].

While not an exhaustive list of transporters and their functions, the above examples illustrate how the selective and regulated movement of substances via transporter proteins helps maintain cellular homeostasis and perform essential cellular functions. By better understanding the exact mechanisms of these machines, we can gain valuable knowledge of critical cellular and physiological processes, as well as targets for therapeutic intervention.

1.1.3 The Mechanistic Complexity of Transporters

Transporter Reaction Pathway Heterogeneity Transport occurs through a series of coupled biochemical reactions involving the binding of a substance, a conformational change in the protein's shape, and the unbinding of a substance. This reaction pathway (or cycle) completely describes the biophysical transport mechanism. Following the seminal work by Mitchell and Jardetzky [166, 116], transporters have traditionally been thought to follow a fixed, often simplistic, reaction pathway. A prime example is the alternating access model that moves substances across the membrane by changing between outward and inward-facing conformations. This is shown in Fig. 4.

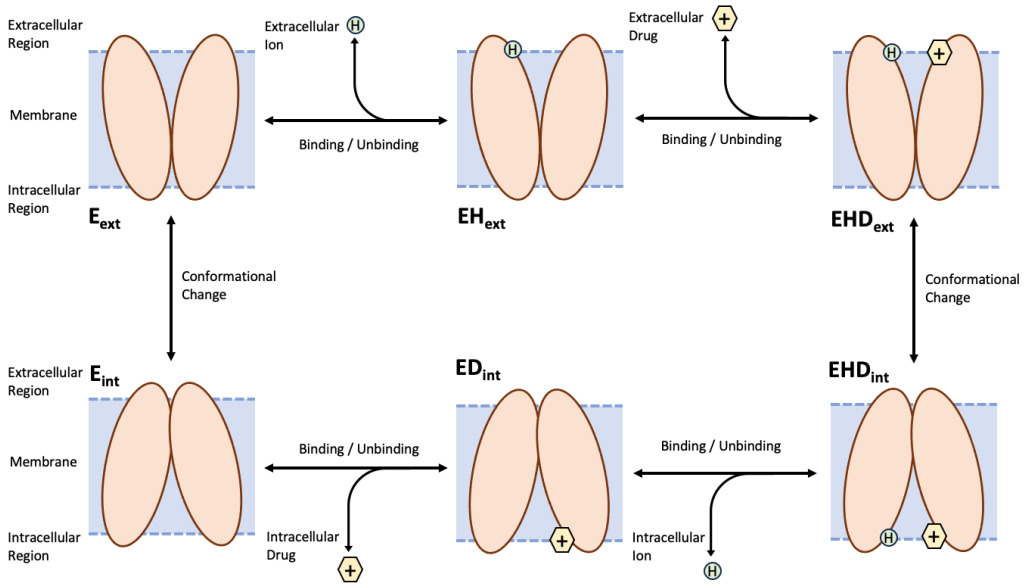


Figure 4: Ideal 1:1 Symporter Reaction Cycle This transport cycle moves an ion, (H) and a drug substrate (hexagon '+') from external to internal regions using a set of six elementary reactions. In the clockwise direction, the transporter is first unbound facing the extracellular region, binds to the ion, binds to the drug, changes conformation to face the intracellular region, unbinds the ion, unbinds the drug, then changes conformation to face the extracellular region. These correspond to the E_{ext}, EH_{ext}, EHD_{ext}, EHD_{int}, ED_{int}, and E_{int} states respectively.

However, there is a growing body of evidence that suggests [160, 14, 95] transporters may exhibit more complex behavior and even utilize multiple reaction pathways - i.e., pathway heterogeneity. For example, electrophysiological characterization of V-ATPases has shown variable coupling ratios of proton and ATP under different environmental conditions [170]. This suggests that additional reaction pathways are utilized that introduce an inefficient 'leak' of the ion depending on the environment (see Fig 5). This would result in both pump and channel mechanisms being available to the V-ATPase. Evidence for similar complexities that challenge the notion of strict integer transport stoichiometry (i.e., 1:1, 2:1) has been found for other transporters and channels such as GLUT and LacY [70, 49]. In addition, there is support that transporters may not have traditional fixed binding sites, such as shown in the leucine transporter LeuT [239]. All of these lines of evidence point toward mechanistic complexity in transporters.

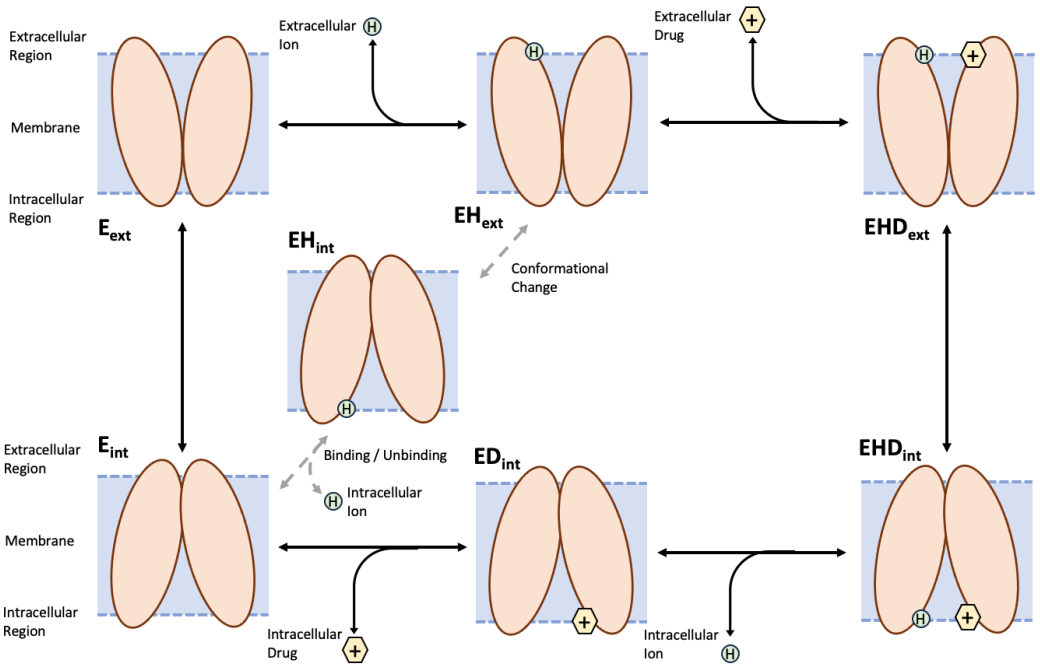


Figure 5: Non-ideal 1:1 Symporter Reaction Cycle with a Leak This transport cycle moves an ion (H) and a drug substrate (hexagon '+') from external to internal regions using a set of six elementary reactions, similar to the ideal transport cycle. However, this model includes an additional conformational state with the ion bound in the inward-facing conformation (EH_{int}). This allows the ion to be transported across the membrane without the drug substrate (grey dashed arrows) - effectively 'leaking' down its electrochemical gradient, similar to a passive transporter.

The complexities of transporters are exemplified in the E. Coli multidrug resistance transporter (EmrE), which motivates much of this dissertation. In-depth nuclear magnetic resonance and biophysical assays by Henzler-Wildman and coworkers [200] have revealed ten conformational states that allow for far more mechanistic diversity than traditional models. This includes pathways for 2:1 antiport, 1:1 symport, and various leak pathways, enabling the protein to use two opposite modes of transport under different external conditions. This new 'free exchange' transport model better accounts for the permissive efflux of drugs by allowing for substrates with multiple valence states (+1 and +2) and large variances in binding affinities[200].

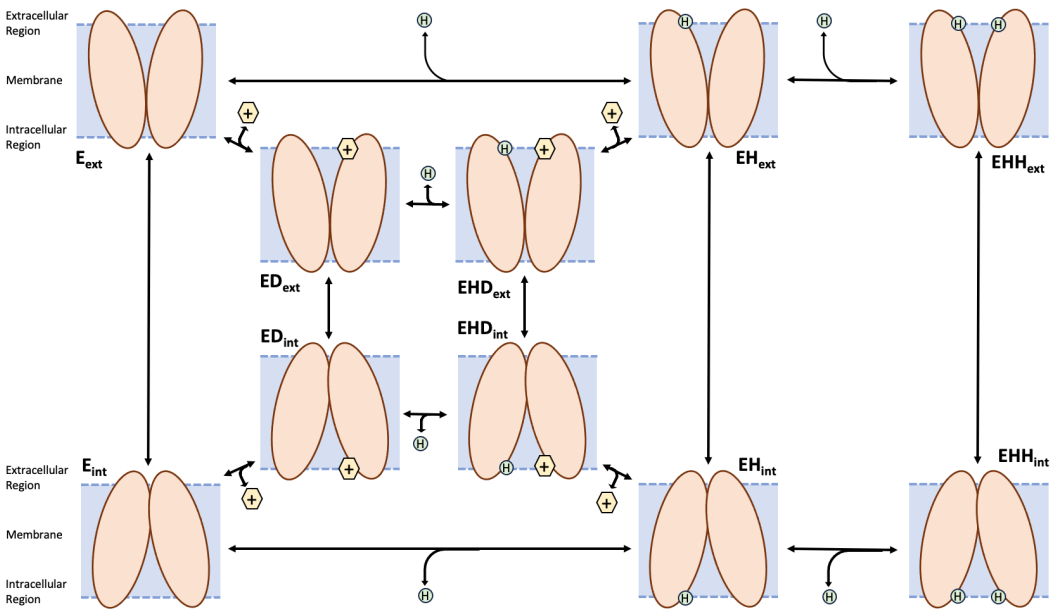


Figure 6: Free Exchange Model of EmrE An expanded transporter model of EmrE. The additional states and reactions enable various mechanistic pathways, including 2:1 antiport and 1:1 symport. Here the protein has two conformations (E_{ext}, E_{int}), can singly or doubly bind with a proton (EH_x, EHH_x), bind with a drug substrate (ED_x), and bind both a proton and drug substrate (EHD_x).

Challenges in Determining Mechanism Due to the dynamic nature of transporters and their varying time and length scales of transporters, it is difficult to characterize their mechanisms precisely. Especially under the paradigm of complex heterogeneous reaction pathways, determining the exact set of conformational states, reactions, and their governing kinetic parameters is exceptionally challenging. To address these challenges, several experimental techniques have been used to aid the study of transporters, each with its own unique strengths and weaknesses. In general, there are trade-offs among the methods between capturing high-resolution structural, and dynamical information, and the suitability for large complexes in a native environment.

X-ray crystallography uses X-ray diffraction of a crystallized protein to generate structural information at an atomic level, and is commonly used to determine key conformational states of membrane transporters [131, 223], such as LeuT, GLUT1, and the sodium-potassium pump [186, 50, 49]. However, this method does not capture dynamic information and can struggle with crystallizing large membrane proteins. A related method, cryo-electron microscopy (cryo-EM)[10, 215] uses electron beams to visualize protein structures in a frozen medium, at nanometer resolution. Unlike X-ray diffraction, cryo-EM can capture multiple conformations in the same sample, as well as examine large proteins in their native environment. However, like X-ray crystallography, dynamic information about the transporter mechanism is not captured.

Another experimental technique, fluorescence resonance energy transfer (FRET), tags specific protein regions with fluorophores, enabling the real-time investigation of conformation changes and dynamics in their native environment [161, 11]. However, this approach does not capture high-resolution spatial information about the conformational states of the studied transporter. Similarly, electrophysiology methods and various biochemical assays (such as patch-clamp

electrophysiology and radioactive assays) can capture dynamic information about the flux and uptake of ions or substrates due to transporters but don't provide direct information regarding the conformational states used during transport [70, 132, 88].

Finally, another important approach is nuclear magnetic resonance (NMR). Briefly, NMR utilizes the response of atomic nuclei to an external magnetic field that undergoes a perturbation from radio frequency pulses. Unlike the previously discussed methods, NMR can provide high-resolution structural and dynamic information about transporter proteins. However, NMR struggles with larger transporters in their native membrane environment. Nonetheless, NMR has successfully been used to probe the mechanistic complexity of small transporters, such as with EmrE [200, 94].

Experimental Method	**Captures Dynamics?**	**High Resolution Structures?**	**Native Environment?**	**Large Complexes?**
X-ray Crystallography	No	Yes	No	Limited
Cryo-Electron Microscopy	No	Yes	Yes	Yes
Fluorescence Resonance Energy Transfer (FRET)	Yes	No	Yes	Yes
Electrophysiology (Patch-clamp)	Yes	No	Yes	Yes
Biochemical Assays	Yes	No	Yes	Yes
Nuclear Magnetic Resonance (NMR)	Yes	Yes	Yes	Limited

Table 1: Overview of experimental methods used to study transporter proteins.

Alternatively, computational methods, including molecular dynamics [125], Markov models [87], and network models [2], can act as complementary approaches to help predict protein dynamics and their potential reaction pathways. These methods (as discussed in subsequent sections) also come with unique challenges, such as the high computational cost required to perform simulations and modeling, the sensitivity to initial conditions and model parameter choices, and the limitations of accurately modeling complex biological systems [179, 35]. So, even with advances in experimental and computational approaches, there is still an ongoing challenge to robustly and precisely characterize the mechanism of transporters - especially in light of the paradigm of heterogeneous (i.e. mixed) reaction pathways.

This dissertation aims to overcome these challenges by developing improved computational tools using network-based models alongside Bayesian inference and synthetic data motivated by electrophysiology experiments.

1.2 Introduction to Key Technical Concepts

This section aims to inform the reader of essential technical details for the subsequent chapters of the dissertation. It starts by explaining biochemical network modeling, then moves on to topics in Bayesian inference and analysis, followed by a discussion of Markov chain Monte Carlo. This section ends with additional strategies to calibrate predictive models from data and an explanation of an emerging electrophysiology experiment type (solid-supported membrane electrophysiology).

1.2.1 Biochemical Network Modeling

Overview Biochemical network models are a flexible and powerful tool to study biological systems, from metabolic networks to signal transduction [6]. These mathematical models describe the connections between various interconnected components via a graph network, with different biochemical states acting as a node of the graph network, and related components connected with an edge. This approach is flexible at multiple levels of detail, from large metabolic networks down to individual protein reactions. The primary advantage of this approach as compared to atomistic simulation methods such as molecular dynamics, is the ability to efficiently examine the behavior of large interconnected systems and their emergent properties across long time scales [6] - with the loss of exact structural detail provided from molecular dynamics.

These network models are often coupled with a representation of the system's dynamics via differential equations. These equations describe how the concentrations of the associated biochemical species change over time. Each equation corresponds to a system state such as a protein or metabolite, and the net difference between the rate of creation and degradation of that protein or metabolite determines the rate of change. These creation and degradation rates are often modeled with a mass action approach, where the rate of a reaction is directly proportional to the concentration of the reactants [34]. This framework assumes well-mixed chemical species with purely random interactions between the reactants. Alternative rate formulations for biochemical rates may be used, such as Michaelis-Menten [65] or Hill kinetics [41, 6], which result in a hyperbolic and sigmoidal dependence on reactant concentrations, respectively.

An illustrative example of a simple 3-state reaction network motivated by the popular Michaelis-Menton model [231] is shown in figure 7. Here an enzyme (E) binds with a substrate (S), forming an enzyme-substrate complex (ES). The enzyme-substrate complex catalyzes the conversion of the substrate into a product, forming an enzyme-product complex (EP). Finally, the product (P) unbinds from the enzyme (E). This is described with the following chemical reactions:

$$E + S \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} ES$$

$$ES \underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}} EP$$

$$EP \underset{k_{13}}{\overset{k_{31}}{\rightleftharpoons}} E + P$$

The change in each biochemical species concentration over time is shown with the following set of ordinary differential equations:

$$\begin{aligned}
\frac{d[E]}{dt} &= k_{21}[ES] + k_{13}[EP] - k_{12}[E][S] - k_{31}[E][P] \\
\frac{d[S]}{dt} &= k_{21}[ES] - k_{12}[E][S] \\
\frac{d[ES]}{dt} &= k_{12}[E][S] + k_{32}[EP] - k_{21}[ES] - k_{23}[ES] \\
\frac{d[EP]}{dt} &= k_{23}[ES] + k_{31}[E][P] - k_{32}[EP] - k_{13}[EP] \\
\frac{d[P]}{dt} &= k_{31}[E][P] - k_{13}[EP]
\end{aligned}$$

As noted above, the change in the biochemical species is equal to the difference between the formation rate of that species minus the degradation rate. For example, the rate of change of the product concentration, $\frac{d[P]}{dt}$, is equal to the creation rate of [P] from $k_{31}[E][P]$ balanced against the destruction rate of [P] from $k_{13}[EP]$.

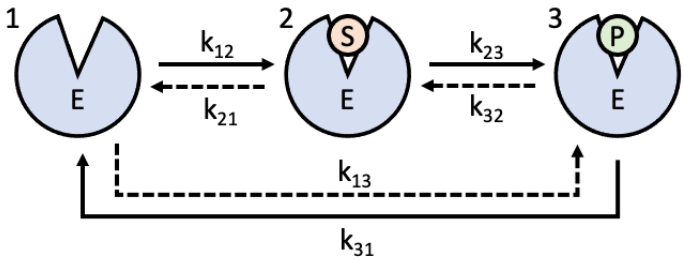


Figure 7: 3 State Network Model An example network describing the process of an enzyme (E) converting a substrate (S) into a product (P), and their governing reaction rate constants (k_x). This process is reversible as noted by the dashed arrows. Here the binding and unbinding events have been shown implicitly for visual clarity.

An important note is that cycles (or loops) within a biochemical network model provide an additional constraint derived from thermodynamic principles. At equilibrium, there is a detailed balance that requires the forward and reverse rates of each reaction to be equal - resulting in no net energy usage along the cycle [101]. This implies that one of the governing rate constants is not independent and can be defined via the remaining rate constants.

Using the above model to illustrate, at equilibrium, the forward and reverse rates of each reaction step are equal:

$$\begin{aligned}
k_{12}[E]^{\text{eq}}[S]^{\text{eq}} &= k_{21}[ES]^{\text{eq}} \\
k_{23}[ES]^{\text{eq}} &= k_{32}[EP]^{\text{eq}} \\
k_{31}[EP]^{\text{eq}} &= k_{13}[E]^{\text{eq}}[P]^{\text{eq}}
\end{aligned}$$

Solving for the equilibrium ratios of $\frac{[ES]^{\text{eq}}}{[EP]^{\text{eq}}}$ and substituting yields:

$$\frac{[ES]^{\text{eq}}}{[EP]^{\text{eq}}} = \frac{k_{12}}{k_{31}} \cdot \frac{k_{21}}{k_{13}} \cdot \frac{[P]^{\text{eq}}}{[S]^{\text{eq}}}$$
$$\frac{[ES]^{\text{eq}}}{[EP]^{\text{eq}}} = \frac{k_{32}}{k_{23}}$$

After further substitution, the expression of the rates along the cycle is given by the equation below:

$$\frac{k_{32}}{k_{23}} = \frac{k_{12}k_{31}[S]^{\text{eq}}}{k_{21}k_{13}[P]^{\text{eq}}}$$

As shown in the above example, one of the rate constants is not independent and is defined by the remaining rate constants - this relation holds true even outside of equilibrium [101]. This constraint is essential for cyclical network models, such as transport cycles, ensuring physical consistency within the model and reducing model complexity, as demonstrated later in this dissertation.

Application to Transporters In the context of mechanistic models of transporters, biochemical reaction networks describe each individual transporter reaction state and reaction, such as the different conformations and binding and unbinding states. Below (Fig. 8) is an illustrative example of a single sodium-glucose transporter cycle motivated by the SGLT family of transporters, which couples the downhill sodium gradient from the Na^+/K^+ ATPase with the transport of glucose uphill. Here, the network consists of 6 conformational states of the protein, representing the different protein orientations (inward-facing OF, and outward-facing IF) and binding conditions for the sodium ion and glucose. Within this network, there are several reaction cycles possible, resulting in different potential reaction mechanisms. For simplicity, only a single transport cycle is shown rather than all possible transport cycles available in the given reaction network. The governing biochemical reactions and their differential equations form are shown below. As before, the rate constants k_{ij} represent the transition rate constants between states i and j, with 'Glu' representing glucose, and 'Na^+' representing sodium ions.

Transport reactions:

$$OF + Na^+_{\text{out}} \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} OF \cdot Na^+$$
$$OF \cdot Na^+ + Glu_{\text{out}} \underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}} OF \cdot Na^+ \cdot Glu$$
$$OF \cdot Na^+ \cdot Glu \underset{k_{43}}{\overset{k_{34}}{\rightleftharpoons}} IF \cdot Na^+ \cdot Glu$$
$$IF \cdot Na^+ \cdot Glu \underset{k_{54}}{\overset{k_{45}}{\rightleftharpoons}} IF \cdot Glu + Na^+_{\text{in}}$$
$$IF \cdot Glu \underset{k_{65}}{\overset{k_{56}}{\rightleftharpoons}} IF + Glu_{\text{in}}$$
$$IF \underset{k_{16}}{\overset{k_{61}}{\rightleftharpoons}} OF$$

Transport differential equations:

$$\frac{d[OF]}{dt} = k_{16}[IF] - k_{61}[OF] + k_{21}[OFNa^+] - k_{12}[OF][Na^+_{\text{out}}]$$
$$\frac{d[OFNa^+]}{dt} = k_{12}[OF][Na^+_{\text{out}}] - k_{21}[OFNa^+] + k_{32}[OFNa^+Glu] - k_{23}[OFNa^+][Glu_{\text{out}}]$$
$$\frac{d[OFNa^+Glu]}{dt} = k_{23}[OFNa^+][Glu_{\text{out}}] - k_{32}[OFNa^+Glu] + k_{43}[IFNa^+Glu] - k_{34}[OFNa^+Glu]$$
$$\frac{d[IFNa^+Glu]}{dt} = k_{34}[OFNa^+Glu] - k_{43}[IFNa^+Glu] + k_{54}[IFGlu][Na^+_{\text{in}}] - k_{45}[IFNa^+Glu]$$
$$\frac{d[IFGlu]}{dt} = k_{45}[IFNa^+Glu] - k_{54}[IFGlu][Na^+_{\text{in}}] + k_{65}[IF] - k_{56}[IFGlu]$$
$$\frac{d[IF]}{dt} = k_{56}[IFGlu] - k_{65}[IF] + k_{61}[OF] - k_{16}[IF]$$
$$\frac{d[Na^+_{\text{out}}]}{dt} = k_{21}[OFNa^+] - k_{12}[OF][Na^+_{\text{out}}]$$
$$\frac{d[Glu_{\text{out}}]}{dt} = k_{32}[OFNa^+Glu] - k_{23}[OFNa^+][Glu_{\text{out}}]$$
$$\frac{d[Na^+_{\text{in}}]}{dt} = k_{45}[IFNa^+Glu] - k_{54}[IFGlu][Na^+_{\text{in}}]$$
$$\frac{d[Glu_{\text{in}}]}{dt} = k_{56}[IFGlu] - k_{65}[IF]$$

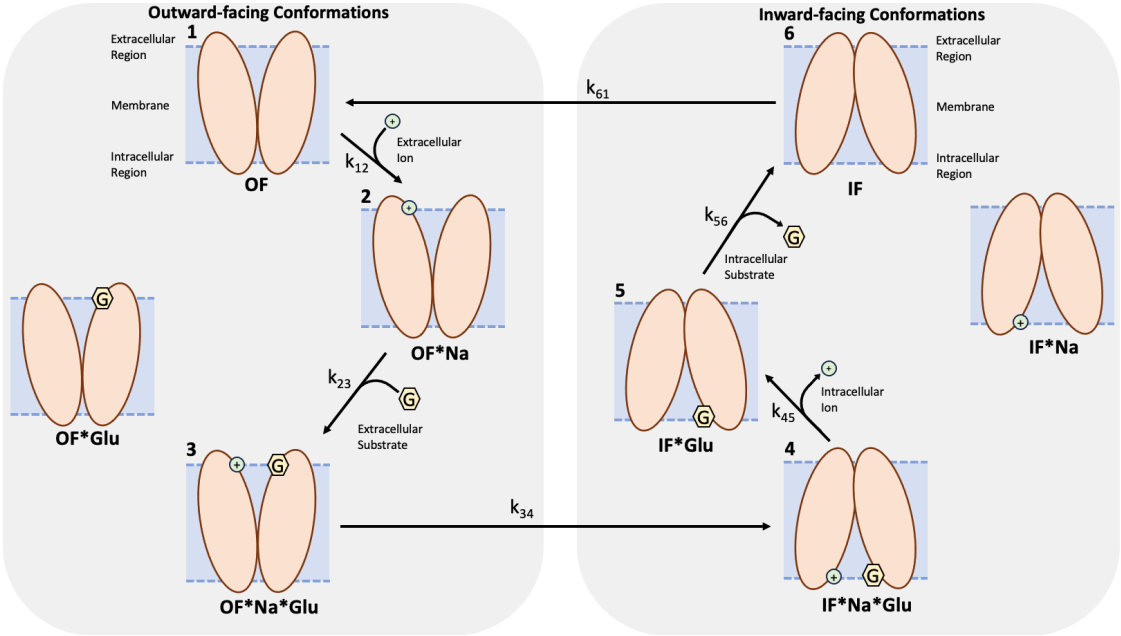


Figure 8: Network and Cycle of SGLT-type Transporters An example network describing the transport mechanism of an SGLT-type transporter. A single reaction cycle in the nominal forward direction is shown as an illustration, but within this set of conformational states, alternative pathways are possible.

The ability to model compartmentalization is crucial for transporter proteins embedded in the cell membrane, separating the extracellular and intracellular regions. Typically either compartment volume terms are added to the differential equations for the relevant species, or the species can be separated explicitly. Either approach ensures the

correct species concentrations in each region, but care is needed to ensure consistent units [106]. In the above example, we explicitly separate the sodium ion and glucose into internal and external concentrations using a X_{in} and X_{out} nomenclature, respectively, which are separate from the bound states, e.g., $OFNa^+$ is separate from Na^+_{out} and Na^+_{in}.

Methods

Given a network model of a system and its associated system of reaction equations, simulations can be performed in various ways. If the number of biochemical species (e.g., transporters and substance concentrations) is low and random fluctuations become significant, stochastic simulations may be used. A common approach is to use the Gillespie algorithm [77] that utilizes a probability distribution for all possible reactions, sampling from the distribution to pick which reaction will occur next and when, and then updating the system. This approach is computationally expensive but generates accurate descriptions of the inherent randomness of microscopic systems.

As an alternative, differential equations may be constructed from the reaction equations, as shown previously. These ordinary differential equations (ODEs) are solved deterministically through numerical integration [163]. Here, the initial concentrations of each species and the governing reaction rate constants are given. The equations are then solved at progressive time points, resulting in the concentrations of each chemical species as a function of time. This method assumes a well-mixed solution of chemical species. It is therefore appropriate when the concentrations are large enough such that the stochastic effects of the molecules interacting can be ignored.

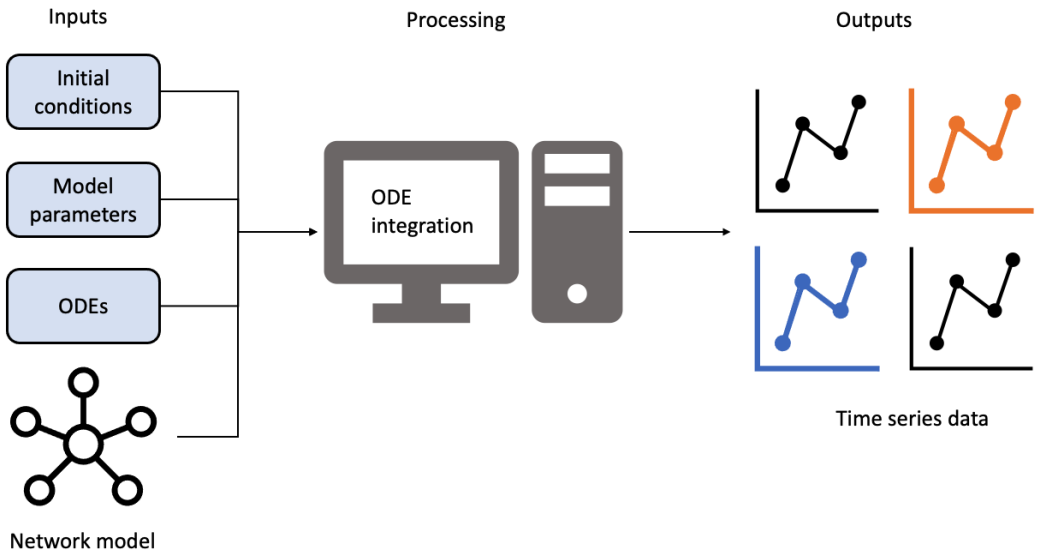


Figure 9: Diagram of Network Simulation Workflow for Ordinary Differential Equations (ODEs) A cartoon visualization of the typical pipeline used to simulate a network using ODEs. Given an initial condition, model parameter values, and the network and ODE model, various numerical methods can be used to integrate the equations. This results in time series data for the different chemical species studied - such as the change in substrate concentration over time.

Integrating differential equations for biochemical networks can prove challenging due to their large size, non-linearity, and varying time scales. In particular, these equations are often known to be 'stiff' [238, 163], causing numerical inconsistencies and errors due to the significant difference in the equation coefficients (i.e. rate constants). Various

algorithmic strategies have been developed to mitigate these issues, utilizing backward differentiation formulas (BDF) and adaptive step-size solvers [39]. Software tools such as Scipy [248], MATLAB [99], DifferentialEquations.jl [194], serve as popular general-purpose methods to integrate differential equations. Similarly, specific tools have been developed for use in systems biology that enable both stochastic and deterministic modeling, including Libroadrunner/Tellurium [230, 37] and COPASI [107]. This dissertation primarily utilizes ordinary differential equation-based modeling and simulation.

Biochemical networks can also be simulated to provide insight into complex biological interactions with spatial resolution. Partial differential equations may be used when spatial heterogeneity is essential to the model. These expensive simulations use similar equations to ordinary differential equations but include additional spatial variables to model how species concentrations change in time and space - with one example being the reaction-diffusion equation [229].

The above approaches are often called 'forward problems' in that they are a class of computational problems that generate a prediction given initial conditions and parameters governing the model. These techniques are valuable in that they generate new hypotheses to test experimentally, support existing experimental data, aid in the design of systems, and are often much quicker and cheaper than experimental techniques. There is another class of computational problems, 'inverse problems', which aim to estimate model parameters and inputs based on data. These techniques are essential to this dissertation and will be discussed later.

Finally, due to the complex nature of these networks, efforts have been made to improve the robustness and reproducibility of these computational methods in the greater systems biology field, although they are not universally adopted. In particular, the Systems Biology Markup Language (SBML) [127, 111], is a crucial standard that enables simulation and analysis of models across different methods and platforms. Similarly, tools to systematically build networks using rules-based approaches (i.e., BioNetGen, NFSim)[60, 227] have been introduced to reduce human error in generating networks that are combinatorically complex. Finally, more human-readable forms of SBML have been developed, such as Antimony [226], to allow for more accessible construction and validation of network models.

1.2.2 Bayesian Inference

Introduction to Bayesian Inference

Bayesian inference [73] is a powerful statistical method used in many fields to compute probabilities utilizing prior knowledge. It is based on Bayes' Theorem, which can be expressed as:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{1}$$

Here each term is defined:

- $P(H|E)$ is the posterior, the probability of the hypothesis H being true given the evidence E.
- $P(E|H)$ is the likelihood, the probability of the evidence E given that hypothesis H is true.

- $P(H)$ is the prior, the probability of hypothesis H being true before having the evidence E.
- $P(E)$ is the marginal likelihood or evidence, the total probability of observing the evidence E.

At its core, Bayesian inference is about learning from evidence. Consider a hypothesis 'H' that one is trying to test, of which you have a prior belief about the likelihood of this hypothesis being true,' P(H)'. After collecting evidence, you can estimate how likely that hypothesis is given the evidence, the posterior. Bayes' theorem shows that the posterior is proportional to your prior beliefs and the likelihood of observing the evidence given your hypothesis is true, $P(E|H)P(H)$. So if you observe new evidence that is very unlikely under the hypothesis, the belief in the hypothesis should decrease. Similarly, if the evidence is very likely under the hypothesis, the belief in the hypothesis should increase.

As a simple example, consider a medical test for a rare disease. Since the disease is rare, the prior belief that a patient has the disease is very low. If the patient tests positive for the disease, even with imperfect tests, this is strong evidence that the person has the disease. This is because a person is much more likely to test positive if you have the disease than if you don't. As a result of the test result, the patient's belief in having the disease significantly increases.

This example showcases the ability of Bayesian inference to update probabilities based on new data and observations, as well as the use of prior beliefs. This can be extended to more general data analysis methods and for analyzing biochemical networks as shown below - a key methodology used in this dissertation.

Bayesian Data Analysis Bayesian inference is a robust method for fitting models to a data set - generating estimates of model parameters and their uncertainty that best fit the data. [73]

Consider a simple linear model where y depends on x :

$$y = ax + b + \epsilon \tag{2}$$

where a is the slope, b is the intercept, and ϵ is an error term, such as noise generated from an experimental apparatus. The goal is to determine the values of a, b, and ϵ given the data D. This can be done by reformulating the problem into Bayes' theorem:

$$P(a, b, \epsilon|D) = \frac{P(D|a, b, \epsilon)P(a, b, \epsilon)}{P(D)} \tag{3}$$

Here the data 'D' serves a similar role as the evidence 'E' in the previous example, and similarly, the model parameters a,b, and ϵ act as the hypothesis 'H'. With a choice (or estimate) of prior and likelihood distribution, this formulation gives the probability of the slope, intercept, and error terms given the data.

An important note is that the evidence term P(D), is a normalization factor that ensures the posterior is a true probability density, summing the probability of the likelihood and prior across the entire parameter space:

$$P(D) = \iiint P(D|a, b, \epsilon)P(a, b, \epsilon), da, db, d\epsilon \tag{4}$$

Since this term scales with the number of dimensions (i.e. parameters), it has important implications for the practical application of Bayesian inference, as discussed in the next section. A key difference between Bayesian inference and other 'frequentist' methods besides using a prior, is that the estimated parameters are treated as random variables. The resulting posterior distribution, which is often multidimensional (e.g. $P(a, b, \epsilon|D)$) can be marginalized to only consider a single variable. This amounts to summing over the other parameter values, for example, $P(a|D) = \int \int P(a, b, \epsilon|D) dbd\epsilon$. In other words, instead of generating a single best parameter estimate, Bayesian inference returns a *distribution* of parameter estimates (see fig. 10).

This means that the posterior distribution contains information regarding parameter means, modes, and variances from the 1D marginal distributions and their correlations and covariances within the multidimensional posterior distributions, detailing with high fidelity how model parameters interrelate. The rich information in the posterior is a primary motivation for its use in this dissertation. **Bayesian Inference for Biochemical Networks** As noted in the previous

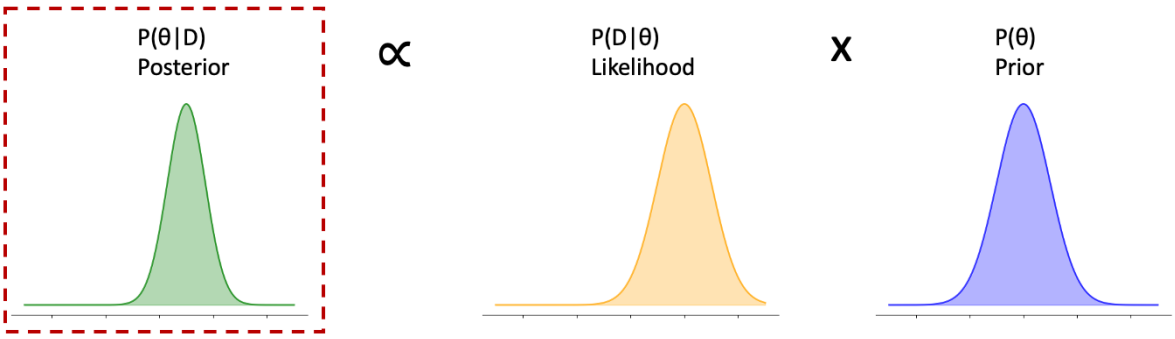


Figure 10: Visual Representation of Bayes' Theorem An illustration showcasing key elements of Bayesian inference: updating beliefs based on new knowledge, incorporating prior beliefs, and modeling parameters θ as random variables. The posterior (left) is proportional to the likelihood (middle) times the prior (right). Here the highly informative prior noticeably shifts the posterior from the likelihood, introducing bias from prior knowledge.

section, Bayes' theorem can be applied to estimate model parameters from data. This is particularly useful in systems biology contexts in which there are key biophysical parameters that may not be precisely known. For biochemical reaction networks defined with ordinary differential equations, the reaction rate constants (k's) control the speed of reactions governing the system's dynamics [98], and are integral to understanding the underlying mechanisms. As such, precise estimation of these parameters and their uncertainties are needed for more accurate predictions as well as increased biological knowledge.

Bayesian inference has been applied to various biological systems to estimate the parameters of computational models [73]. Still, it remains an active area of research due to the significant computational cost associated with medium to large-scale networks. In practice, Bayes' theorem cannot be solved directly, as the posterior contains a multidimensional integral (P(D)) that is prohibitively expensive to calculate as the dimension of the system increases.

To address this issue, many alternative approaches have been developed to estimate the posterior distribution, each with its own trade-offs. Markov chain Monte Carlo (MCMC)[117, 92]are related methods that are often used to estimate the

posterior. These methods provide a high level of accuracy with a high computational cost and were used extensively in this dissertation. MCMC methods are described in greater detail in a later section. Other approaches include variational inference [24, 80, 52] which fits an approximate distribution (i.e. Gaussian) to the posterior distribution. This approach requires a careful choice of distributions for accurate results and tends to underestimate the posterior variance. Similarly, approximate Bayesian computation [16, 152, 234] simulates data and compares it to the observed data, rejecting parameter sets that don't closely fit the data. These methods may be helpful when the likelihood function is costly to calculate, but data is easy to simulate. Both variational inference and approximate Bayesian methods trade accuracy for improved performance, which may be necessary for large complex biological systems.

In addition to these computational issues, Bayesian inference requires the choice of a prior, which may introduce bias into your estimates. If the prior distribution is too informative it will strongly bias the posterior (see fig 10). Still, if the prior is too uninformative it may be challenging to efficiently estimate the posterior due to the large uninformative problem space.

In conclusion, Bayesian inference is a powerful technique to reckon with uncertainty and update predictions based on new data - generating a multidimensional distribution covering the model parameters based on the data. In the context of membrane transporters, this can help unlock more robust estimates of parameters governing transporter mechanisms, such as reaction rate constants. However, Bayesian inference is computationally challenging due to the size and complexity of most problems of interest, spurring the development and use of numerical methods such as MCMC to estimate the posterior.

1.2.3 Markov Chain Monte Carlo (MCMC) and Variants

An Introduction to Markov Models A Markov model [192] describes how a system changes over time using a probabilistic approach and is commonly used in many computational fields to model the behavior of stochastic systems. For example, biochemical networks are generally Markovian, with reaction rate constants acting as unnormalized transition probabilities [192]. In a Markov model, the system is broken into different states that are connected by a transition probability. At any given time, the next state is determined only by the current state of the system and its associated transition probabilities.

Consider a system occupying two discrete states, such as a protein with two different conformations. An abstract representation of this system is shown in the diagram below, with associated transition probabilities for each state:

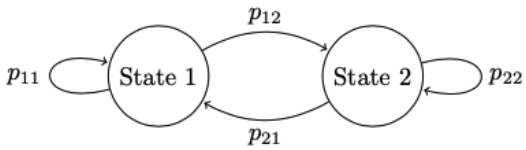


Figure 11: Two State Markov Model A Markov model consisting of two states, and the p_{ij} transition probabilities from state i to state j.

Once the transition probabilities are determined, the system can be simulated at successive time points, generating a trajectory of the system's state over time, as shown below using $p_{11} = 0.7$, $p_{12} = 0.3$, $p_{21} = 0.4$, and $p_{22} = 0.6$. After many timesteps, the system will reach an equilibrium - allowing for determining average quantities such as the rate of conformational change in a protein [84].

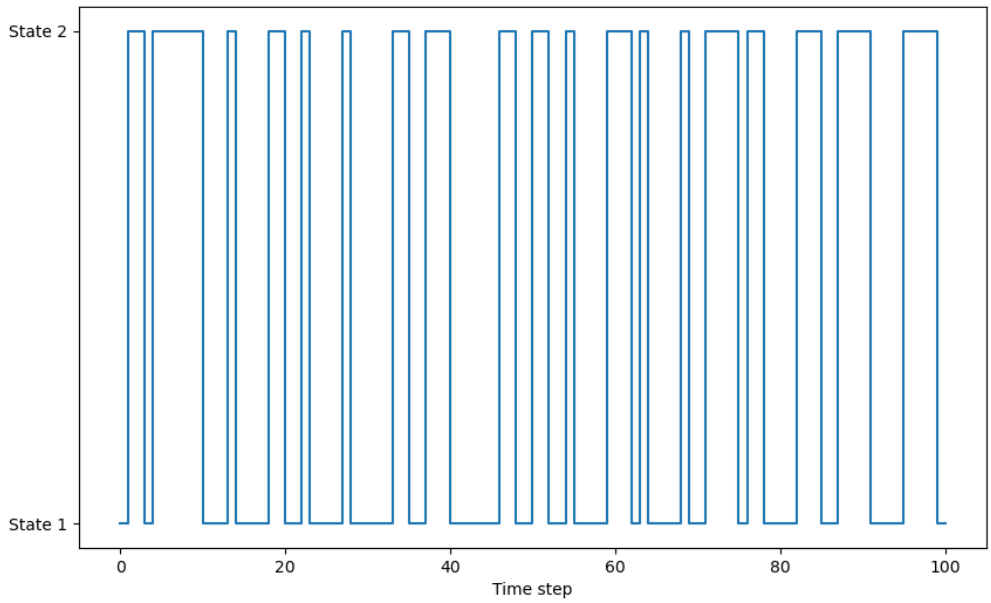


Figure 12: Two State Markov Model Trajectory A trajectory of Markov model states at different time points. Over many time steps, average behavior can be observed.

It is important to note that this trajectory of Markov states at different iterations is also called a Markov chain, the fundamental object used in Markov chain Monte Carlo sampling methods.

An Introduction to Markov chain Monte Carlo

In Markov chain Monte Carlo sampling [92, 117], the Markov model is abstracted from representing the different states of a physical system to representing different values for a random variable or sets of random variables. In a Bayesian inference and parameter estimation context, this corresponds to model parameters, such as reaction rate constants. This methodology was used extensively for this dissertation to explore the space of transporter reaction mechanisms and estimate reaction rates from data via Bayesian inference.

Markov chain Monte Carlo uses elements of Monte Carlo, random sampling to estimate a quantity of interest, with a Markov chain where new samples are generated based on the current sample (i.e. the Markov property). At each iteration, a new sample is proposed based on a proposal distribution centered on the current sample position. This new proposal is either accepted or rejected based on acceptance criteria, such as the ratio of probabilities of the new and old sample position [92]. The chain of samples forms a random walk through the sample space, and after many iterations is guaranteed to visit all the non-zero probability regions of the sample space and therefore estimate the target distribution - such as the posterior distribution. The adaptive random proposal approach contrasts with the fixed method used in

traditional Monte Carlo, which is inefficient for complex multidimensional distributions in which the probability covers a small region of the sample space [203, 42].

Figure 13 (shown below) illustrates the Markov chain Monte Carlo method in practice. Here the posterior distribution of a 1D problem $P(\theta|D)$ is approximated using MCMC, with the Markov chain shown on the left, and the true and estimated distribution (from the Markov chain values) shown on the right.

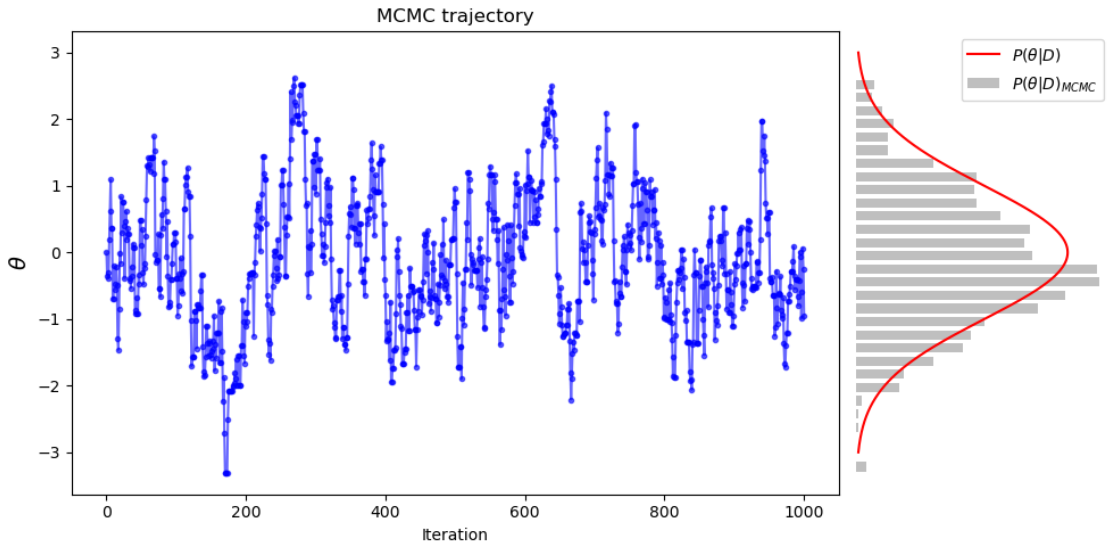


Figure 13: Markov Chain Monte Carlo Trajectory A 1D posterior distribution $P(\theta|D)$ is approximated using Markov chain Monte Carlo. The left panel shows parameter values over several iterations forming a chain of θ (parameter) values. This chain can be used to generate a histogram of parameter values $P(\theta|D)_{MCMC}$ that approximates the true posterior $P(\theta|D)$

A seminal method for Markov chain Monte Carlo is the Metropolis-Hastings algorithm [92], which is described below:

1. Start with an initial parameter guess, $\theta^{(0)}$.
2. Generate a new candidate parameter value θ^* through a random proposal centered at the current parameter value $q(\theta^*|\theta^{(t)})$.
3. Evaluate the probability of this candidate parameter value $p(\theta^*)$
4. Decide whether to keep this candidate parameter value or not. This is done by computing the ratio of the likelihood of the proposed candidate state compared to the current state, given by $r = \min\left(1, \frac{p(\theta^*)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)})q(\theta^*|\theta^{(t)})}\right)$. If the candidate has a higher likelihood than the current state, keep that candidate and add it to the Markov chain. If the candidate has a lower likelihood, then there is still a chance to accept the new parameter value based on the probability r. If a candidate is not accepted, then the current parameter value is stored for that iteration of the chain
5. Repeat steps 2-4 for the desired number of steps

In the next section, we will discuss the challenges facing the use of MCMC to estimate distributions.

Challenges in MCMC Methods: Alternatives and Insights

While the Metropolis-Hastings (MH) algorithm has been used successfully in many problems, it faces significant challenges in estimating the complex multidimensional distributions typically found in systems biology problems. MH requires careful tuning of the proposal distribution. If the proposal distribution is too narrow, it will take a very long time to explore the parameter space. If the proposal distribution is too wide, new proposals may not be accepted often. Similarly, since samples in MCMC are chosen based on previous values, they may have strong correlations, resulting in a low number of 'effective samples' that are independent, requiring significantly more iterations to converge [42, 203].

This is related to the issue of "trapping" and complex distribution geometries, which can be illustrated using a physical interpretation of the MCMC sampling process. In this analogy, the probability distribution is related to a hypothetical "energy landscape" (not true physical energy) by a Boltzmann factor: $p(x) \propto e^{-\beta U(x)}$ where β is an inverse temperature term and U(x) is the energy term. Here a high probability corresponds to a low energy basin, and low probabilities correspond to a high energy peak. In this context, trapping occurs when the samples are in a low-energy state and unable to climb out due to the high-energy barrier. As a result, the samples will not explore the neighboring region, so the full distribution will not be sampled. This is illustrated in figure 14.

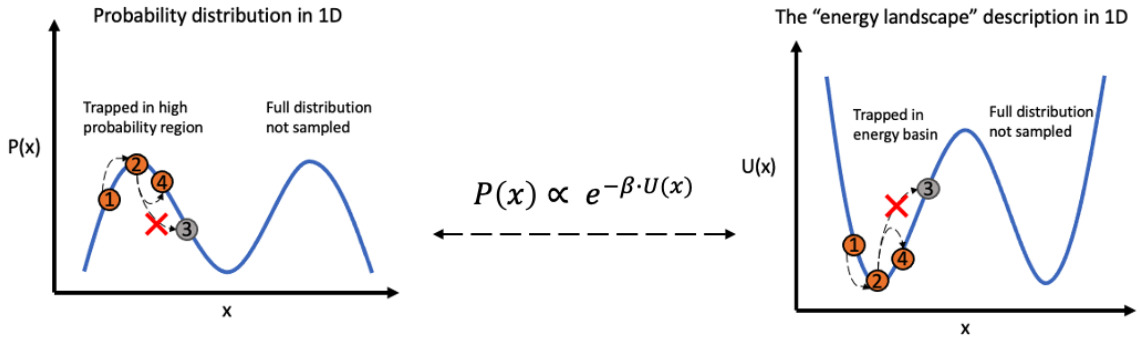


Figure 14: An Energy Perspective for MCMC Sampling A 1D parameter distribution is shown on the left, with the corresponding energy landscape on the right. Due to an inadequate proposal distribution, the high energy barrier traps samples in the leftmost basin, unable to sample the neighboring region.

As the complexity of the distribution increases, these sampling issues become more significant, effectively preventing convergence to the target distribution in feasible time scales. This is particularly relevant in biochemical network models with large dimensions and highly correlated parameters. As a result of the aforementioned challenges, many different methods have been developed to address the poor performance of MH, including:

1. **Hamiltonian Monte Carlo methods** These methods incorporate additional information about the shape (i.e. derivative) of the sampled distribution, using a physics-informed 'momentum' term to move samples along the surface of the sampled distribution, improving the efficiency compared to MH. However, this method requires careful tuning of additional hyper-parameters and access to gradient information which may be challenging to compute[20, 105]

2. **Sequential Monte Carlo methods** These methods sample a sequence of distributions from an easier-to-sample distribution to the (complex) target distribution. This is done via multiple independent chains (i.e. walkers or particles) which can capture complex multi-modal distributions but requires many chains to adequately cover the sampling space [53, 38].

3. **Annealed importance sampling** Is a special case of the sequential Monte Carlo method that follows a specific “cooling" schedule from high temperature (i.e. uniform distribution) to low temperature (i.e. the target distribution). This is modeled by $p_\beta(x) = p_0(x)^{1-\beta}p(x)^\beta$, where $p_\beta(x)$ is the probability at a given inverse temperature β and p_0 is the initial (high temperature) distribution sampled. Adjusting β changes the mixture of $p_0(x)$ and $p(x)$ distributions used for sampling. [174]

4. **Parallel Tempering** This method runs multiple Markov chains in parallel at different “temperatures", with high-temperature chains able to jump across energy barriers and avoid trapping. The low and high-temperature chains periodically swap which enables further exploration of the target distribution. This approach is sensitive to the number of parallel chains with poor scaling for large dimensional systems [57].

5. **Affine invariant ensemble methods** This method uses a collection of Markov chains (i.e. walkers) that explore the sampling space and leverage the position of their neighbors to make new step proposals. This approach is well suited for ill-shaped and skewed distributions but requires many walkers and tuning of other hyper-parameters [79].

6. **Preconditioned Monte Carlo methods** This method simplifies the target distribution into a simpler distribution that is easier to sample by other means, such as Metropolis-Hastings. It requires tuning of hyper-parameters to get an appropriate simplification of the target distribution [122].

Both Metropolis-Hastings and Hamiltonian Monte Carlo samplers are implemented in most general-purpose probabilistic modeling tools, including Stan, PyMC, and Turing.jl, [32, 72, 184]. Sequential Monte Carlo and annealed importance sampling are implemented in PyMC, and variants of parallel tempering and affine invariant methods are implemented in the emcee package. More niche algorithms usually are packaged in separate tools such as the preconditioned Monte Carlo that is available in pocoMC [123].

Poor sampler convergence was *the* major challenge faced throughout this dissertation. Over the course of the study, variations of all of the above methods (and some novel methods developed for this work) were trialed with limited success until using a preconditioned Monte Carlo sampler. This implementation used normalizing flows [183] to learn a transformation from the complex target distribution to a simple Gaussian distribution via an autoencoder neural network architecture [182, 75]. In addition, this implementation utilized an annealed importance sampler to sequentially sample from the simple to the target distribution, updating the neural network during each stage. Essentially, this approach transforms the target distribution (i.e. posterior) into the same Gaussian distribution used to generate proposals - dramatically reducing the issues arising from the more complex distribution shapes typical of systems biology problems.

These methods generally require expert knowledge for tuning based on the system being studied, and as such it is recommended that robust diagnostic measures are used such as comparing multiple independent sampling runs, measuring the auto-correlation time, and visually inspecting the MCMC trajectories and their estimated distributions. Doing so will help ensure reliable parameter estimates and more accurate predictions.

1.2.4 Model Calibration, Selection, and Experimental Optimization

Introduction to Model Calibration

Model calibration is the process of determining the parameters of a given model based on observations and data - also known as parameter estimation or the “inverse problem" of computational modeling. Here the goal is to reliably estimate model parameters to make more accurate predictions with the model, as well as insights into the model structure.

Bayesian inference, as previously discussed, is one method for model calibration, but there are other approaches, such as least squares and maximum likelihood estimation (MLE) [73]. MLE aims to find the single set of parameters that best fit the data given the model. This is done by maximizing the likelihood function $L(D|\theta)$ through various numerical optimization methods. For example, global optimization methods such as differential evolution [233] and particle swarm optimization [130], can be used to find the set of parameters that gives the highest likelihood value. These approaches are generally more efficient than Bayesian approaches, however, they do not necessarily generate information about parameter uncertainty, which must be done with alternative methods such as bootstrapping [58] or profile likelihood calculations [144]. Also, similarly to Markov chain Monte Carlo methods, optimization strategies may struggle to converge for complex, non-linear, and high dimensional likelihood functions that are typical in biological models.

Nonetheless, maximum likelihood estimation techniques form an essential benchmark for model calibration studies. In this dissertation, the development of a robust model calibration pipeline is informed by both Bayesian inference and maximum likelihood estimation methods. In this way, the validity of the transporter reaction rate parameter estimates and resulting mechanistic insights are strengthened by the inclusion of multiple calibration methods.

Parameter Identifiability, Information Quantification, and Model Selection

Related to the estimation of model parameters and their variance is the determination of parameter identifiability and sensitivity. Here identifiability describes the ability to precisely determine the model parameters, either based on the underlying structure of the differential equation model or from a practical perspective based on the variance of the parameter estimates [255]. For the purposes of this dissertation, practical identifiability of model parameters is considered - where estimated parameters that have an extremely large variance are considered practically non-identifiable. This consideration is helpful because it can reduce the number of free parameters by fixing the unidentifiable parameters. Similarly, practically unidentifiable parameters can serve as targets for future experiments, which may reduce their estimated variance through more informative data. The practical identifiability of parameters is

a primary motivation of the model calibration methods used in this dissertation - to have all reaction rate constants identifiable, enabling the precise understanding of the reaction rate constants and their respective reaction pathways used by transporters.

Similarly, the variance in parameter estimates (and therefore practical identifiability) directly relates to the data quality used. As noted earlier, the Bayesian posterior distribution provides a rich source of information about parameter variances and co-variances. The information contained in the posterior can be quantified by comparing the posterior to the prior, measuring how much information was gained based on the divergence of the distributions [146, 110]. By quantifying the information content in the datasets, it is possible to examine which experimental protocols generate the highest quality data. This can be achieved by generating a set of optimal protocols as priori that maximizes the information gain [33, 245, 162] via Bayesian experiment recommendation. By recommending more informative experiments in an automated way, this methodology can help uncover the complex nature of transporter reaction pathways.

A simple illustration of the quantification of information in data and the resulting distributions is shown below in figure 15. Here an informative dataset results in a posterior distribution that significantly overlaps with the uniform prior - signifying a negligible amount of information was introduced in the data. Similarly, an informative dataset is shown that results in a posterior that is noticeably different than the prior, implying a significant amount of information in the data.

Finally, the process of model calibration depends on the choice of model, but often in biological systems, the exact model is unknown. In the context of transporter research, this corresponds to different potential reaction pathway models, or mixtures of pathways, as indicated by the paradigm of pathway heterogeneity. As such, the comparison and selection of competing mechanistic models is a core goal of this dissertation. Some typical strategies include using the Bayes factor to calculate the odds of one model against another [136, 152], to compare the likelihood distributions of the different models, or to compute an information criterion that balances model fitting against the number of parameters [73, 176]. More sophisticated approaches like hierarchical Bayesian model selection introduce a comprehensive framework to examine models and mixtures of models but come at a more significant computational cost. This dissertation primarily uses likelihood comparison methods readily calculated as part of the Bayesian inference process.

1.2.5 Solid-Supported Membrane Electrophysiology (SSME)

Introduction to SSME and Modeling Considerations Solid-supported membrane-based electrophysiology (SSME) is an emerging form of electrophysiology. Traditional electrophysiological techniques such as patch clamping [206] and voltage clamping [104] have performed an essential role in many important discoveries related to neuroscience and physiology. Notable examples include elucidating the kinetics of ion channels [102] and neuronal transmission [104]. As such traditional electrophysiology methods have a long-standing history within the transporter research community [246]. They provide dynamic information regarding the transport of ions across a membrane which can be used to estimate specific kinetic parameters such as net transport rates. Despite their wide use, these methods have limitations

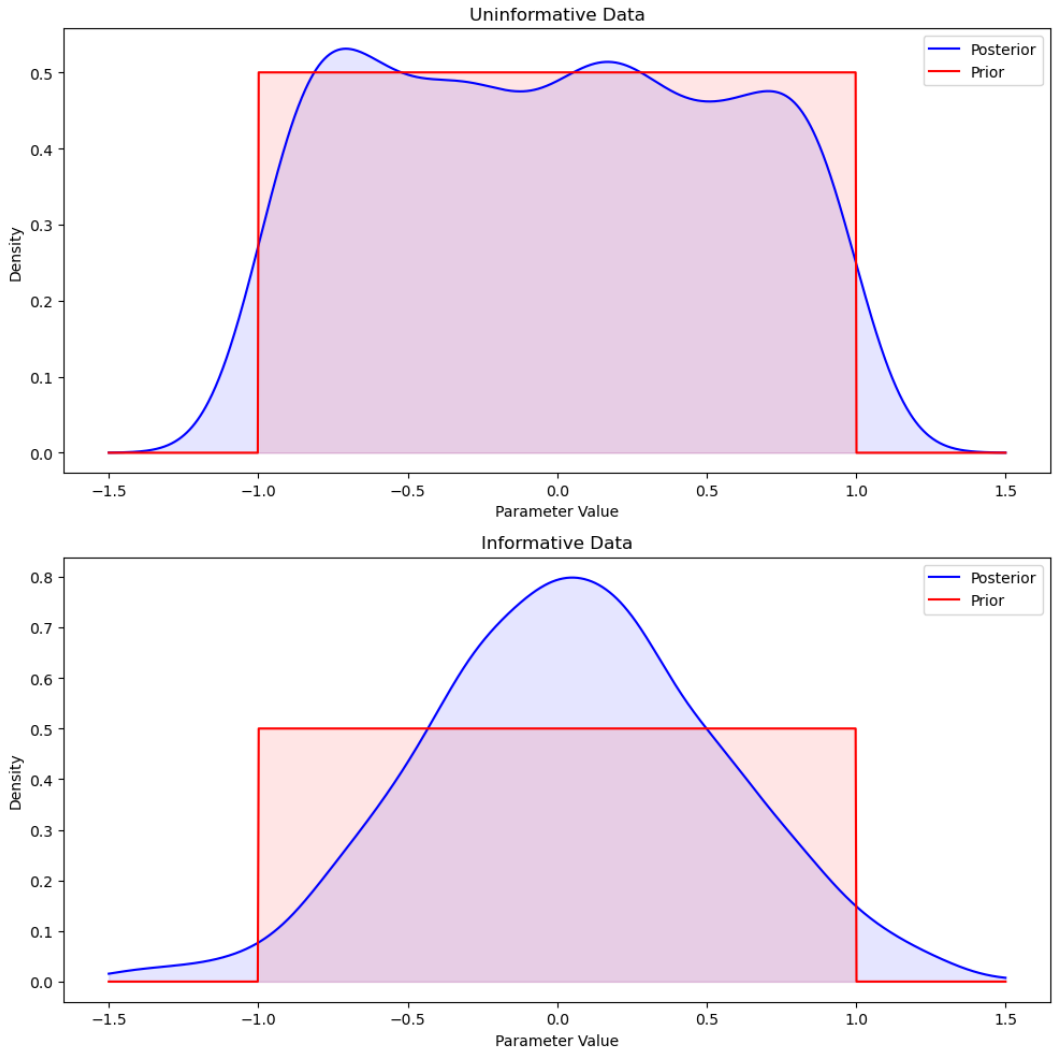


Figure 15: Bayesian Information Quantification Visualization of the quantification of data using the divergence between the Bayesian posterior and prior. In the top panel, a low-information dataset is used, resulting in significant overlap between the prior and posterior. In the bottom panel, a high-information dataset is used, resulting in a significant divergence between the prior and posterior. The Kullback–Leibler divergence of these probability distributions provides a metric to recommend experimental protocols that yield higher information content. Note that this is a simplified diagram for illustrative purposes, in practice, probability should not extend past the prior range.

such as low throughput, technical difficulty, and relatively low spatial resolution [12]. SSME provides a robust alternative method to measure ion transport [12, 13].

In SSME, transporter proteins are reconstituted into proteoliposomes and placed onto a solid-supporting membrane [12, 13]. The liposomes and solid-supporting membrane are bath in a solution of the chemical species transported via the protein of interest. Once a steady-state is reached, the batch solution is perturbed to induce a gradient that drives the transport of the ions. The resulting current across the liposome membrane is capacitively coupled to the solid-supporting membrane, detected from an external sensor. The observed current is transient and results from a large population of transporters. Recent assay developments motivated by EmrE have emerged that include a reversal perturbation [[240]] which improves the richness of the data. Below in figure 16 a simplified cartoon diagram of an SSME experimental setup and an idealized trace is shown.

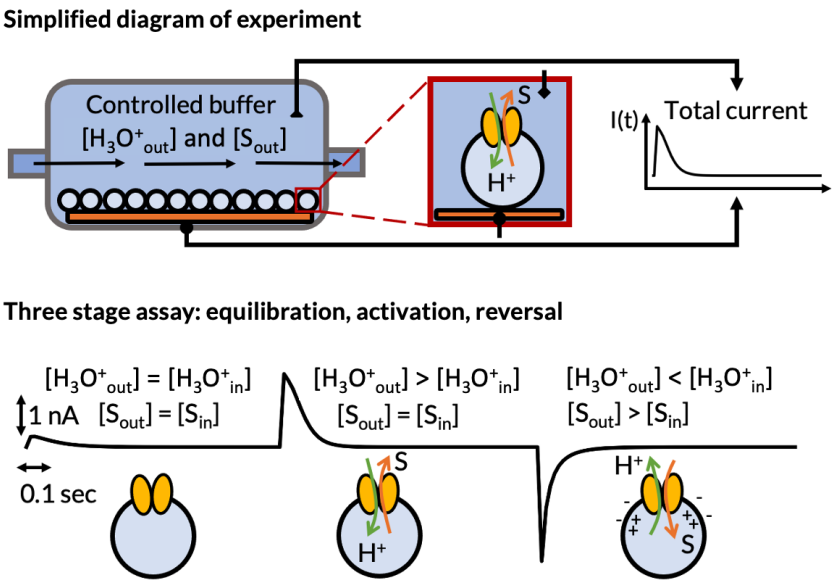


Figure 16: Diagram of SSME Experiments Cartoon representation of an SSME experiment. The top section illustrates the essential apparatus and working conditions of the experiment. Here multiple liposomes with embedded transporters are adsorbed to the solid membrane. An external concentration of ion and substrate species is perturbed, creating a gradient detected via capacitive coupling of the solid-supported membrane to the liposome membrane. The resulting transient current is the aggregate of all the transporters. In the bottom section, an ideal three-stage assay is presented, with an equilibration, activation, and reversal state. During the activation and reversal stage, the gradient directions switch causing transport to occur in opposite directions.

Compared to traditional electrophysiology methods, SSME has a better signal-to-noise ratio, improved stability, better time resolution, and higher throughput[12]. However, SSME also has notable drawbacks related to the reconstitution of proteins, their orientation, deposition, and uniformity of liposomes, and the potential mechanical effects from fluid exchange [12]. Despite these challenges, electrophysiology has been used to study a wide range of transporter proteins and channels.

Analyzing the data generated by these experiments is an ongoing challenge. While experiments can be performed under different sequences of perturbation concentrations, the current is only transient and relatively brief. Depending on the perturbation amount, these simple time series curves have an exponential-like relaxation and a moderate signal-to-noise.

It is unknown what the best strategies are to analyze these datasets and how much microscopic information about transporters can be extracted. These questions were a primary motivation to develop a robust Bayesian inference pipeline for estimating microscopic reaction rate parameters and their uncertainty. Given the sparse nature of the data relative to the number of microscopic rate constants, this task is significantly challenging, requiring a detailed computational model to simulate the necessary data for parameter estimation accurately.

In summary, SSME provides an alternative approach to capture dynamic information about transporters. However, it is unknown how much microscopic insight can be gained from this type of experiment that captures net aggregate transporter measurements. This dissertation addresses this issue directly by implementing a detailed synthetic SSME model and Bayesian inference methods, enabling a deeper understanding of these experiments and their limits on detailed mechanistic information.

1.3 Overview of the Dissertation and Contributions

Reiterating the goals of this study, the original research contributions of this dissertation are focused on the following areas:

- Developing robust computational tools for modeling and analysis of transporter proteins.
- Investigating mechanistic heterogeneity of transporters in silico.

This work is broken down into three key research chapters, followed by a discussion of the results and their implications.

1.3.1 Chapter 2: Exploring Mechanistic Heterogeneity and Kinetic Proofreading of Transporter Proteins

What new mechanisms are possible in the full conformational space of transporters?

Can transporter mechanisms be engineered in silico that enable enhanced selectivity in a competitive environment?

This chapter addresses these questions by developing a computational tool that systematically constructs and explores the space of reactions for transporters. Applying this to study an SGLT-like transporter, new classes of theoretical mechanisms exhibited extreme discrimination against similar (decoy) substrates, providing a kinetic scheme for enhanced selectivity.

This work has broader implications in improved computational modeling for systems biology, design of more effective therapeutics targeting transporters, insight into transporters' evolutionary process, and biotechnology applications.

1.3.2 Chapter 3: Constructing a Robust Pipeline for Calibrating Mechanistic Transporter Models

Can reaction rate constants be estimated from synthetic data based on solid supporting membrane-based electrophysiology (SSME)?

What parameter estimation strategies are the most robust for these data types and models?

This chapter addresses these questions by implementing a pipeline that uses Bayesian inference and maximum likelihood to estimate rate constants from in silico SSME experiments. Few reaction rate constants could be precisely determined for the data and models studied. Bayesian inference using Preconditioned Monte Carlo was the most robust strategy for parameter estimation.

This work has broader implications for improved accuracy of predictive biological network models, biological software development, and advances in parameter estimation methods.

1.3.3 Chapter 4: Model Selection and Experiment Recommendation for Transporters

How much information is contained in SSME-like data sets?

Can a mechanistic pathway be selected from competing mechanistic models?

This chapter addresses these questions by applying the Bayesian inference pipeline developed in the previous section. The information content of several synthetic experiments was quantified, and multiple mechanistic models were compared for selection. This resulted in a ranking (recommendation) of experiments to perform that yield the most information and the accurate selection of a mechanistic model from similar competing models.

This work has broader implications in experiment recommendation, investigation of pathway heterogeneity in biological systems, and data-driven decision-making.

1.3.4 Additional Contributions

In addition to the above research, efforts were made to the development of additional computational methods aimed at better understanding protein complexity:

Development of Bayesian Inference Pipeline for Isothermal Titration Calorimetry (ITC) Motivated by the difficulty in determining the mechanisms of disordered proteins, a Bayesian inference pipeline was developed for use with Isothermal Titration Calorimetry through external collaboration. This method leveraged a more complete model of ITC experimental uncertainty and provided more robust parameter estimates than previous Bayesian methods. By applying this approach to empirical studies of the intrinsically disordered protein LC8, new insights into its cooperative mechanism were gained [59]. This work has broader implications in the research of disordered proteins, which are involved in numerous biological processes and diseases [241].

Development and Survey of Sampling Methods Motivated by challenges in estimating parameters for biochemical network models from noisy and sparse data sets, existing Markov chain Monte Carlo (and related) methods were surveyed, and new approaches were developed. Numerous algorithms, including nested sampling, Metropolis-Hastings, no-u-turn Hamiltonian MCMC, annealed importance sampling, sequential Monte Carlo, sequential Monte Carlo with data tempering, affine invariant ensemble, and affine invariant ensemble with parallel tempering, were found to have poor performance compared to the preconditioned Monte Carlo implementation of PocoMC. In addition, new methods

were developed and found to have more unsatisfactory performance than preconditioned Monte Carlo, including an adaptive grid-based sampler, top-heavy annealed importance sampler, and parallel affine invariant ensemble sampler (see Appendix). This work has implications for future parameter estimation methodological development for systems biology.

1.3.5 Conclusion and Future Directions

This final chapter synthesizes the key results from the previous chapters, considering their implications for understanding transport proteins, the computational methods used, and their potential broader impact and applications. Also, this chapter examines the limitations of the dissertation findings and describes future research areas in the study of transporter mechanisms and biological systems at large. This chapter emphasizes the iterative process of scientific investigation, with new discoveries leading to new questions and avenues for further research.

2 Exploring Mechanistic Heterogeneity and Kinetic Proofreading of Transporter Proteins

Note: This work has already been published. Here we have made further adjustments for improved clarity, denoted with [[double square brackets]].

A systems-biology approach to molecular machines: Exploration of alternative transporter mechanisms

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

1 Dept. of Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA

2 Dept. of Pharm. Chem., University of California San Francisco, San Francisco, CA, USA

3 Dept. of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA

* zuckermd@ohsu.edu

2.1 Abstract

Motivated by growing evidence for pathway heterogeneity and alternative functions of molecular machines, we demonstrate a computational approach for investigating two questions: (1) Are there multiple mechanisms (state-space pathways) by which a machine can perform a given function, such as cotransport across a membrane? (2) How can additional functionality, such as proofreading/error correction, be built into machine function using standard biochemical processes? Answers to these questions will aid both the understanding of molecular-scale cell biology and the design of synthetic machines. Focusing on transport in this initial study, we sample a variety of mechanisms by employing the Metropolis Markov chain Monte Carlo. Trial moves adjust transition rates among an automatically generated set of conformational and binding states while maintaining fidelity to thermodynamic principles and a user-supplied fitness/functionality goal. Each accepted move generates a new model. The simulations yield both single and mixed reaction pathways for cotransport in a simple environment with a single substrate along with a driving ion. In a "competitive" environment including an additional decoy substrate, several qualitatively distinct reaction pathways are found which are capable of extremely high discrimination coupled to a leak of the driving ion, akin to proofreading. The array of functional models would be difficult to find by intuition alone in the complex state spaces of interest.

2.2 Author summary

Molecular machines, which operate on the nanoscale, are proteins/complexes that perform remarkable tasks such as the selective absorption of nutrients into the cell by transporters. These complex machines are often described using a fairly simple set of states and transitions that may not account for the stochasticity and heterogeneity generally expected at the nanoscale at body temperature. New tools are needed to study the full array of possibilities. This study presents a novel in silico method to systematically generate testable molecular-machine kinetic models and explore alternative mechanisms, applied first to membrane transport proteins. Our initial results suggest these transport machines may

contain mechanisms that 'detoxify' the cell of an unwanted toxin, as well as significantly discriminate against the import of the toxin. This novel approach should aid the experimental study of key physiological processes such as renal glucose re-absorption, rational drug design, and potentially the development of synthetic machines.

2.3 Introduction

The proteins and protein complexes known as molecular machines perform essential functions in the cell, including transport, locomotion, energy production, and gene expression [5]. Secondary active transporters, the focus of the present study, move ions and small molecules across a membrane driven by an electrochemical gradient of an ion [5]. For example, sodium-glucose transporters (SGLT) are of biomedical interest due to the vital role that SGLT1 and SGLT2 play in the uptake of glucose in the small intestines and reabsorption in the kidneys, respectively [189], which in turn has prompted biophysical scrutiny of their mechanisms [48, 49, 260, 22]. Numerous other transporters have also been assayed on a quantitative basis [160, 14, 95, 132, 200, 222].

The biological mechanisms of transporters as well as other molecular machines can be modeled using chemical reaction networks, typically along with mass action kinetics [101]. In a chemical reaction network, the system process is decomposed into discrete states connected by transition rates between states [101] forming a network of interconnected reactions in the state-space (Fig. 17). 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 [210]. Biochemical networks are generally Markovian [84], have a number of different control patterns [69], and typically adhere to specific design principles [153].

Despite the complex state-spaces accessible to molecular machines such as transporters, their mechanisms are often described using single-pathway, highly machine-like cartoon-like models [5, 228, 19], building on the seminal suggestions of Mitchell [166] and Jardetzky [116] (see Fig. 17). While such models are helpful for a qualitative understanding of complex protein behavior and chemical networks, simple models 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' cartoon models [160, 14]. Recent experimental studies have shown that certain traditional model assumptions such as fixed stoichiometry [95, 132], homogeneous pathways [200], and unique binding sites [222] may be incorrect.

As an example of mechanistic alternatives within a simple state-space, consider a hypothetical cotransporter motivated by the SGLT symporter which transports a single substrate and is driven by an ion gradient. The state-space is constructed using three state 'dimensions': conformational state, ion binding state, and substrate binding state. For this hypothetical transporter there are two conformations, each permitting four ion/substrate binding states: fully unbound, ion bound, substrate bound, and fully bound. Within this relatively simple state-space, we can construct four ideal kinetic pathways (Fig. 17) that connect the minimum number of states to produce a symport cycle (i.e., intracellular transport of the substrate coupled to ion flow). However, there are numerous additional mechanistic possibilities:

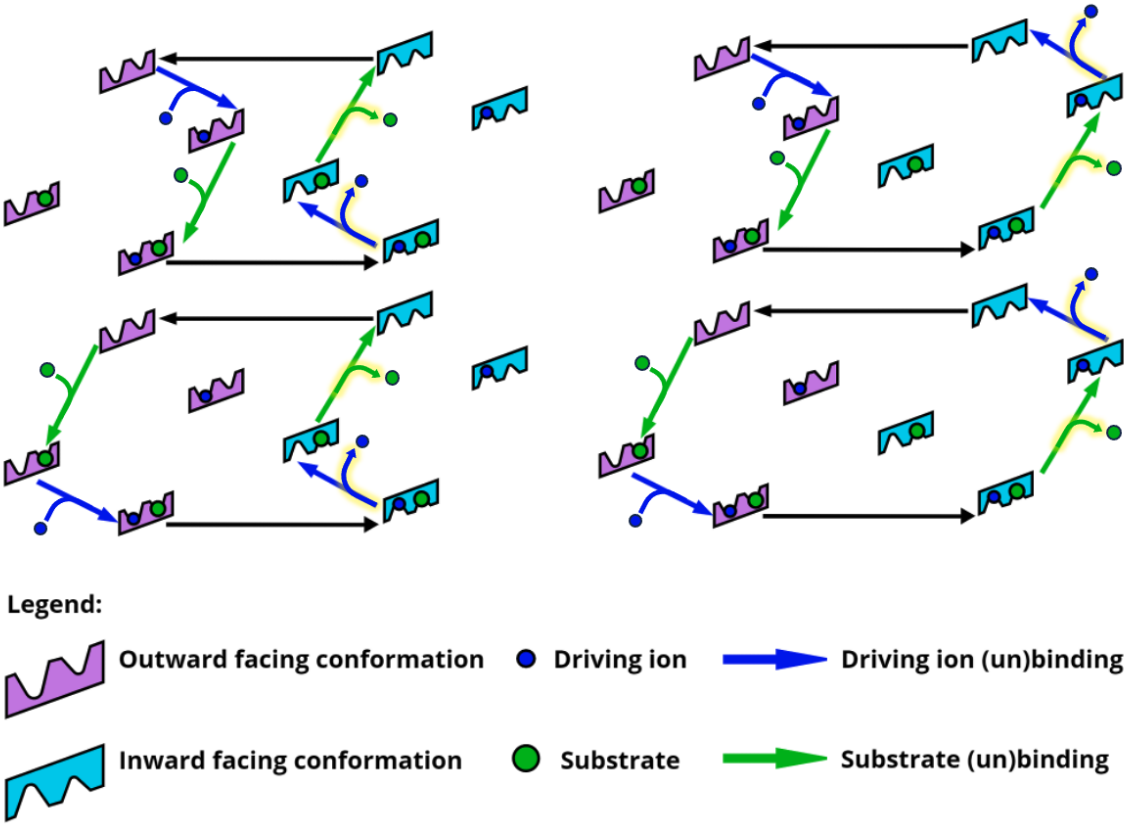


Figure 17: Multiple Mechanisms for Ideal Symport. The four "ideal" kinetic pathways of a hypothetical symporter that transports substrate using the available free energy of the driving ion are shown. This state-space contains eight states, and symport models include at least six connecting transitions. [[The ion and substrate binding and dissociation events are shown explicitly here, and the highlighted edges correspond to the pathways used to calculated the net fluxes for the substrate and ion.]]

combinations of the ideal pathways, or even non-ideal pathways including, e.g., an ion leak. Note that antiport cycles can be similarly constructed using this same state-space [265].

Beyond the nominal functions of transporters, we also take note of one of the most remarkable properties of some molecular machines: the ability to perform "proofreading" or error-correction [108, 178, 61]. Specifically, certain network topologies promote enhanced selectivity (i.e., reduced error) in systems with a competing substrate [108, 178, 61]; this enhancement in selectivity incurs a free energy cost, typically paid via hydrolysis of a phosphodiester bond. While some aspects of proofreading networks have been examined – such as the speed, accuracy, and dissipation trade-offs [9, 159], as well as non-equilibrium proofreading regimes [172] – the possibility that *transporters* might exhibit proofreading has not been explored to our knowledge.

Here we pursue a systematic exploration of mechanistic and functional diversity, building on strategies developed largely within the field of systems biology. Due to the challenges of modeling complex biochemical networks, such as enumerating combinatorically large state-spaces, new approaches have been developed to keep these systems tractable. Genetic algorithm sampling has been used to evolve complex biochemical networks such as metabolic pathways [45, 82].

In related work applied to ion channels, models have been fit to experimental data using genetic algorithms and simulated annealing [91, 164, 237]. These reverse-engineering studies primarily sought optimal individual models, not the model-sets we pursue here, and also did not account for non-equilibrium constraints on transition rates [101].

Motivated by the possibility of discovering new potential mechanisms and the limitations of current "manual" approaches for analyzing molecular machines, we systematically explore the biochemical network model space for a given molecular machine or function. 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 proofreading driven by an ion leak.

2.4 Methods

We have developed a custom software prototype, ModelExplorer, to study molecular machine behavior. Although this study is focused primarily on membrane cotransporters, 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 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.

2.4.1 Model specification

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 S1 Text). The Monte Carlo sampling generates a trajectory in model space (Fig. 18 and see below) 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. Only elementary transitions are allowed, i.e., single (un)binding transitions or single conformational changes. To ensure thermodynamic consistency among all rate constants [101], we use an energy-based formulation [180, 30, 216] where a free energy value is assigned to each state and transition state. To simplify equations, we use reduced units, with all energies expressed in units of $k_B T$. For a conformational transition from state $i \rightarrow j$, the Arrhenius-like first-order rate constant is expressed using Hill's notation [101] as:

$$\alpha_{ij} = k_0 \, e^{-(E_{ij}^{\text{bar}} - E_i)} \tag{5}$$

The user-defined prefactor k_0 is set arbitrarily to $10^{-3}\ s^{-1}$, while E_{ij}^{bar} corresponds to the transition state (free) energy and $E_i < E_{ij}^{\text{bar}}$ 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, but rather it influences the overall rates as a scaling factor.

For a process $i \to j$ involving binding or unbinding, the ion or substrate concentration is built into an effective first-order rate constant [101] given by:

$$\alpha_{ij} = e^{\Delta\mu_{ij}} k_0\, e^{-(E_{ij}^{\text{bar}} - E_i)} \tag{6}$$

with the same parameters as above, and with $\Delta\mu_{ij}$ being the non-equilibrium difference in the chemical potential for the $i \to j$ state transition:

$$\Delta\mu_{ij} = \begin{cases} \frac{-\Delta\mu_x}{2}, & \text{if the } i \to j \text{ transition is a binding event for species } x \\ \frac{+\Delta\mu_x}{2}, & \text{if the } i \to j \text{ transition is an unbinding event for species } x \end{cases} \tag{7}$$

Here, $\Delta\mu_x$ is the difference in the chemical potential across the membrane for species x (e.g. ion or substrate) [101]:

$$\Delta\mu_x = \mu_x^{\text{in}} - \mu_x^{\text{out}} = \ln \frac{[x_{\text{in}}]}{[x_{\text{out}}]} \tag{8}$$

where $[x_{\text{in}}]$ and $[x_{\text{out}}]$ are the intracellular and extracellular concentrations of species x. Note that by construction only one species can have a binding/unbinding event during an $i \to j$ state transition. The factor of $1/2$ in eq. 7 is an arbitrary choice for dividing the driving force and results in a symmetrical splitting of the binding chemical potential difference. Although in principle such splitting factors can affect driven processes [63], we found empirically that changing the factor had a minimal effect on the models discovered for the transporter systems of interest. Note that eq. 8 does not include the membrane potential for charged species (e.g., a sodium ion), which is excluded from this study to focus on the simplest cases.

We use a novel string-based approach to construct the state-space for a molecular machine in a combinatorial fashion, filtering out user-defined exclusions. 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, with further details given in S1 Text.

In order to investigate Hopfield-like enhanced discrimination [108], the cotransport system with a decoy substrate has additional parameters and constraints. The decoy-bound state free 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 ($\Delta\Delta G$), and not “internal proofreading”, which is consistent with Hopfield’s kinetic proofreading network [108]. Note that the difference in binding affinities in our Hopfield-like scheme effectively changes the activation barrier height, as seen via eq. 6. See Discussion. We note that occluded states (open to neither inside nor outside) are omitted for simplicity and that decoy and substrate are mutually exclusive binders.

[[Note: Here we use the terms 'decoy substrate' and 'unbinding', which are analogous to the biochemistry nomenclature of 'competitive substrate' and 'dissociation', respectively.]]

2.4.2 Model sampling

We use the Monte Carlo (MC) Metropolis-Hastings algorithm [165, 92] with tempering to sample model space, where a trial move consists of randomly adjusting a single state or transition energy, E_i or E_{ij}, along with the energies of its equivalent tied members. As noted in the SI, the tied members consist of states which are physically equivalent given a steady state where the outside and inside concentrations are fixed. Physically equivalent transitions are those occurring between physically equivalent states; see S1 Text. Adjusting all the "tied" states during a trial move prevents physically equivalent states in the model from having different free energies, maintaining thermodynamic consistency.

For each trial model generated using MC, we evaluate its MC "energy" or fitness based on its steady-state characteristics. To do so, we construct a rate matrix, $\mathbf{K}$, for the new model using the equilibrium and non-equilibrium (binding) energies for each state/transition along with the modified Arrhenius equations, (eq. 5 and 6). The rate matrix, $\mathbf{K}$, can be written with the state probabilities column vector, $\vec{P}$, to form the chemical master equation [210]:

$$\frac{d\vec{P}}{dt} = -\mathbf{K}\vec{P} \tag{9}$$

where $K_{ij} = -\alpha_{ji}$ for each transition with $i \neq j$, and K_{ii} is the sum over outgoing rate constants, $\sum_{j \neq i} \alpha_{ij}$. The chemical master equation can be solved under steady-state conditions using the definition of steady-state, $\frac{dP_i^{\text{ss}}}{dt} = 0$, 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 [185]:

$$j_{ij} = P_i^{\text{ss}} \alpha_{ij} \tag{10}$$

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 α_{ij} is the transition rate between state i and j. The overall flux, J_x, 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):

$$J_x = \sum_{ij \in \{x \text{ unbinding in cell}\}} j_{ji} \tag{11}$$

[[For example, for a 1:1 cotransporter model without a competitive substrate, the ions and substrate fluxes are calculated from the sum of the net flows of their (un)binding transitions, IF-Nb-Sx $\rightarrow$ IF-Ni-Sx, and IF-Nx-Sb $\rightarrow$ IF-Nx-Si, respectively. These transitions are highlighted in Fig 17.]]

Each model (set of rate constants between states) will have a Monte Carlo 'energy' E_{MC} assigned (i.e. fitness score) based on a user-defined general function. [[To be clear, each model contains the same graph of reaction states with edges determined by their respective, and same set of reaction rate constants. The difference is the values of those rate constants, which are explored during the Markov Chain Monte Carlo sampling. Once the biochemical network is

defined initially, no states or edges are removed during sampling.]] In this study we use substrate flows into the cell, as specified in the Results section for different choices of the fitness function.

Each Monte Carlo step results in a model, and thus each simulation yields a trajectory in "model space" (Fig. 18) – i.e., a sequence of models. By convention, lower fitness scores are more fit. The current model's 'energy' is compared to the previous model's energy and accepted/rejected based on the Metropolis-Hastings selection criterion [165, 92]. That is, the probability of accepting a trial move is the usual $p = \min[1, e^{-\beta \Delta E_{MC}}]$, where ΔE_{MC} is the change in the Monte Carlo 'energies' of the two models, and β is the effective inverse thermal-energy parameter. We emphasize, however, that our MC procedure does not generate a true Boltzmann-distributed thermal ensemble, and the MC β parameter is used only to aid sampling. In order to avoid trapping in deep "energy" (high fitness) basins we employ a tempering [254] procedure, described in S1 Text.

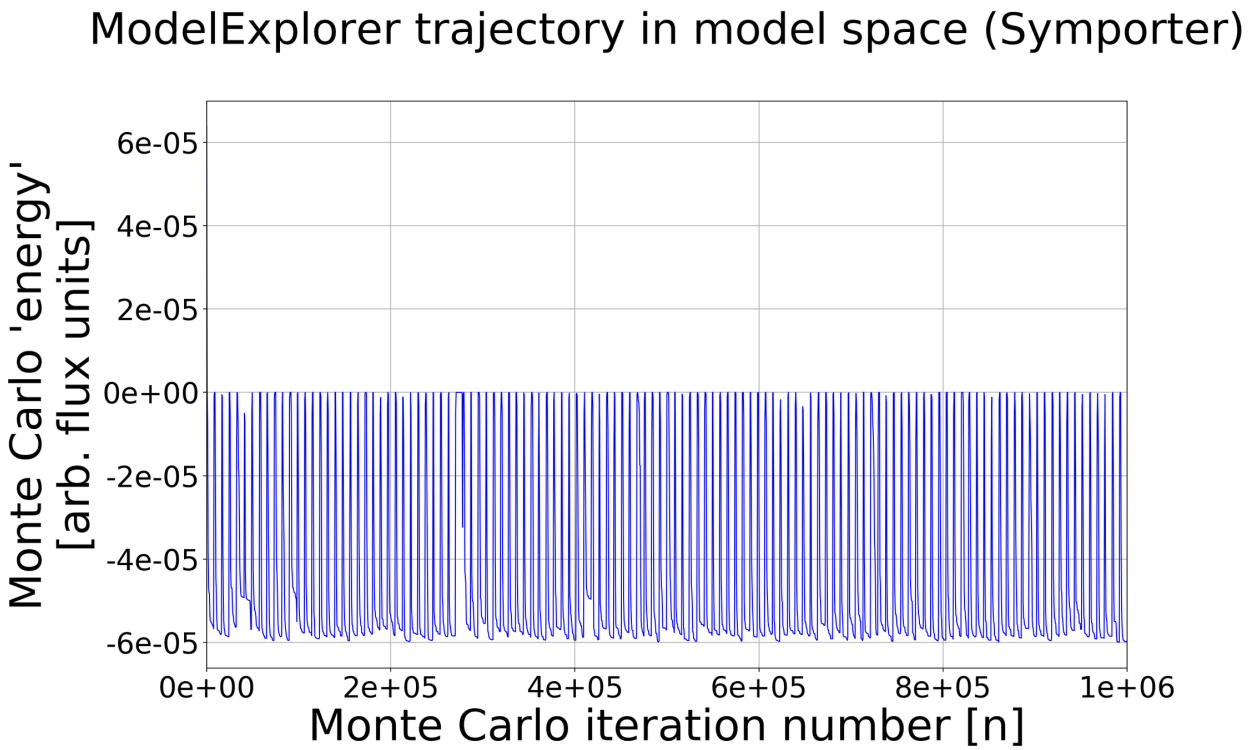


Figure 18: Exploration of Model Space Using Markov-chain Monte Carlo. The plot shows a ModelExplorer trajectory based on a symporter energy function during a 1e6 MC step simulation. Note that each point represents a different fully specified model. Energy minima correspond to models that are more fit, using $E_{\text{MC}} = -J_{\text{substrate}}$ as a fitness function in this case, where $J_{\text{substrate}}$ is the flux of substrate. Models are initially at a high energy but quickly find local minima. A tempering schedule (see S1 Text) of alternating temperature increases and decreases prevents the simulation from being trapped in local minima.

2.4.3 Model analysis

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 diagram of the network pathways can be made for a model at user-defined chemical potential differences. [[For visual clarity of the networks, edges with a negligible net flow relative to the other edges, were omitted from the diagrams.]] Note that the absolute flow (and flux) values are not physically relevant due to an arbitrary overall rate constant prefactor: all flows can be scaled by an arbitrary constant and remain consonant with the governing equations.

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 (Fig. 19) provides a means to investigate the dynamic behavior of molecular machines and provide testable mechanistic hypotheses. Since the flow and flux values are arbitrary (see above) the ion and substrate flux have been scaled by the maximum flux (i.e. ion flux at largest chemical potential difference).

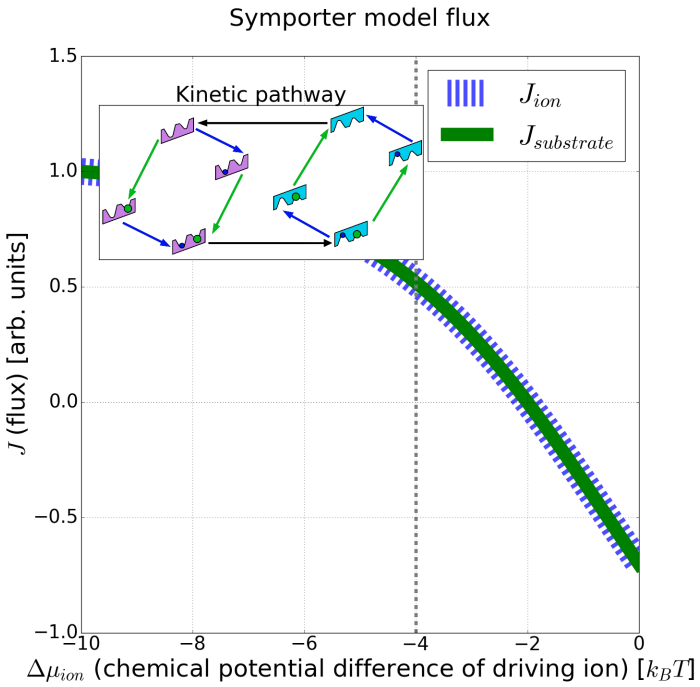


Figure 19: Ideal Symport in a Non-ideal (Mixed) Model. The fluxes, J, of the driving ion and substrate of an example symporter model are plotted over a range of ion chemical potential differences. Note the 1:1 stoichiometry of the substrate to ion flux, indicating an ideal symporter with no leaks. The ion and substrate flux have been scaled by the maximum ion flux for visual clarity. Inset: kinetic pathway of the same symporter model at an ion chemical potential difference of $-4k_BT$, as indicated by the vertical line. [[The ion and substrate binding and dissociation events are shown implicitly here for visual clarity.]]

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 unique models found under different simulation conditions. Models may also be clustered hierarchically [201], allowing for the discovery of different model classes. The model differences are calculated using the Euclidean distance between the vectors of the scaled (by vector magnitude) net flows between the states of a given model.

2.4.4 Computing details

The current prototype of ModelExplorer was written in Perl 5 with additional scripts for analysis and visualization written using Python 2.7. The software was used on a desktop PC running Windows 10 64-bit OS with an Intel i7-6700 CPU. Each simulation was run for 1e6 MC steps, storing models every 500 MC steps to reduce the run-time and memory requirements. Each run yielded 2000 models. For simulation parameters see S2 Table. All scripts and data for the manuscript are available at: https://github.com/ZuckermanLab/ModelExplorer.

2.5 Results

We present results for a range of systems demonstrating the ability of the computational approach to discover both expected and surprising mechanisms in single and competing substrate environments. For a simple cotransporter system without a decoy, the results include validation of ideal symporter/antiporter mechanisms, selected network topologies, and substrate/ion fluxes 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.

2.5.1 Ideal and mixed-mechanism cotransport with a single substrate

We first studied a transporter system with a single substrate driven by an ion gradient. To generate symport models, the substrate chemical potential difference ($\Delta\mu_{\text{substrate}}$) was set to $2k_BT$ and the ion chemical potential difference ($\Delta\mu_{\text{ion}}$) was set to $-4k_BT$; see (8). The fitness goal as embodied in the Monte Carlo energy was set to:

$$E_{\text{MC}}^{\text{symport}} = -J_{\text{substrate}} \tag{12}$$

where $J_{\text{substrate}}$ is the flux of the substrate into the cell (with units of s^{-1} that are scaled out by the Monte Carlo β parameter) as given by (11). This promotes intracellular substrate flow against its own gradient but down 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 of Fig. 17, as well as combinations of the idealized symporter cycles (Fig. 19 and see S3 Figure). All of these models exhibit a 1:1 ratio of substrate and ion flux into the cell - consistent with the idealized predictions of an optimized symporter (Fig. 17).

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

$$E_{\text{MC}}^{\text{antiport}} = J_{\text{substrate}} \tag{13}$$

where $J_{\text{substrate}}$ is the flux of the substrate (with units of s^{-1} that are scaled out by the Monte Carlo β parameter). Because MC favors lower energy, the use of (13) promotes the extracellular flow of the substrate opposite to the ion gradient and flow. The simulation of 1e6 MC steps yielded antiporter models with a 1:1 ratio of substrate flux out of the cell to ion flux into the cell (see S4 Figure and S5 Figure) – consistent with theoretical expectations for an optimized antiporter.

2.5.2 Discriminative models in the presence of a competing substrate

A primary motivation for this work was the challenge of generating models with particular functions in complex state-spaces where a combinatorial number of possibilities preclude guessing of mechanisms based on intuition. Specifically, motivated by hints that vSGLT exhibited non-productive reversal events (substrate unbinding to the extracellular side) [1], we wanted to investigate whether slippage events [101] might be able to enhance selectivity in the presence of a "decoy" substrate.

To seek models capable of enhancing selectivity for one substrate over another (*beyond* that generated by their differing affinities, importantly), a competing decoy substrate was added to a transport state-space that included a driving ion gradient. The decoy substrate was set to have a weaker affinity by $\Delta\Delta G = 1k_BT$, but otherwise the substrates are treated identically. The substrate and decoy chemical potential differences ($\Delta\mu_{\text{substrate}}$, $\Delta\mu_{\text{decoy}}$) were both set to $2k_BT$, and the ion chemical potential difference ($\Delta\mu_{\text{ion}}$) was set to $-4k_BT$. The fitness function (Monte Carlo energy) was set to:

$$E_{\text{MC}}^{\text{competitive}} = -J_{\text{substrate}} \cdot \frac{|J_{\text{substrate}}| + \epsilon}{|J_{\text{decoy}}| + \epsilon} \tag{14}$$

where, $\epsilon = 1\text{e}{-15}$ improves numerical stability, and $J_{\text{substrate}}$ and J_{decoy} are the fluxes of the substrate and decoy, respectively (with units of s^{-1} that are scaled out by the Monte Carlo β parameter) as defined in (11). The first factor of equation 14 promotes the intracellular flux of the substrate (negative sign used by convention), while the second factor promotes a high ratio of substrate to decoy flux (i.e. enhanced selectivity). Note that while the models are optimized at an ion chemical potential difference of $-4k_BT$, we are also able to find enhanced discrimination at larger ion chemical potential difference (i.e., concentration) values.

2.5.3 Analysis of a single discriminative model

To highlight the power of the sampling strategy to discover non-trivial mechanisms, we examine the kinetic pathways of a model (Fig. 20A) with enhanced selectivity. (Below this is referred to as "Model B" - MC index 29000.1, see S8 Text). This model can be described as a combination of several pathways: two pathways which symport both the ion and substrate (Fig. 20C), and two ion leak pathways which only transport the ion (Fig. 20B, D). Here we have defined

an ion leak as a "futile" cycle in the state-space leading solely to dissipation of the ion gradient. We can intuitively understand the discrimination mechanism: the ion leak pathways (which include the central horizontal arrow in Figure 20B, D) drive the substrate and decoy to unbind in the outward facing conformation (on the left side of the diagrams). Due to the $1k_BT$ difference in binding affinities, the substrate is more likely to rebind and be transported into the intracellular region. In fact, under the conditions which lead to the flows shown in Figure 20A, there is negligible decoy flow into the cell, which is why no flow arrow is shown connecting the two decoy-and-ion bound states. The mechanism of this model directly echoes the driven tRNA unbinding from the ribosome analyzed by Hopfield [108].

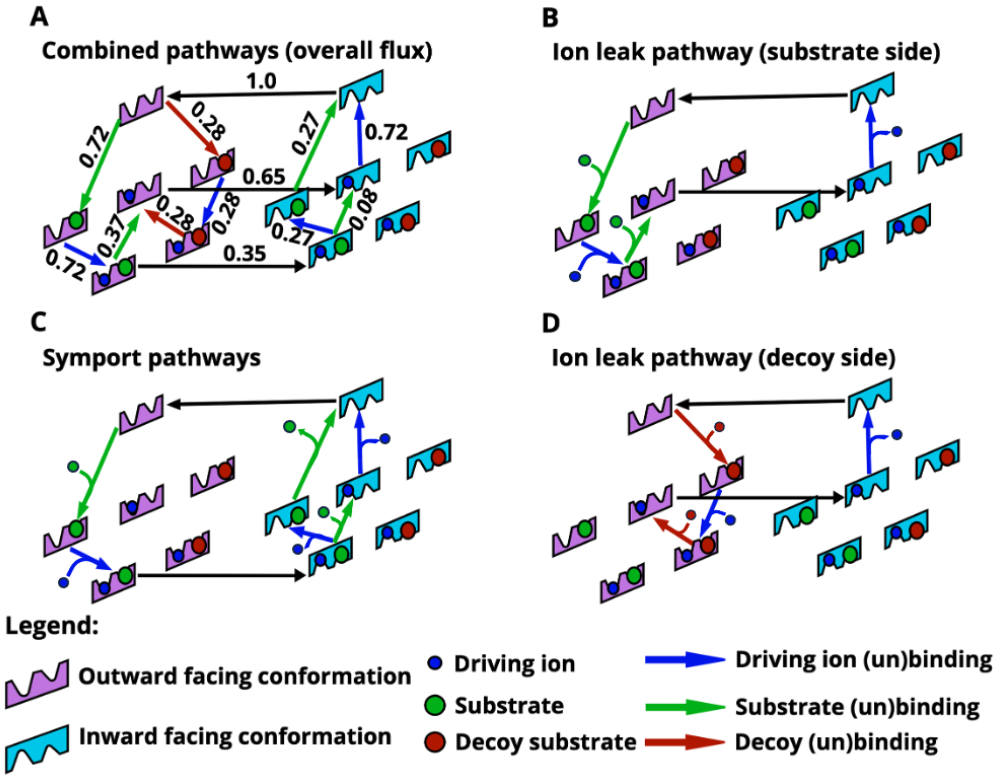


Figure 20: Dissection of a Single Model Exhibiting Enhanced Selectivity Into Component Pathways. The full model is shown in **(A)** with the net probability flows scaled by the largest edge flow, while panels (B) - (D) are various continuous cycles abstracted from the full model. **(B)** An ion leak pathway in which the substrate and ion bind extracellularly, but only the ion is transported into the cell because the substrate unbinds on the extracellular side. **(C)** A (split) cotransport pathway in which the substrate and ion both are transported into the cell. **(D)** A second ion leak pathway, mirroring (B), in which the decoy and ion bind extracellularly, but only the ion is transported into the cell. Overall, the substrate and decoy are both driven to unbind in the outward-facing conformation, shown on the left, due to ion leak pathways. However, due to the difference in binding affinities between decoy and substrate, the substrate is more likely to rebind and be transported into the intracellular region. The full process employs the ion leaks to enhance selectivity. [[The ion and substrate binding and dissociation events are shown explicitly here.]]

Selectivity can be examined over a range of driving ion chemical-potential differences, $\Delta\mu_{\rm ion}$, by examining the substrate and decoy substrate fluxes (Fig.21). The discrimination ratio of substrate to decoy, $J_{\rm substrate}/J_{\rm decoy}$, is essentially perfect at the value used to perform the MC sampling, namely $\Delta\mu_{\rm ion} = -4k_BT$, where the decoy flux vanishes while the substrate flux remains significant. Interestingly, in the range $-4k_BT < \Delta\mu_{ion} \lesssim -3k_BT$, substrate is pumped into the cell, while the decoy flows out, down its gradient.

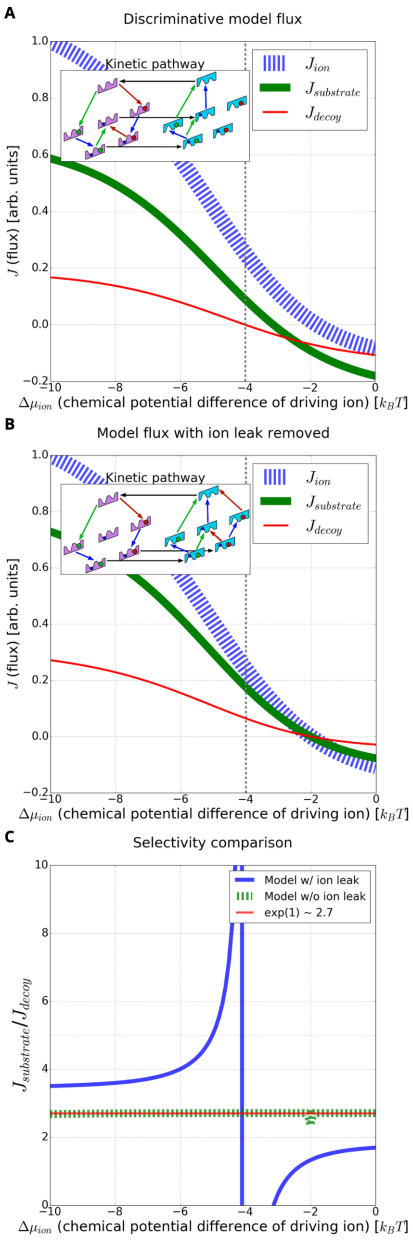


Figure 21: Enhanced discrimination driven by an ion leak. **(A)** For the model of Fig. 20, the substrate, ion, and decoy flux are shown for a varying chemical potential difference of the driving ion. Note the negligible decoy flux relative to the substrate flux near $-4k_BT$. **Inset** is the kinetic pathway diagram of the model at a specific chemical potential difference ($-4k_BT$, vertical line) of the ion. **(B)** The same discriminative model as in (A), but with the energy barrier between the ion-only bound states in the inward and outward conformations raised by $100k_BT$, effectively shutting off the ion leak. Both the substrate and decoy fluxes increase. [[Because there is no driving force from the ion leak pathway that pushes both the substrate and decoy to unbind (see inset A vs B), the substrate and decoy are transported based on their binding affinities, resulting in a larger flux in both species but lower discrimination. This is shown in panel C.]] **Inset** is the kinetic pathway diagram of the model with the leak removed, resulting in two symmetrical pathways for substrate and decoy transport. (**C**) Comparing the ratio of substrate to decoy flux (selectivity) for the same model with and without an ion leak. With the ion leak, the selectivity approaches infinity due to the negligible decoy flux. In contrast, removing the ion limits the selectivity to the expected equilibrium-like value of $e^{\Delta\Delta G=1}$. Note that the sign change of the selectivity is due to the change in substrate and decoy flux direction. [[For visual clarity, the binding and dissociation events are shown implicitly. We note that the y-axis of panel **C** has a different range than panels **A** and **B**. From $\Delta\mu_{ion}$ -4 to -3 k_BT, the ratio of substrate to decoy fluxes for the discriminative model is negative, and is positive elsewhere, as expected from the panel **A** flux traces.]]

To further investigate the mechanism of enhanced selectivity, we compared the original model B (Fig. 21A) to the same model with the ion leak removed (Fig. 21B, note absence of central horizontal flow). The leak-free model – generated by increasing the ion-only-bound conformational transition barrier energy by $100k_BT$ – has symmetric pathways for both substrate and toxin transport (Fig. 21B inset). Removing the leak dramatically decreases discrimination, as shown in Figure 21C, especially near the ion chemical potential difference $\Delta\mu_{\rm ion} = -4k_BT$ at which MC sampling was performed. In particular, the leak-free model exhibits a decrease in selectivity, approaching the expected equilibrium-like value of $e^{\Delta\Delta G} \approx 2.7$ with $\Delta\Delta G = 1$ (see Fig. 21B and S6 Figure) [108]. This suggests that the ion leak is the prime mechanism driving enhanced selectivity for this model.

2.5.4 Meta-analysis of discriminative models

Our model-sampling approach yields numerous models. We examined them on a "systems basis" using a filtering and clustering procedure.

To start, we filtered for the highest-performing models found during four separate runs of 1e6 MC steps (see S8 Text). Specifically, models were filtered based on the ion-to-substrate flux ratio being greater than 0.1, and the substrate-to-decoy ratio being greater than $10e^{\Delta\Delta G}$ (ten times the expected discrimination ratio at equilibrium) [108], where $\Delta\Delta G = 1$ is the difference in binding affinities between the substrate and decoy substrate. These filtering constraints ensured that the analysis only contained models with a sufficient stoichiometry of sodium ions to the substrate, as well as a minimum baseline for enhanced selectivity.

Filtering for performance resulted in 1783 of the aggregate 8000 models exhibiting enhanced selectivity, and we probed these via clustering based on a similarity metric. Clustering revealed four different mechanistic model classes (clusters) with differing kinetic pathways leading to enhanced selectivity (Fig. 22). Procedurally, for each model in the filtered set, the flows on all edges were scaled by the Euclidean norm of edge flows in that model. The normalized edge flows define a flow vector used to calculate Euclidean distances and perform complete-linkage clustering [201] based on a distance threshold of 0.65, which we found empirically to reveal qualitatively different pathways. [[The flow vector represents the normalized edge flows of each network edge and has the same length and order for each model (i.e. the same edges are used, none are removed). For visual clarity, edges with negligible normalized flow were omitted from the kinetic diagrams. The clustering distance threshold was iteratively adjusted until the kinetic pathway diagrams (and flux traces) of several randomly selected models were similar within each cluster, yet distinct from the other pathway diagrams of the other clusters. Representative models were chosen randomly from this set of random models from each cluster.]]

All model clusters, which importantly were filtered for enhanced selectivity of substrate over decoy, contain an ion leak. As discussed previously (see Fig. 20), the models shown in Figure 23 contain "futile" ion-leak cycles, which consist of all the paths that include the central arrow connecting the ion-only-bound outward-facing to inward-facing conformations. In many, but not all models (see models in clusters A, B, D), this ion leak is coupled to substrate and decoy unbinding in the outward-facing state. In other words, the molecules bound to the OF state of the transporter are effectively ejected back to the extracellular space in a process that expends free energy from the ion gradient.

To further confirm the role of the futile ion cycle in promoting selectivity, the ion leak was removed for each of the representative cluster models (Fig. 23A-D). This was done by raising the corresponding energy barriers as described previously for model B. The resulting leak-free models again exhibited a decrease in selectivity to the expected equilibrium value ($e^{\Delta\Delta G}$ with $\Delta\Delta G = 1$), indicating that the ion leak is indeed the driving mechanism for enhanced selectivity in our models as in Hopfield-like kinetic proofreading [108]. We note that in our formulation, the difference in binding affinities is effectively a kinetic parameter that adjusts the reaction rate, as seen from (6). Our results are thus consistent with recent theoretical arguments which suggest that only kinetic parameters adjust the ratio of stationary fluxes [9].

2.5.5 Model class analysis

It is instructive to examine all the model classes further to understand their differences. Models corresponding to clusters A and B (Fig. 23A,B) share a similar network structure where the substrate (or decoy) tends to bind before the ion, followed by the unbinding of both substrate and decoy, which favors the stronger-binding substrate for transport. A and B differ slightly in that models in cluster A contain an ion-only binding transition in the outward conformation and have a single unbinding pathway for the substrate and ion in the inward-facing conformation (i.e. one symporting pathway). The models in cluster B have two unbinding pathways for the substrate and ion in the inward-facing conformation (i.e. two symporting pathways, see Fig. 20C).

The model class of cluster C (Fig. 23) embodies a much less intuitive mechanism. First, the model is fully connected: each allowed state transition has a non-zero flow. Unlike the other three model classes, class C does not exhibit a net

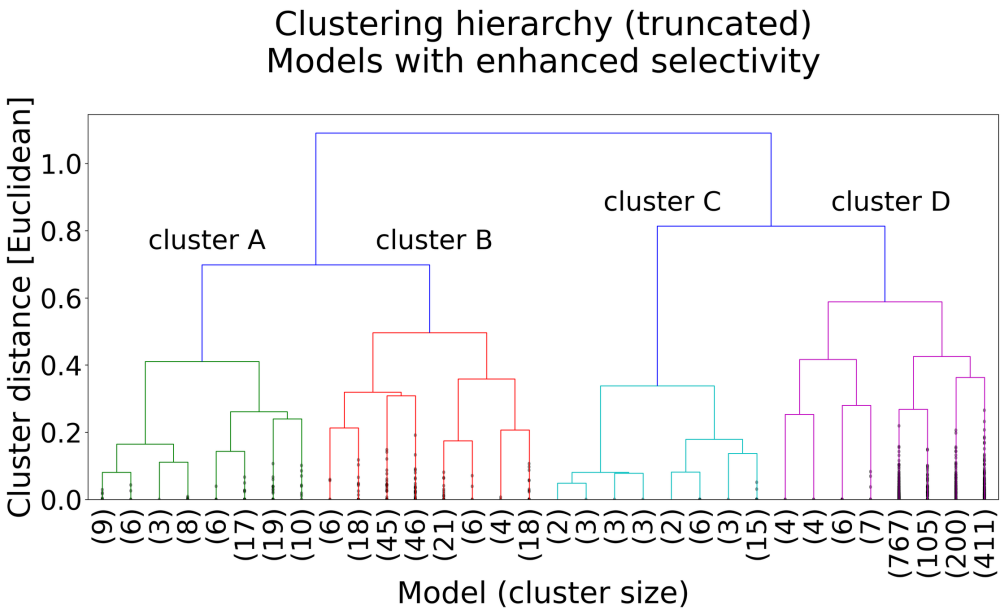


Figure 22: Pathway Meta-analysis: Clustering analysis of the best-performing models based on model similarity. The dendrogram was truncated to four levels for visual clarity, with the number of models below the truncation shown in parenthesis. Dendrogram created using Python/SciPy and Matplotlib.

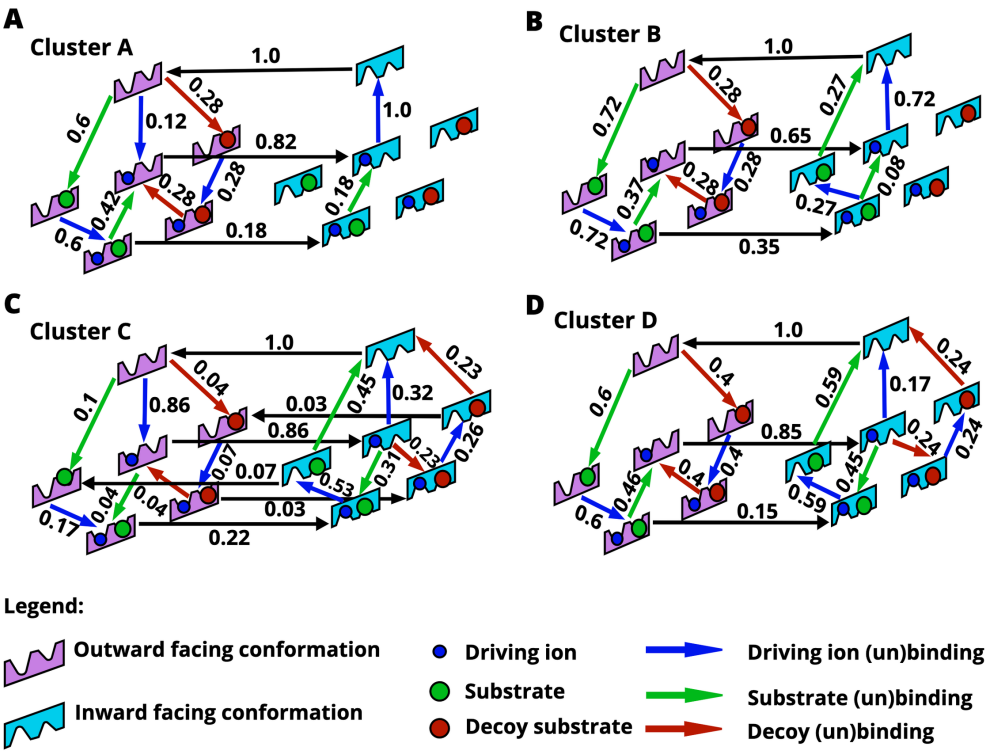


Figure 23: Kinetic Pathways of the Four Model Classes (A-D) Found From Clustering Analysis (Fig. 22). Each of these models exhibits a high level of selectivity due to an ion leak, as discussed in the text. Note that the net flow values shown along edges are scaled by the maximum flow edge of the individual model. [[For visual clarity, the binding and dissociation events are shown implicitly.]]

flow of substrate unbinding in the outward-facing conformation. Model C also contains parallel futile cycles with no net transport for either the substrate and decoy in which the ion dissipates its gradient. In the inward-facing conformation, both the substrate and decoy are driven by the ion leak to bind and then unbind again, resulting in more substrate than decoy flux due to their different binding affinities.

Models in cluster D (Fig. 20) contain unbinding steps for both the substrate and decoy in both OF and IF conformations, driven by an ion leak. On the OF side, ion driving appears to force all the decoy to unbind, since there are no horizontal transitions from outward-facing on the left to the corresponding inward-facing states on the right, whereas a fraction of the native substrate remains bound (stronger affinity by $\Delta\Delta G = 1k_BT$) during the OF-to-IF conformational transition; more precisely, there is an equal and opposite flow of conformational transitions for the decoy-bound states, while the substrate exhibits a net productive flow into the cell. The ion leak also drives the substrate and decoy to bind and unbind in the inward-facing conformation, but these processes 'cancel out' and do not lead to net flux of the decoy substrate.

For all four models, at low ion chemical potential differences (similar to the magnitude of the substrate/decoy chemical potential difference), there is a regime where the ion and substrate are transported into the cell, and the decoy is transported out of the cell (see Fig. 21A and S7 Figure). This suggests an alternative mode of transport in which the cell 'detoxifies' by exporting the decoy substrate.

We can also probe the costs and restrictions on enhanced selectivity. Although our Monte Carlo sampling optimized substrate-to-decoy selectivity at a particular set of chemical-potential differences, the enhanced selectivity generally occurs over a range of thermodynamic driving forces (S12 Figure). However, some models exhibit high selectivity at a low cost (i.e., low stoichiometric ratio of ion to substrate flux), without any added constraints on the flux ratios. Analysis of the representative models suggests that uniformly high discrimination over a range of thermodynamic conditions occurs with a correspondingly uniformly high cost (see S11 Figure), although this is not necessarily a representative sample. This issue will be of interest in future work.

2.6 Discussion

Overall we have seen that Monte Carlo sampling of model space can find diverse, testable models that provide insight into complicated cotransport systems. Unlike prior related work for ion channels [91, 164, 237], we have not sought to fit or optimize a single model to a set of experimental data. Rather, we have taken a *systems* approach in asking for *sets of models capable of performing a given function* and applied this to discover multiple transporter models with a discriminatory capability not previously envisioned in the literature, to our knowledge. Our main point is not that proofreading surely occurs in transport, but rather that an automated approach is required to consider mechanistic possibilities in a systematic way.

First, the model-sampling approach systematically identified ideal and mixture pathways for both simple symporters and antiporters, validating the approach. Subsequent study of a more complex state space including the very realistic possible binding of a "decoy" substrate indicates the possibility of cotransporters that exhibit enhanced selectivity similar to Hopfield's and Ninio's kinetic proofreading models [108, 178]. The enhnanced-selectivity models all exhibit an ion leak which, when removed, prevents the enhanced discrimination. This apparently unreported mechanism for secondary active transporters can enhance selectivity to a remarkable degree in a limited range of conditions. Clustering analysis of all models sampled shows a fairly diverse group of model classes that exhibit enhanced selectivity using different kinetic pathways. [[However, we acknowledge that our sampled models may not be representative of all possible kinetic pathways for enhanced selectivity and that there may be more diversity within our clusters than presented in our selected representative models. Future work will explore enhanced sampling and classification methods for a more robust characterization of transporter mechanisms.]]

Every kinetic model that is generated is also thermodynamically consistent with both equilibrium and non-equilibrium constraints. 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 of reaction rates α_{ij}, including non-equilibrium effects where appropriate, combined with an automated state-space construction distinguishes our method from similar approaches. We believe the steady-state conditions studied here provide the best simple model for many cellular conditions which may change very slowly – over timescales of minutes, hours, or days – but undoubtedly transient effects are important for some physiological conditions such as release from

starvation conditions. Also many experiments study transient phenomena by construction [160, 14, 132, 200] and so it will be valuable to use our models accordingly in future work.

Although our approach was not developed for the study of evolutionary molecular biology, the method "discovers" working models starting from unproductive models akin to earlier work [45, 82, 91, 164, 237]. The changes to rate constants observed in some trajectories may be related to structurally feasible changes. In the future, a more sophisticated approach could attempt to build in additional structural constraints to develop candidate feasible pathways for transporter evolution. We note that the present study already included constraints on the similarity of substrate and decoy, as well as their mutually exclusive binding.

In addition, by utilizing additional 'base states' (i.e. Ni, No, Nb, see S1 Text) our method may be extended to include an arbitrary number of additional binding sites for a given species. This will generate a larger state-space to search for models with a range of possible substrate/solute binding sites and pathways. This capability is particularly useful to study systems which may exhibit variable (or unknown) substrate/solute stoichiometries [132] and is a direction for active work.

All the enhanced-selectivity models sampled in this study, by construction, were of the 'Hopfield type' [108] 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. Note that by Eqs (1) and (2), our barrier energies E^{bar} are absolute energy levels and not the conventional "barrier heights," which are energy *differences* $E^{\text{bar}}_{ij} - E_i$. Thus, in our Hopfield-like model with equal E^{bar}_{ij} values for substrate (S) and decoy (W), the effective barrier heights differ between S and W because the stabilities (affinities) E_i differ. In other words, the kinetic parameters are coupled to the affinity/stability parameters, although formulations can be constructed in which they are independent [9].

Hopfield-type proofreading differs from what might be called "internal proofreading" where additional discrimination results from differing barrier heights [61]. Biological proofreading can be expected to mix both mechanisms to some degree [9], 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 has a negligible on-rate for the transporter. Our approach differs somewhat from Hopfield's and Nino's [108, 178] in that no irreversible steps enter our models. The full reversibility appears to be a necessary ingredient in the ideal discrimination (unbounded ratio of substrate to decoy flux) that occurs in some models for specific concentrations.

As emphasized decades ago by Hill [100], in complex networks such as those explored here, it is the network as a whole rather than key steps which define the mechanism. The current networks are simple enough that we can point to intuitive processes coupling ion leaks to unbinding, but the mechanism is defined by the overall process. Undoubtedly, in more realistic networks including a fuller set of conformational states, it will become more difficult to describe an

intuitive mechanism. Nevertheless, the principles uncovered in the simple systems can provide useful guidance for conceptualizing complex systems.

2.7 Conclusion

Motivated by evidence for the alternative behavior of molecular machines, we have developed a thermodynamically consistent approach to systematically explore a range of mechanisms, generating multiple experimentally testable kinetic models based on a predetermined fitness function. Prior work has focused on developing individual sets of optimal parameters [91, 164, 237] and not on generating model sets, which we believe is essential for developing precise mechanistic hypotheses. The approach was designed for complex state spaces which can be automatically generated, and which would be difficult to analyze by intuition alone. To our knowledge, the platform is the first to enable model sampling building in both equilibrium and non-equilibrium thermodynamic rules. This 'systems biology' approach to analyzing mechanisms of molecular machines was applied to a cotransporter state-space with and without a decoy substrate. After validating the method against 'textbook' symporter/antiporter models, we 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 on an *ad hoc* basis starting from a limited set of experimental structures.

2.8 Supporting information

[[See the attached appendices for the supporting information on this chapter.]]

S1 Text. Detailed methods. Details regarding the string-based creation of states, state definitions, equivalent states/transitions, tempering procedures, and flux calculations

S2 Table. Simulation parameter values. Table containing the parameter values used for each of the simulations in this study.

S3 Figure. Symporter model pathway (without decoy substrate). Pathway of a symporter model found at MC step = 1800. This model exhibits a combination of all four ideal symporter pathways which result in the intracellular transport of one substrate per ion. Note that the flows are scaled by the largest flow edge.

S4 Figure Antiporter model pathway (without decoy substrate). Pathway of an antiporter model found during the antiporter simulation run at MC step = 845000. This model exhibits a combination of two ideal antiporter pathways which result in the extracellular transport of one substrate per ion. Note that the flows are scaled by the largest flow edge

S5 Figure. Antiporter flux diagram (without decoy substrate). Flux of an antiporter model found during the antiporter simulation run at MC step = 845000, analyzed over a range of ion chemical potential differences. This model

exhibits a 1:1 ratio of ion influx to substrate efflux over a wide range of ion chemical potential differences. Note that the fluxes are scaled by the largest flux value.

S6 Figure. Symporter model pathway with ion leak removed (and decoy substrate present). Pathway of the model with enhanced selectivity representing cluster B, with the futile ion cycle removed. The energy barrier between the ion-only bound states in the inward and outward conformations was raised by 100 k_BT , effectively shutting off the ion leak. Note the two symmetrical pathways for substrate and decoy transport. The net flows have been scaled by the maximum flow edge.

S7 Figure. Flux diagrams of the representative models for each cluster. Flux of the representative model for each cluster, scaled by the maximum flux, over a range of ion chemical potential differences. Each model has a narrow regime where the toxin flows down its gradient out of the cell, while the substrate is driven into the cell by the ion. Near the optimized conditions for the simulation, $\Delta\mu_{\text{ion}} = -4k_BT$, these models have negligible decoy flux, resulting in an unbounded substrate to decoy discrimination ratio.

S8 Text. Model clustering method. Details on the clustering procedure, parameters, and the chosen representative models.

S9 Figure. Trajectory in model space. Trajectory in the model space of four different 1e6 Monte Carlo (MC) step simulations at different sampling settings. Simulations were run for transporters in a 'competitive' environment with a decoy using the MC energy function: $-J_{\text{substrate}} \frac{|J_{\text{substrate}}|+\epsilon}{|J_{\text{decoy}}|+\epsilon}$ where $J_{\text{substrate}}$ and J_{decoy} are the fluxes of the substrate and decoy respectively, and ϵ = 1e-15. Lower MC energy values denote models that are more fit, by convention. As shown in the figures, the tempering schedule aids in avoiding low energy basins. Note that each point on the trajectory is a kinetic model.

S10 Figure. Trajectory in cluster space. Trajectory in the cluster space of four different 1e6 MC step simulations. Simulations were run for transporters in a 'competitive' environment with a decoy. Models were filtered based on a cost (ion to substrate flux ratio) below 10, and selectivity (substrate to decoy ratio) above $10e^{\Delta\Delta G=1}$. Clusters were determined using hierarchical clustering with complete-linkage and the Euclidean distance between the scaled flows of each model. The threshold of 0.65 was determined empirically to produce qualitatively different kinetic pathways. These graphs indicate that each run only finds a few model classes during the simulation – implying the need for improved sampling methods. Note that in run 4, models meeting the selection criteria (i.e. cost and selectivity) were not found until approximately 1.5e5 MC steps.

S11 Figure. Cost of the representative models. The cost of the representative model for each cluster, over a range of ion chemical potential differences. All of these models exhibit a cost above the ideal 1:1 stoichiometric ratio for a

wide range of chemical potential differences. The extra ions transported relative to the substrate suggest a futile ion transport cycle - i.e. an ion leak. Note that the cost was not included as a constraint in the energy function

S12 Figure. Selectivity of the representative models. The stoichiometric ratio of the substrate to ion flux (selectivity), over a range of ion chemical potential differences. All the models demonstrate enhanced selectivity over a range of chemical potential differences, and unbounded selectivity at the optimized condition (at $\Delta\mu_{ion} = -4k_BT$). Models A and D exhibit enhanced discrimination over a wide range of conditions. The expected equilibrium value ($e^{\Delta\Delta G=1}$) is shown as a reference. Note that this stoichiometric ratio was used as a primary constraint in our energy function.

Acknowledgments

We thank Barmak Mostofian and Jeremy Copperman for their helpful discussions.

3 Constructing a robust pipeline for calibrating mechanistic transporter models

Constructing a Robust Pipeline for Calibrating Mechanistic Transporter Models

August George[1], Daniel M. Zuckerman[1*],

1 Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon, USA

* zuckermd@ohsu.edu

3.1 Abstract

Mechanistic modeling holds out the promise of providing valuable insights into the intricate processes of proteins, such as membrane transporters implicated in numerous diseases. A significant challenge in this field is ensuring accurate estimation of model parameters consistent with the data. Issues such as model construction errors, poor convergence of parameter estimates, inadequate validation, and sparse data often result in biased parameter estimates which yield inaccurate predictions. Our research introduces a Bayesian inference pipeline designed for solid-supported membrane electrophysiology-based (SSME) experiments. We propose strategies for model construction, simulation, and calibration suitable and present a high-performing, reproducible pipeline that encompasses user-friendly model construction, optimized model simulation, and efficient model calibration using Bayesian inference. Our results indicate that we can reliably estimate microscopic biophysical properties of complex protein systems from small, sparse, and noisy datasets while maintaining unbiased prior assumptions of the model. This work advocates for a shift towards more robust model calibration methods and sets the stage for studying more complex systems as well as more traditional systems biology models.

3.2 Author summary

This research aims to enhance our understanding of proteins, specifically transporters, which perform essential biological processes. Our method uses mechanistic models, mathematical descriptions of the underlying biochemical reactions, to discern the complex behavior of these proteins. Determining the exact set of biochemical reactions and their governing parameters is challenging experimentally due to the inability to directly observe all the microscopic processes that occur during transport. To help overcome these obstacles, we have developed a computational approach to determine model parameters and their uncertainties that use data generated from an emerging experimental method. We utilize state-of-the-art techniques to construct mechanistic models, simulate experimental assays used in transporter research, and statistically infer the model parameters. Our approach is validated using synthetic data with known parameter values, demonstrating the ability for unbiased model parameter estimation. These promising results pave the way for a more comprehensive study of transporter mechanisms.

3.3 Introduction

Proteins are nanometer-scale biological assemblies that perform specific functions that are required for life [4, 5, 19]. They are involved in critical processes such as signal transduction, structural support, and transport [5, 93, 128, 83]. Of interest in this study are secondary active transporters which use the energy stored in electrochemical gradients to selectively move nutrients across a membrane [5, 25, 64]. These proteins have substantial relevance to biomedical research given their connection to several diseases and role as therapeutic targets for diseases such as diabetes and depression [257, 236, 154, 76, 145]. A notable example is the use of selective serotonin reuptake inhibitors which target serotonin transporters and are used to treat anxiety and depression [157, 207]. In addition, certain active transporters have shown complex transport mechanisms, such as the ability to efflux a wide range of dissimilar drugs, contributing to multidrug resistance as found in some infectious diseases and cancers [97, 199, 115, 86, 200].

While transporters have been extensively studied, determining their biophysical mechanisms and kinetics, i.e. the exact set of biochemical reactions and their respective reaction rate constants, is an ongoing challenge [116, 54, 239, 261, 264]. Due to the time and length scales of many membrane transporters, their mechanism can only be partially observed with experimental methods, and traditional atomistic simulations are computationally intractable [124, 55]. Instead, biochemical reaction networks are commonly used to describe complex biological processes [249, 6]. Mathematically, each biochemical reaction is represented with a rate equation that determines how the concentrations of products and reactants change over time, typically governed by a reaction rate constant. These equations are combined to form a system of ordinary differential equations (ODEs), which when given the rate constants and initial values, can be integrated to fully predict the time evolution of the model. Alternatively, the differential equations can be simulated stochasticly, such as with the Gillespie algorithm [77]. This approach is typically used when random fluctuations of the underlying chemical reactions are significant, such as with a small number of molecules [98, 235].

These types of network models can scale from the behavior of proteins involved in transport to the complex processes involved in cell metabolism [81, 56, 68]. Given the complexity of these types of models, it is important to use systematic approaches to construct, validate, and calibrate each model to reduce bias and the chances of human errors. Great strides have been made in the greater systems biology field to develop best practices [155, 167, 196] and robust tools, such as interoperable models using systems biology markup language (SBML) [111, 127], efficient model simulation with Libroadrunner [230], rules-based construction of networks using BioNetGen [60], and all-in-one packages such as COPASI [107]. However, there have been limited applications towards membrane transport proteins, and accurately and efficiently determining model parameters (i.e. reaction rate constants) and their uncertainty from data is an ongoing challenge [196]. Unbiased parameter estimates are needed to improve our understanding of the underlying biological processes, as well as ensure that the model predictions are accurate.

Two common approaches for parameter estimation are maximum likelihood estimation (MLE) [134] and Bayesian inference (BI) [73, 256]. Briefly, MLE returns the model parameters that best fit the data given a likelihood (or cost) function. Optimization methods such as simulated annealing [133] or genetic algorithms [126] are used to determine

the optimal set of parameters given the data, but the high-dimensional, non-convex, and non-linear nature of many biological models makes convergence challenging in practical timescales. While MLE only returns a single value, methods utilizing the profile likelihood, bootstrapping, and the curvature near the maximum likelihood function (i.e. Fischer information matrix) can be used to quantify the uncertainty of the model parameter estimates [144, 58, 221, 67].

In contrast, Bayesian inference, generates a distribution of model parameters that fit the data, given some prior assumptions about the parameters and data, as well as a likelihood function. Bayesian inference is based on Bayes' theorem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \propto p(D|\theta)p(\theta) \tag{15}$$

where $p(\theta|D)$ (i.e. posterior) is the probability of a parameter set given the data, $p(D|\theta)$ (i.e. likelihood) is the probability of the data given a parameter set, $p(\theta)$ (i.e. prior) is the probability of the parameters, and $p(D)$ (i.e. evidence) is the probability of the data. One challenge in BI is the choice of priors and likelihood, which may introduce strong biases into your model, and should be carefully chosen based on domain knowledge. In addition, the posterior requires the integration of a complex multidimensional integral which is generally intractable for problems of interest.

Markov chain Monte Carlo (MCMC) [220, 36, 117] and variational inference (VI) [24, 263, 52] are two methods that numerically approximate the posterior distribution. Variational inference trades accuracy for efficiency by optimizing tractable distributions (like Gaussians) to approximate the true posterior distribution. For complex models, the distributions used in VI may not be sufficient to accurately capture the complexity of the posterior. As an alternative, Markov chain Monte Carlo methods are statistical sampling methods that generate a random list of parameter values, a Markov chain, that after convergence will approximate the posterior distribution. Briefly, a new parameter value is randomly selected based on its current value, and either accepted or rejected based on the probability of that candidate parameter value [165, 92], creating a long chain of parameter values that eventually approximate the target distribution. This accuracy comes at an increased computational cost, and convergence may still be intractable due to issues arising from large dimensions, complex posterior geometries, and strong parameter correlations [203, 42]. Many different software tools [32, 184, 72] and algorithms have been developed to address these issues such as Hamiltonian Monte Carlo [20], and sequential Monte Carlo [38], but choosing the appropriate algorithm and its associated hyperparameters is non-trivial for biochemical reaction network models, and great care should be used when checking for convergence.

Here we develop a parameter estimation pipeline using Bayesian inference for secondary active transporters that combines current best practices and methodologies in systems biology with an emerging experimental technique for transporter research, solid-supported membrane electrophysiology (SSME). SSME involves a series of external concentration perturbations that yield a net aggregate measurement of transient current, which is generated by the movement of ions across the membrane via transporter proteins. SSME experiments can have higher signal-to-noise, higher time resolution, and improved stability compared to traditional electrophysiology methods [214, 12, 13]. Like other systems biology datasets such as those used to study gene expression dynamics [71, 112] or signal transduction

pathway models [173, 3], SSME generates noisy time series data that changes over time due to perturbations and is sparse compared to the number of model parameters.

While related work using model calibration and Bayesian inference has been done in many areas within systems biology [155, 151, 152, 26], including ion channel transporters [169, 31, 224], to our knowledge no robust pipelines have been developed focused on secondary active membrane transporters or SSME-type experiments. Motivated by this gap in methodology and application, we present a tool to efficiently simulate SSME-like assays for a given transporter model and estimate its parameters via machine learning accelerated Bayesian inference. We validate our method using synthetic data with known synthetic ground truth and show that precise estimation of microscopic rate constants is indeed possible for transporter models and SSME-like datasets - even with extremely uninformative priors.

3.4 Materials and methods

3.4.1 Pipeline overview

Our methodology involves the following steps as shown in Fig 24: We select a dataset representing an SSME assay current trace. The Bayesian inference probabilistic model (priors and likelihood) is defined as well as an ODE model describing the biochemical network for the transporter. Assay conditions for the simulation are also defined in alignment with the data. Once the model calibration problem has been specified, we do Bayesian inference. This process iteratively selects model parameter values, simulates the assay based on the parameter values and problem specification, computes the likelihood, and then accepts or rejects that parameter candidate. After convergence, the Bayesian inference pipeline will output the posterior, a rich set of parameter distributions of the model based on the given data.

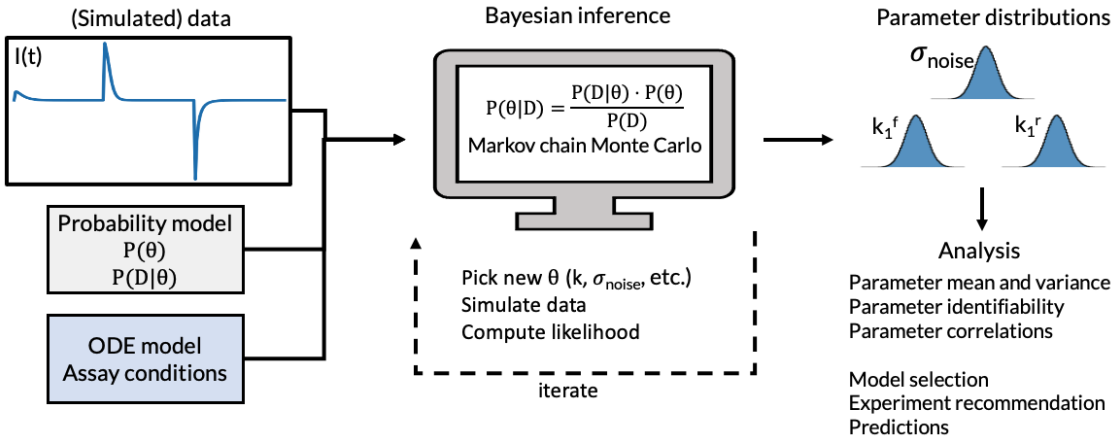


Figure 24: Model Calibration Pipeline for Transporter Research The different elements of the process to estimate transporter reaction rate constants and assay nuisance parameters from SSME-like data.

3.4.2 Transporter model specification

We use an idealized tightly coupled 1:1 antiporter transport cycle (see Fig 25) based on an alternating access model which is the prevailing mechanism used for secondary active transporters [116, 64, 147]. Here, the transporter is assumed to pivot between an outward-facing and an inward-facing conformation, 'OF', and 'IF' respectively. In the forward direction, one ion (H) is transported from outside to inside, while the substrate (S) is transported from inside to outside (see figure). The transporter reaction cycle consists of a set of coupled elementary biochemical reactions such as species binding, unbinding, and protein conformational changes, where each reaction is governed by reaction rate constants. Note each reaction is reversible and that the transport cycle is therefore reversible under certain electrochemical conditions.

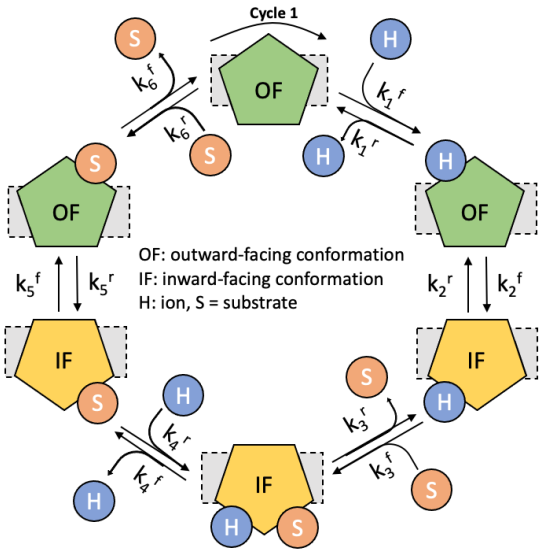


Figure 25: Antiporter Transport Cycle A cartoon representing the biochemical reaction network of coupled transport of an ion (H) and substrate (S) across a membrane via an alternating access mechanism, with outward-facing (OF) and inward-facing (IF) protein conformations.

The change in each biochemical species over time is defined with a set of ordinary differential equations (ODEs) motivated by the law of mass action [250, 251, 129]. Here the rate of a reaction is assumed to be the product of the reactant concentrations. That proportionality constant is the reaction rate constant, k_i^j, where i is the reaction step and j denotes the forward ('f') or reverse ('r') reaction.

The governing differential equations are shown below (see the SI for specific reaction rates):

$$\begin{aligned}
\frac{dH_{\text{in}}}{dt} &= k_4^f \cdot IF_Hb_Sb - k_4^r \cdot IF_Sb \cdot H_{\text{in}}, \\
\frac{dS_{\text{in}}}{dt} &= -(k_3^f \cdot IF_Hb \cdot S_{\text{in}} - k_3^r \cdot IF_Hb_Sb), \\
\frac{dOF}{dt} &= (k_6^f \cdot OF_Sb - k_6^r \cdot OF \cdot S_{\text{out}}) - (k_1^f \cdot OF \cdot H_{\text{out}} - k_1^r \cdot OF_Hb), \\
\frac{dOF_Hb}{dt} &= (k_1^f \cdot OF \cdot H_{\text{out}} - k_1^r \cdot OF_Hb) - (k_2^f \cdot OF_Hb - k_2^r \cdot IF_Hb), \\
\frac{dIF_Hb}{dt} &= (k_2^f \cdot OF_Hb - k_2^r \cdot IF_Hb) - (k_3^f \cdot IF_Hb \cdot S_{\text{in}} - k_3^r \cdot IF_Hb_Sb), \\
\frac{dIF_Hb_Sb}{dt} &= (k_3^f \cdot IF_Hb \cdot S_{\text{in}} - k_3^r \cdot IF_Hb_Sb) - (k_4^f \cdot IF_Hb_Sb - k_4^r \cdot IF_Sb \cdot H_{\text{in}}), \\
\frac{dIF_Sb}{dt} &= (k_4^f \cdot IF_Hb_Sb - k_4^r \cdot IF_Sb \cdot H_{\text{in}}) - (k_5^f \cdot IF_Sb - k_5^r \cdot OF_Sb), \\
\frac{dOF_Sb}{dt} &= (k_5^f \cdot IF_Sb - k_5^r \cdot OF_Sb) - (k_6^f \cdot OF_Sb - k_6^r \cdot OF \cdot S_{\text{out}}).
\end{aligned}$$

Here the outward or inward-facing conformation state is denoted with an 'OF' or 'IF' respectively, and the ion and substrates are compartmentalized as inside ('X_{in}'), outside ('X_{out}'), or bound to the protein ('X_b'), with the rate constants as specified earlier. Due to the nature of the SSME experiment, the outside concentrations are held fixed for a given assay stage (i.e. $\frac{dH_{\text{out}}}{dt} = \frac{dS_{\text{out}}}{dt} = 0$).

To account for the effects of the membrane voltage on the transport of charged chemical species we use a modified Arrhenius rate formulation using Eyring rate theory [23, 85]:

$$k_i^j(V) = k_i^j(0) \exp(-\epsilon_i^j VF/(RT)) \tag{16}$$

where $k_i^j(0)$ is the rate constant at zero voltage, F, R, and T are their usual biophysical meanings, V is the membrane voltage, and ϵ_i^j is the fractional charge transported during reaction step i, in the j (forward or reverse) direction. In our model, it is assumed that all the charge is transported when an ion, H^+, binds or unbinds inside the liposome. In other words, $\epsilon_4^f = 1$, $\epsilon_4^r = -1$, and all other $\epsilon = 0$.

The membrane voltage, V, is time-dependent and is approximated using an idealized capacitor model of the membrane [104, 27, 187]:

$$V_m(t) = Q(t)/C_m = \int I(t)dt/C_m \tag{17}$$

where $I(t)$ is the current of the transported ion, derived from the change in concentration of the ion inside the liposome, and C_m is the capacitance of the liposome membrane, which is known empirically. Empirical results for our model suggested a negligible effect of the membrane voltage on the rate constants and observables (see SI).

At equilibrium, the products of the forward reaction rate constants must equal the products of the reverse reaction rate constants for each step to preserve microscopic reversibility [101, 15]. This necessarily means that one rate constant is

not independent. In our implementation, we explicitly set k_6^- as the non-independent rate constant defined by the ratio of remaining forward and reverse reaction rate constants. See the SI for more details on the model specification.

3.4.3 Synthetic assay specification

We simulate three-stage SSME assays (see Fig 26) motivated by Gdx and EmrE transporter research [240, 114] to test and validate our model calibration strategies, although our implementation is flexible to accommodate other assays. In short, transporter proteins are reconstituted into liposomes which are deposited on a solid-supported membrane. The system is in a controlled external bath solution and allowed to equilibrate to a steady-state. After an equilibration phase, the external concentrations are perturbed to create a chemical gradient that drives solution exchange and transport until a new steady-state is reached. The transport of ions across the membrane creates a charge differential. Since the liposome membranes and the solid-supported membrane are coupled, the charge differential across many liposomes and transporters generates a net current that can be measured externally (see Fig 26). We refer to the guide by Bazzone et al [13] for a more in-depth discussion on SSME for transporter research.

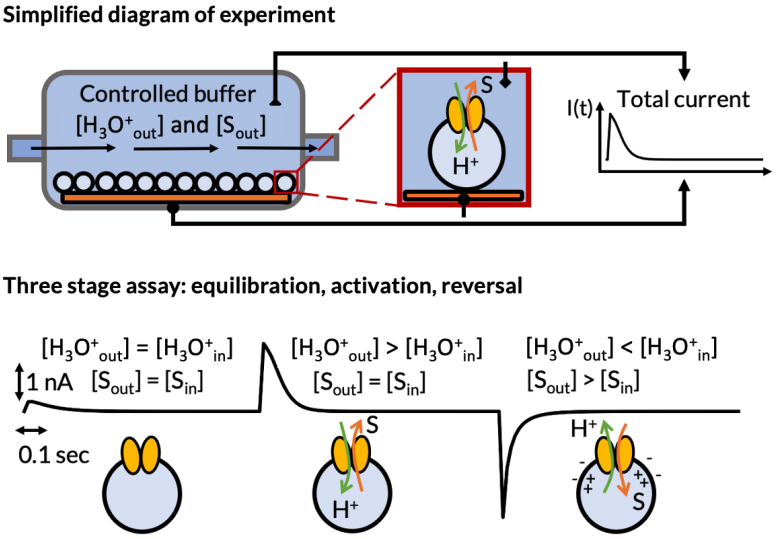


Figure 26: SSME Assay Diagram Simplified diagram of an SSME experiment with a three-stage assay (top) and observed output (bottom). A collection of liposomes with transporters is deposited on a solid-supported membrane in an external bath solution. The concentrations of the external species is perturbed after equilibration which induces a transient current via the transport of ions across a membrane.

We numerically simulate the assay by integrating the ODEs describing the system in three discrete sections corresponding to the equilibration, activation, and reversal stages at t=0, 1, and 2 seconds. During each assay stage, the external bath solution concentrations are adjusted and held at a fixed value (see concentration tables in SI). We use a stiff ODE solver (CVODES) [217] with a low tolerance to improve numerical stability. The net current is calculated from the change in internal ion concentrations of a single liposome – converting from the change in molar concentration to current, and multiplying by the number of liposomes in the experiment.

$$I_{\text{net}}(t) = I_{\text{liposome}}(t) N_{\text{liposomes}} = \left(\frac{d[H_{in}^+]}{dt} \text{Vol}_{\text{in}} N_{Av} z\right) N_{\text{liposomes}} \tag{18}$$

where $\mathrm{Vol}_{\mathrm{in}}$ is the internal volume of a single liposome embedded with transporters, N_{Av} is Avogadro's constant, z is the elementary charge of an H^+ ion, and $N_{\mathrm{liposomes}}$ is the total number of liposomes in the SSME assay. Here we assume uniformity across the aggregate of liposomes. We do not explicitly model water or the anions of the bath solution mixture, which are assumed to be controlled during SSME-like experiments to have a negligible effect on the (synthetic) observed current. Furthermore, for improved readability, we notate using protons (H^+) as the active transported ion rather than hydronium ions (H_3O^+) which are generally found in aqueous solutions. For more information on the exact assay simulation settings, see the SI.

We note that while our data generating function in Eq. 18 is relatively simple with few parameters, the underlying dynamics are determined by the set of ordinary differential equations defining the model. In particular, the change in the ion concentration $\frac{d[H^+_{in}]}{dt}$ used to generate the data is coupled to the other differential equations. These differential equations are governed by the reaction rate constants that we are estimating but are hidden from our data, so we only have partial observations of our model. So in effect, we are doing parameter estimation for latent variables describing the transporter reaction network - i.e. the ODE model.

3.4.4 Probabilistic model specification and Bayesian inference

Our probabilistic model used for Bayesian inference depends on a) a choice of priors and b) a choice of the likelihood function. For transporter biochemical reaction network models we use a log scale for the model parameters, with extremely uninformative (i.e. uniform) priors covering six orders of magnitude for the rate constants. This allows for unbiased prior assumptions but greatly increases the computational cost. We assume the noise in our data has a mean of zero with unknown variance σ^2, and that the errors are normally distributed. As such we use a Normal (i.e. Gaussian) log-likelihood function [21, 204]:

$$L(\theta, \sigma^2|D) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(D_i - D_{pred}(\theta)_i)^2 \tag{19}$$

Here our synthetic observation data, D, is generated from Eq. 18 by solving the governing ODEs with a set of known reaction rate constants, compartment volumes, number of liposomes, and initial concentrations, and then adding Gaussian white noise. The predicted data, D_{pred}, is similarly generated by using estimated reaction rate constants (θ vector) from the sampling (or optimization) process, with the Gaussian noise variance estimated as a separate nuisance parameter (σ^2). We assume a known compartment volume, number of liposomes, and the initial concentrations of transporters, ions, and substrates - unless specified otherwise (see algorithm comparison and test models below).

For Bayesian inference, we use a cutting-edge method developed for astronomy, Preconditioned Monte Carlo (PMC) [122]. This is a sequential Monte Carlo algorithm that uses multiple independent Markov chains (i.e. walkers or particles) to iteratively transition from sampling a tractable distribution (the prior) to the target distribution (the posterior).

$$p(\theta|D)^\beta \propto p(D|\theta)^\beta p(\theta) \tag{20}$$

where β=0,...,1. Initially, at β=0, the walkers are assigned parameter values based on the prior distribution. Then each walker is independently propagated using a proposal distribution. After this propagation, the walkers are assigned a weight based on their likelihood, resampled, and β is incremented. This process is continued until β=1, where the walker positions now approximate the target posterior distribution.

Sequential Monte Carlo methods have several advantages: they do not require derivative information (unlike Hamiltonian Monte Carlo methods [20]), are inherently parallelizable, can approximate multi-model distributions, and can estimate normalizing constants of the target density (i.e. model evidence in a Bayesian context) [7, 46, 53]. However, a careful choice of proposal distribution, number of walkers, and sequence of β values is required, which may be difficult due to complex distribution geometries arising from highly correlated parameters. PMC helps alleviate this issue by leveraging normalizing flows [183] via specialized autoencoder neural networks [182, 75, 148] to learn a transformation of the sampled distribution into a Standard Normal distribution.

In short, for each β stage after the initial sampling of the priors at $\beta = 0$, the samples (i.e. walker positions) from the previous stage are used to train a deep neural network that encodes a reversible transformation from the current posterior to a Normal distribution with zero mean and unit variance. After training, the Markov chain Monte Carlo sampling proceeds in the *transformed* sampling space, using the Metropolis-Hastings random walk criterion [92] with a Normal distribution with zero mean and unit variance. If the learned transformation is accurate, then the transformed posterior and MCMC proposal distribution will be approximately identical, resulting in efficiently drawn independent samples - a challenge for many complex biological models that have highly correlated parameters.

The primary gap that Bayesian methods fill as compared to standard methods such as maximum likelihood estimation is the ability to incorporate prior knowledge and the estimation of the full posterior distribution which provides detailed parameter uncertainty quantification that we can use to precisely identify mechanisms. In other words, we can get unbiased point estimates as well as robust confidence intervals for transporter reaction rate constants. However, the challenge of Bayesian methods is the convergence of the sampler due to sampling space geometry - something that is also a difficulty for optimization problems (i.e. maximum likelihood estimation). Our pipeline using preconditioned Monte Carlo combines both sequential tempering and machine learning methods (as described above) to address this gap in sampling and optimization capabilities. Furthermore, this work specifically aims to bring these tools to SSME-like data sets for active transporter research, which to our knowledge has not been done before.

3.4.5 Algorithm comparison

To validate our pipeline using precondition Monte Carlo we examine several different parameter estimation strategies for comparison. For Bayesian inference we use the affine invariant ensemble sampler (AIES) [79] as an alternative method for comparison. Briefly, AIES uses multiple MCMC chains (i.e. walkers), each sampling the posterior space via correlated jumps based on the position of other walkers and the posterior geometry. This method is well suited for multi-modal posteriors of moderate dimensions with unknown complex geometries and requires minimal tuning. We contrast these Bayesian methods using several maximum likelihood estimation [21] approaches. For these algorithms,

the goal is to minimize the negative likelihood (i.e., maximize the likelihood) using numerical optimization strategies that generate a single estimate of the parameters that are the most likely. We use ten different optimization algorithms to minimize the negative log-likelihood function: differential evolution[233], basin hopping[252], dual annealing [133], Nelder Mead [177], Powell [190], conjugate gradient [96], limited-memory Broyden–Fletcher–Goldfarb–Shanno with box constraints (L-BFGS-B) [156], constrained optimization by linear approximations (COBYLA) [191], direct search [139] and sequential least squares programming (SLSQP) [142]. These methods, while not exhaustive, provide a survey of local and global optimizers [198, 137] covering a wide breadth of algorithmic strategies, and serve as useful baseline models to compare our pipelines' performance.

Test models and data To explore the scalability of these methods with increasing complexity we use two transporter models. We first use a 1:1 antiporter model (see Fig 25) with 12 free parameters: 11 reaction rate constants (one for each reaction, minus one constraint) and an unknown noise standard deviation. Synthetic data is generated via a simulated assay (described in Methods) with Gaussian white noise added. We use reference rate constants and assay conditions motivated by Gdx and EmrE transporter studies [240, 114].

Building off the previously described 1:1 antiporter model (see Fig 25) with 12 free parameters we create another model with 16 free parameters, with 4 additional nuisance parameters: scaling factors for the initial external ion and substrate concentrations, a scaling factor for the known number of transporters in the assay, and a scaling factor for the observed net current. These additional nuisance parameters scale using a value between 0.8 to 1.2, allowing for an uncertainty of +/- 20% of their associated parameter's respective nominal value - e.g. the initial ion concentration could be adjusted by up to 20%. This increases the complexity of the probabilistic model in order to better represent the uncertainty of experimental conditions, such as the exact amount of transporter proteins present or a systemic bias in the observations. We use the same synthetic data as described previously. Both of these test models are detailed in the SI.

Simulation conditions We use the preconditioned Monte Carlo and affine invariant ensemble under their default hyperparameter settings unless otherwise noted. Three replicas are run for each Bayesian inference algorithm and model, generating a total of twelve likelihood and posterior distributions from the sampling. We use the default hyperparameters unless otherwise noted.

We conduct three independent runs from random starting points for the Bayesian inference pipeline using pocoMC [123] implementation of the preconditioned Monte Carlo sampler [122]. As described in the Methods and SI, we use broad uninformative priors and a Normal log-likelihood function for the probabilistic model. Each replica run generates 3000 samples of the posterior after convergence. For the 16D model, we adjusted the correlation coefficient γ in pocoMC [123] from its default value of 0.75 to 0.70, in order to increase the number of intermediate sampling steps (i.e. the number of β values) and improve convergence.

For AIES we also conduct three independent runs from random starting points with the same broad uninformative priors and a Normal log-likelihood function for the probabilistic model. Each replica run generates uses 1000 walkers and 10,000 steps, for 1e7 total samples - with the first 5e6 (i.e. half) of the samples removed.

Each maximum likelihood estimation algorithm was tuned using randomized hyperparameter optimization [262]. We then conducted ten independent runs from random starting points for each algorithm, resulting in one hundred maximum likelihood estimation trials.

3.4.6 Implementation

We use the Tellurium package [37] in python [244] to build human-readable systems biological markup language (SBML) files using Antimony [226] and simulate the ODEs using libroadrunner [230]. For improved reproducibility, we use a .yaml [17] configuration file to specify the relevant model and data files, as well as the simulated assay conditions and model calibration settings. Bayesian inference with preconditioned Monte Carlo is done using the pocoMC package [123], affine invariant ensemble sampling using the emcee package [79] and maximum likelihood estimation is done using the Scipy Optimize package [248]. Graphs are generated using Matplotlib [113], and numerics are done using numpy [243]. The code is available on GitHub: $\mathrm{github.com/ZuckermanLab/Bayesian\_Transporter}$

3.5 Results

To validate our Bayesian inference pipeline using preconditioned Monte Carlo we use synthetic data with known synthetic ground truth values for the model parameters. Our study examines two levels of model complexity for the same transporter protein and dataset and contrasts these findings to those from an alternative Bayesian inference algorithm (AIES) [79] and various maximum likelihood estimation algorithms. Our outcomes demonstrate the strength of our pipeline in reliably determining reaction rate constants, nuisance parameters, and their uncertainties with minimal adjustments or fine-tuning. Our approach, while computationally expensive, finds the most likely models of all the methods tried, in addition to generating a posterior with better convergence performance than an alternative Bayesian method. Finally, we find that for our model and dataset, only a limited number of model parameters can be precisely and practically identified.

3.5.1 Log-Likelihood Comparisons

We begin our comparison by examining the log-likelihoods generated by the MLE algorithm replicas and Bayesian inference replicas (Fig. 27). As compared to the reference log-likelihood generated from known reference values, the MLE values are generally many orders of magnitude lower than the reference value which suggests poor convergence by the MLE algorithms. We also observe a large variance in the parameter estimates between most algorithms and between replicas for a given algorithm - much lower than the range of estimates generated from our pipeline with PMC. The exception is the differential evolution algorithm which is consistently close to the reference value. We note that

some of the MLE algorithms perform better for the more complex model, which is likely an artifact from the randomized hyperparameters used across models.

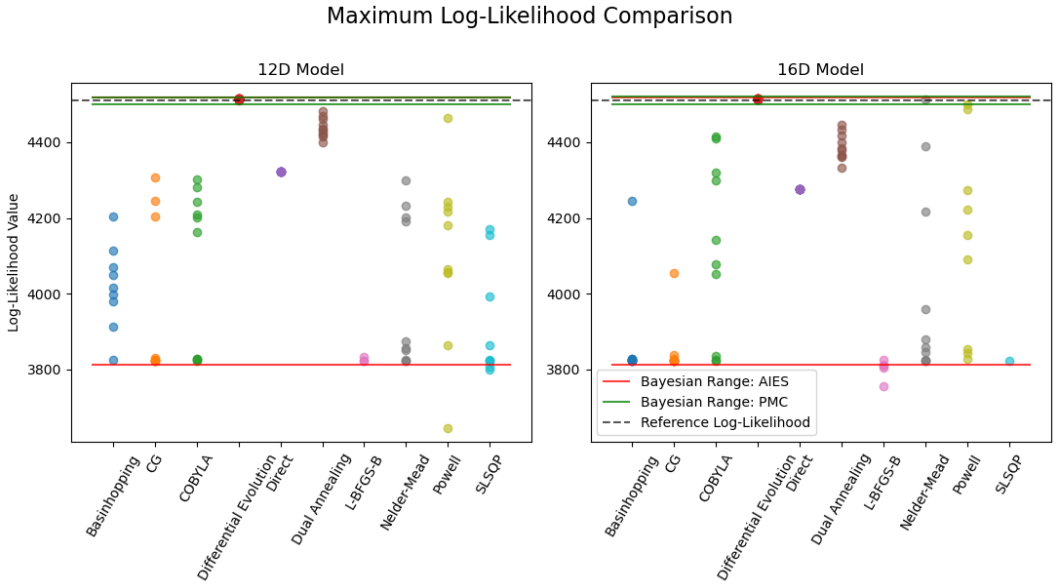


Figure 27: Maximum Log-Likelihood Comparison - MLE The maximum values found for each MLE algorithm and replica. The reference log-likelihood using the known synthetic model parameters is shown as a dashed line. Also, the Bayesian likelihood minimum and maximum values are shown for comparison. Most of the MLE algorithms fail to consistently estimate values near the reference value, or within the bounds of our PMC pipeline. The exception is the differential evolution algorithm. We note the wide range of the AIES due to poor sampling convergence.

Next, we investigate the log-likelihood distributions generated from Bayesian inference methods (Fig. 28) and compare that to the best log-likelihood estimates from each algorithm. We find that AIES and PMC find higher log-likelihood estimates than any of the MLE methods, and that the differential evolution algorithm performs the best from the MLE methods. The AIES algorithm appears to be partially stuck in local minimal as there are multiple modes at low log-likelihood values that correspond to MLE values from other algorithms such as L-BFGS-B and Basin-hopping.

3.5.2 Predicted Current Traces

These results can be further explored through predictive analysis (Fig. 29 and Fig. 30). Here we take the parameter estimates from a given algorithm and plot the predicted SSME-like current trace (see Methods). These traces can be compared against the synthetic observed data to check for goodness of fit. We plot predicted traces for each MLE algorithm and replica. As compared to the observed data, most MLE algorithms did not consistently generate well-fitting curves. While differential evolution had the best fits, direct search and dual annealing also had fairly close fits compared to the remaining algorithms. We also note that L-BFGS-B, failed to have a single replica with a close fit to the data. These results are consistent with the log-likelihoods shown previously.

For the Bayesian inference methods, we do predictive posterior analysis. Here we select 100 random samples from the posterior corresponding to 100 model parameters (i.e. rate constants) and generate predicted current traces. AIES does not consistently predict well-fitting curves from its posterior but our method utilizing PMC does. Amongst all of the

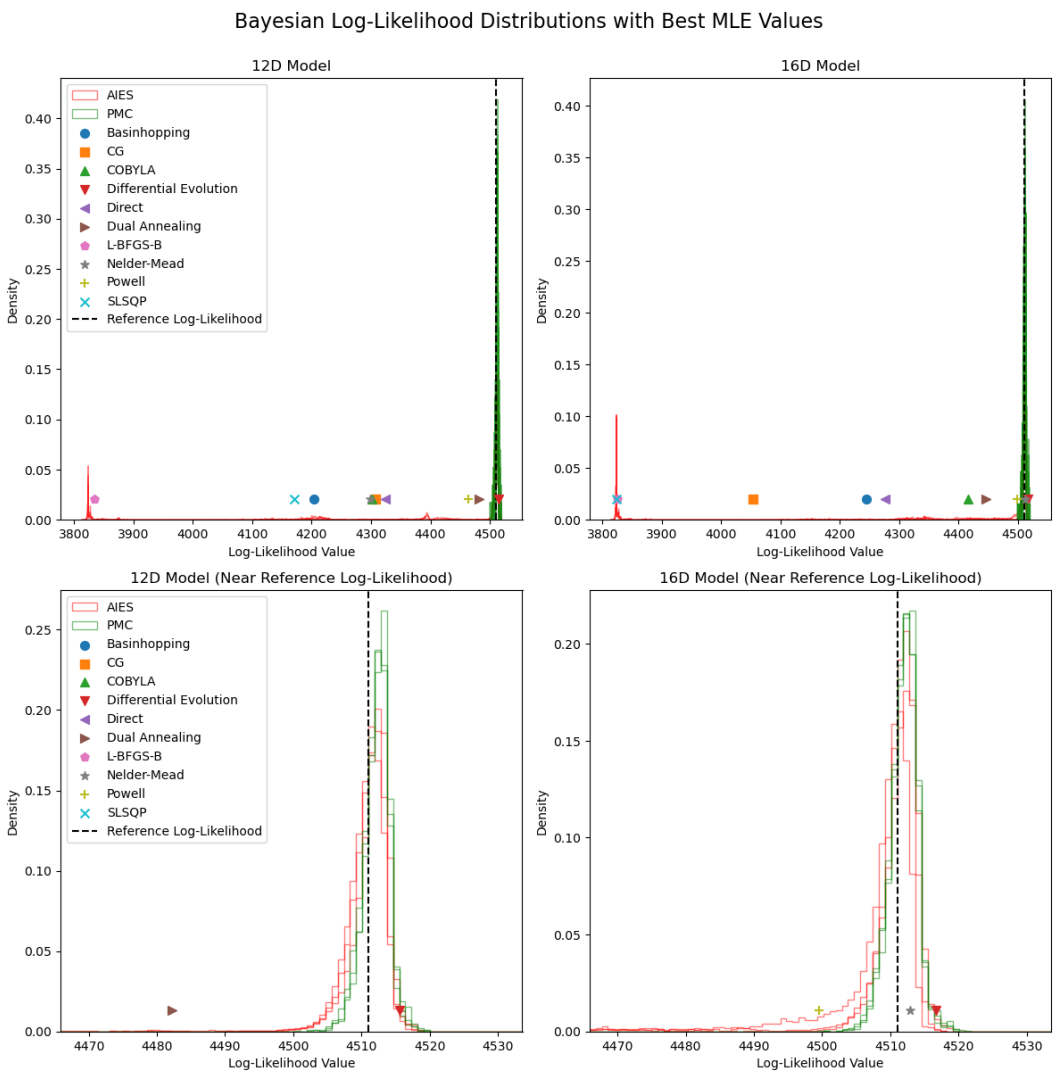


Figure 28: Log-Likelihood Comparison - Bayes The log-likelihood distributions of each Bayesian algorithm and replica are shown, with the full range on the top row, and a reduced x-scale on the bottom row. The reference log-likelihood is shown as a dashed line, with markers denoting the best MLE estimates overlaid on the distributions. The AIES algorithm appears to be stuck in local minimal as there are multiple modes at low log-likelihood values. Also, the Bayesian inference methods find higher log-likelihood values than the MLE methods, with PMC sampling the maximum value.

MLE and Bayesian methods, differential evolution and PMC consistently generate predicted data sets that fit the data well across both models.

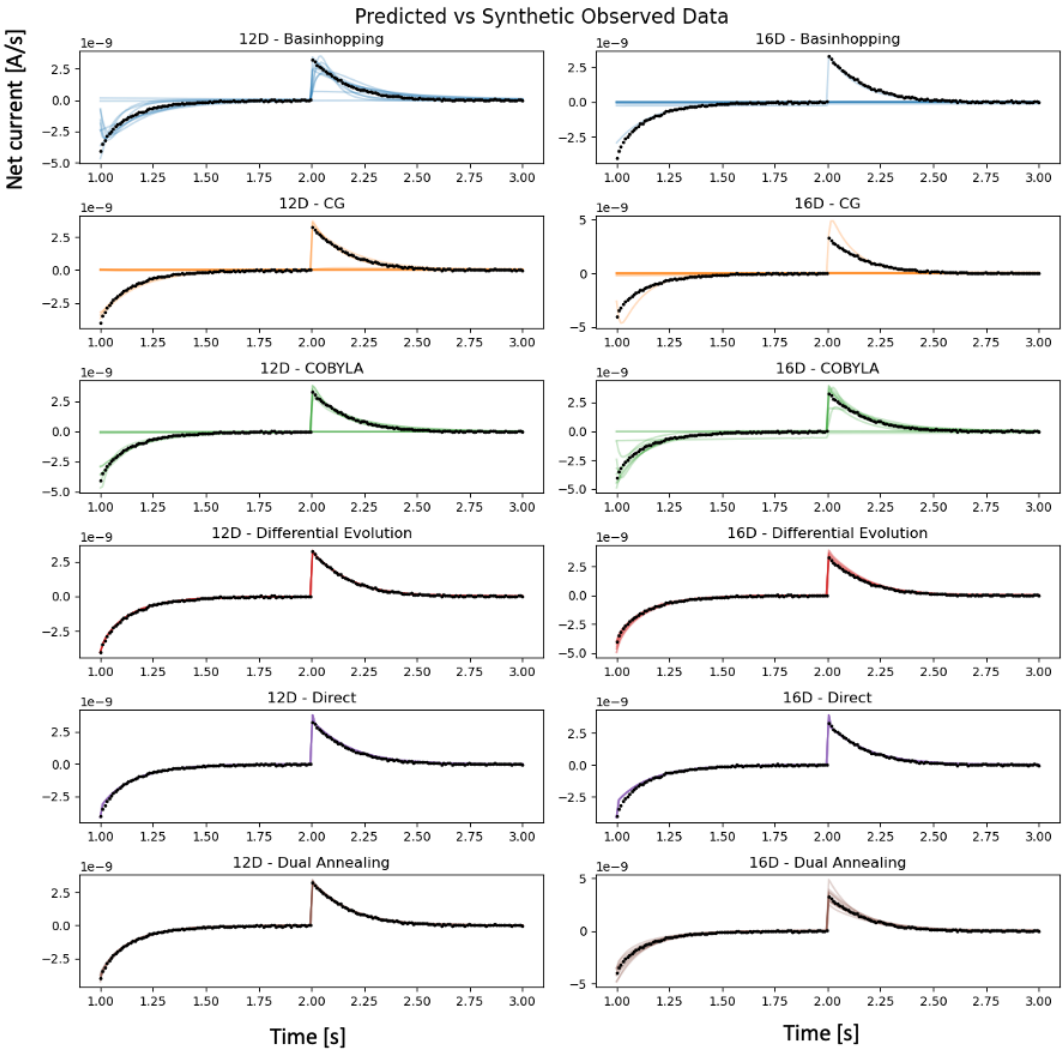


Figure 29: Predicted Current Traces I The predicted current traces using the parameter sets from each MLE replica run plotted against the synthetic observed data (dots), for selected algorithms. Differential evolution and direct algorithms generate closely fit traces across both models.

3.5.3 Computational Cost

In addition to evaluating the accuracy and consistency of each algorithm, we also examine the computational cost. We track the wall-clock run time in seconds (per replica), the number of log-likelihood evaluations used during the algorithm per replica, and the average number of log-likelihood evaluations per second (per replica). This is shown in Fig. 31. As expected the Bayesian inference methods incur the highest cost both in terms of run time and number of calculations used. The third most expensive algorithm was differential evolution, which used 1-2 orders of magnitude fewer log-likelihood calculations and wall clock time. However, we note that the Bayesian inference methods generate

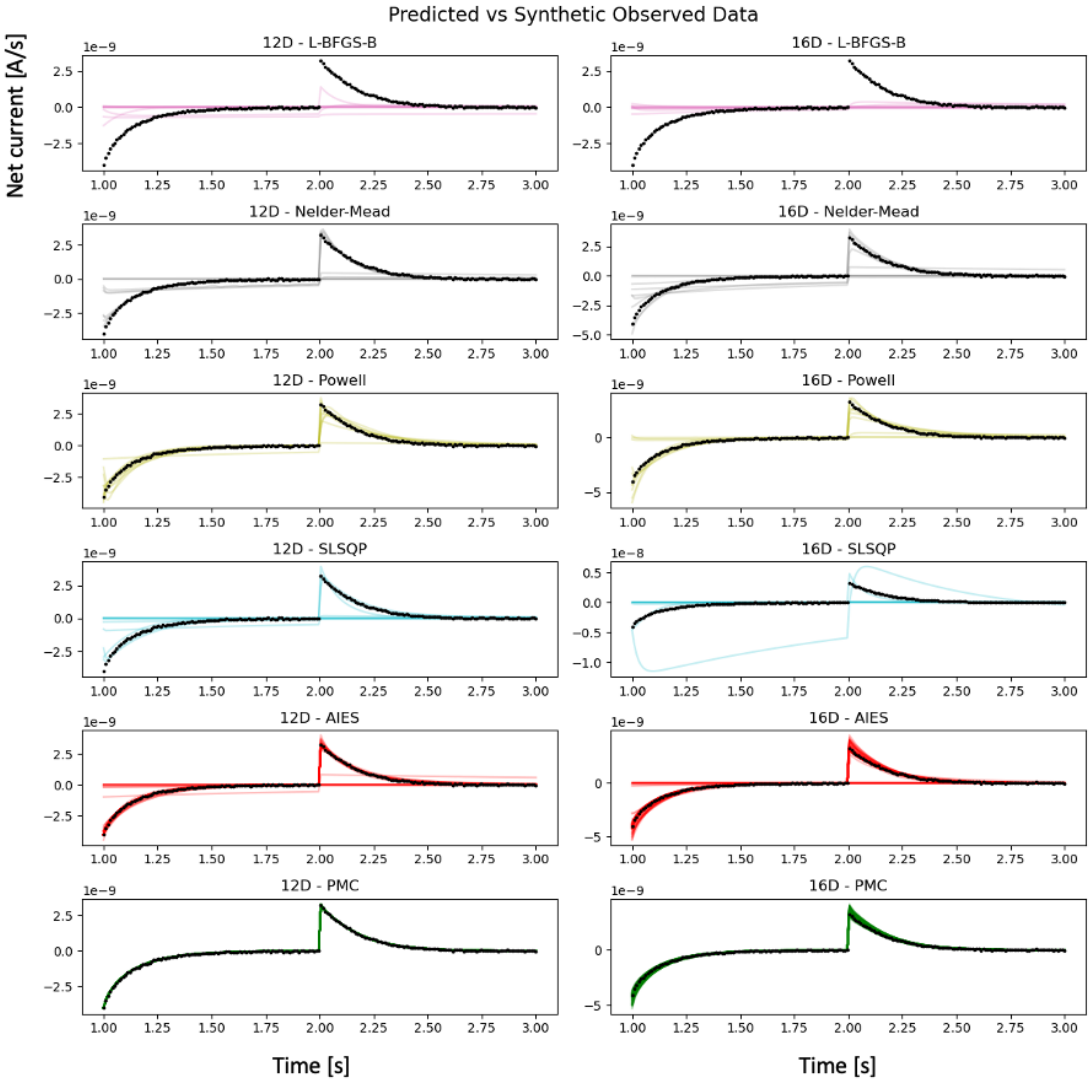


Figure 30: Predicted Current Traces II The predicted current traces using the parameter sets from each MLE replica run plotted against the synthetic observed data (dots), for the remaining algorithms. For the Bayesian inference algorithms, 100 random parameter sets are chosen from the posterior and used to plot the current trace. Our pipeline using PMC generates closely fit parameter sets across both models.

a full posterior that contains information for both parameter estimates as well as uncertainty quantification and parameter correlations. Furthermore, we find that the Bayesian inference methods have a comparable efficiency (log-likelihood calculations per second) as differential evolution.

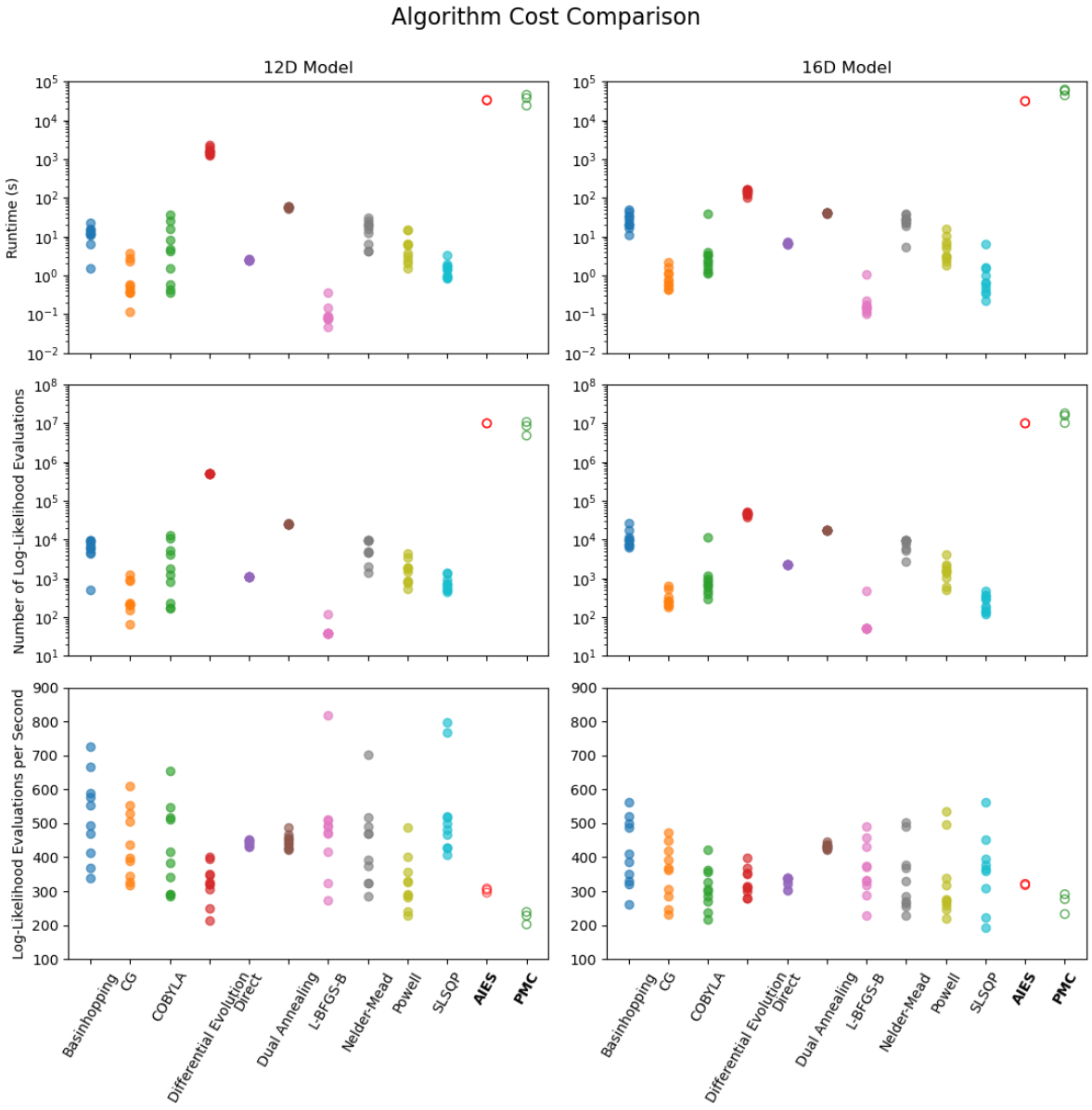


Figure 31: Computational Cost Comparison The run time (top row), number of log-likelihood evaluations (middle row), and number of log-likelihood evaluations per second (bottom row) are shown for all the algorithms tested. The Bayesian inference methods AIES and PMC have the longest run time and number of log-likelihood evaluations, followed by differential evolution. We note that the Bayesian inference methods do not only estimate parameters, but also generate a full posterior distribution, and have a similar efficiency (log-likelihood calculations/sec as differential evolution).

3.5.4 Bayesian Marginal Posterior Analysis

Finally, we examine the marginalized posterior distributions generated from our Bayesian inference replica runs for both the 12D (Fig. 32) and 16D (Fig. 32) models. Unlike MLE methods, Bayesian inference generates a posterior

which contains detailed uncertainty and correlation information of the model parameters. This posterior can be leveraged via marginalization to determine parameter means, modes, variances, and confidence intervals. For the 12D model, we find the PMC replicas are all in agreement suggesting convergence. On the other hand, the AIES replicas do not appear to be fully converged as there is a divergence between them for certain parameters (i.e. $\log10k_5^r$). However, we do find a general agreement between both of the Bayesian methods, and they both cover the synthetic reference values. For this data set and model, only one rate constant (k_2^f) is precisely determined within less than one order of magnitude, with a few other rate constants within 2-3 orders of magnitude (k_2^r and k_5^f), and the rest covering many orders of magnitude. The extremely broad distributions covering many orders of magnitude make several parameters effectively unidentifiable for biophysical purposes (k_4^f and k_4^r).

For the 16D model, we find that as with the 12D model, all three PMC replicas are in agreement (suggesting convergence), but not all three of the AIES replicas are in agreement (as shown for $\log10k_2^r$, $\log10k_3^r$, and $\log10k_5^r$ parameters). Also, we see that both Bayesian inference algorithms generate marginal distributions that cover similar ranges, including the synthetic reference values. As before, for the given data set and model, only one rate constant (k_2^f) is precisely determined within less than one order of magnitude, with a few other rate constants within 2-3 orders of magnitude (k_2^r and k_5^f), and the rest covering many orders of magnitude. Due to the increased uncertainty, the rate constants distributions are broadened when compared to the results for the 12D model. Examples of this are $log_{10}k_3^f$ and $log_{10}k_3^r$ which widen. The scaling factor nuisance parameters have extremely wide distributions which suggest they are insensitive - e.g. do not change the model predictions when varied.

3.6 Discussion

Overall we have seen that our Bayesian inference pipeline can reliably converge on the posterior in order to determine the biophysical properties of a biochemical reaction network and characterize uncertainty in the data. While related work has been done using Bayesian inference in systems biology contexts and ion channels, we introduce a novel application for solid-supported electrophysiology experiments. In addition, in our implementation we utilize a machine learning accelerated inference algorithm and prioritize reproducibility and compatibility with existing systems biology software tools. We also enforce broad and unbiased priors to allow for equal consideration of all physically realizable rate constants. In order to emphasize the importance of method validation when performing parameter estimation, we use synthetic data and carefully check for convergence of the algorithms used. Finally, we find that our Bayesian inference pipeline utilizing preconditioned Monte Carlo can reliably converge while maintaining a similar efficiency as the best-performing MLE algorithm of differential evolution.

First, we developed a Python pipeline, combining existing systems biology modeling tools with a state-of-the-art Bayesian inference algorithm and commonly used maximum likelihood estimation algorithms. We use SBML to encode the biochemical reaction network which enables the transfer of models to different modeling tools. We emphasize using a high-performing and robust ODE integrator (such as Libroadrunner) to solve systems biology problems which are often stiff non-linear systems that prove challenging to integrate without special consideration. Our

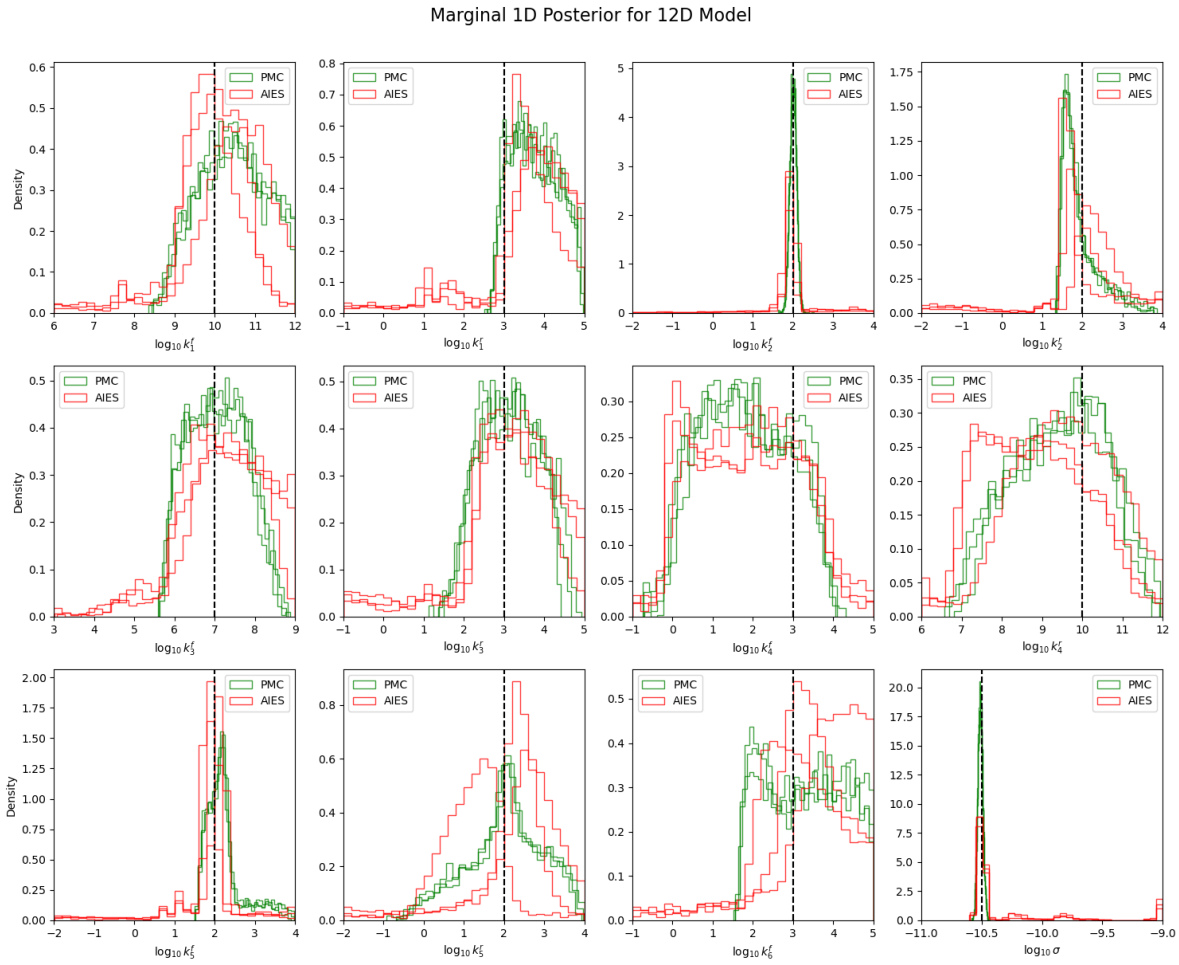


Figure 32: Marginal Posteriors for 12D Model The 1D marginal posteriors are shown for each replica run for both the PMC and AIES Bayesian methods. The posteriors from both methods cover the synthetic reference values (dashed lines) and mostly agree with each other. However, the AIES replica distributions have increased variance between themselves (i.e. $\log 10 k_5^r$) which suggests that the sampler hasn't fully converged. In contrast, all of the PMC replicas are in agreement with each other.

implementation also enables the user definition of the SSME assay protocol, ODE simulation settings, as well as Bayesian inference and maximum likelihood settings. This configuration can be further developed to integrate with developing parameter estimation standards and methods such as PEtab [209] and pyPESTO [208]. While the focus of the pipeline has been using the preconditioned Monte Carlo algorithm for Bayesian inference, it is easily extendable for use with other Bayesian inference methods such as the affine invariant ensemble sampler [79], or maximum likelihood estimation methods - as used in this study. Finally, while we focused on SSME and membrane transporters, our pipeline could be expanded to work with traditional electrophysiology experiments and ion channels.

Next, we examined a 12D and 16D transporter model and found that both Bayesian inference with PMC and differential evolution are able to reliably estimate the rate constant parameters, yielding predictions that closely fit the observed data across multiple samples and replicas. These methods generated log-likelihoods near the maximum value consistently, although we note that both PMC and AIES log-likelihood distributions contained the largest log-likelihood values of all the algorithms. However, the other MLE algorithms tested as well as AIES failed to reliably converge to

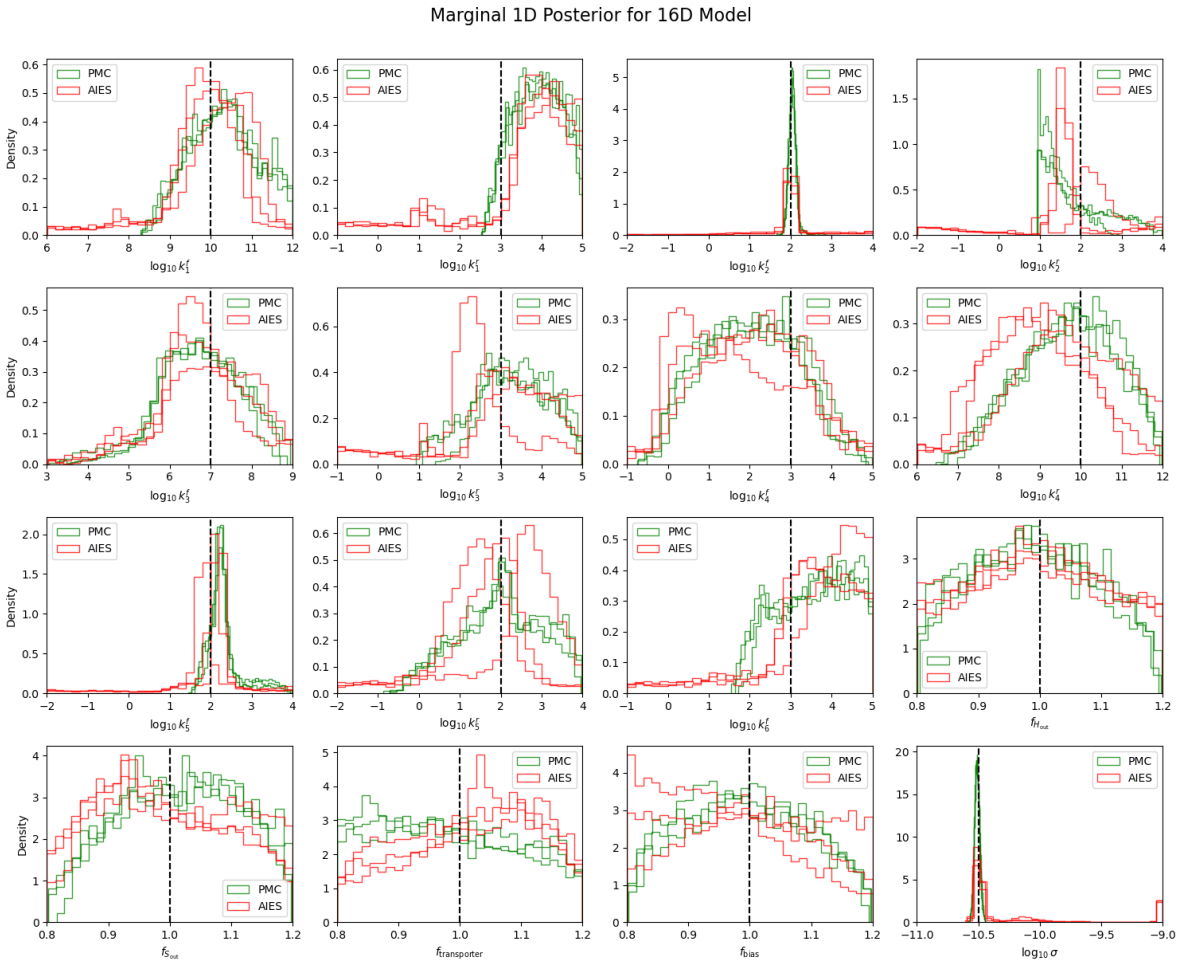


Figure 33: Marginal Posteriors for 16D Model The 1D marginal posteriors are shown for each replica run for both the PMC and AIES Bayesian methods. The posteriors from both methods cover the synthetic reference values (dashed lines) and mostly agree with each other. However, the AIES replica distributions have increased variance between themselves (i.e. $\log 10 k_2^r$, $\log 10 k_3^r$, and $\log 10 k_5^r$) which suggests that the sampler hasn't fully converged. In contrast, all of the PMC replicas are in agreement with each other.

the maximum log-likelihood values and generate accurate predictions. The main point here is not that these methods are inherently worse than our pipeline using PMC, but rather that a careful choice of algorithm and hyperparameters is needed to ensure meaningful results, and that these results should be validated for convergence - this is true whether using Bayesian inference or maximum likelihood estimation. In particular, for typical problems in systems biology, many 'out-of-the-box' solutions, as we have demonstrated with AIES and other MLE algorithms such as L-BFGS-B and SLSQP. We suggest that any model calibration should include a validation step to ensure convergence, and that when possible multiple algorithms and replicas should be examined.

In addition, the computational cost of the algorithm used is important to consider when doing parameter estimation and uncertainty quantification. Here we examined both the run time and number of function (log-likelihood) evaluations used. The Bayesian inference methods were the most expensive, 1-2 orders of magnitude greater than differential evolution, but this cost enables the estimation of a posterior distribution which is much richer with information than the

point estimates provided by maximum likelihood estimation. Similarly, we find that both differential evolution and Bayesian methods have a comparable efficiency in terms of number of log-likelihood calculations per second. Based on the results of this work, we suggest differential evolution for limited computational budgets, or if only a single most likely parameter estimate is needed. However, for larger computational budgets, when robust parameter uncertainties or correlations are needed along with parameter estimates, we suggest the preconditioned Monte Carlo method.

We note that there are many alternative MLE and Bayesian inference algorithms, but it was outside the scope of this work to perform an exhaustive survey. We used MLE algorithms found in the standard Scipy library as a representative set, but this does not include a particle swarm optimization algorithm which may perform better for non-linear and non-convex problems [130, 247]. Similarly, we did not examine approximate Bayesian computation [43, 16, 234] or variational inference [80, 181] methods for Bayesian inference which may provide better performance and scaling than the preconditioned Monte Carlo algorithm. During the development of the pipeline, Markov chain Monte Carlo methods such as the No-U-Turn sampler [105] were found empirically to have poor convergence. Future work will examine these alternative model calibration strategies in more depth.

Our results indicate that our pipeline can reliably converge and provide estimates for model parameters and their uncertainty using Bayesian inference with preconditioned Monte Carlo. For the validation data, we found that the several parameter estimate variances increased with the model complexity, and that only a limited number of reaction rate constants could be identified within a narrow biophysical range - i.e. were practically identifiable. This suggests that our data does not have sufficient information to adequately characterize the reaction rate constants of our model. Potentially, data from other experimental methods like nuclear magnetic resonance [94] could be used to generate more informative priors and improve the practical identifiability of the rate constants. Further, this raises an interesting question of what, if any, experimental SSME protocol could be used to improve the identifiability of the rate constant estimates. And further, how can we determine if our model is correct to begin with? One advantage of using Bayesian inference is that the posterior contains a comprehensive description of the model and data uncertainty (and information content) which can be leveraged to explore these questions. We aim to further investigate this matter in future work.

In the future, we aim to incorporate experimental data from a range of experimental conditions and transporter proteins. This will require the development of more complex models to better capture sophisticated transport processes and experimental conditions. While we used a tightly coupled 1:1 cycle model, many transporters of interest exhibit multiple possible reaction pathways in a loosely coupled 2:1 cycle which increases the dimensionality of the computational model. Experimental data will likely also include additional sources of uncertainty and artifacts that are not seen in the data. In particular, issues of transporter polarity and uniformity, time delays from the mechanics of fluid mixing, and membrane capacitive coupling should be closely examined [13]. Despite these potential challenges, our results are a promising first step toward a complete modeling and analysis platform for transporter research.

3.7 Conclusion

In conclusion, we have developed a pipeline incorporating Bayesian inference and maximum likelihood estimation methods, motivated by solid-supported membrane electrophysiology experiments used for transporter research. Under the conditions studied, we found that the preconditioned Monte Carlo and differential evolution algorithms had superior convergence performance. Preconditioned Monte Carlo generated the maximum log-likelihood among all the algorithms, along with full posterior distribution for a comparable relative efficiency as differential evolution but higher absolute cost. We emphasize that any algorithm selected for parameter estimation should be carefully validated to ensure meaningful results and reliable model predictions. Our initial findings suggest a relatively low amount of information in the synthetic SSME data which yields a large variance for most rate constant estimates. Future work will focus on exploring strategies to reduce the uncertainty in model parameters, as well as methodological improvements for Bayesian inference to enable better scalability. Finally, our insights inform future model calibration studies within the broader systems biology field, and our pipeline provides a versatile tool to aid in transporter research - an essential biophysical process.

3.8 Supporting information

See the attached appendices for the supporting information on this chapter.

S1 Text Detailed method specifications

S2 Text and Fig Electrogenic characterization of simulated SSME assay

S3 Additional figures MLE Tuning and Algorithm Performance

4 Model Selection and Experiment Recommendation for Transporters

Model Selection and Experiment Recommendation for Transporters

August George[1], Daniel M. Zuckerman[1*],

1 Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon, USA

* zuckermd@ohsu.edu

4.1 Abstract

This study develops tools to address gaps in our knowledge in the mechanistic diversity of transporters, such as the multi-drug resistance transporters, including EmrE, that exhibit impressive adaptability and are of notable biomedical interest for therapeutics. Despite their significance, the exact characterization of transporter proteins is an ongoing challenge. Here we utilize data motivated by solid supporting membrane-based electrophysiology, an important biophysical technique that generates high signal dynamical data, and apply information theoretical and Bayesian methods to quantify the information contained within these datasets. Leveraging the Kullback-Leibler divergence between Bayesian posterior and prior distributions, we provide insight into the information contained in SSME-type datasets. We extend this approach to compare the information content of several different assay protocols, providing a ranking of the most informative assay conditions, which is used to perform model selection against alternative transporter mechanisms. Our results suggest that assays containing large perturbations and their combinations provide the most information and are able to distinguish between potential transport cycle mechanisms. The findings support the utility of Bayesian inference as a robust framework for uncertainty, offer recommendations for experimental design, and suggest how SSME may be able to fully determine the reaction pathways of transporters.

4.2 Author summary

In our study, we focus on the complexity of transporter proteins, such as multidrug resistance transporters. These transporters are able to remove a wide range of drugs from inside the cell and are implicated with resistance to chemotherapeutics and antibiotics. Our challenge is to try and figure out exactly how these proteins work, particularly since their behavior can significantly vary. To address this we utilize concepts from information theory and Bayesian inference, along with (simulated) data from solid-supported membrane electrophysiology, to determine how much useful information can be derived from this data. In particular, we examined how the amount of information changed across different simulated experiments. We found that experiments that generated a large signal, or combinations of different experiments, provided the most information to aid with the identification of the mechanism. These results help guide the design of future experiments and how these advanced statistical methods can better understand important biological systems.

4.3 Introduction

Transporters are a type of molecular machine that helps regulate cellular homeostasis by pumping substances across a membrane and maintaining ion gradients [5, 213]. As such these transporters play in essential role in the uptake of nutrients and expelling of waste, among other roles. These proteins operate in a stochastic molecular environment which suggests they exhibit some degree of stochasticity in their mechanism, a concept recently emerging as 'pathway heterogeneity'. In essence, evidence suggests that these proteins may exhibit more complexity than previously thought [239, 200, 95, 132]. In particular, multidrug resistance transporters present the remarkable ability to export a wide range of drugs that are structurally dissimilar. These proteins are found across a wide range of organisms, from humans to bacteria. In humans, the multidrug resistance transporters such as the P-gp [199, 119] protein exports a wide range of substances due to a low specificity [140]. However, this allows for the export of chemotherapeutic drugs when in a tumor and cancer environment. As such these proteins are of significant biomedical interest as a therapeutic drug target. Similarly in bacteria, the EmrE protein [211] is part of the small multidrug resistance family of transporters. Studies have shown that this protein could potentially exhibit a wide range of mechanistic behaviors including 2:1 and 1:1 transport stoichiometry [86, 200]. The low specificity of these proteins suggests an adaptability which may be a result of pathway heterogeneity. These multidrug resistance transporters are clearly of biomedical significance, but their exact mechanisms are poorly understood and are challenging to observe directly due to the challenges in directly observing the transport process at sufficiently high time and length scales.

An emerging method, solid-supported membrane electrophysiology (SSME) [12, 214] generates dynamic information at a high time resolution but without any spatial resolution regarding the protein conformational states during transport. Three-stage reversal assays [240] provide transient current traces across an aggregate of transporters, and can be used to determine select kinetic properties [12]. However, it is unknown how well SSME can recapture microscopic information from the generated macroscopic dataset that includes noise and sparsity. In other words, it is not known how much information is contained in these datasets, which is an important gap in knowledge that we aim to address in this study.

To investigate the information content of these datasets, information measures of some kind are needed. Information quantification has roots in information theory, stemming back to Shannon's seminal work relating entropy to information [219]. Originally developed for telecommunications and signal processing applications, a host of quantitative measures have been developed that quantify the information contained in a noisy signal [78]. This includes entropy, mutual information, Kullback-Leibler divergence (KL divergence), and cross-entropy measures. Briefly, entropy serves as a fundamental measure of information, with high entropy suggesting less predictability and therefore more information. Mutual information [143] measures how reducing uncertainty in one variable reduces the information of the other variable. KL divergence measures the dissimilarity between a reference and non-reference distribution, and the related cross-entropy which is the KL divergence [158] plus entropy of the reference distribution. These methods provide different, but related insight into the information content in a signal, and the choice of information quantification strategy depends on the specific needs of the problem at hand.

Relatedly, biological modeling has continued to adopt these ideas in various capacities to study a wide range of systems. Entropy has been used to quantify the complexity of coding and non-coding DNA sequences [89] and for protein structural analysis [171]. Mutual information has been used to study the information in the cardiorespiratory system [188] and gene regulatory networks [149]. KL divergence has been used for evolutionary studies [225] and comparative genomics [253]. And cross-entropy has been used in genomics [29] and protein prediction [193] problems. These selected examples demonstrate the wide breadth of application that these quantitative approaches have within computational biology. However, to our knowledge, these techniques have not been applied to secondary active transport proteins (i.e. EmrE) in the context of SSME experiments and Bayesian inference.

Furthermore, the information quantity of a dataset provides important insight into the design of experiments. Here information criteria like KL divergence can be used to determine how much uncertainty can be reduced from the experiment[146]. This can then be used as part of a larger model calibration pipeline to infer model parameters with greater precision, yielding better predictive accuracy in the associated model [90]. This approach can be done 'online', picking the optimal experiment one at a time, or 'offline' to provide a ranking of experiments [218, 245]. The improved precision from optimal experiment design also relates to the potential multiple reaction pathways that some transporters are thought to exhibit. With more informative datasets, the estimated model parameter variances will be lower which can be leveraged to potentially distinguish between competing mechanistic models.

Motivated by the gap in the knowledge about these multidrug transporters and electrophysiology datasets, as well as recent the use of posterior information analysis in related systems biology problems [110], we extend our previous work on Bayesian inference for SSME data and transporter research. The Bayesian inference paradigm naturally fits with the notion of information quantification via the prior and posterior which effectively describe the information contained in the data set. More specifically, we use the KL divergence between the prior and posterior distributions to quantify the information from different synthetic SSME datasets in order to provide a ranking of optimal experiment protocols. Using these datasets we then use a Bayesian model selection strategy to compare four idealized 1:1 transport cycles.

We find that the Bayesian inference pipeline is well suited for this task, efficiently generating posterior distributions across the various models and synthetic assays studied. The experiment optimization ranking found assays with large perturbations as well as combinations of different assays to contain the most information. Also, we found that high-information datasets could distinguish between the four possible mechanistic models, whereas low-information datasets could not. These results underpin the utility of Bayesian inference and information theory in studying transporter mechanisms, giving insight into optimal experimental design, and suggesting how SSME may be able to fully determine the reaction pathways of multidrug resistance proteins.

4.4 Materials and methods

4.4.1 Quantifying Information Content and Experiment Recommendation

Bayesian Inference Method We use a Bayesian inference pipeline to model transporters in SSME experiments, as explained in previous work. Briefly, SSME experiments[12] consist of many proteoliposomes with reconstituted proteins deposited on a solid membrane. This system is placed in a bath of chemical species and when perturbed creates a gradient that drives ion transport across the membrane that is measured. After an initial equilibration stage, the concentrations are perturbed, letting the system relax to a new steady-state condition. Then the external concentrations are adjusted back to the initial values, switching the gradient and driving transport in the opposite direction. This is the three-stage reversal assay. An important note is that due to the stability of SSME [12] multiple assays can be done in sequence under different perturbation amounts - an important consideration of this work. Figure 34 shows an idealized diagram of an SSME experiment.

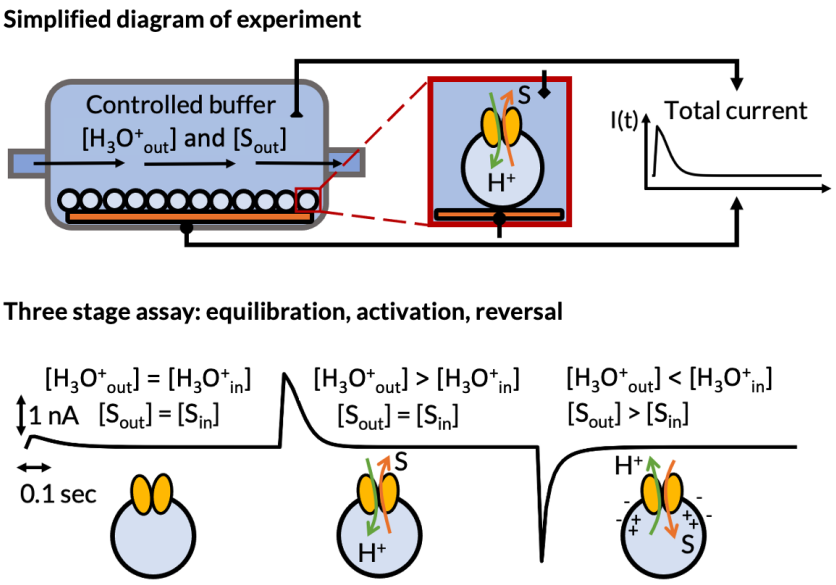


Figure 34: A Simplified Diagram of SSME Experiments The top panel illustrates the apparatus and experimental setup used for SSME, where transporters embedded on liposomes transfer ions under a gradient, inducing a current. The bottom panel shows a three-stage reversal assay. The concentrations of the external species are perturbed after equilibration which induces a transient current via the transport of ions across a membrane.

Ordinary differential equations are used to model the dynamics of a biochemical network describing the transporter system [74] (see SI). Our implementation uses Systems Biology Markup Language (SBML) [111, 127] to encode the model and its governing equations, and simulation is done using Libroadrunner [230]. The three-stage SSME assay is simulated by adjusting external concentrations of the transporter SBML at discrete time points reflecting the different external bath perturbations typical of a reversal assay. The integration of the ordinary differential equation results in concentrations over time, where the rate of change of the transported ion on a single liposome is converted into a net current, using a uniformity assumption for the liposomes.

This process generates predicted data which can then be used in the Bayesian pipeline to estimate the probability of the model parameters given the data, as given by Bayes' theorem:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} \tag{21}$$

where:

$P(\theta|D)$ is the posterior probability of the model parameters θ given the observed data D.

$P(D|\theta)$ is the likelihood of observing the data D given the model parameters θ.

$P(\theta)$ is the prior probability of the model parameters θ.

$P(D)$ is the marginal probability of the data D.

As in our previous work, we use a Normal log-likelihood function and wide priors spanning several size orders of magnitude to reduce bias:

$$L(\theta, \sigma^2|D) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(D_i - D_{pred}(\theta)_i)^2 \tag{22}$$

We use the preconditioned Monte Carlo sampler implementation pocoMC [123] which is accelerated via machine learning [122].

Information Quantification Information is quantified using KL divergence [146], which for discrete values is defined as:

$$D_{KL}(P||Q) = \sum P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{23}$$

where:

$D_{KL}(P||Q)$ is the Kullback-Leibler divergence from Q to P,

$P(x)$ is the probability distribution for P

$Q(x)$ is the probability distribution for Q

When estimated with a large number of samples this becomes :

$$D_{KL}(P||Q) \approx \frac{1}{N}\sum \log\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{24}$$

where:

$D_{KL}(P||Q)$ is the Kullback-Leibler divergence from Q to P,

$P(x_i)$ is the probability of the i-th sample under P,

$Q(x_i)$ is the probability of the i-th sample under Q,

N is the total number of samples.

In our Bayesian context, we let $P(x_i)$ be the posterior distribution and $Q(x_i)$ be the prior distribution, giving the divergence between our updated beliefs after data has been observed, to our prior beliefs before data was observed. We assign this divergence as our information score criteria to rank different data sets.

An important consideration is that the number of samples generated by the preconditioned Monte Carlo method may not be large enough to satisfy the large number approximation of the KL divergence. To overcome the issue of low sample number and the 'noise' in the posterior distribution, we utilize a Gaussian mixture model (GMM) [197] to estimate a smooth approximation of the estimate posterior.

GMMs are a probabilistic model that aim to fit a collection of weighted multidimensional Gaussians to the target (in this case posterior) distribution. They generate a smooth analytical function that can easily generate a large number of samples, or be used directly for calculations.

$$p(\mathbf{x}) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{25}$$

where:

- $\mathbf{x}$ is the data vector.
- π_i is the mixture weight for the i-th Gaussian, and all weights sum to 1.
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the i-th Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$.
- K is the total number of Gaussian distributions in the mixture model.

A key hyper-parameter governing GMMs is the number of Gaussians, K. We use an elbow plot method [138] to select K. Here we iterate from small to larger K values, calculate an information criteria value that balances the goodness of fit against the number of parameters, and stop when a minimum has been reached.

For this study, we used the Bayesian Information Criterion (BIC)[176]:

$$BIC = \ln(n)k - 2\ln(\hat{L}) \tag{26}$$

where n is the number of observations, $\hat{L}$ is the computed likelihood and k is the number of free parameters in the model.

This entire process of information quantification is outlined in figure 35 below.

Experiment Recommendation Building off the information quantification method described above, we implement an experiment recommendation system. Multiple assays can be simulated (or performed via experiment) generating multiple datasets, under different assay conditions. For each dataset, Bayesian inference is performed and the posterior is estimated. The information score from the KL divergence is calculated for that particular dataset (and model). The scores from each experiment are then ranked, giving the assay protocols with the most informative data. This is illustrated in Figure 36.

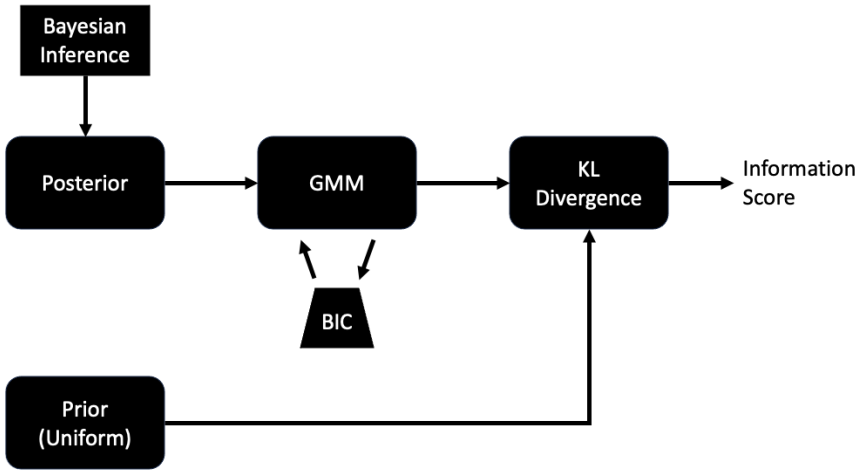


Figure 35: Information Quantification Workflow Bayesian inference generates an estimated posterior which is approximated using a Gaussian mixture model (GMM). The number of Gaussians is optimized using the Bayesian information criterion that penalizes over-fitting. The GMM, representing the posterior, can then calculate the KL divergence with the prior distribution.

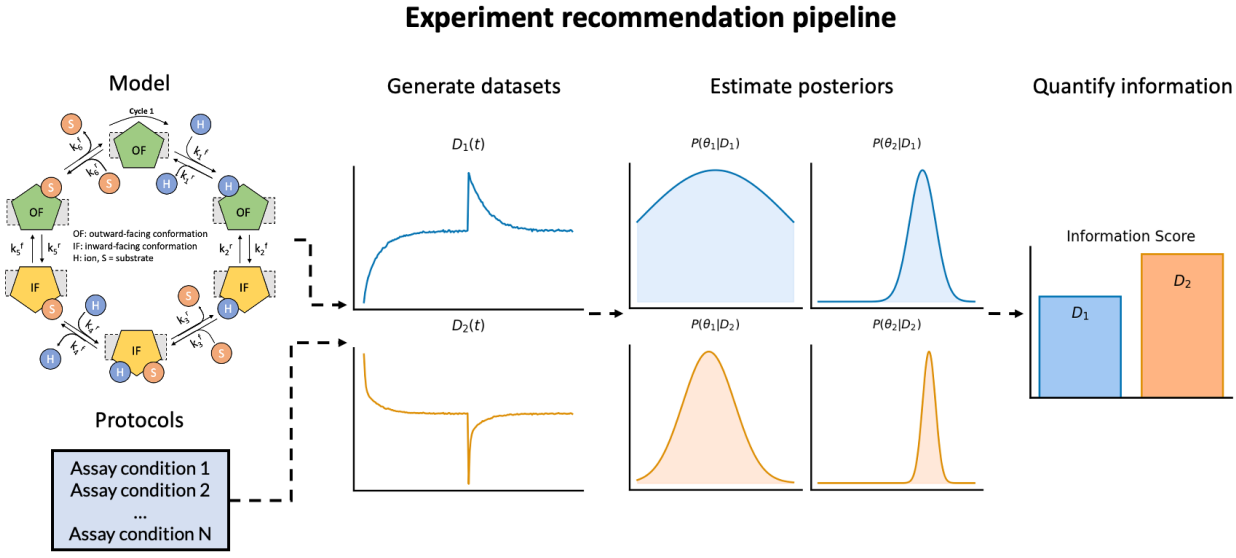


Figure 36: Pipeline for Experiment Recommendation Different datasets are generated which result in different posterior distributions, that in turn give different information scores. This approach enables the ranking of datasets and assay protocols based on their information quality.

To validate this approach, we compare eight different experimental protocols corresponding to four different perturbation conditions (experiments 1-4) and four different combinations of assays. Experiment 1 is repeated x2 and x4, experiment 1 is combined with experiment 2, and all the experiments are combined (see Fig. 37). The buffer concentrations and assay protocols used for each data set are specified in Tables 2 and 3. All experiments contain the same level of noise and are generated using the same transporter model (cycle 1), shown in Fig. 36. Further details on the simulation settings are found in the SI.

Protocol	**Synthetic Data**	H_{out} **Concentration in M**
1	Experiment 1	1.e-7, 0.5e-7, 1.e-7
2	Experiment 2	1.e-7, 0.5e-7, 1.e-7
3	Experiment 3	1.e-7, 0.5e-7, 1.e-7
4	Experiment 4	1.e-7, 0.5e-7, 1.e-7
5	Experiment 1x2	1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7
6	Experiment 1+2	1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7
7	Experiment 1x4	1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7
8	Experiment 1+2+3+4	1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7, 0.5e-7, 1.e-7

Table 2: Buffer concentration sequence for experiment recommendation data sets - H_{out} concentrations

Protocol	**Synthetic Data**	S_{out} **Concentration in M**
1	Experiment 1	1.e-3, 1.0e-3, 1.e-3
2	Experiment 2	1.e-3, 0.353e-3, 1.e-3
3	Experiment 3	1.e-3, 0.5e-3, 1.e-3
4	Experiment 4	1.e-3, 0.25e-3, 1.e-3
5	Experiment 1x2	1.e-3, 1.0e-3, 1.e-3, 1.0e-3, 1.e-3
6	Experiment 1+2	1.e-3, 1.0e-3, 1.e-3, 0.353e-3, 1.e-3
7	Experiment 1x4	1.e-3, 1.0e-3, 1.e-3, 1.0e-3, 1.e-3, 1.0e-3, 1.e-3, 1.0e-3, 1.e-3
8	Experiment 1+2+3+4	1.e-3, 1.0e-3, 1.e-3, 0.5e-3, 1.e-3, 0.353e-3, 1.e-3, 0.25e-3, 1.e-3

Table 3: Buffer concentration sequence for experiment recommendation data sets - S_{out} concentrations

Model Selection The Bayesian inference pipeline is utilized in order to compare different mechanistic models. To validate our method, we use a reference model to generate synthetic data. Then using that dataset, the 'ground truth' reference model and three other 1:1 antiporter models are sampled using Bayesian inference. This results in four posterior distributions and likelihood distributions. Comparing the likelihood shows which models best explain or fit the data [73]. Additionally, if a sequential Monte Carlo sampler [66] is used, the model evidence can be determined and compared for another comparison between models [136]. This workflow is shown in Figure 38. The synthetic reference data is the same as previously used for experiment recommendation, with the same underlying transporter cycle used to generate all the synthetic 'ground truth' data.

The four ideal 1:1 antiporter models all transport a single ion (H) and substrate (S) in opposite directions and consist of six reaction states, with a unique set of reactions and conformational states between the models. These differences amount to distinct reaction pathways (i.e. mechanisms) of transport, as shown in Fig. 39. For example, in cycles 1 and 3, k_1^f corresponds to an ion binding, but in cycles 2 and 4, k_1^f corresponds to a substrate unbinding. These different

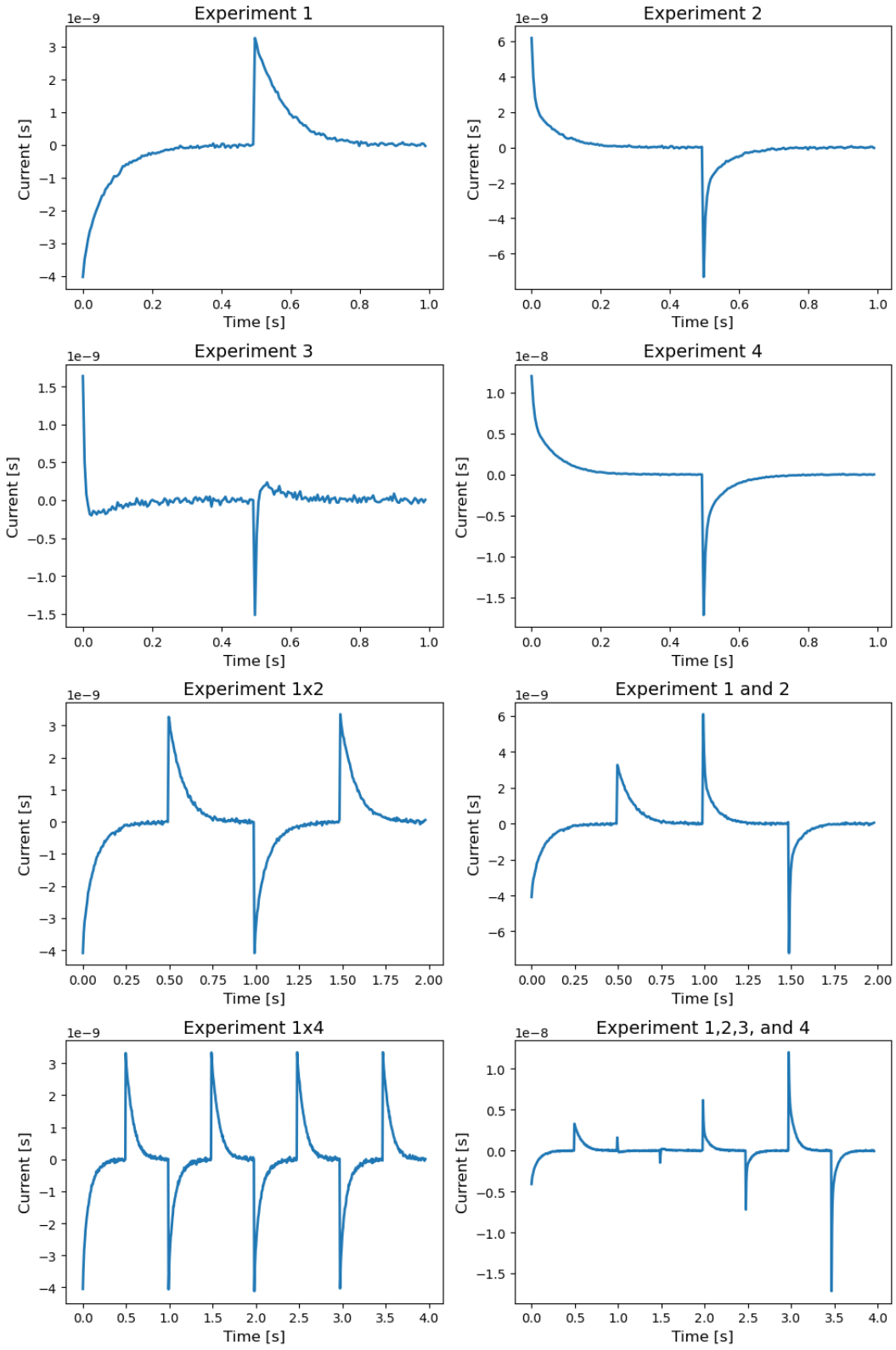


Figure 37: Synthetic SSME Assay Datasets Eight different datasets corresponding to four different perturbation amounts and four different combinations of experiments. Experiment 1 is repeated x2 and x4, experiment 1 is combined with experiment 2, and all the experiments are combined. All experiments contain the same level of noise and are generated from the same model (cycle 1).

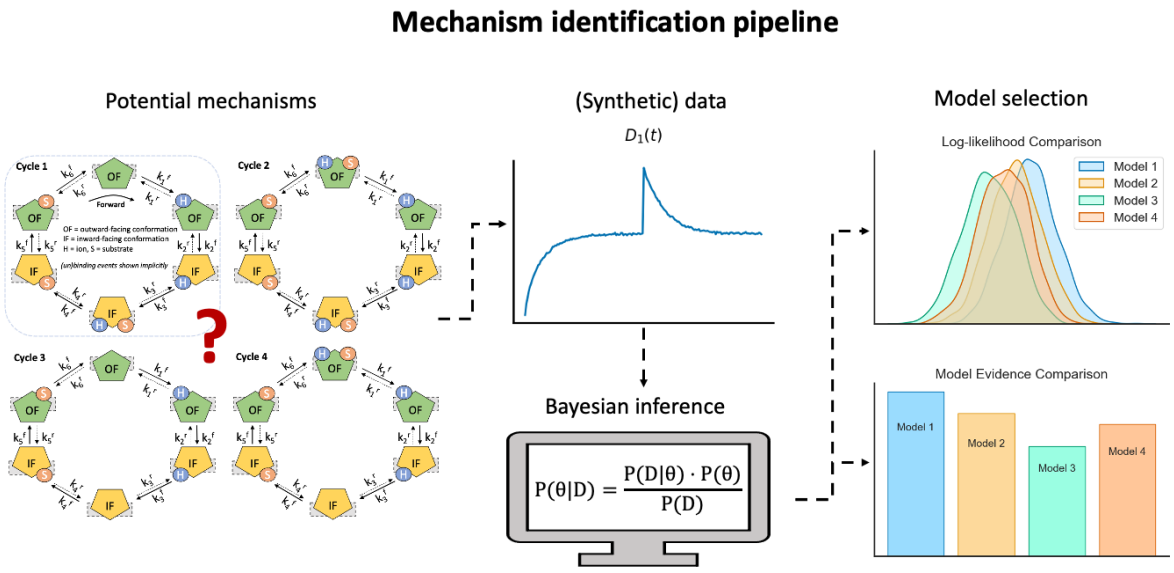


Figure 38: Pipeline for Mechanism Identification A method to compare alternative mechanistic models using Bayesian inference. Here four ideal 1:1 transporters are compared against a dataset generated by one of the four models as a reference point. Bayesian inference generates the likelihood which can be used to compare models. Similarly, the model evidence can be estimated as an alternative comparison point.

physical processes arise from the different states used in the model, such as with an unbound outward-facing state used for cycles 1 and 3, and a doubly bound outward-facing conformation in cycles 2 and 4. Their governing differential equations are shown below, with further simulation and parameter details described in the SI.

Cycle 1 differential equations:

$$\frac{d[H_{\text{in}}]}{dt} = k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}]$$
$$\frac{d[S_{\text{in}}]}{dt} = -(k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb])$$
$$\frac{d[OF]}{dt} = -(k_1^f[OF][H_{\text{out}}] - k_1^r[OF\_Hb]) + k_6^f[OF\_Sb] - k_6^r[OF][S_{\text{out}}]$$
$$\frac{d[OF\_Hb]}{dt} = k_1^f[OF][H_{\text{out}}] - k_1^r[OF\_Hb] - (k_2^f[OF\_Hb] - k_2^r[IF\_Hb])$$
$$\frac{d[IF\_Hb]}{dt} = k_2^f[OF\_Hb] - k_2^r[IF\_Hb] - (k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb])$$
$$\frac{d[IF\_Hb\_Sb]}{dt} = k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb] - (k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}])$$
$$\frac{d[IF\_Sb]}{dt} = k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}] - (k_5^f[IF\_Sb] - k_5^r[OF\_Sb])$$
$$\frac{d[OF\_Sb]}{dt} = k_5^f[IF\_Sb] - k_5^r[OF\_Sb] - (k_6^f[OF\_Sb] - k_6^r[OF][S_{\text{out}}])$$

Cycle 2 differential equations:

$$\frac{d[H_{\text{in}}]}{dt} = k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}]$$
$$\frac{d[S_{\text{in}}]}{dt} = -(k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb])$$
$$\frac{d[OF\_Hb\_Sb]}{dt} = -(k_1^f[OF\_Hb\_Sb] - k_1^r[OF\_Hb][S_{\text{out}}]) + k_6^f[OF\_Sb][H_{\text{out}}] - k_6^r[OF\_Hb\_Sb]$$
$$\frac{d[OF\_Hb]}{dt} = k_1^f[OF\_Hb\_Sb] - k_1^r[OF\_Hb][S_{\text{out}}] - (k_2^f[OF\_Hb] - k_2^r[IF\_Hb])$$
$$\frac{d[IF\_Hb]}{dt} = k_2^f[OF\_Hb] - k_2^r[IF\_Hb] - (k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb])$$
$$\frac{d[IF\_Hb\_Sb]}{dt} = k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb] - (k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}])$$
$$\frac{d[IF\_Sb]}{dt} = k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}] - (k_5^f[IF\_Sb] - k_5^r[OF\_Sb])$$
$$\frac{d[OF\_Sb]}{dt} = k_5^f[IF\_Sb] - k_5^r[OF\_Sb] - (k_6^f[OF\_Sb][H_{\text{out}}] - k_6^r[OF\_Hb\_Sb])$$

Cycle 3 differential equations:

$$\frac{d[H_{\text{in}}]}{dt} = k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}]$$
$$\frac{d[S_{\text{in}}]}{dt} = -(k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb])$$
$$\frac{d[OF]}{dt} = -(k_1^f[OF][H_{\text{out}}] - k_1^r[OF\_Hb]) + k_6^f[OF\_Sb] - k_6^r[OF][S_{\text{out}}]$$
$$\frac{d[OF\_Hb]}{dt} = k_1^f[OF][H_{\text{out}}] - k_1^r[OF\_Hb] - (k_2^f[OF\_Hb] - k_2^r[IF\_Hb])$$
$$\frac{d[IF\_Hb]}{dt} = k_2^f[OF\_Hb] - k_2^r[IF\_Hb] - (k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}])$$
$$\frac{d[IF]}{dt} = k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}] - (k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb])$$
$$\frac{d[IF\_Sb]}{dt} = k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb] - (k_5^f[IF\_Sb] - k_5^r[OF\_Sb])$$
$$\frac{d[OF\_Sb]}{dt} = k_5^f[IF\_Sb] - k_5^r[OF\_Sb] - (k_6^f[OF\_Sb] - k_6^r[OF][S_{\text{out}}])$$

Cycle 4 differential equations:

$$\begin{aligned}
\frac{d[H_{\text{in}}]}{dt} &= k_3^f[IF_Hb] - k_3^r[IF][H_{\text{in}}] \\
\frac{d[S_{\text{in}}]}{dt} &= -(k_4^f[IF][S_{\text{in}}] - k_4^r[IF_Sb]) \\
\frac{d[OF_Hb_Sb]}{dt} &= -(k_1^f[OF_Hb_Sb] - k_1^r[OF_Hb][S_{\text{out}}]) + k_6^f[OF_Sb][H_{\text{out}}] - k_6^r[OF_Hb_Sb] \\
\frac{d[OF_Hb]}{dt} &= k_1^f[OF_Hb_Sb] - k_1^r[OF_Hb][S_{\text{out}}] - (k_2^f[OF_Hb] - k_2^r[IF_Hb]) \\
\frac{d[IF_Hb]}{dt} &= k_2^f[OF_Hb] - k_2^r[IF_Hb] - (k_3^f[IF_Hb] - k_3^r[IF][H_{\text{in}}]) \\
\frac{d[IF]}{dt} &= k_3^f[IF_Hb] - k_3^r[IF][H_{\text{in}}] - (k_4^f[IF][S_{\text{in}}] - k_4^r[IF_Sb]) \\
\frac{d[IF_Sb]}{dt} &= k_4^f[IF][S_{\text{in}}] - k_4^r[IF_Sb] - (k_5^f[IF_Sb] - k_5^r[OF_Sb]) \\
\frac{d[OF_Sb]}{dt} &= k_5^f[IF_Sb] - k_5^r[OF_Sb] - (k_6^f[OF_Sb][H_{\text{out}}] - k_6^r[OF_Hb_Sb])
\end{aligned}$$

In these equations, [X] denotes the concentration of state [X], OF and IF represent the outward-facing and inward-facing conformations. The transported ion (H) and substrate (S) are compartmentalized into either outside the liposome (X_{out}), inside the liposome (X_{in}), or bound to the transporter protein (X_b). The reaction rate constant k_x^f corresponds to the rate constant for reaction x, in the forward clockwise direction 'f', with the counterclockwise direction denoted with 'r'.

Quantitative metrics for distribution analysis and comparison For experiment optimization and recommendation, we are primarily interested in screening for protocols that yield high information data and reduce the variance of our parameter estimates. As discussed previously we use the KL divergence between posterior and prior as a metric for the information gained for a given data set. Additionally, there are many possible methods to quantify the reduction in variance between two posterior distributions, such as comparing the Bayesian credible intervals, computing the KL divergence, computing the overlapping coefficient, or the computing sum of standard deviations. For this work, we opt to compare the sum of standard deviations across the marginalized posterior:

$$\text{Marginalized Total Standard Deviation for } D_i = \sum_{i=1}^{N} \sigma_{D_i,x_j} \tag{27}$$

where x_j is the jth parameter (N total) and D_i is the ith data set, σ is the standard deviation.

Further, for model comparison and selection, we primarily examine the log-likelihoods, which in our formulation essentially describe the scaled mean-squared error of the residuals (plus additional terms as shown previously). With Bayesian inference, we generate a distribution of log-likelihoods, which correspond to the log probability of the data given the model parameters. The ability to discern the most likely model depends on the separation of these log-likelihood distributions - if all models are equally likely then their log-likelihood distributions will be overlapping, and if one model is more likely then its log-likelihood distribution will contain the maximum and be separated from the

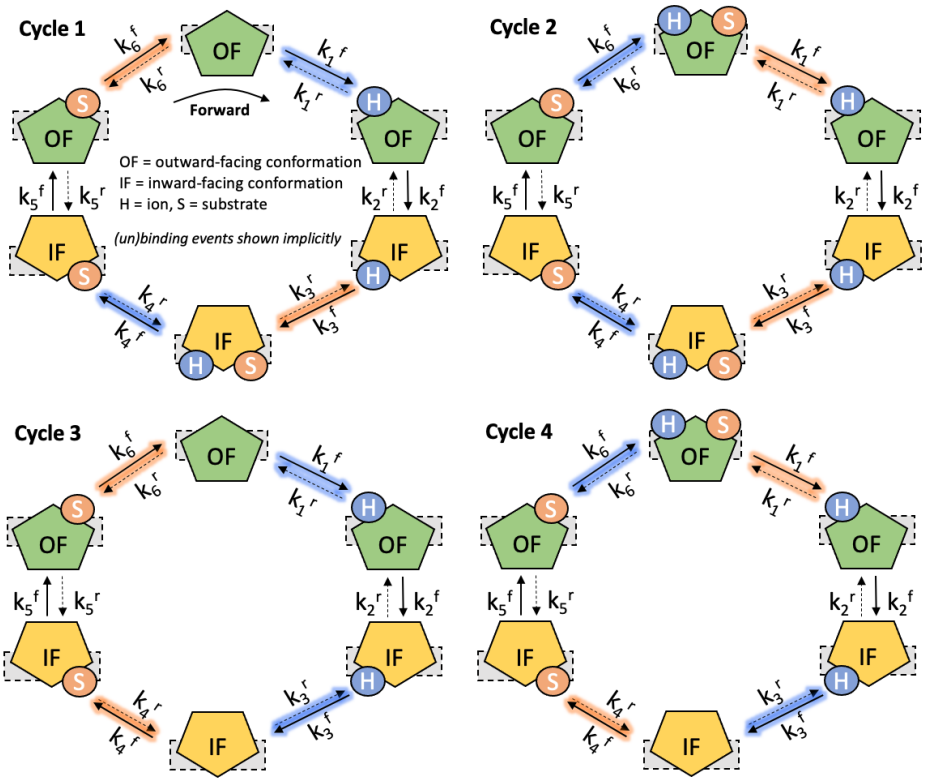


Figure 39: Four Tightly Coupled 1:1 Antiporter Reaction Cycles Each cycle transports an ion (H) and a substrate (S) in opposite directions via alternating access (outward-facing, OF, and inward-facing, IF) conformations. Each network has six unique reaction states and reaction pathways resulting from different outward and inward-facing binding states. For example, cycle 1 contains an unbound outward-facing conformation, with a k_1^f corresponding to the reaction rate of an ion (H) binding. In contrast, cycle 2 has a doubly bound outward-facing conformation, with a k_1^f corresponding to the reaction rate of a substrate (S) dissociating. The ion (un)binding reactions are highlighted in blue, with the substrate (un)binding reactions are highlighted in orange, with each cycle having a unique combination of ion and substrate reactions. The binding and unbinding (i.e. dissociation) events are shown implicitly for improved visual clarity.

others. We can quantify this by comparing the difference between the modes of the distributions and the 95% confidence interval overlap. Additional metrics for model comparison include evaluating the model evidence and Bayes factor which are presented in the SI.

4.5 Results

4.5.1 Recommendation of Informative Experimental Protocols

We examine the 1D marginal posterior distributions of all eight tested data sets (Fig 40). Here the combination of all assay conditions in protocol 8 has the lowest variance in parameter estimates. This combined data set has significantly lower variance in k_1, k_2, and k_6, than the other data sets. The data from only experiment 3 had noticeably higher variance than the other data sets. This is corroborated when examining the total marginalized standard deviation for the data sets (Fig. 42). We see that the data set with only experiment 3 has the largest total standard deviation and that the data set with experiments 1, 2, 3, and 4 has the lowest standard deviation. This data set has significantly lower standard

deviation compared to both single experiment data sets, and replica experiments of the same size. The standard deviations for each marginal are examined further in the SI.

Similarly, the KL divergence is approximated for all eight tested datasets (Fig 41). The data set from experiments 1, 2, 3, and 4 combined had the most information (protocol 8), followed by the data from the combination of experiments 1 and 2 (protocol 6), then the data from only experiment 4 (protocol 4). The data from only experiment 3 has noticeably lower information than the other datasets. Also, the (synthetic) replicas for experiment 1, as shown in protocols 5 and 7, do not have noticeably more information compared to the single experiment in protocol 1.

Since experiment 4 single experiment (protocol 4) has the largest magnitude relative to the noise, data sets containing that experiment were expected to have the highest information score (protocol 8). However, we do see that the additional experiments are providing some information, as the combination of all experiments in protocol 8 has a higher score than just the experiment 4 dataset alone. Similarly, protocol 6, containing data from experiments 1 and 2, has the second highest score, while the replica runs (protocols 5 and 7) have marginally higher scores compared to their single experiment data set (protocol 1). The disparity of information between the different protocols suggests that certain assay conditions are significantly better at exciting the underlying transporter, yielding a higher current signal-to-noise ratio.

4.5.2 Model Selection Outcomes

We compare the four ideal 1:1 antiporter cycles described previously, using a moderate and high information dataset from our ranking method. We use experiment 1 (protocol 1) and experiment 1+2+3+4 (protocol 8) datasets for Bayesian inference, which are generated from the same model, cycle 1.

First, we examine the log-likelihood distributions (Fig. 43) for the protocol 1 data set, comparing their modes and 95% confidence intervals (from 2.5 to 97.5%). We see that the distributions for cycles 1 and 2 are nearly identical, having similar modes and overlapping confidence intervals. This suggests that cycles 1 and 2 cannot be distinguished from each other based on this data set. Cycles 3 and 4 have similar overlapping distributions with each other, and a slight separation from cycles 1 and 2. So, based on the given data set, it is not possible to clearly separate the different reaction mechanisms. We note that the synthetic ground truth model (cycle 1) does find the maximum log-likelihood as expected, although it is not significantly larger than the maximum of an alternative model (cycle 2).

Next, we examine the log-likelihood distributions (Fig. 44) for the protocol 8 data set combining the data from experiments 1, 2, 3, and 4. As before, we compare their modes and 95% confidence intervals (from 2.5 to 97.5%). In stark contrast to the previous data set, we see that the distributions for all of the cycles are clearly separated, as indicated by the confidence intervals not overlapping. Furthermore, we see that there is a large difference between the distribution modes with cycle 1 (the reference model) having the largest mode and maximum log-likelihood value. The next most likely model (cycle 3), has a mode that is approximately 38 less than the cycle 1 mode, corresponding to a e^{38} (1e16 orders of magnitude) less likely model. These results suggest that the reference model can be correctly identified based on this data set, an important validation step for this pipeline.

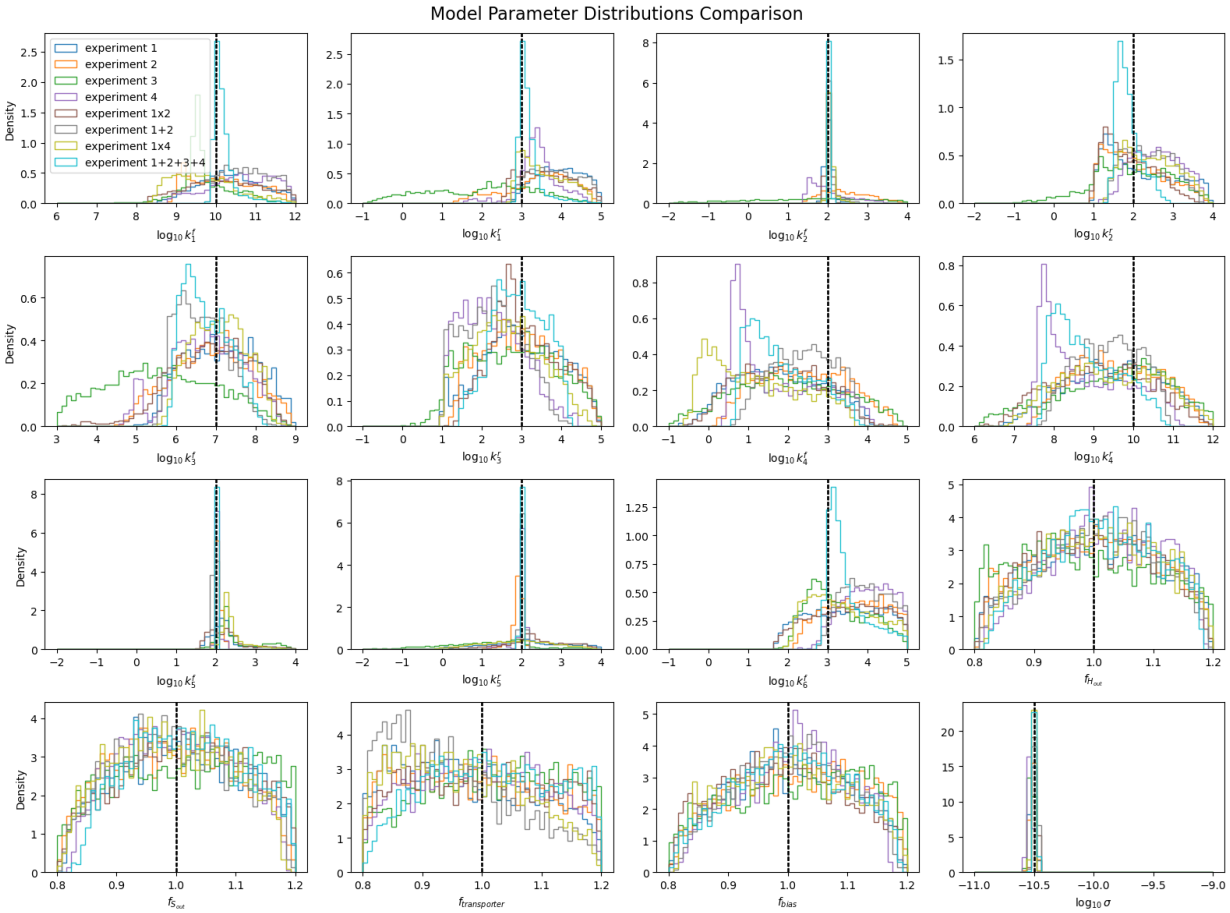


Figure 40: 1D Marginal Posteriors Across Datasets The 1D marginal distributions for each synthetic assay are shown. Here the combination of all assay conditions (experiments 1+2+3+4) has the lowest variance in parameter estimates. This dataset has significantly lower variance in k_1, k_2, and k_5, than the other datasets.

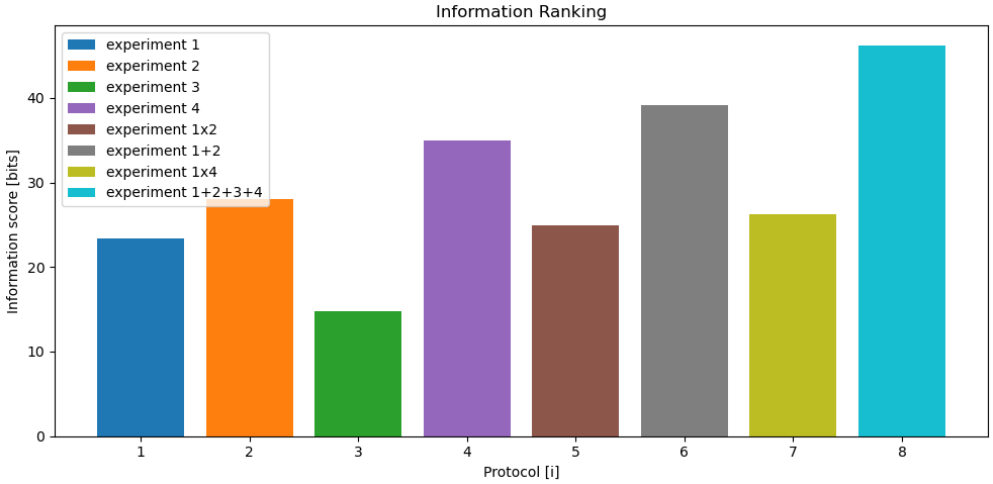


Figure 41: Information Ranking The ranking of information content based on the KL divergence between prior and posterior distributions across eight datasets. The data set from experiments 1, 2, 3, and 4 combined had the most information, followed by the data from the combination of experiments 1 and 2, then the data from only experiment 4. Technical replicas do not significantly increase the information content for the data studied.

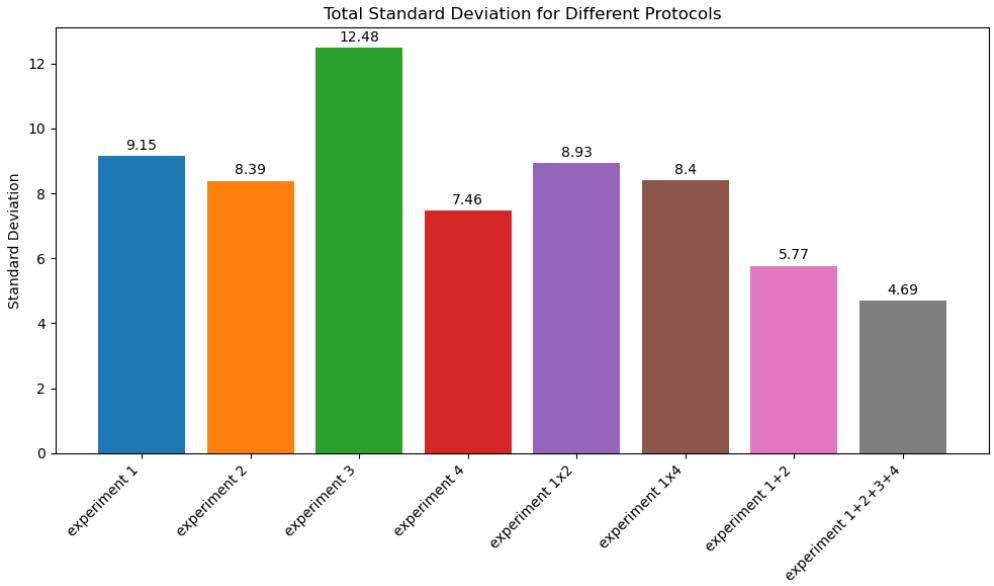


Figure 42: Total Standard Deviation Comparison The comparison of the total standard deviation from the marginal 1D posteriors generated from Bayesian inference of different data sets. The data set from experiments 1, 2, 3, and 4 combined had the lowest total standard deviation, with a 2.5x reduction in the standard deviation as compared to the experiment 3 dataset.

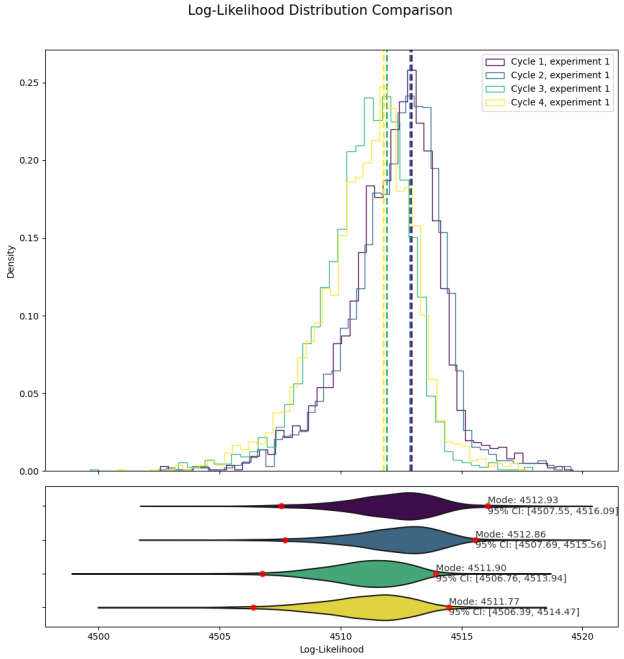


Figure 43: Log-Likelihood Distributions I Distribution of the log-likelihoods generated from Bayesian inference of four different models using a less informative data set (protocol 1), with the violin plots shown on the bottom panel. Here transporter cycle 1 (reference model) and cycle 2 models have overlapping likelihood distributions, with close modes and confidence intervals. Cycles 3 and 4 have similar overlaps with each other, with a small separation between cycles 1 and 2. Overall, all these distributions have a significant overlap, making model selection infeasible using the given data set.

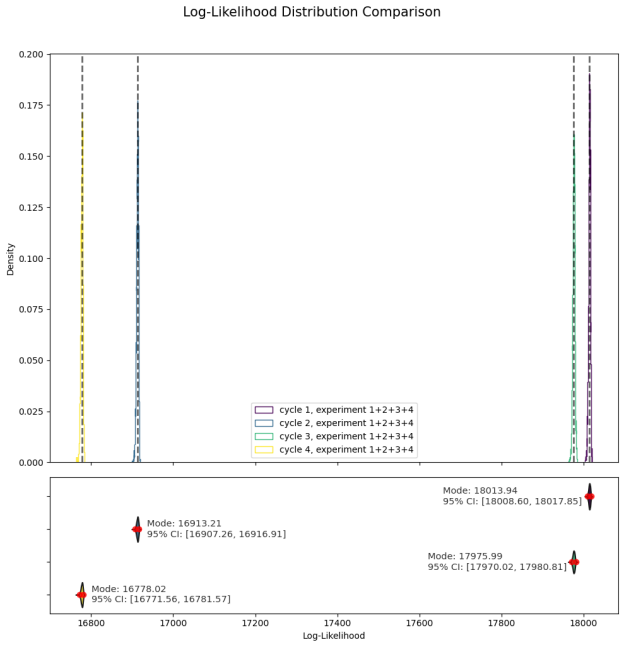


Figure 44: Log-Likelihood Distributions II Distribution of the log-likelihoods generated from Bayesian inference of four different models using a very informative data set (protocol 8), with the violin plots shown on the bottom panel. Here transporter cycle 1 (reference model) has the maximum log-likelihood and largest log-likelihood mode. The remaining models for cycles 2, 3, and 4 are greatly separated from cycle 1 and each other, as demonstrated by the non-overlapping confidence intervals and large differences in modes. Here the difference between the mode of cycle 1 and the next most likely model (cycle 3) is approximately 38, suggesting a e^{38} (1e16 orders of magnitude) more likely model.

Overall, we found that as expected, information-poor data sets make model selection challenging. However with a careful choice of experimental protocol more informative data sets can be generated that enable model selection, and therefore the identification of transporter reaction mechanisms. These general findings are also corroborated when using other quantitative metrics for model selection such as model evidence and the Bayes factor (see SI).

It's worth pausing to consider why certain models performed better than others: due to different physical processes occurring in each model - i.e. different binding or unbinding event orders during transport. As the data becomes more informative these differences are likely to become more noticeable, resulting in the clear separation between models.

In this study, we have investigated the information content under different experimental conditions and used this to distinguish between multiple transporter models. We found that combining all assay conditions significantly reduced parameter variance and increased the information in the data. In contrast, technical replicas did not significantly improve the information content. When using an informative dataset, we were able to accurately identify cycle 1 as the most likely model. These findings with synthetic data validate our approach but also provide insight into experimental design and mechanism identification, which we will discuss next.

4.6 Discussion

To summarize, this study investigated how much information was contained in noisy SSME-type datasets, leveraging Bayesian inference.

An important finding of our work was that the most significant reduction in parameter estimate variance and the maximum amount of information was when all assay conditions were combined into a single dataset. This shows how integrating multiple experiments generates more informative data. Our results also indicate that repeated experiments under the same conditions do not significantly contribute to the increased amount of information. These results suggest a shift in experimental design to focus on diverse assay conditions instead of repetition. Including additional pH adjustments is another potential way to add more diverse assay parameters.

We note that it is expected that more information can be extracted from data sets containing larger signal-to-noise ratios and increasing amounts of data. However, our results suggest an interesting balance between large magnitude perturbations, data amount, and protocol design. For example, we find that a single experiment (experiment 4) is more informative than a data set containing four times as much data containing lower magnitude perturbations. Future work will examine this in greater detail.

This work could pave the way for automated experimental design in SSME. By quantifying the data in a given experiment, sophisticated algorithms [150, 162, 33] could choose the most informative experiment to do next. Or more simply, synthetic experiments could be run first to pick the most informative ones to run given limited resources. Either of these approaches could dramatically improve the efficiency of experimental design and data analysis.

Our investigation of four 1:1 antiporter cycles demonstrated the power of this method. By using high-information datasets from a combination of assay conditions, we were able to identify the most probable mechanism, illustrating a proof of concept for this approach to estimate complex mechanisms and potentially disentangle ambiguous reaction pathways. Future work could explore how well this method is able to precisely determine mechanistic heterogeneity from within a single model. That is to say, if there are multiple pathways within a single model, how well would this approach hold?

These results lay the foundation for future work to advance the understanding of the true mechanisms of multidrug resistance transporters such as EmrE. Our results show that with informative data sets (i.e. optimized experimental protocols), the uncertainty of parameter estimates is significantly reduced and the separation between model likelihoods is significantly increased. More precise estimates open the door to more robust and exact modeling of transporter behavior, which could eventually lead to better therapeutic treatments. However, to study more complex transporters, more efficient computation methods would be needed to reduce the run time, as well as stronger assumptions on the network to reduce the complexity.

In addition, the informative nature of diverse assay conditions suggests that complementary or even orthogonal data sources may be beneficial to better understanding transporters. For example, single-molecule methods like FRET

[202, 212] or fluorescence tracking [168, 118] may be a rich source of information alongside the macroscopic data generated by SSME. However, appropriately incorporating this data into a comprehensive model will be a significant challenge.

It is worth noting again that we used synthetic datasets in our current work and that applying this method to real experimental SSME data will likely introduce additional complexities. Experimental artifacts from issues in sample prep, contamination from the environment, and non-homogeneity, along with potentially more complex noise variability, are potential issues that could arise [12]. Sample prep issues and contamination can likely be accounted for with the scaling bias term - but adjustments may be required such as expanding the prior range to allow for more extreme effects. Non-homogenous mixtures are a greater obstacle and would likely require a different probabilistic model - such as hierarchical Bayes[184].

4.7 Conclusion

In summary, this research sheds new light on the study of transporter proteins, offering a systematic method to compare mechanisms and guide experiment design, through the use of Bayesian inference and information theory. Not only does this method improve the precision of our parameter estimates, but it also guides the design of the experiment to generate the most informative data. Future work will focus on refining our methods, expanding this approach to include real SSME data, and extending our techniques to more complex models of transporters - such as 2:1 transporters. Ultimately, we aim for this pipeline to be useful for empirical transporter research. By enabling effective model comparison, precise parameter estimation, and informed experimental design, this method could help move the field toward a more complete understanding of transporter mechanisms and their far-reaching implications.

4.8 Supporting information

See the attached appendices for the supporting information on this chapter.

S1 Text **Additional Methods Used:** Detailed explanation of the network and assay model, model comparison strategies, and Gaussian mixture model validation.

5 Discussion

5.1 Synopsis of the Dissertation and Key Findings

5.1.1 Restatement of Research Objectives and Outcomes

The primary research question guiding this work was, “Can the development of robust computational methods provide more accurate modeling of transporter proteins and their mechanistic complexities?"

This was addressed across two major research areas in the dissertation:

1. Developing robust computational tools to improve biochemical network modeling and data analysis techniques targeting transporter proteins
2. Utilizing these computational tools to explore the mechanistic heterogeneity and complexity of transporter proteins in silico

The fundamental hypothesis underlying this work was that the use of novel and robust computational methods would provide more detailed insights into transporter protein complexities than previous methodologies.

The following sections provide an overview of how these objectives were met during this research.

Developing robust computational tools The development of robust computational tools was at the heart of this dissertation, and each main chapter highlighted the development or implementation of novel computational tools for transporter research. In the first research chapter, a brand new computational approach was implemented, ModelExplorer, that enabled the study of alternative transport mechanisms. This modeling method was demonstrated to be robust through its validation with different model types, as well as the sheer number of different ‘engineered’ models that were generated and clustered into classes. Importantly it improved on existing methods by forcing physical constraints into the optimization process, and by expanding the search space to include all possible conformational (reaction) states.

Similarly, in the second research chapter an implementation of a powerful statistical method, Bayesian inference [73], was developed in order to analyze data from electrophysiology-type datasets. This enabled the determination of microscopic rate constants from noisy and sparse macroscopic data. This approach was shown to be robust by outperforming competing algorithms and demonstrating convergence without the presence of biasing prior knowledge. Finally, this pipeline is the first to target SSME datasets, improving the current modeling capabilities in that way.

In addition, in the third research chapter, information quantification and model selection strategies were implemented to extend the capabilities of the previously discussed Bayesian inference pipeline. The robustness of these methods was demonstrated across the analysis of different transporter models and information criteria. This approach provides a foundation to improve upon current ad-hoc experiment design strategies in SSME-like experiments by giving an

automated workflow to explore protocol design, and enables a systematic and efficient determination of transporter reaction pathways from a collection of models.

Aside from methodological robustness, this study emphasized robust implementation. Engineering best practices such as thorough documentation, version control, and unit testing were employed whenever possible to ensure reliability and accuracy. Further, the robustness of the software was improved through the use of more efficient algorithms and parallelization - enabling faster analysis of models and datasets.

In summary, each research project provided a new set of computational methods in the context of transporter research, that were robust in their ability to create, analyze, and infer from biochemical networks - with data inspired by solid-supported membrane-based electrophysiology (SSME) [12] as a motivating force. In each case, the methods provided improvements against alternative approaches by ensuring physical consistency and limited bias, as well as new modes for automated transporter study through model exploration, selection, and decision support.

Exploring the mechanistic complexity of transporters The other primary goal of this work was to investigate the theoretical complexity of reaction mechanisms for transporter models, and, in a broader sense, help change the paradigm of overly simplistic models for molecular machine research. This was achieved through each major research effort in this dissertation. First, ModelExplorer was explicitly developed to find different reaction pathways and engineer optimal networks based on an arbitrary fitness function. This approach yielded new theoretical models that exhibit a novel function of enhanced selectivity, utilizing a complex set of reaction steps to achieve this goal. This result highlights the potential for a new behavior to emerge from simple sets of reactions and supports the notion of alternative functionality in transporters on a theoretical basis.

In the second major research thrust, integrating Bayesian inference methods for transporters, the mechanistic complexity was examined in a more subtle way. Here, instead of directly searching for new reaction pathways as above, a platform is built that can be used to inspect pathway information from data. Through the selection and comparison of competing algorithms for inference, the Bayesian posterior is estimated, which can provide detailed information about protein mechanisms and whether alternative mechanisms may be used.

To that end, the third and final research focus augments the Bayesian inference pipeline to better reckon with potential alternative mechanisms. Here an automated approach to design and recommend experiments is developed with the expressed purpose of improving the mechanistic determination from data. By selecting the proper set of experiments, mechanistic information can be inferred with greater precision, and therefore, complex or alternative mechanisms can be disentangled from the data. This is demonstrated through the selection of a particular transport cycle from similar competing models that may be present in a complex transporter system. In short, the transport mechanisms and potential alternative pathways can be more precisely determined from data through the use of automated decision support from Bayesian inference.

Below, each major research focus is examined in more depth.

5.1.2 Overview of ModelExplorer

The primary motivation behind the development of ModelExplorer was to address a key challenge in the study of transporter proteins: the understanding and identification of diverse potential reaction pathways. Existing methods did not have the ability to fully account for the wide spectrum of possible mechanisms, or groups of mechanisms while maintaining detailed balance constraints.

ModelExplorer was developed as a critical part of this dissertation. It presents a unique computational tool designed to probe this space of chemical possibilities. Its robust design principles and functionalities enabled it to generate and explore various optimized models and cluster them into classes. A notable feature was the enforcement of physical constraints into the optimization process, thereby ensuring greater accuracy in its modeling.

In the context of this research, ModelExplorer proved invaluable in discovering new mechanisms. As an example, it revealed new theoretical models that exhibited enhanced selectivity. This enhanced selectivity was made possible through complex sets of elementary reaction steps without biasing the network model. As with kinetic proofreading [108, 178], these mechanisms all featured a reversible unbinding step to enhance the selectivity of a desired substrate vs. a decoy. These theoretical discoveries highlight the potential for novel behaviors to emerge from seemingly simple biochemical reactions, thereby supporting the notion of alternative mechanistic pathways and behaviors in transporters.

Not only was a single mechanism discovered, but classes of mechanisms with divergent mechanisms were all found to have this enhanced selectivity capacity. This suggests a certain robustness of the enhanced selectivity function, as it is possible across a range of different mechanisms. These results were validated through repeated independent trials of the ModelExplorer tool, demonstrating the robustness of our approach.

5.1.3 Overview of the Bayesian Inference Pipeline

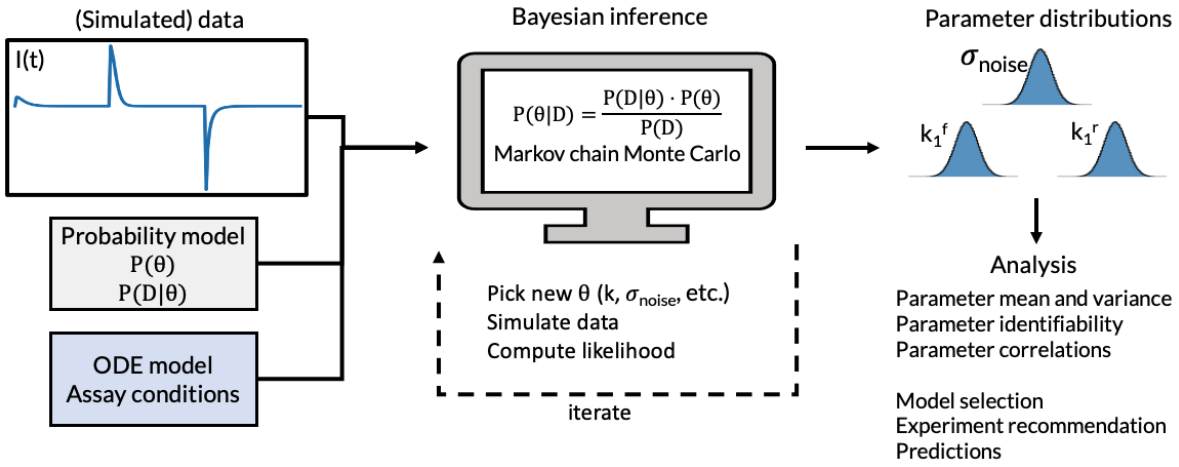


Figure 45: Bayesian Inference Pipeline - Revisited Schematic of the Bayesian inference pipeline implemented in this dissertation

The decision to use Bayesian inference was made to bolster the robustness of transporter model calibration. This statistical method brings the advantage of incorporating prior knowledge and a general probabilistic interpretation of

results. These characteristics make Bayesian inference an ideal strategy for dealing with noisy and sparse data sets, such as electrophysiology experiments.

The Bayesian inference pipeline (fig. 45) was specifically targeted toward solid-supported membrane electrophysiology datasets. It utilized a state-of-the-art sampling method enhanced with machine learning, which enabled the Bayesian inference pipeline to estimate parameters more reliably than alternative methods, including maximum likelihood estimation.

The robustness, accuracy, and adaptability of the Bayesian inference pipeline were validated using synthetic data where a synthetic ground truth is known for comparison. In these validation efforts, the Bayesian inference method was able to accurately estimate parameters for a complex model, whereas competing methods could not. Even with a small data set and extremely wide and uninformative prior assumption, certain microscopic quantities were able to be precisely determined from a macroscopic measurement.

The ability of Bayesian inference to accurately estimate model parameters and their uncertainty gives it an advantage over alternative methods and enables better scrutiny of the mechanistic details of transporters. In doing so, a robust Bayesian inference pipeline as presented in this dissertation, can offer unparalleled insight into the mechanistic complexity of transporters.

5.1.4 Overview of the Decision Support and Selection Pipeline

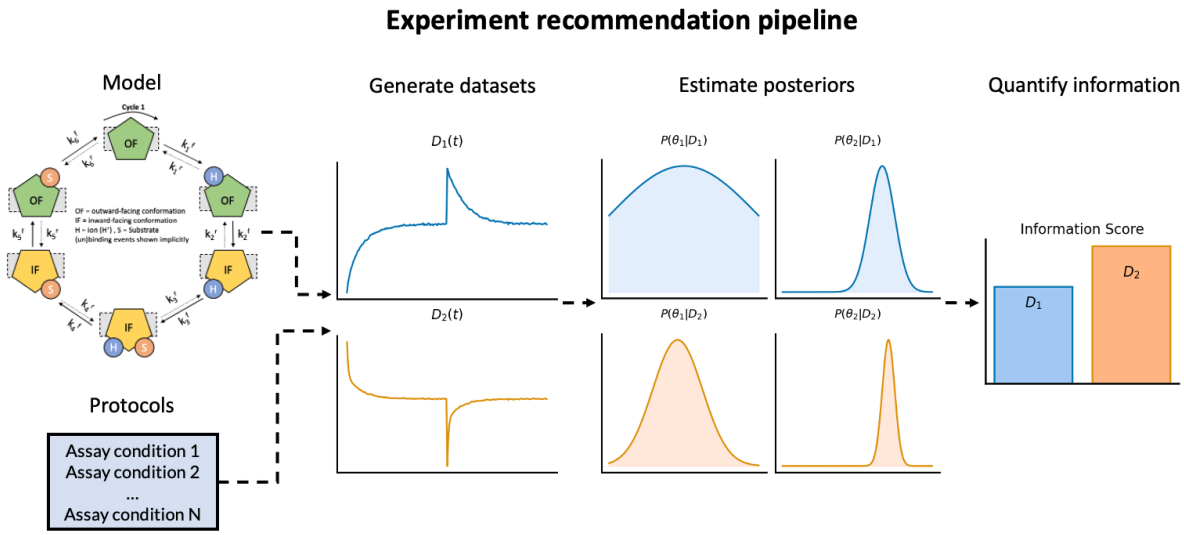


Figure 46: Experiment Recommendation Pipeline - Revisited Schematic of the experiment recommendation pipeline implemented in this dissertation

The implementation of a decision support and model selection pipeline (fig. 46) was a critical step in the overall Bayesian inference methodology. This process was designed to increase the robustness of findings and streamline decision-making processes involved in transporter research. This method complemented the Bayesian inference pipeline by utilizing the posterior distribution to provide an automated workflow for solid-supported membrane

electrophysiology experiment design, along with the systematic and efficient determination of transporter reaction rates and pathways.

This approach was utilized in order to determine the most suitable experiments needed to gather mechanistic information. The recommendations showed a dramatic improvement in the precision of parameter insights. Furthermore, the precise estimates enabled the differentiation of reaction mechanisms based on the data without prior knowledge enforced.

The results demonstrate the value of both the Bayesian inference method and the experiment recommendation system, which, when coupled together, provide unparalleled insight into transporter mechanisms from SSME-type datasets.

5.1.5 Key Findings and their Implications

The key findings of this dissertation are shown in figure 47 presented below and expanded on in the following sections.

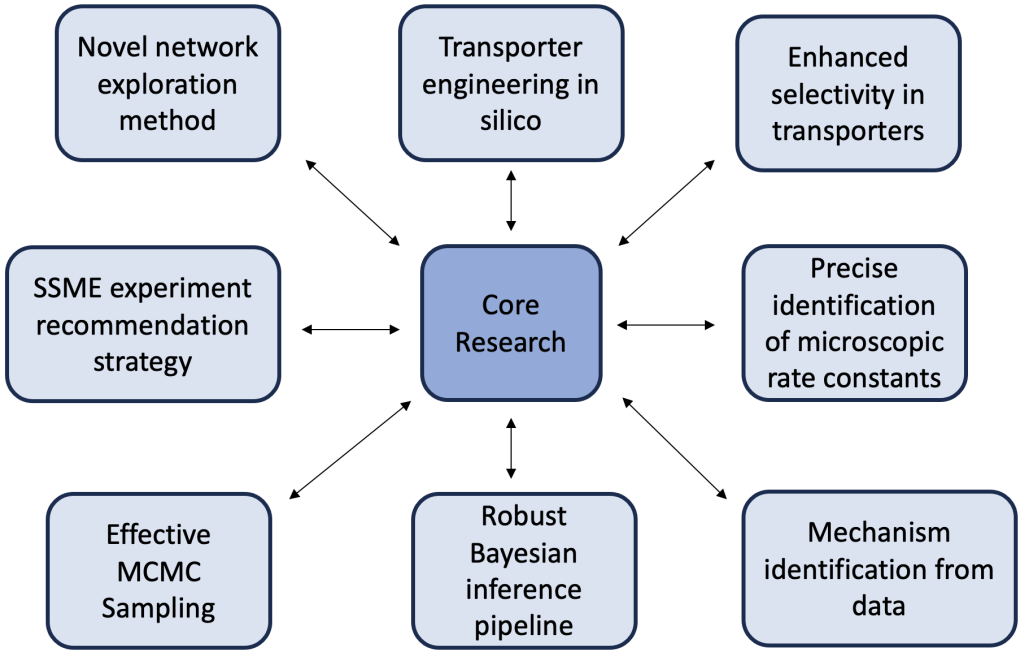


Figure 47: Key Findings

Innovative Exploration Method: Our research was instrumental in developing a pioneering method to stochastically and adaptively traverse the complete reaction space of molecular machines or transporters. This approach facilitates the comprehensive investigation of potential reactions within the biochemical reaction space, significantly enhancing the depth and breadth of our understanding.

Engineering In Silico Transporter Function: We successfully applied our novel exploration method to optimize new transporter functions in silico. By leveraging elementary biochemical reactions available in the entire reaction space of

an idealized transporter, we could simulate and study novel transporter behaviors, marking a notable theoretical advancement in transporter protein research.

Enhanced Selectivity in Transport: Our study discovered enhanced selectivity in transport processes, the ability to discriminate against substrates with similar structures at an extremely high level. We identified different classes of network structures exhibiting this selectivity, all of which showcased an additional unbinding 'leak' step similar to Hopfield and Ninio's [108, 178] kinetic proofreading. This finding suggests transporters could have more complexity than originally thought, enabling a more nuanced understanding of selective transport processes. Similarly, these mechanisms may give insight into evolutionary forces acting on the single-celled organisms in a hostile environment, where enhanced selectivity is (potentially) extremely beneficial to life

Robust Bayesian Inference Pipeline: We implemented a robust Bayesian inference pipeline specifically designed for solid supporting membrane electrophysiology [12, 13], (SSME) type data, commonly used in transporter research. This pipeline offers a powerful tool for data analysis to complement experimental methods, enhancing the accuracy and resiliency of data interpretation through the use of a Bayesian posterior.

Effective MCMC Sampling: Our research revealed a machine-learning accelerated Preconditioned Monte Carlo (PMC) sampler as the best-performing algorithm for Bayesian inference among Markov chain Monte Carlo (MCMC) and its related methods. This finding streamlines future computational workflows by identifying the most proficient tools for data sampling in complex biological systems.

Precise Identification of Microscopic Reaction Rate Constants: We precisely identified a set of microscopic rate constants for a synthetic data set and idealized transporter. This advancement enhances the precision of our model predictions and improves our understanding of the underlying mechanism.

Experiment Recommendation through KL Divergence: This research demonstrated an experiment recommendation system via ranking of information gained from the posterior and prior (Kullback-Leibler divergence) under different assay protocols, using synthetic data. We found that larger perturbations along with combinations of experiments provided more insightful data, indicating that assays producing large perturbations or a sequence of unique perturbations may be the most informative.

Model Selection for 1:1 Transporters: We implemented model selection by comparing four different 1:1 transporter mechanisms using a synthetic dataset with a known synthetic ground truth mechanism. We found that with a low-information dataset, it was not possible to distinguish all the models, but with a high-information dataset, the models could be differentiated via likelihood comparison. Moreover, high-information datasets yielded much more precise parameter estimates, indicating the importance of high-quality data in understanding transporter mechanisms.

Generalizable Methods: While this dissertation focused on secondary active transporter research, it could also be applied to other molecular machines to varying degrees. The approach used in the ModelExplorer project could be extended to other machines, such as ATPases or motor proteins, because of its generic ability to build reaction networks

from a set of user-defined states and then optimize based on a given cost function. Similarly, the Bayesian inference (and model selection) pipeline could be adapted to work with any dynamic reaction network system since it uses general systems biology tools such as SBML [127] and libroadrunner [230] for model specification and simulation, and generic optimizers/samplers for parameter estimation. The primary adjustment would be creating a custom data-generating function for the molecular machine of interest that can simulate the experimental data - as our approach was customized for electrophysiology-type experiments used for transporter research. These robust methods could help the broader fields of systems biology and protein research.

5.1.6 Applications and Broader Implications:

The outcomes of this work can be leveraged in several different ways.

The ModelExplorer tool provides a framework for future studies to engineer new functions in biology or artificial molecular machines. The applications of these engineered nanomachines are numerous, from improved drug delivery to technologies for biochemical processing and bioremediation, and also greatly improve our understanding of biological systems. This type of nano-scale precision and design could enable more advanced materials and potentially revolutionize the pharmaceutical and biotechnologies fields.

The theoretical mechanism of enhanced selectivity could be utilized to inform the design of drug delivery systems that leverage enhanced selectivity and boost the biological processes that rely on selective transport. This could result in novel therapeutics, improved drug designs, and optimal biochemical reactions.

The computational tools developed in this work could improve data analysis and modeling across many scientific areas, such as systems biology, computer science, and physics. Future work that builds on these tools could make them more efficient, reliable, and accessible, accelerating the pace of scientific discovery.

Finally, experiment recommendation and model selection systems can aid in the design of experiments, making them more efficient and lowering the cost. Future researchers could leverage these tools to pick the optimal experimental conditions and to interrogate competing reaction mechanisms more precisely. A possible direction of this work in the future could be to make the prototype system into a user-friendly software package online, expanding the impact of these tools.

5.2 Reflections on the Research Process

The overarching research methodology used in this dissertation is described below:

5.2.1 Research Design

The initial motivation for this work was the emerging idea of alternative mechanisms for proteins [97, 200]. Since proteins operate at the nanoscale where randomness has a strong effect on movement[185], it seemed plausible that proteins did not have rigid mechanisms but instead had some level of flexibility. This raised the question of how these

proteins might change behaviors under different conditions or external conditions. Due to the vast space of environmental conditions and reaction states that any given protein could have, computer simulations of the reaction networks seemed the most appropriate approach to examine different mechanisms.

Existing tools did not have the combination of flexibility, performance, or accessibility that this work required. For example, network engineering strategies had been used to probe a reaction space via evolutionary processes, but they did not consider the entire reaction space or include physical constraints on the energy of the system [237, 164, 91]. As such, a hypothesis was formed that more robust tools were needed to study the possible alternative functions of proteins in various environments.

This hypothesis was further boosted by the continued emergence of solid-supporting membrane electrophysiology experiments that generated high-resolution time series data under different environmental conditions. This datatype seemed well suited for biochemical networks, which abstract out most structural details and instead focus on kinetics. This realization helped motivate the development of the Bayesian inference pipeline.

5.2.2 Data Collection and Analysis

For ModelExplorer, the goal was to explore mechanisms in silico, so no experimental data was necessary. All data were simulated and then analyzed for novel mechanisms. With the Bayesian inference paradigm, the primary motivation was to use experimental data. However, due to technical challenges, synthetic data was used instead. Throughout this dissertation, different software tools and programming languages have been used, with Perl being used initially, then Julia and Python exclusively.

Using synthetic data has many trade-offs. For methods development, synthetic data provides exact control of model parameters so that there is a known ground truth that can be used for method testing and validation (e.g. Bayesian inference methods development). In addition, control over the model allows for the exploration of rare or new theoretical processes, as shown with the proof-reading transporter models. Synthetic data is also generally less expensive than biophysical experiments such as electrophysiology and is reproducible. However, synthetic data may be generated from oversimplified models that make strong assumptions, leading to misleading results. Because of this, experimental data should be used in combination with synthetic data when possible. Importantly for this project, experimental data could be used to confirm the existence of proof-reading transporters, and corroborate the parameter estimates from the Bayesian inference pipeline using synthetic data.

5.2.3 Methodological Decisions

In general, well-tested existing methods were preferred when possible. However, due to the challenging nature of computational biology research, novel approaches are often required. Novel algorithms became necessary due to the struggle with sampling (as discussed in the next section) from the posterior during Bayesian inference. The decision to

use Bayesian inference was due to its more robust treatment of uncertainty and bias [73]. However, this choice proved to be a major obstacle that impacted the trajectory of this research.

5.2.4 Challenges and Issues Encountered

The primary challenge of this dissertation was statistical sampling. This obstacle forced flexibility in order to adopt new methodologies. Over the course of this dissertation, numerous alternative sampling methods were tested: Metropolis-Hastings [165], Parallel Tempering [57], Hamiltonian Monte Carlo [20], Nested sampling [8], Annealed Importance sampling [174], Data Tempering methods [38], Variational methods [24], approximate Bayesian computation [16], slice sampling [175], sequential Monte Carlo [38], affine invariant ensemble sampling [79], and finally preconditioned Monte Carlo [122]. The mixed results of most of these methods led to the development of novel algorithms and extensions, including an adaptive grid with data tempering, a parallelized affine invariant sampler, and a 'top heavy' annealed importance sampler. The challenge of sampling that occurred during this dissertation cannot be understated.

An additional challenge was the standard issue that arises from software development and programming - bugs. Occasionally, small mistakes in the implementation of an algorithm or model went undetected for extended periods of time. This obstacle was exacerbated by the issues with sampling, making it difficult to disentangle the root causes of failure. The adoption of stricter testing and validation methods helped overcome this challenge.

Finally, a critical obstacle was the computational cost of simulating and fitting biochemical networks. Even for relatively simple networks associated with transporter proteins, the problem's dimensionality gets prohibitively large as new conformational states or biochemical species are added. This obstacle necessitates a careful choice of simulation engine (i.e., ordinary differential equation solver) and parallelization strategies to remain computationally feasible.

5.2.5 Evolution of Research

The direction of research changed dramatically throughout this dissertation. The most significant shift was the change from forward modeling with ModelExplorer to inverse modeling with Bayesian inference. This change was reflected on many levels. First, the core algorithms and tools are much different, and second, the datatype is different. As previously discussed, statistical sampling was a major obstacle to using Bayesian inference effectively. Consequently, the direction and scope of this dissertation moved from biological modeling and insight to focus on sampling methods development and evaluation more heavily. This new direction, while intriguing, resulted in less biologically relevant research than was initially intended.

5.3 Implications and Future Directions of the Research

5.3.1 Potential Impact

In this dissertation, new computational tools were developed and applied to better study transporter proteins. This may have far-reaching impacts on biological research and more.

The computational tools developed in this dissertation can greatly enhance the ability to study and predict the mechanistic complexity of transporters. However, by providing robust, efficient, and accurate methods that model biochemical networks at large, these tools could be utilized in a broader systems biology context to study a wide array of systems and gain valuable insight.[242, 2]

Also, the new knowledge from these models could potentially advance drug discovery and design methods. Since transporters play an essential role in drug transport within the body, a precise understanding of their mechanism could help with targeted drug design that aims to modulate transporter behavior or test their behavior under different conditions.

As an example, consider the mechanistic complexity of vSGLT and EmrE-like transporters. From the study of potential kinetic proofreading and multi-drug efflux mechanisms, we can gain insight into the survival strategies of bacteria in the presence of antibacterial drugs. This knowledge could be leveraged to improve the performance of antibacterial treatments, which may be less effective due to enhanced selectivity or multi-drug efflux pathways that are poorly understood. More specifically, the reaction rate constants and pathways that are predicted from our computational models could be used to fine-tune, and guide computer-aided drug design - providing key kinetic parameter ranges to search and test potential drug substances, which would then be tested and verified experimentally before translational trials.

Finally, this dissertation could potentially influence broader policy decisions. For example, this research could inform regulatory guidelines regarding drug testing or approval processes for drugs that interact with transporters.

5.3.2 Future Research Opportunities

This research has created several interesting areas for future work.

The immediate direction of research should incorporate experimental SSME data into the Bayesian inference and model selection pipeline. In collaboration with experimentalists, our synthetic data and models should be validated and updated as required based on this experimental data. Then, using Bayesian inference, the EmrE protein free-exchange model [200] could be robustly quantified from the data, providing even stronger evidence for the complexity of transporter mechanisms.

In addition, the current trajectory of research exploring transporter mechanisms could be extended to larger transporter models and experimental datasets, possibly from other sources besides EmrE. This would require fine-tuning the

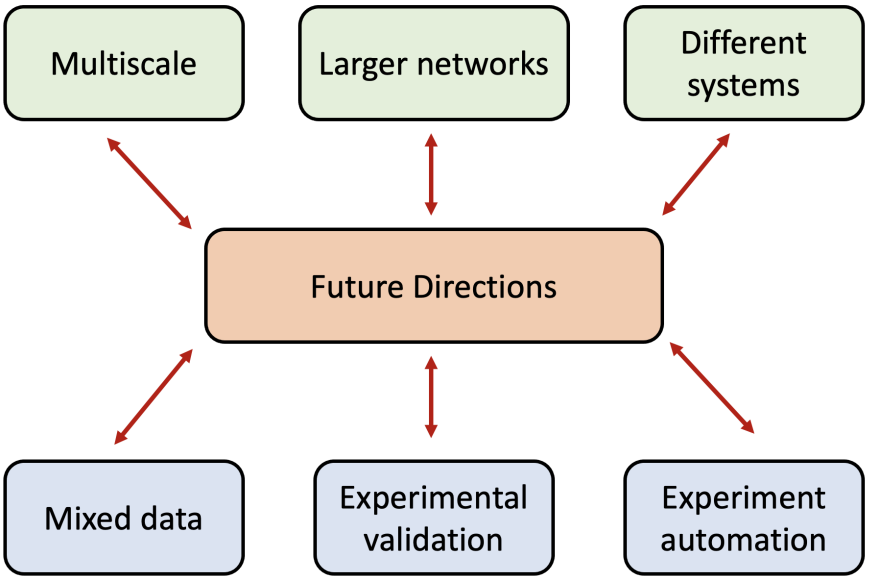


Figure 48: Future Directions

implementation to handle more extensive networks efficiently and potentially more detailed models to better account for experimental uncertainty.

Similarly, we could introduce orthogonal data from different experimental methods to improve the precision and identifiability of our model parameters (i.e., rate constants). This could be done initially through priors based on data from other experiments of the same system - such as with nuclear magnetic resonance experiments for transporters [200, 94]. Afterward, the orthogonal datasets could be integrated directly into the data-generating function and log-likelihood formulation, allowing for statistical inference as new data from either experimental method is added.

On a related note, the computational tools developed for this research could also likely be applied to study other biological systems that can be framed as a biochemical network. This could be, for example, a similar molecular machine such as a motor protein [5] kinesin or a more extensive metabolic network [56]. This would again require refining and exploring the methods for the particular system studied, aiming to balance precision and efficiency in the selected model and algorithms.

Also, future studies could focus on experimental validation of the theoretical predictions generated by this dissertation. While our tools were validated with synthetic data, they have not been well-tested with experimental data. Doing so would lend more credibility to the approaches and theoretical models discussed in this work - such as the kinetic proofreading models with enhanced selectivity.

Finally, the modeling could be expanded into other areas and levels of abstraction to provide a more holistic view of transporters, or other systems studied. For example, transporters could be integrated into a whole cell model in a more coarse-grained way but preserve their essential characteristics. Or different data sources could be integrated, such as

genomics or metabolic data. These combined analysis tools and datasets could allow for greater biological insight, such as metabolic data in light of transporter activity.

In summary, the tools and insights gained from this research contain many exciting possibilities for future work that advances biological understanding.

5.3.3 Limitations and Challenges

The research presented in this dissertation is not without its limitations and challenges, which may present opportunities for further research.

One of the primary limitations is the lack of relevant experimental data to validate the findings. While the synthetic results have had interesting implications, they require validation from experimental data. Future work should focus on coordinating computational predictions with expert experimental design and validation. This could be achieved through increased collaboration with experimental laboratories, and doing so would further the field.

Next, larger and more complex models are not tractable with the current sampling limitations. This limits the study of larger transporter systems or hierarchical models of transporters. Future work should focus on developing more efficient computational methods that can estimate the posterior (or similar metrics), potentially leveraging acceleration through machine learning techniques. This will enable the study of full reaction spaces of transporters in a Bayesian framework.

Additionally, in order to move towards a more complete multi-scale model of molecular machines, the challenge of integrating different levels of detail becomes apparent. There needs to be an efficient strategy to combine a highly detailed individual model of a transporter with more coarse-grained models that treat transporters as a node with a net flow. Future work could focus on developing multi-scale modeling and model integration methods.

Finally, the experiment recommendation system is too inefficient to use in an ‘online’ or real-time use, and it currently lacks the capacity to fully automate the process of experimental design. Efficiencies aside, future work could focus on enhancing the experiment recommendation system by incorporating the Bayesian inference methodology more directly into the experimental apparatus with a controller device. This could enable a fully automated and closed system to start, analyze, recommend, and repeat.

In summary, while these are notable challenges, they also serve as exciting opportunities for innovation and research. Implementing the above goals could dramatically advance the field as well as our understanding of transporters.

5.4 Conclusion

The central research theme of this dissertation is the detailed complexity of molecular machines, and specifically transporters, that compete against the notion of simple machine behavior. If these machines are as complex as emerging evidence suggests, how can improved computational approaches better model this behavior? This motivated the overarching objectives of the dissertation: the development of robust computational pipelines and the in silico

exploration of diverse transporter mechanisms. To realize these goals, a rigorous research methodology was used to develop new computational methods and implement pipelines for transporter research. These tools enabled the efficient and accurate investigation of alternative transporter mechanisms and the analysis of electrophysiology-type data. A key methodology was the newly developed ModelExplorer which provided a powerful engine to optimize and engineer transporter biochemical networks in silico. The implementation of ModelExplorer demonstrated resilience, flexibility, and the ability to identify novel transporter functions.

Concurrently, the implementation of advanced Bayesian methods delivered a statistical framework for the analysis of solid-supported membrane electrophysiology used for transporter research. This approach permitted the precise determination of microscopic rate constants from macroscopic data containing noise and sparsity - a novel development for this type of experimental assay and transporter system. In addition, experiment recommendation and model selection strategies were integrated to further enhance the power of the Bayesian inference pipeline. These contributions enabled greater precision for mechanistic differentiation and systematic experimental decision support.

The research presented in this dissertation aligns with the emerging perspective [200, 86, 239] that transporters exhibit complex mechanisms outside of the simple 1:1 transporter shown in textbooks often [5]. The exploration of transporter mechanisms and model classes of enhanced selectivity found by ModelExplorer strongly support this notion. In addition, the new computational methods developed offer more robust and powerful approaches than alternatives used for transporter research.

The future of molecular machine research seems very promising. Potential areas of work include the development and utilization of machine learning / deep learning methods to deal with more complex models and larger datasets, as well as the integration of multilevel models to capture a more comprehensive picture of these machines. More efforts to streamline the experiment recommendation and model comparison methods could significantly enhance the efficiency and performance of experimental data collection. Also, improved collaboration between computational and experimental scientists could lead to exceptionally better insights into the detailed functioning of transporters.

In conclusion, this dissertation displays an improvement in the computational exploration of transporter proteins. The methodologies and tools developed, along with the sights gained from their use, all significantly contribute to transporter research and systems biology overall. This progress provides many areas for future research and development, and it is hoped that the efforts of this dissertation will motivate that effort.

6 References

[1] Joshua L. Adelman, Chiara Ghezzi, Paola Bisignano, Donald D. F. Loo, Seungho Choe, Jeff Abramson, John M. Rosenberg, Ernest M. Wright, and Michael Grabe. Stochastic steps in secondary active sugar transport. *Proceedings of the National Academy of Sciences*, 113(27):E3960–E3966, July 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1525378113. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1525378113`.

[2] Alan Aderem. Systems biology: its practice and challenges. *Cell*, 121(4):511–513, 2005.

[3] John G Albeck, Gordon B Mills, and Joan S Brugge. Frequency-modulated pulses of erk activity transmit quantitative proliferation signals. *Molecular cell*, 49(2):249–261, 2013.

[4] Bruce Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 1998.

[5] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, New York, NY, sixth edition edition, 2015. ISBN 9780815344322 9780815344643 9780815345244.

[6] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2019.

[7] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342, 2010.

[8] Greg Ashton, Noam Bernstein, Johannes Buchner, Xi Chen, Gábor Csányi, Andrew Fowlie, Farhan Feroz, Matthew Griffiths, Will Handley, Michael Habeck, et al. Nested sampling for physical scientists. *Nature Reviews Methods Primers*, 2(1):39, 2022.

[9] Kinshuk Banerjee, Anatoly B. Kolomeisky, and Oleg A. Igoshin. Elucidating interplay of speed and accuracy in biological error correction. *Proceedings of the National Academy of Sciences*, 114(20):5183–5188, May 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1614838114. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1614838114`.

[10] Stefanie A Baril, Tomoka Gose, and John D Schuetz. How cryo-em has expanded our understanding of membrane transporters. *Drug Metabolism and Disposition*, 51(8):904–922, 2023.

[11] Kim Bartels, Tanya Lasitza-Male, Hagen Hofmann, and Christian Löw. Single-molecule fret of membrane transport proteins. *Chembiochem*, 22(17):2657–2671, 2021.

[12] Andre Bazzone, Wagner Steuer Costa, Markus Braner, Octavian Călinescu, Lina Hatahet, and Klaus Fendler. Introduction to solid supported membrane based electrophysiology. *JoVE (Journal of Visualized Experiments)*, 75:e50230, 2013.

[13] Andre Bazzone, Maria Barthmes, and Klaus Fendler. Ssm-based electrophysiology for transporter research. In *Methods in enzymology*, volume 594, pages 31–83. Elsevier, 2017.

[14] Andre Bazzone, Annas J. Zabadne, Anastasia Salisowski, M. Gregor Madej, and Klaus Fendler. A Loose Relationship: Incomplete H + /Sugar Coupling in the MFS Sugar Transporter GlcP. *Biophysical Journal*, 113: 2736–2749, December 2017. ISSN 00063495. doi: 10.1016/j.bpj.2017.09.038. URL `https://linkinghub.elsevier.com/retrieve/pii/S0006349517311190`.

[15] Daniel A Beard and Hong Qian. *Chemical biophysics: quantitative analysis of cellular systems*, volume 126. Cambridge University Press Cambridge, 2008.

[16] Mark A Beaumont. Approximate bayesian computation. *Annual review of statistics and its application*, 6: 379–403, 2019.

[17] Oren Ben-Kiki, Clark Evans, and Brian Ingerson. Yaml ain't markup language (yaml™) version 1.1. *Working Draft 2008*, 5:11, 2009.

[18] Howard C Berg. The rotary motor of bacterial flagella. *Annual review of biochemistry*, 72(1):19–54, 2003.

[19] Jeremy M. Berg, John L. Tymoczko, Lubert Stryer, and Lubert Stryer. *Biochemistry*. W.H. Freeman, New York, 5th ed edition, 2002. ISBN 9780716730514.

[20] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[21] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

[22] Paola Bisignano, Chiara Ghezzi, Hyunil Jo, Nicholas F. Polizzi, Thorsten Althoff, Chakrapani Kalyanaraman, Rosmarie Friemann, Matthew P. Jacobson, Ernest M. Wright, and Michael Grabe. Inhibitor binding mode and allosteric regulation of Na+-glucose symporters. *Nature Communications*, 9(1):5245, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07700-1. URL `http://www.nature.com/articles/s41467-018-07700-1`.

[23] Paola Bisignano, Michael A Lee, August George, Daniel M Zuckerman, Michael Grabe, and John M Rosenberg. A kinetic mechanism for enhanced selectivity of membrane transport. *PLoS computational biology*, 16(7): e1007789, 2020.

[24] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[25] Patrick D Bosshart and Dimitrios Fotiadis. Secondary active transporters. *Bacterial Cell Walls and Membranes*, pages 275–299, 2019.

[26] Richard J Boys, Darren J Wilkinson, and Thomas BL Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18:125–135, 2008.

[27] C Brosseau and E Sabri. Resistor–capacitor modeling of the cell membrane: A multiphysics analysis. *Journal of Applied Physics*, 129(1), 2021.

[28] Nia J Bryant, Roland Govers, and David E James. Regulated transport of the glucose transporter glut4. *Nature reviews Molecular cell biology*, 3(4):267–277, 2002.

[29] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

[30] John C. Mason and Markus W. Covert. An energetic reformulation of kinetic rate laws enables scalable parameter estimation for biochemical networks. *Journal of Theoretical Biology*, 461:145–156, January 2019. ISSN 00225193. doi: 10.1016/j.jtbi.2018.10.041. URL `https://linkinghub.elsevier.com/retrieve/pii/S002251931830523X`.

[31] Ben Calderhead, Michael Epstein, Lucia Sivilotti, and Mark Girolami. Bayesian approaches for mechanistic ion channel modeling. *In silico systems biology*, pages 247–272, 2013.

[32] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76, 2017.

[33] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

[34] Vijaysekhar Chellaboina, Sanjay P Bhat, Wassim M Haddad, and Dennis S Bernstein. Modeling and analysis of mass-action kinetics. *IEEE Control Systems Magazine*, 29(4):60–78, 2009.

[35] Zaineb Chelly Dagdia, Pavel Avdeyev, and Md Shamsuzzoha Bayzid. Biological computation and computational biology: survey, challenges, and discussion. *Artificial Intelligence Review*, 54:4169–4235, 2021.

[36] Siddhartha Chib. Markov chain monte carlo methods: computation and inference. *Handbook of econometrics*, 5: 3569–3649, 2001.

[37] Kiri Choi, J Kyle Medley, Matthias König, Kaylene Stocking, Lucian Smith, Stanley Gu, and Herbert M Sauro. Tellurium: an extensible python-based modeling environment for systems and synthetic biology. *Biosystems*, 171:74–79, 2018.

[38] Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

[39] Scott D Cohen, Alan C Hindmarsh, and Paul F Dubois. Cvode, a stiff/nonstiff ode solver in c. *Computers in physics*, 10(2):138–143, 1996.

[40] GM Cooper. *The Cell: A Molecular Approach*. Sinauer Associates, 2000.

[41] Athel Cornish-Bowden. One hundred years of michaelis–menten kinetics. *Perspectives in Science*, 4:3–9, 2015.

[42] Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[43] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[44] Svein G Dahl, Ingebrigt Sylte, and Aina Westrheim Ravna. Structures and models of transporter proteins. *Journal of Pharmacology and Experimental Therapeutics*, 309(3):853–860, 2004.

[45] Anastasia Deckard and Herbert M Sauro. Preliminary Studies on the In Silico Evolution of Biochemical Networks. *ChemBioChem*, 5(10):1423–1431, October 2004. ISSN 14394227. doi: 10.1002/cbic.200400178. URL `http://doi.wiley.com/10.1002/cbic.200400178`.

[46] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

[47] Natalia Demeshkina, Lasse Jenner, Eric Westhof, Marat Yusupov, and Gulnara Yusupova. A new understanding of the decoding principle on the ribosome. *Nature*, 484(7393):256–259, 2012.

[48] Oleg Demin Jr., Tatiana Yakovleva, Dmitry Kolobkov, and Oleg Demin. Analysis of the efficacy of SGLT2 inhibitors using semi-mechanistic model. *Frontiers in Pharmacology*, 5, October 2014. ISSN 1663-9812. doi: 10.3389/fphar.2014.00218. URL `http://journal.frontiersin.org/article/10.3389/fphar.2014.00218/abstract`.

[49] Dong Deng and Nieng Yan. GLUT, SGLT, and SWEET: Structural and mechanistic investigations of the glucose transporters: Structural Investigations of the Glucose Transporters. *Protein Science*, 25(3):546–558, March 2016. ISSN 09618368. doi: 10.1002/pro.2858. URL `http://doi.wiley.com/10.1002/pro.2858`.

[50] Dong Deng, Chao Xu, Pengcheng Sun, Jianping Wu, Chuangye Yan, Mingxu Hu, and Nieng Yan. Crystal structure of the human glucose transporter glut1. *Nature*, 510(7503):121–125, 2014.

[51] Nita Deshpande, Kenneth J Addess, Wolfgang F Bluhm, Jeffrey C Merino-Ott, Wayne Townsend-Merino, Qing Zhang, Charlie Knezevich, Lie Xie, Li Chen, Zukang Feng, et al. The rcsb protein data bank: a redesigned query system and relational database based on the mmcif schema. *Nucleic acids research*, 33(suppl_1):D233–D237, 2005.

[52] Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael R Andersen, Jonathan Huggins, and Aki Vehtari. Challenges and opportunities in high dimensional variational inference. *Advances in Neural Information Processing Systems*, 34:7787–7798, 2021.

[53] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. *Sequential Monte Carlo methods in practice*, pages 3–14, 2001.

[54] David Drew and Olga Boudker. Shared molecular mechanisms of membrane transporters. *Annual review of biochemistry*, 85:543–572, 2016.

[55] Ron O Dror, Robert M Dirks, JP Grossman, Huafeng Xu, and David E Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41:429–452, 2012.

[56] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007.

[57] David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[58] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[59] Aidan B Estelle, August George, Elisar J Barbar, and Daniel M Zuckerman. Quantifying cooperative multisite binding in the hub protein lc8 through bayesian inference. *PLoS computational biology*, 19(4):e1011059, 2023.

[60] James R Faeder, Michael L Blinov, and William S Hlavacek. Rule-based modeling of biochemical systems with bionetgen. *Systems biology*, pages 113–167, 2009.

[61] Alan R. Fersht. Editing mechanisms in protein synthesis. Rejection of valine by the isoleucyl-tRNA synthetase. *Biochemistry*, 16(5):1025–1030, March 1977. ISSN 0006-2960, 1520-4995. doi: 10.1021/bi00624a034. URL `https://pubs.acs.org/doi/abs/10.1021/bi00624a034`.

[62] Stephen K. Field. Bedaquiline for the treatment of multidrug-resistant tuberculosis: great promise or disappointment? *Therapeutic Advances in Chronic Disease*, 6(4):170–184, July 2015. ISSN 2040-6223. doi: 10.1177/2040622315582325.

[63] M. E. Fisher and A. B. Kolomeisky. The force exerted by a molecular motor. *Proceedings of the National Academy of Sciences*, 96(12):6597–6602, June 1999. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.96.12.6597. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.96.12.6597`.

[64] Lucy R Forrest, Reinhard Krämer, and Christine Ziegler. The structural basis of secondary active transport mechanisms. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1807(2):167–188, 2011.

[65] Steven A Frank. Input-output relations in biological systems: measurement, information and the hill equation. *Biology direct*, 8(1):1–25, 2013.

[66] Nial Friel and Anthony N Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(3):589–607, 2008.

[67] Fabian Fröhlich, Fabian J Theis, and Jan Hasenauer. Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more. In *Computational Methods in Systems Biology: 12th International Conference, CMSB 2014, Manchester, UK, November 17-19, 2014, Proceedings 12*, pages 61–72. Springer, 2014.

[68] James E Galagan, Kyle Minch, Matthew Peterson, Anna Lyubetskaya, Elham Azizi, Linsday Sweet, Antonio Gomes, Tige Rustad, Gregory Dolganov, Irina Glotova, et al. The mycobacterium tuberculosis regulatory network and hypoxia. *Nature*, 499(7457):178–183, 2013.

[69] Vahe Galstyan, Luke Funk, Tal Einav, and Rob Phillips. Combinatorial Control through Allostery. *The Journal of Physical Chemistry B*, 123(13):2792–2800, April 2019. ISSN 1520-6106, 1520-5207. doi: 10.1021/acs.jpcb.8b12517. URL `https://pubs.acs.org/doi/10.1021/acs.jpcb.8b12517`.

[70] Juan J Garcia-Celma, Irina N Smirnova, H Ronald Kaback, and Klaus Fendler. Electrophysiological characterization of lacy. *Proceedings of the National Academy of Sciences*, 106(18):7373–7378, 2009.

[71] Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241–4257, 2000.

[72] Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *International conference on artificial intelligence and statistics*, pages 1682–1690. PMLR, 2018.

[73] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[74] August George, Paola Bisignano, John M Rosenberg, Michael Grabe, and Daniel M Zuckerman. A systems-biology approach to molecular machines: Exploration of alternative transporter mechanisms. *PLoS computational biology*, 16(7):e1007884, 2020.

[75] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015.

[76] Ulrik Gether, Peter H Andersen, Orla M Larsson, and Arne Schousboe. Neurotransmitter transporters: molecular function of important drug targets. *Trends in pharmacological sciences*, 27(7):375–383, 2006.

[77] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[78] Amos Golan and John Harte. Information theory: A foundation for complexity science. *Proceedings of the National Academy of Sciences*, 119(33):e2119089119, 2022.

[79] Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010.

[80] Nico S Gorbach, Stefan Bauer, and Joachim M Buhmann. Scalable variational inference for dynamical systems. *Advances in neural information processing systems*, 30, 2017.

[81] Marko Gosak, Rene Markovič, Jurij Dolenšek, Marjan Slak Rupnik, Marko Marhl, Andraž Stožer, and Matjaž Perc. Network science of biological systems at different scales: A review. *Physics of life reviews*, 24:118–135, 2018.

[82] Willi Gottstein, Stefan Müller, Hanspeter Herzel, and Ralf Steuer. Elucidating the adaptation and temporal coordination of metabolic pathways using in-silico evolution. *Biosystems*, 117:68–76, March 2014. ISSN

03032647. doi: 10.1016/j.biosystems.2013.12.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0303264713002475.

[83] Eric Gouaux and Roderick MacKinnon. Principles of selective ion transport in channels and pumps. *science*, 310(5753):1461–1465, 2005.

[84] J. Goutsias and G. Jenkinson. Markovian dynamics on complex reaction networks. *Physics Reports*, 529(2): 199–264, August 2013. ISSN 03701573. doi: 10.1016/j.physrep.2013.03.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0370157313001014.

[85] Michael Grabe, Harold Lecar, Yuh Nung Jan, and Lily Yeh Jan. A quantitative assessment of models for voltage-dependent gating of ion channels. *Proceedings of the National Academy of Sciences*, 101(51): 17640–17645, 2004.

[86] Michael Grabe, Daniel M Zuckerman, and John M Rosenberg. Emre reminds us to expect the unexpected in membrane transport. *Journal of General Physiology*, 152(1), 2020.

[87] Jasleen K Grewal, Martin Krzywinski, and Naomi Altman. Markov models—markov chains. *Nat. Methods*, 16 (8):663–664, 2019.

[88] Christof Grewer, Armanda Gameiro, Thomas Mager, and Klaus Fendler. Electrophysiological characterization of membrane transport proteins. *Annual review of biophysics*, 42:95–120, 2013.

[89] Ivo Grosse, Hanspeter Herzel, Sergey V Buldyrev, and H Eugene Stanley. Species independence of mutual information in coding and noncoding dna. *Physical Review E*, 61(5):5624, 2000.

[90] Hoshin V Gupta, Keith J Beven, and Thorsten Wagener. Model calibration and uncertainty estimation. *Encyclopedia of hydrological sciences*, 2006.

[91] Meron Gurkiewicz and Alon Korngreen. A Numerical Approach to Ion Channel Modelling Using Whole-Cell Voltage-Clamp Recordings and a Genetic Algorithm. *PLoS Computational Biology*, 3(8):e169, 2007. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.0030169. URL https://dx.plos.org/10.1371/journal.pcbi.0030169.

[92] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109, April 1970. ISSN 1464-3510, 0006-3444. doi: 10.1093/biomet/57.1.97. URL https://academic.oup.com/biomet/article/57/1/97/284580.

[93] Reinhart Heinrich, Benjamin G Neel, and Tom A Rapoport. Mathematical models of protein kinase signal transduction. *Molecular cell*, 9(5):957–970, 2002.

[94] Ute A Hellmich and Clemens Glaubitz. Nmr and epr studies of membrane transporters. *Biological Chemistry*, 2009.

[95] Ryan K Henderson, Klaus Fendler, and Bert Poolman. Coupling efficiency of secondary active transporters. *Current Opinion in Biotechnology*, 58:62–71, August 2019. ISSN 09581669. doi: 10.1016/j.copbio.2018.11.005. URL https://linkinghub.elsevier.com/retrieve/pii/S095816691830185X.

[96] Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.

[97] Christopher F Higgins. Multiple molecular mechanisms for multidrug resistance transporters. *Nature*, 446 (7137):749–757, 2007.

[98] Desmond J Higham. Modeling and simulating chemical reactions. *SIAM review*, 50(2):347–368, 2008.

[99] Desmond J Higham and Nicholas J Higham. *MATLAB guide*. SIAM, 2016.

[100] Terrell L. Hill. *Free energy transduction in biology: the steady-state kinetic and thermodynamic formalism*. Academic Press, New York, 1977. ISBN 9780123482501.

[101] Terrell L. Hill. *Free Energy Transduction and Biochemical Cycle Kinetics*. Springer New York, New York, NY, 1989. ISBN 9780387968360 9781461235583. doi: 10.1007/978-1-4612-3558-3. URL http://link.springer.com/10.1007/978-1-4612-3558-3.

[102] Bertil Hille et al. Ion channels of excitable membranes (vol. 507). *Sunderland, MA: Sinauer*, 2001.

[103] Nobutaka Hirokawa. Kinesin and dynein superfamily proteins and the mechanism of organelle transport. *Science*, 279(5350):519–526, 1998.

[104] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

[105] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[106] Martin Holz and Alfred Fahr. Compartment modeling. *Advanced Drug Delivery Reviews*, 48(2-3):249–264, 2001.

[107] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. Copasi—a complex pathway simulator. *Bioinformatics*, 22(24): 3067–3074, 2006.

[108] J. J. Hopfield. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proceedings of the National Academy of Sciences*, 71(10):4135–4139, October 1974. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.71.10.4135. URL http://www.pnas.org/cgi/doi/10.1073/pnas.71.10.4135.

[109] Shaohui Huang and Michael P Czech. The glut4 glucose transporter. *Cell metabolism*, 5(4):237–252, 2007.

[110] Holly A Huber, Senta K Georgia, and Stacey D Finley. Systematic bayesian posterior analysis guided by kullback-leibler divergence facilitates hypothesis formation. *Journal of Theoretical Biology*, 558:111341, 2023.

[111] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4): 524–531, 2003.

[112] Timothy R Hughes, Matthew J Marton, Allan R Jones, Christopher J Roberts, Roland Stoughton, Christopher D Armour, Holly A Bennett, Ernest Coffey, Hongyue Dai, Yudong D He, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[113] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[114] Grant A Hussey, Nathan E Thomas, and Katherine A Henzler-Wildman. Highly coupled transport can be achieved in free-exchange transport models. *Journal of General Physiology*, 152(1):e201912437, 2019.

[115] Soojin Jang. Multidrug efflux pumps in staphylococcus aureus and their clinical implications. *Journal of Microbiology*, 54:1–8, 2016.

[116] Oleg Jardetzky. Simple Allosteric Model for Membrane Pumps. *Nature*, 211(5052):969–970, August 1966. ISSN 0028-0836, 1476-4687. doi: 10.1038/211969a0. URL `http://www.nature.com/articles/211969a0`.

[117] Galin L Jones and Qian Qin. Markov chain monte carlo in practice. *Annual Review of Statistics and Its Application*, 9:557–578, 2022.

[118] Chirlmin Joo, Hamza Balci, Yuji Ishitsuka, Chittanon Buranachai, and Taekjip Ha. Advances in single-molecule fluorescence methods for molecular biology. *Annu. Rev. Biochem.*, 77:51–76, 2008.

[119] Iman Imtiyaz Ahmed Juvale, Azzmer Azzar Abdul Hamid, Khairul Bariyyah Abd Halim, and Ahmad Tarmizi Che Has. P-glycoprotein: new insights into structure, physiological function, regulation and alterations in disease. *Heliyon*, 2022.

[120] Sanjay Kalra. Sodium Glucose Co-Transporter-2 (SGLT2) Inhibitors: A Review of Their Basic and Clinical Pharmacology. *Diabetes Therapy*, 5(2):355–366, December 2014. ISSN 1869-6953. doi: 10.1007/s13300-014-0089-4. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4269649/`.

[121] Jack H Kaplan. Biochemistry of na, k-atpase. *Annual review of biochemistry*, 71(1):511–535, 2002.

[122] Minas Karamanis, Florian Beutler, John A Peacock, David Nabergoj, and Uroš Seljak. Accelerating astronomical and cosmological inference with preconditioned monte carlo. *Monthly Notices of the Royal Astronomical Society*, 516(2):1644–1653, 2022.

[123] Minas Karamanis, David Nabergoj, Florian Beutler, John A Peacock, and Uros Seljak. pocomc: A python package for accelerated bayesian inference in astronomy and cosmology. *arXiv preprint arXiv:2207.05660*, 2022.

[124] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature structural biology*, 9(9):646–652, 2002.

[125] Martin Karplus and Gregory A Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631–639, 1990.

[126] Sourabh Katoch, Sumit Singh Chauhan, and Vijay Kumar. A review on genetic algorithm: past, present, and future. *Multimedia tools and applications*, 80:8091–8126, 2021.

[127] Sarah M Keating, Dagmar Waltemath, Matthias König, Fengkai Zhang, Andreas Dräger, Claudine Chaouiya, Frank T Bergmann, Andrew Finney, Colin S Gillespie, Tomáš Helikar, et al. Sbml level 3: an extensible format for the exchange and reuse of biological models. *Molecular systems biology*, 16(8):e9110, 2020.

[128] Carolyn Keenan and Dermot Kelleher. Protein kinase c and the cytoskeleton. *Cellular signalling*, 10(4): 225–232, 1998.

[129] Terry Kenakin. The mass action equation in pharmacology. *British Journal of Clinical Pharmacology*, 81(1): 41–51, 2016.

[130] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE, 1995.

[131] Ali A Kermani. A guide to membrane protein x-ray crystallography. *The FEBS journal*, 288(20):5788–5804, 2021.

[132] Carsten Kettner, Adam Bertl, Gerhard Obermeyer, Clifford Slayman, and Hermann Bihler. Electrophysiological Analysis of the Yeast V-Type Proton Pump: Variable Coupling Ratio and Proton Shunt. *Biophysical Journal*, 85 (6):3730–3738, December 2003. ISSN 00063495. doi: 10.1016/S0006-3495(03)74789-4. URL `https://linkinghub.elsevier.com/retrieve/pii/S0006349503747894`.

[133] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220 (4598):671–680, 1983.

[134] David G Kleinbaum, Mitchel Klein, David G Kleinbaum, and Mitchel Klein. Maximum likelihood techniques: An overview. *Logistic regression: A self-learning text*, pages 103–127, 2010.

[135] Amira Klip, Timothy E McGraw, and David E James. Thirty sweet years of glut4. *Journal of Biological Chemistry*, 294(30):11369–11381, 2019.

[136] Kevin H Knuth, Michael Habeck, Nabin K Malakar, Asim M Mubeen, and Ben Placek. Bayesian evidence and model selection. *Digital Signal Processing*, 47:50–67, 2015.

[137] Mykel J Kochenderfer and Tim A Wheeler. *Algorithms for optimization*. Mit Press, 2019.

[138] Trupti M Kodinariya, Prashant R Makwana, et al. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[139] Tamara G Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM review*, 45(3):385–482, 2003.

[140] Roman V Kondratov, Pavel G Komarov, Yigal Becker, Ariel Ewenson, and Andrei V Gudkov. Small molecules that dramatically alter multidrug resistance phenotype by modulating the substrate specificity of p-glycoprotein. *Proceedings of the National Academy of Sciences*, 98(24):14078–14083, 2001.

[141] Arnost Kotyk. *Cell membrane transport: principles and techniques*. Springer Science & Business Media, 2012.

[142] Dieter Kraft. A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[143] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[144] Clemens Kreutz, Andreas Raue, Daniel Kaschek, and Jens Timmer. Profile likelihood in systems biology. *The FEBS journal*, 280(11):2564–2571, 2013.

[145] Anders S Kristensen, Jacob Andersen, Trine N Jørgensen, Lena Sørensen, Jacob Eriksen, Claus J Loland, Kristian Strømgaard, and Ulrik Gether. Slc6 neurotransmitter transporters: structure, function, and regulation. *Pharmacological reviews*, 63(3):585–640, 2011.

[146] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[147] Christopher J Law, Peter C Maloney, and Da-Neng Wang. Ins and outs of major facilitator superfamily antiporters. *Annu. Rev. Microbiol.*, 62:289–305, 2008.

[148] Quoc V Le et al. A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. *Google Brain*, 20:1–20, 2015.

[149] Kuo-Ching Liang and Xiaodong Wang. Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008:1–14, 2008.

[150] Qiaohao Liang, Aldair E Gongora, Zekun Ren, Armi Tiihonen, Zhe Liu, Shijing Sun, James R Deneault, Daniil Bash, Flore Mekki-Berrada, Saif A Khan, et al. Benchmarking the performance of bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, 7(1):188, 2021.

[151] Juliane Liepe, Sarah Filippi, Michał Komorowski, and Michael PH Stumpf. Maximizing the information content of experiments in systems biology. *PLoS computational biology*, 9(1):e1002888, 2013.

[152] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.

[153] Wendell A. Lim, Connie M. Lee, and Chao Tang. Design Principles of Regulatory Networks: Searching for the Molecular Algorithms of the Cell. *Molecular Cell*, 49(2):202–212, January 2013. ISSN 10972765. doi:

10.1016/j.molcel.2012.12.020. URL https://linkinghub.elsevier.com/retrieve/pii/S109727651300004X.

[154] Lawrence Lin, Sook Wah Yee, Richard B Kim, and Kathleen M Giacomini. Slc transporters as therapeutic targets: emerging opportunities. *Nature reviews Drug discovery*, 14(8):543–560, 2015.

[155] Nathaniel J Linden, Boris Kramer, and Padmini Rangamani. Bayesian parameter estimation for dynamical models in systems biology. *PLOS Computational Biology*, 18(10):e1010651, 2022.

[156] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[157] Stephen M Stahl, Clara Lee-Zimmerman, Sylvia Cartwright, and Debbi Ann Morrissette. Serotonergic drugs for depression and beyond. *Current drug targets*, 14(5):578–585, 2013.

[158] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[159] Joel D. Mallory, Anatoly B. Kolomeisky, and Oleg A. Igoshin. Trade-Offs between Error, Speed, Noise, and Energy Dissipation in Biological Processes with Proofreading. *The Journal of Physical Chemistry B*, 123(22): 4718–4725, June 2019. ISSN 1520-6106, 1520-5207. doi: 10.1021/acs.jpcb.9b03757. URL https://pubs.acs.org/doi/10.1021/acs.jpcb.9b03757.

[160] Kathryn E. Mangold, Brittany D. Brumback, Paweorn Angsutararux, Taylor L. Voelker, Wandi Zhu, Po Wei Kang, Jonathan D. Moreno, and Jonathan R. Silva. Mechanisms and models of cardiac sodium channel inactivation. *Channels*, 11(6):517–533, November 2017. ISSN 1933-6950, 1933-6969. doi: 10.1080/19336950.2017.1369637. URL https://www.tandfonline.com/doi/full/10.1080/19336950.2017.1369637.

[161] Hisham Mazal and Gilad Haran. Single-molecule fret methods to study the dynamics of proteins at work. *Current opinion in biomedical engineering*, 12:8–17, 2019.

[162] Robert D McMichael, Sean M Blakley, and Sergey Dushenko. Optbayesexpt: Sequential bayesian experiment design for adaptive measurements. *Journal of Research of the National Institute of Standards and Technology*, 126:1–5, 2021.

[163] Pedro Mendes, Stefan Hoops, Sven Sahle, Ralph Gauges, Joseph Dada, and Ursula Kummer. Computational modeling of biochemical networks using copasi. *Systems Biology*, pages 17–59, 2009.

[164] V. Menon, N. Spruston, and W. L. Kath. A state-mutating genetic algorithm to design ion-channel models. *Proceedings of the National Academy of Sciences*, 106(39):16829–16834, September 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0903766106. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0903766106.

[165] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):

1087–1092, June 1953. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.1699114. URL http://aip.scitation.org/doi/10.1063/1.1699114.

[166] Peter Mitchell. A General Theory of Membrane Transport From Studies of Bacteria. *Nature*, 180(4577): 134–136, July 1957. ISSN 0028-0836, 1476-4687. doi: 10.1038/180134a0. URL http://www.nature.com/articles/180134a0.

[167] Eshan D Mitra and William S Hlavacek. Parameter estimation and uncertainty quantification for systems biology models. *Current opinion in systems biology*, 18:9–18, 2019.

[168] WE Moerner and David P Fromm. Methods of single-molecule fluorescence spectroscopy and microscopy. *Review of Scientific instruments*, 74(8):3597–3619, 2003.

[169] Jan L Münch, Fabian Paul, Ralf Schmauder, and Klaus Benndorf. Bayesian inference of kinetic schemes for ion channels by kalman filtering. *Elife*, 11:e62714, 2022.

[170] Dennis L Murphy, Alicja Lerner, Gary Rudnick, and Klaus-Peter Lesch. Serotonin transporter: gene, genetic disorders, and pharmacogenetics. *Molecular interventions*, 4(2):109, 2004.

[171] Kenneth P Murphy, Dong Xie, Kelly S Thompson, L Mario Amzel, and Ernesto Freire. Entropy in biological binding processes: estimation of translational entropy loss. *Proteins: Structure, Function, and Bioinformatics*, 18(1):63–67, 1994.

[172] Arvind Murugan, David A Huse, and Stanislas Leibler. Discriminatory Proofreading Regimes in Nonequilibrium Systems. *Physical Review X*, 4(2):021016, April 2014. doi: 10.1103/PhysRevX.4.021016. URL https://link.aps.org/doi/10.1103/PhysRevX.4.021016.

[173] Takashi Nakakuki, Marc R Birtwistle, Yuko Saeki, Noriko Yumoto, Kaori Ide, Takeshi Nagashima, Lutz Brusch, Babatunde A Ogunnaike, Mariko Okada-Hatakeyama, and Boris N Kholodenko. Ligand-specific c-fos expression emerges from the spatiotemporal control of erbb network dynamics. *Cell*, 141(5):884–896, 2010.

[174] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

[175] Radford M Neal. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

[176] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[177] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4): 308–313, 1965.

[178] Jacques Ninio. Kinetic amplification of enzyme discrimination. *Biochimie*, 57(5):587–595, July 1975. ISSN 03009084. doi: 10.1016/S0300-9084(75)80139-8. URL https://linkinghub.elsevier.com/retrieve/pii/S0300908475801398.

[179] Ruth Nussinov. Advancements and challenges in computational biology. *PLoS computational biology*, 11(1): e1004053, 2015.

[180] Julien F. Ollivier, Vahid Shahrezaei, and Peter S. Swain. Scalable Rule-Based Modelling of Allosteric Proteins and Biochemical Networks. *PLoS Computational Biology*, 6(11):e1000975, November 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000975. URL https://dx.plos.org/10.1371/journal.pcbi.1000975.

[181] Manfred Opper and Guido Sanguinetti. Variational inference for markov jump processes. *Advances in neural information processing systems*, 20, 2007.

[182] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[183] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[184] Anand Patil, David Huard, and Christopher J Fonnesbeck. Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, 35(4):1, 2010.

[185] Rob Phillips, Jane Kondev, and Julie Theriot. *Physical biology of the cell*. Garland Science, New York, 2009. ISBN 9780815341635. OCLC: 234234136.

[186] Chayne L Piscitelli and Eric Gouaux. Insights into transport mechanism from leut engineered to transport tryptophan. *The EMBO journal*, 31(1):228–235, 2012.

[187] Michael Plaksin, Eitan Kimmel, and Shy Shoham. Correspondence: Revisiting the theoretical cell membrane thermal capacitance response. *Nature communications*, 8(1):1431, 2017.

[188] Bernd Pompe, Pierre Blidh, Dirk Hoyer, and Michael Eiselt. Using mutual information to measure coupling in the cardiorespiratory system. *IEEE Engineering in Medicine and Biology Magazine*, 17(6):32–39, 1998.

[189] Søren Brandt Poulsen, Robert A. Fenton, and Timo Rieg. Sodium-glucose cotransport:. *Current Opinion in Nephrology and Hypertension*, 24(5):463–469, September 2015. ISSN 1062-4821. doi: 10.1097/MNH.0000000000000152. URL http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00041552-201509000-00011.

[190] Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[191] Michael JD Powell. *A direct search optimization method that models the objective and constraint functions by linear interpolation*. Springer, 1994.

[192] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, 134(17), 2011.

[193] Yu-Hui Qu, Hua Yu, Xiu-Jun Gong, Jia-Hui Xu, and Hong-Shun Lee. On the prediction of dna-binding proteins only from primary sequences: A deep learning approach. *PloS one*, 12(12):e0188129, 2017.

[194] Christopher Rackauckas and Qing Nie. Differentialequations. jl–a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of open research software*, 5(1), 2017.

[195] Venki Ramakrishnan. Ribosome structure and the mechanism of translation. *Cell*, 108(4):557–572, 2002.

[196] Andreas Raue, Marcel Schilling, Julie Bachmann, Andrew Matteson, Max Schelke, Daniel Kaschek, Sabine Hug, Clemens Kreutz, Brian D. Harms, Fabian J. Theis, Ursula Klingmüller, and Jens Timmer. Lessons learned from quantitative dynamical modeling in systems biology. *PLOS ONE*, 8(9):1–17, 09 2013. doi: 10.1371/journal.pone.0074335. URL `https://doi.org/10.1371/journal.pone.0074335`.

[197] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[198] Luis Miguel Rios and Nikolaos V Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56:1247–1293, 2013.

[199] Robert W Robey, Kristen M Pluchino, Matthew D Hall, Antonio T Fojo, Susan E Bates, and Michael M Gottesman. Revisiting the role of abc transporters in multidrug-resistant cancer. *Nature Reviews Cancer*, 18(7): 452–464, 2018.

[200] Anne E. Robinson, Nathan E. Thomas, Emma A. Morrison, Bryan M. Balthazor, and Katherine A. Henzler-Wildman. New free-exchange model of EmrE transport. *Proceedings of the National Academy of Sciences*, 114(47):E10083–E10091, November 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1708671114. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1708671114`.

[201] Lior Rokach and Oded Maimon. Clustering Methods. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 321–352. Springer-Verlag, New York, 2005. ISBN 9780387244358. doi: 10.1007/0-387-25465-X_15. URL `http://link.springer.com/10.1007/0-387-25465-X_15`.

[202] Rahul Roy, Sungchul Hohng, and Taekjip Ha. A practical guide to single-molecule fret. *Nature methods*, 5(6): 507–516, 2008.

[203] Vivekananda Roy. Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 7:387–412, 2020.

[204] David Ruppert. The elements of statistical learning: data mining, inference, and prediction, 2004.

[205] G Sachs, JM Shin, and CW Howden. The clinical pharmacology of proton pump inhibitors. *Alimentary pharmacology & therapeutics*, 23:2–8, 2006.

[206] Bert Sakmann and Erwin Neher. Patch clamp techniques for studying ionic channels in excitable membranes. *Annual review of physiology*, 46(1):455–472, 1984.

[207] Katrin Sangkuhl, Teri Klein, and Russ Altman. Selective serotonin reuptake inhibitors (ssri) pathway. *Pharmacogenetics and genomics*, 19(11):907, 2009.

[208] Yannik Schälte, Fabian Fröhlich, Paul J Jost, Jakob Vanhoefer, Dilan Pathirana, Paul Stapor, Polina Lakrisenko, Dantong Wang, Elba Raimúndez, Simon Merkt, et al. pypesto: A modular and scalable tool for parameter estimation for dynamic models. *arXiv preprint arXiv:2305.01821*, 2023.

[209] Leonard Schmiester, Yannik Schälte, Frank T Bergmann, Tacio Camba, Erika Dudkin, Janine Egert, Fabian Fröhlich, Lara Fuhrmann, Adrian L Hauber, Svenja Kemmer, et al. Petab—interoperable specification of parameter estimation problems in systems biology. *PLoS computational biology*, 17(1):e1008646, 2021.

[210] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9):093001, March 2017. ISSN 1751-8113, 1751-8121. doi: 10.1088/1751-8121/aa54d9. URL `http://stacks.iop.org/1751-8121/50/i=9/a=093001?key=crossref.d8b950432e1fa1388f3a8600ab8ff70e`.

[211] Shimon Schuldiner, Dorit Granot, Sonia Steiner Mordoch, Shira Ninio, Dvir Rotem, Michael Soskin, Christopher G Tate, and Hagit Yerushalmi. Small is mighty: Emre, a multidrug transporter as an experimental paradigm. *Physiology*, 16(3):130–134, 2001.

[212] Benjamin Schuler and William A Eaton. Protein folding studied by single-molecule fret. *Current opinion in structural biology*, 18(1):16–26, 2008.

[213] Stanley G Schultz. *Basic principles of membrane transport*. Cambridge University Press, Cambridge, MA, 1980.

[214] Patrick Schulz, Juan J Garcia-Celma, and Klaus Fendler. Ssm-based electrophysiology. *Methods*, 46(2):97–103, 2008.

[215] Markus A Seeger. Membrane transporter research in times of countless structures. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1860(4):804–808, 2018.

[216] John A.P. Sekar, Justin S. Hogg, and James R. Faeder. Energy-based modeling in BioNetGen. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1460–1467, Shenzhen, China, December 2016. IEEE. ISBN 9781509016112. doi: 10.1109/BIBM.2016.7822739. URL `http://ieeexplore.ieee.org/document/7822739/`.

[217] Radu Serban and Alan C Hindmarsh. Cvodes: the sensitivity-enabled ode solver in sundials. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 47438, pages 257–269, 2005.

[218] Ben Shababo, Brooks Paige, Ari Pakman, and Liam Paninski. Bayesian inference and online experimental design for mapping neural microcircuits. *Advances in Neural Information Processing Systems*, 26, 2013.

[219] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

[220] Sanjib Sharma. Markov chain monte carlo methods for bayesian data analysis in astronomy. *Annual Review of Astronomy and Astrophysics*, 55:213–259, 2017.

[221] Jesse A Sharp, Alexander P Browning, Kevin Burrage, and Matthew J Simpson. Parameter estimation and uncertainty quantification using information geometry. *Journal of the Royal Society Interface*, 19(189): 20210940, 2022.

[222] Lei Shi, Matthias Quick, Yongfang Zhao, Harel 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*, 30(6):667–677, June 2008. ISSN 10972765. doi: 10.1016/j.molcel.2008.05.008. URL `https://linkinghub.elsevier.com/retrieve/pii/S1097276508003596`.

[223] Brian H Shilton. Small-angle x-ray scattering studies of membrane transporter peripheral components and soluble domains. *Membrane Transporters: Methods and Protocols*, pages 281–306, 2003.

[224] Ivo Siekmann, Larry E Wagner, David Yule, Colin Fox, David Bryant, Edmund J Crampin, and James Sneyd. Mcmc estimation of markov models for ion channels. *Biophysical journal*, 100(8):1919–1929, 2011.

[225] Adam Siepel, Katherine S Pollard, and David Haussler. New methods for detecting lineage-specific selection. In *Annual International Conference on Research in Computational Molecular Biology*, pages 190–205. Springer, 2006.

[226] Lucian P Smith, Frank T Bergmann, Deepak Chandran, and Herbert M Sauro. Antimony: a modular model definition language. *Bioinformatics*, 25(18):2452–2454, 2009.

[227] MW Sneddon, James R Faeder, and T Emonet. Efficient modeling, simulation and coarse-graining of biological complexity with nfsim. *Nat Methods*, 8(2):177–183, Feb 2011. doi: 10.1038/nmeth.1546.

[228] Iwona Sobczak and Juke S. Lolkema. The 2-Hydroxycarboxylate Transporter Family: Physiology, Structure, and Mechanism. *Microbiology and Molecular Biology Reviews*, 69(4):665–695, December 2005. ISSN 1092-2172, 1098-5557. doi: 10.1128/MMBR.69.4.665-695.2005. URL `https://mmbr.asm.org/content/69/4/665`.

[229] Siowling Soh, Marta Byrska, Kristiana Kandere-Grzybowska, and Bartosz A Grzybowski. Reaction-diffusion systems in intracellular molecular transport and control. *Angewandte Chemie International Edition*, 49(25): 4170–4198, 2010.

[230] Endre T Somogyi, Jean-Marie Bouteiller, James A Glazier, Matthias König, J Kyle Medley, Maciej H Swat, and Herbert M Sauro. libroadrunner: a high performance sbml simulation and analysis library. *Bioinformatics*, 31 (20):3315–3321, 2015.

[231] Bharath Srinivasan. A guide to the michaelis–menten equation: steady state and beyond. *The FEBS journal*, 289 (20):6086–6098, 2022.

[232] Thomas A Steitz. A structural understanding of the dynamic ribosome machine. *Nature reviews Molecular cell biology*, 9(3):242–253, 2008.

[233] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11:341–359, 1997.

[234] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.

[235] Tamás Székely Jr and Kevin Burrage. Stochastic simulation in systems biology. *Computational and structural biotechnology journal*, 12(20-21):14–25, 2014.

[236] Abd A Tahrani, Anthony H Barnett, and Clifford J Bailey. Sglt inhibitors in management of diabetes. *The lancet Diabetes & endocrinology*, 1(2):140–151, 2013.

[237] Zachary R. Teed and Jonathan R. Silva. A computationally efficient algorithm for fitting ion channel parameters. *MethodsX*, 3:577–588, 2016. ISSN 22150161. doi: 10.1016/j.mex.2016.11.001. URL `https://linkinghub.elsevier.com/retrieve/pii/S2215016116300395`.

[238] Nuno Tenazinha and Susana Vinga. A survey on methods for modeling and analyzing integrated biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):943–958, 2010.

[239] Daniel S Terry, Rachel A Kolster, Matthias Quick, Michael V LeVine, George Khelashvili, Zhou Zhou, Harel Weinstein, Jonathan A Javitch, and Scott C Blanchard. A partially-open inward-facing intermediate conformation of leut is associated with na+ release and substrate transport. *Nature communications*, 9(1):230, 2018.

[240] Nathan E Thomas, Wei Feng, and Katherine A Henzler-Wildman. A solid-supported membrane electrophysiology assay for efficient characterization of ion-coupled transport. *Journal of Biological Chemistry*, 297(4), 2021.

[241] Peter Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends in biochemical sciences*, 37(12):509–516, 2012.

[242] Anthony Trewavas. A brief history of systems biology: "every object that biology studies is a system of systems." francois jacob (1974). *The Plant Cell*, 18(10):2420–2430, 2006.

[243] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in science & engineering*, 13(2):22–30, 2011.

[244] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[245] Joep Vanlier, Christian A Tiemann, Peter AJ Hilbers, and Natal AW van Riel. A bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, 2012.

[246] Alexei Verkhratsky and Vladimir Parpura. History of electrophysiology and the patch clamp. *Patch-clamp methods and protocols*, pages 1–19, 2014.

[247] Jakob Vesterstrom and Rene Thomsen. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In *Proceedings of the 2004 congress on evolutionary computation (IEEE Cat. No. 04TH8753)*, volume 2, pages 1980–1987. IEEE, 2004.

[248] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[249] Eberhard O Voit. *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.

[250] Eberhard O Voit, Harald A Martens, and Stig W Omholt. 150 years of the mass action law. *PLoS computational biology*, 11(1):e1004012, 2015.

[251] Peter Waage and Cato Maximilian Gulberg. Studies concerning affinity. *Journal of chemical education*, 63(12): 1044, 1986.

[252] David J Wales and Jonathan PK Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

[253] Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson LS Tang, and Weichuan Yu. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, 2010.

[254] Stephen Whitelam and Phillip L. Geissler. Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles. *The Journal of Chemical Physics*, 127(15):154101, October 2007. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.2790421. URL `http://aip.scitation.org/doi/10.1063/1.2790421`.

[255] Franz-Georg Wieland, Adrian L Hauber, Marcus Rosenblatt, Christian Tönsing, and Jens Timmer. On structural and practical identifiability. *Current Opinion in Systems Biology*, 25:60–69, 2021.

[256] Darren J Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116, 2007.

[257] EM Wright, BA Hirayama, and DF Loo. Active sugar transport in health and disease. *Journal of internal medicine*, 261(1):32–43, 2007.

[258] Ernest M Wright and Eric Turk. The sodium/glucose cotransport family slc5. *Pflügers Archiv*, 447:510–518, 2004.

[259] Ernest M Wright, Donald DF Loo, and Bruce A Hirayama. Biology of human sodium glucose transporters. *Physiological reviews*, 91(2):733–794, 2011.

[260] Ernest M. Wright, Chiara Ghezzi, and Donald D. F. Loo. Novel and Unexpected Functions of SGLTs. *Physiology*, 32(6):435–443, November 2017. ISSN 1548-9213, 1548-9221. doi: 10.1152/physiol.00021.2017. URL `https://www.physiology.org/doi/10.1152/physiol.00021.2017`.

[261] Nieng Yan. Structural biology of the major facilitator superfamily transporters. *Annual review of biophysics*, 44: 257–283, 2015.

[262] Tong Yu and Hong Zhu. Hyper-parameter optimization: A review of algorithms and applications, 2020.

[263] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

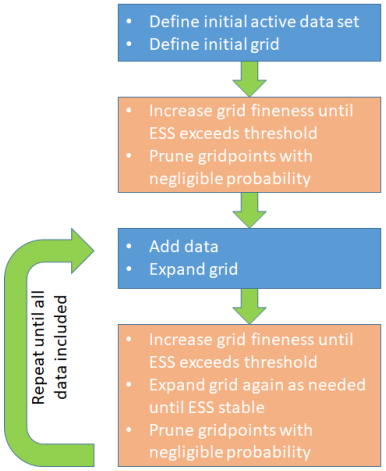
[264] Xiaoming Zhou, Elena J Levin, Yaping Pan, Jason G McCoy, Ruchika Sharma, Brian Kloss, Renato Bruni, Matthias Quick, and Ming Zhou. Structural basis of the alternating-access mechanism in a bile acid transporter. *Nature*, 505(7484):569–573, 2014.

[265] Daniel Zuckerman. Physical Lens on the Cell | Basic Principles Underlying Cellular Processes, 2023. URL `http://physicallensonthecell.org/`.

7 Appendix of Additional Sampling Efforts

7.1 Adaptive Grid-Based Sampler

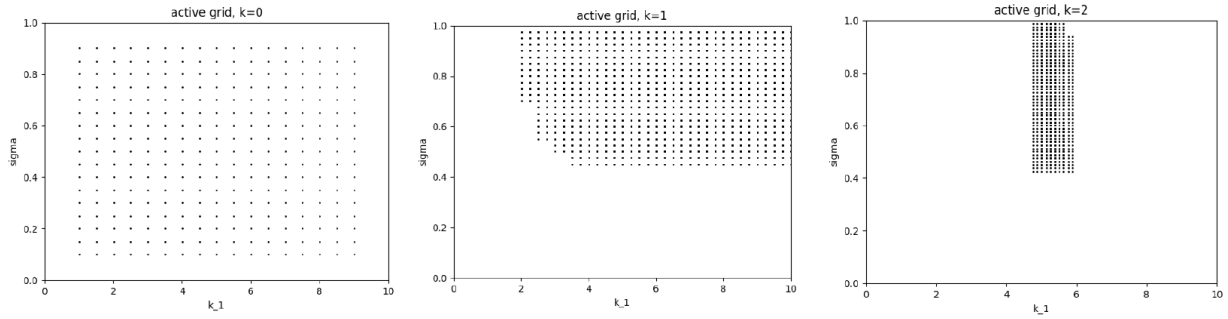
Motivated by data tempering methods by Chopin et al. in IBIS, this algorithm performs non-stochastic data tempering by gradually increasing the amount of data included and updating the estimated posterior using an adaptive grid. The grid starts uniform and relatively coarse for a small amount of data but increases fineness in a small subset of regions seen to be important. The important regions are updated as more data are added. The procedure avoids the uncertainties of stochastic (e.g., MCMC) sampling even if MCMC is used for exploring the parameter space because all likelihood/posterior computations are grid-based. A schematic is shown in the figure below:



Schematic of multigrid data tempering algorithm. In this non-stochastic approach, grid-based calculation is used as increasing amounts of data are added for Bayesian posterior calculation. The grid is adapted to each amount of data, both by becoming finer and by growing to accommodate nearby regions with significant probability as well as pruning away regions with negligible probability.

We have obtained highly promising preliminary results with the method for a toy system. Synthetic time-course data is used based on the function $y(t) = y_0\, e^{-k_1 t}$ for $0 \leq t \leq 2$ with $y_0 = 50$ and $k_1 = 5.0$. The function is sampled using ten evenly sampled time points with Gaussian noise added based on $\sigma = 0.5$. As shown in the figure below, the procedure does indeed narrow the grid down to a region surrounding the true values as more data is added.

An implementation of this algorithm can be found here: https://github.com/ZuckermanLab/pyGridSampler



Adaptive grid. The initial grid (left) covers the full space of possible parameters. Based on the first two data points, the set of active points is significantly pruned, and the grid is made finer (middle). The grid adaptively occupies an increasingly smaller volume with increasing grid fineness (right) as more data is added. The grids shown represent the first iterations for a system with true parameters of $k_1 = 5$ and $\sigma = 0.5$.

7.2 Parallel Affine Invariant Ensemble Sampler

Motivated by the ability of the affine invariant ensemble sampler (by Goodman and Weare and others) to sample complex distribution, an extension was developed for parallelization. Here ensembles are run independently in a trivially parallel fashion, and then after some time point, walkers are mixed between all the ensembles. This mixing improves the resiliency against trapping and may improve the overall sampling efficiency. Similarly, parallelization allows for the utilization of many more walkers than just a single ensemble. A schematic is shown below:

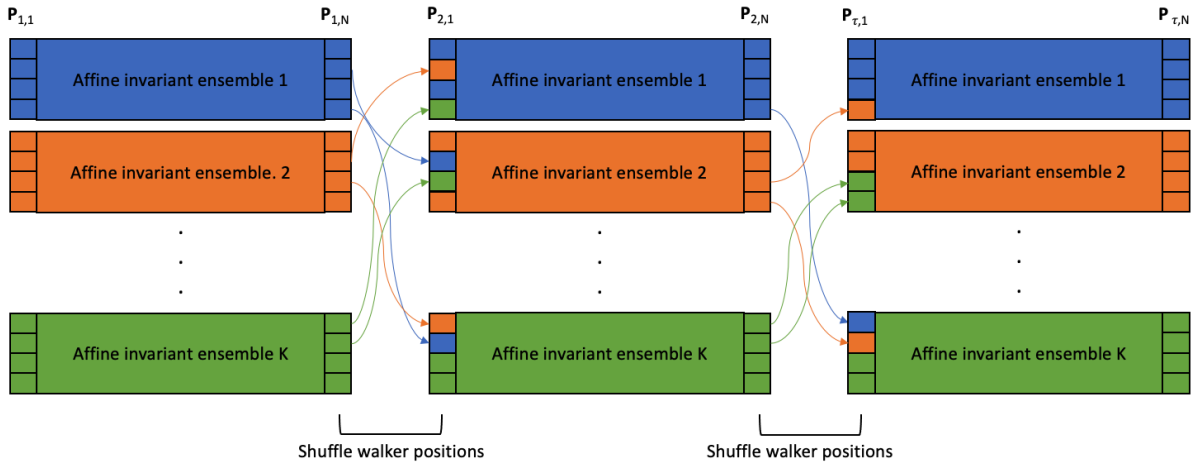


Diagram of the parallel affine ensemble sampler with mixing. Parallel ensembles are launched and then periodically mixed among themselves until a termination point is reached.

An implementation of this algorithm can be found here: https://github.com/ZuckermanLab/pyPAM

8 Chapter 2 Appendix

8.1 Detailed methods

8.1.1 String-based state-space specification

The state string consists of a set of user-defined ‘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. The inside and outside designations are necessary to define the direction of a transition (e.g., N binding from outside) but note that under the steady-state conditions analyzed here, species concentrations are held fixed but with differing values inside and outside. That is, transitions do not change the steady-state concentrations.

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). These states are defined in an analogous manner: e.g., OF-Nb-Wi. By choice, S and W cannot both be bound simultaneously.

Ultimately, steady-state populations are calculated for distinguishable states. For example, the “OF-Nb-Si" and “OF-Nb-So” states are not distinguishable in steady state because inside and outside substrate concentrations are held fixed. (As noted above, the inside and outside designations are necessary to assign directionality to binding events.) In contrast, the “OF-Nb-Si" and “OF-Ni-Sb” states differ in their binding state making them distinguishable. Likewise the “OF-Nb-Si" and “IF-Nb-Si” states are distinguishable. Hence, as detailed below, the full set of states is mapped to a smaller set of physically unique states for determining populations. Furthermore, any set of indistinguishable states must be energetically equivalent with the same state energy, E_i, value. These indistinguishable states are “tied” together for the Monte Carlo procedure. Note that state populations cannot be directly inferred from the equilibrium E_i values because we are studying driven, non-equilibrium steady states.

8.1.2 State name definitions

OF	Outward-facing conformation
IF	Inward-facing conformation
No	Extracellular sodium ion
Ni	Intracellular sodium ion
Nb	Bound sodium ion
So	Extracellular substrate
Si	Intracellular substrate
Sb	Bound substrate
Wo	Extracellular decoy substrate
Wi	Intracellular decoy substrate
Wb	Bound decoy substrate

8.1.3 Equivalent states and transitions

We have defined groups of states that are physically equivalent at steady state with fixed extracellular and intracellular concentrations. After a species (e.g. ion or substrate) is transported, the physical state will remain the same because of the steady-state assumption. As an example, consider a hypothetical transporter of a substrate (S) driven by a sodium ion (N). The state 'OF-Nb-So' describes the extracellular (outward) facing conformation (OF) with sodium-bound (Nb) and substrate unbound and in the extracellular region (So). This is physically equivalent to the state 'OF-Nb-Si', which only differs by the "location" of the unbound substrate (Si, substrate inside the cell). The location of a substrate or ion is needed in order to fully identify transitions – i.e., the origin (inside or outside) of the substrate or ion in a binding process.

Transitions are similarly grouped based on states that are equivalent under the conditions stated above. Considering the same hypothetical transporter: an extracellular-to-bound sodium transition (No $\rightarrow$ Nb) would be physically equivalent for extracellular (So) and intracellular (Si) substrate in the outward-facing conformation (OF). The equivalent state and transition groups are constrained to share the same state or transition energy during the Monte Carlo (MC) energy perturbations for self-consistency.

To investigate the Hopfield kinetic proofreading model of transport, we have an additional constraint that groups 'equivalent' transitions for the substrate and decoy substrate. As an example, an inward-to-outward facing conformational transition with only the decoy bound (e.g. OF-No-So-Wb to IF-No-So-Wb) would be equivalent to an inward-to-outward facing conformational transition with only the substrate-bound (e.g. OF-No-Sb-Wo to IF-No-Sb-Wo). This extra constraint prohibits a difference in transition energies between equivalent substrate and decoy transitions; effectively removing "internal proofreading" models from our search.

List of equivalent states/transitions for a cotransporter without decoy substrate

Equivalent states:

- IF-Ni-So, IF-No-Si, IF-No-So, IF-Ni-Si
- OF-No-So, OF-Ni-Si, OF-No-Si, OF-Ni-So
- IF-Ni-Sb, IF-No-Sb
- IF-Nb-So, IF-Nb-Si
- OF-No-Sb, OF-Ni-Sb
- OF-Nb-So, OF-Nb-Si

Equivalent transitions:

- OF-Ni-Sb⟷OF-Ni-Si, OF-No-Sb⟷OF-No-Si
- OF-Ni-So⟷OF-Ni-Sb, OF-No-So⟷OF-No-Sb
- IF-Nb-Si⟷IF-Ni-Si, IF-Nb-So⟷IF-Ni-So
- OF-No-So⟷OF-Nb-So, OF-No-Si⟷OF-Nb-Si
- OF-No-So⟷IF-No-So, OF-Ni-Si⟷IF-Ni-Si, OF-Ni-So⟷IF-Ni-So, OF-No-Si⟷IF-No-Si
- OF-Ni-Sb⟷IF-Ni-Sb, OF-No-Sb⟷IF-No-Sb
- IF-No-So⟷IF-Nb-So, IF-No-Si⟷IF-Nb-Si,
- OF-Nb-Si⟷OF-Ni-Si, OF-Nb-So⟷OF-Ni-So,
- OF-Nb-So⟷IF-Nb-So, OF-Nb-Si⟷IF-Nb-Si
- IF-No-So⟷IF-No-Sb, IF-Ni-So⟷IF-Ni-Sb,
- IF-No-Sb⟷IF-No-Si, IF-Ni-Sb⟷IF-Ni-Si

List of equivalent states/transitions for a cotransporter with decoy substrate

Equivalent states:

- OF-Nb-So-Wb, OF-Nb-Sb-Wo, OF-Nb-Sb-Wi, OF-Nb-Si-Wb
- IF-Ni-So-Wb, IF-No-Sb-Wo, IF-Ni-Sb-Wo, IF-No-Sb-Wi, IF-Ni-Si-Wb, IF-No-Si-Wb, IF-Ni-Sb-Wi, IF-No-So-Wb
- OF-No-So-Wb, OF-No-Si-Wb, OF-Ni-So-Wb, OF-Ni-Sb-Wo, OF-No-Sb-Wo, OF-Ni-Si-Wb, OF-Ni-Sb-Wi, OF-No-Sb-Wi
- IF-Nb-So-Wo, IF-Nb-Si-Wo, IF-Nb-Si-Wi, IF-Nb-So-Wi
- IF-Ni-So-Wo, IF-Ni-Si-Wo, IF-No-Si-Wi, IF-No-So-Wo, IF-No-Si-Wo, IF-Ni-So-Wo, IF-Ni-So-Wi, IF-No-So-Wi, IF-Ni-Si-Wi
- IF-Nb-So-Wb, IF-Nb-Sb-Wi, IF-Nb-Si-Wb, IF-Nb-Sb-Wo
- OF-Nb-So-Wo, OF-Nb-Si-Wo, OF-Nb-Si-Wi, OF-Nb-So-Wi
- OF-No-So-Wo, OF-Ni-So-Wi, OF-No-Si-Wi, OF-Ni-Si-Wo, OF-Ni-Si-Wi, OF-No-Si-Wo, OF-No-So-Wi, OF-Ni-So-Wo

Equivalent transitions:

- IF-No-So-Wb⟷IF-Nb-So-Wb, IF-No-Sb-Wi⟷IF-Nb-Sb-Wi, IF-No-Sb-Wo⟷IF-Nb-Sb-Wo, IF-No-Si-Wb⟷IF-Nb-Si-Wb
- OF-Nb-So-Wi⟷IF-Nb-So-Wi, OF-Nb-Si-Wo⟷IF-Nb-Si-Wo, OF-Nb-Si-Wi⟷IF-Nb-Si-Wi, OF-Nb-So-Wo⟷IF-Nb-So-Wo
- IF-Nb-So-Wb⟷IF-Nb-So-Wi, IF-Nb-Si-Wb⟷IF-Nb-Si-Wi, IF-Nb-Sb-Wo⟷IF-Nb-Si-Wo, IF-Nb-Sb-Wi⟷IF-Nb-Si-Wi
- OF-Nb-Si-Wb⟷IF-Nb-Si-Wb, OF-Nb-So-Wb⟷IF-Nb-So-Wb, OF-Nb-Sb-Wo⟷IF-Nb-Sb-Wo, OF-Nb-Sb-Wi⟷IF-Nb-Sb-Wi
- IF-Nb-Sb-Wi⟷IF-Ni-Sb-Wi, IF-Nb-So-Wb⟷IF-Ni-So-Wb, IF-Nb-Si-Wb⟷IF-Ni-Si-Wb, IF-Nb-Sb-Wo⟷IF-Ni-Sb-Wo
- OF-No-Sb-Wo⟷OF-No-Si-Wo, OF-No-Sb-Wi⟷OF-No-Si-Wi, OF-Ni-Si-Wb⟷OF-Ni-Si-Wi, OF-No-Si-Wb⟷OF-No-Si-Wi, OF-Ni-Sb-Wo⟷OF-Ni-Si-Wo, OF-Ni-So-Wb⟷OF-Ni-So-Wi, OF-No-So-Wb⟷OF-No-So-Wi, OF-Ni-Sb-Wi⟷OF-Ni-Si-Wi

- OF-Ni-Sb-Wo⟷IF-Ni-Sb-Wo, OF-Ni-So-Wb⟷IF-Ni-So-Wb, OF-No-Sb-Wi⟷IF-No-Sb-Wi, OF-Ni-Sb-Wi⟷IF-Ni-Sb-Wi, OF-No-Sb-Wo⟷IF-No-Sb-Wo, OF-Ni-Si-Wb⟷IF-Ni-Si-Wb, OF-No-Si-Wb⟷IF-No-Si-Wb, OF-No-So-Wb⟷IF-No-So-Wb
- IF-No-Si-Wb⟷IF-No-Si-Wi, IF-Ni-Sb-Wo⟷IF-Ni-Si-Wo, IF-No-So-Wb⟷IF-No-So-Wi, IF-No-Sb-Wi⟷IF-No-Si-Wi, IF-Ni-So-Wb⟷IF-Ni-So-Wi, IF-Ni-Sb-Wi⟷IF-Ni-Si-Wi, IF-No-Sb-Wo⟷IF-No-Si-Wo, IF-Ni-Si-Wb⟷IF-Ni-Si-Wi
- OF-Ni-Si-Wi⟷IF-Ni-Si-Wi, OF-No-So-Wi⟷IF-No-So-Wi, OF-Ni-Si-Wo⟷IF-Ni-Si-Wo, OF-No-So-Wo⟷IF-No-So-Wo, OF-Ni-So-Wo⟷IF-Ni-So-Wo, OF-No-Si-Wo⟷IF-No-Si-Wo, OF-No-Si-Wi⟷IF-No-Si-Wi, OF-Ni-So-Wi⟷IF-Ni-So-Wi
- OF-Nb-Sb-Wo⟷OF-Nb-Si-Wo, OF-Nb-So-Wb⟷OF-Nb-So-Wi, OF-Nb-Sb-Wi⟷OF-Nb-Si-Wi, OF-Nb-Si-Wb⟷OF-Nb-Si-Wi
- IF-No-Si-Wi⟷IF-Nb-Si-Wi, IF-No-Si-Wo⟷IF-Nb-Si-Wo, IF-No-So-Wi⟷IF-Nb-So-Wi, IF-No-So-Wo⟷IF-Nb-So-Wo
- OF-Nb-So-Wi⟷OF-Nb-Sb-Wi, OF-Nb-Si-Wo⟷OF-Nb-Si-Wb, OF-Nb-So-Wo⟷OF-Nb-So-Wb, OF-Nb-So-Wo⟷OF-Nb-Sb-Wo
- IF-Nb-So-Wo⟷IF-Ni-So-Wo, IF-Nb-Si-Wi⟷IF-Ni-Si-Wi, IF-Nb-Si-Wo⟷IF-Ni-Si-Wo, IF-Nb-So-Wi⟷IF-Ni-So-Wi
- OF-No-Si-Wi⟷OF-Nb-Si-Wi, OF-No-Si-Wo⟷OF-Nb-Si-Wo, OF-No-So-Wi⟷OF-Nb-So-Wi, OF-No-So-Wo⟷OF-Nb-So-Wo
- OF-Nb-Si-Wb⟷OF-Ni-Si-Wb, OF-Nb-So-Wb⟷OF-Ni-So-Wb, OF-Nb-Sb-Wo⟷OF-Ni-Sb-Wo, OF-Nb-Sb-Wi⟷OF-Ni-Sb-Wi
- OF-No-Sb-Wo⟷OF-Nb-Sb-Wo, OF-No-So-Wb⟷OF-Nb-So-Wb, OF-No-Si-Wb⟷OF-Nb-Si-Wb, OF-No-Sb-Wi⟷OF-Nb-Sb-Wi
- IF-No-Si-Wo⟷IF-No-Si-Wb, IF-Ni-Si-Wo⟷IF-Ni-Si-Wb, IF-Ni-So-Wi⟷IF-Ni-Sb-Wi, IF-Ni-So-Wo⟷IF-Ni-Sb-Wo, IF-No-So-Wo⟷IF-No-So-Wb, IF-No-So-Wo⟷IF-No-Sb-Wo, IF-Ni-So-Wo⟷IF-Ni-So-Wb, IF-No-So-Wi⟷IF-No-Sb-Wi
- OF-Nb-So-Wi⟷OF-Ni-So-Wi, OF-Nb-Si-Wo⟷OF-Ni-Si-Wo, OF-Nb-Si-Wi⟷OF-Ni-Si-Wi, OF-Nb-So-Wo⟷OF-Ni-So-Wo
- OF-No-So-Wi⟷OF-No-Sb-Wi, OF-No-So-Wo⟷OF-No-So-Wb, OF-Ni-So-Wo⟷OF-Ni-Sb-Wo, OF-No-Si-Wo⟷OF-No-Si-Wb, OF-No-So-Wo⟷OF-No-Sb-Wo, OF-Ni-Si-Wo⟷OF-Ni-Si-Wb, OF-Ni-So-Wo⟷OF-Ni-So-Wb, OF-Ni-So-Wi⟷OF-Ni-Sb-Wi

- IF-Nb-Si-Wo⟷IF-Nb-Si-Wb, IF-Nb-So-Wo⟷IF-Nb-Sb-Wo, IF-Nb-So-Wo⟷IF-Nb-So-Wb, IF-Nb-So-Wi⟷IF-Nb-Sb-Wi

8.1.4 Tempering

In order to avoid trapping in deep "energy" (high fitness) basins we employ a tempering procedure. This procedure cyclically raises and lowers the inverse-temperature β in the Metropolis-Hastings acceptance criterion, facilitating the exploration of different areas in the model space because of the increased likelihood of acceptance. ModelExplorer allows for both adaptive and fixed-cycle tempering. Adaptive 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 below). Fixed-cycle 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 2 (main text) is an example of the Monte Carlo "energy" trajectory in model space produced in ModelExplorer for a simple symporter system using fixed-cycle tempering. 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.

Automated 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. An initial β is set and then checked at fixed Monte Carlo step intervals. At these intervals, the fractional Monte Carlo energy difference from that previous checkpoint is calculated: $E_{MC}^{\text{frac}} = 2\frac{E_{MC}^{\text{new}} - E_{MC}^{\text{old}}}{|E_{MC}^{\text{new}}| - |E_{MC}^{\text{old}}|}$ where E_{MC}^{new} is the current Monte Carlo energy and E_{MC}^{old} is the Monte Carlo energy at the previous checkpoint. If the fractional energy is approximately constant, $|E_{MC}^{frac}| <$ tolerance, β decreases (heats) by a user-constrained scale factor. If the fractional energy difference has decreased, β increases (cools) by a user-constrained scale factor, subject to a user-defined probability to stay at the current beta, P_{β}^{stay}. If the fractional energy has increased, β decreases (heats) by a user-constrained scale factor, subject to a user-defined probability to stay at the current beta, P_{β}^{stay}.

Manual tempering Manual tempering uses a fixed schedule that adjusts β by a set amount at each Monte Carlo step interval. The user defines the minimum and maximum β (inverse temperature) and also defines how many Monte Carlo steps remain at the minimum and maximum β, decrease β from maximum to minimum, and to increase β from minimum to maximum. This schedule can then be scaled by a user-defined factor for further sampling optimization. For enhanced selectivity simulations, the default tempering schedule (found empirically) is 125 MC steps at the minimum β, 100 MC steps increasing β, 1450 MC steps at the maximum β, and 325 MC steps decreasing β.

8.1.5 Flux calculation

The flux of a given species x, J_x, is calculated by summing the net flows along a user-defined set of transitions for that species. For the ion, this is the set of transitions from an ion-bound state to an ion-inside state (Nb→Ni). The substrate flux is calculated from the set of transitions from a substrate-bound state to a substrate-inside state (Sb→Si). Similarly, decoy flux is calculated from the set of transitions from a decoy-bound state to a decoy-inside state (Wb→Wi). For simplicity, we have removed 'backdoor' transport which would allow a species to be transported in the opposing

direction of the conformational state (e.g. OF-Nb→OF-Ni). Due to these added constraints, the flux for a given species is calculated using only the net flows of the (un)binding transitions for that species in the *inward-facing conformation*.

8.2 Simulation parameters

8.2.1 Cotransporter without decoy substrate

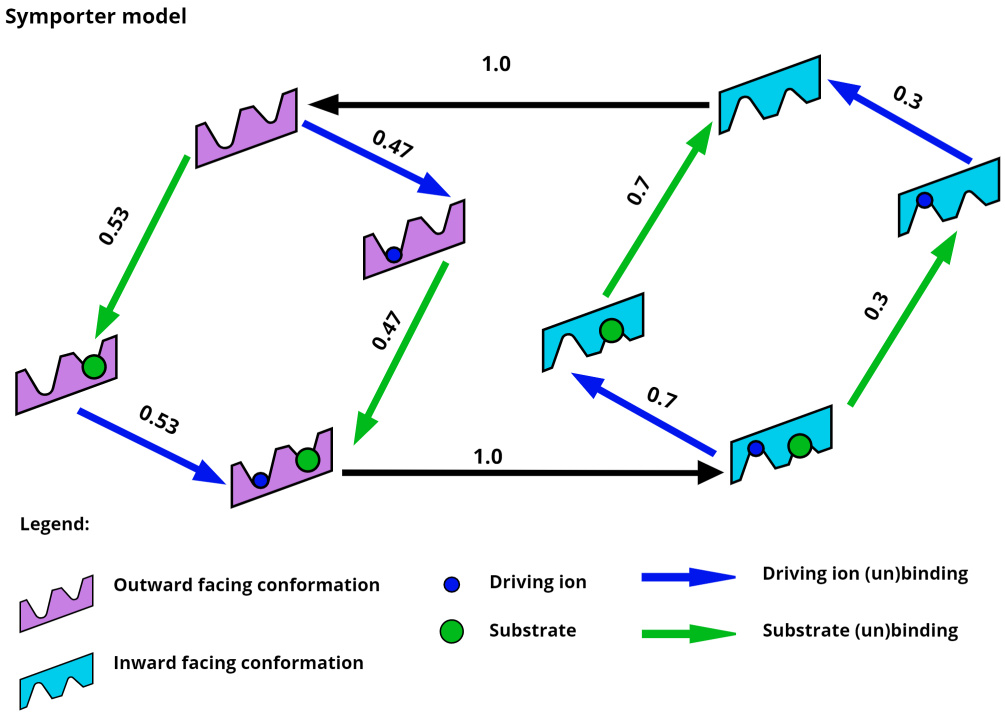
	Symporter	**Antiporter**
MC steps	1e6	1e6
Random seed	123456	123456
Maximum ΔE for state/transition $[k_B T]$	1.0	1.0
Tempering schedule	Automatic	Automatic
Tempering tolerance	0.3	0.3
β initial $[k_B T]^{-1}$	1e1	1e1
β scale factor	1e3	1e3
P_β^{stay}	0.2	0.2
MC steps to change β	2e2	2e2
$\Delta\mu_{\text{ion}}$ $[k_B T]$	-4	-4
$\Delta\mu_{\text{substrate}}$ $[k_B T]$	+2	-2
Rate prefactor, k_0 $[\text{s}^{-1}]$	1e-3	1e-3
Energy function, E_{MC}	$-J_{\text{substrate}}$	$+J_{\text{substrate}}$

8.2.2 Cotransporter with decoy substrate

	Run 1	**Run 2**	**Run 3**	**Run 4**
MC steps	1e6	1e6	1e6	1e6
Random seed	456789	456789	456789	456789
Maximum ΔE for state/transition $[k_B T]$	1.0	0.5	0.2	1.0
Tempering schedule	Manual	Manual	Manual	Manual
Tempering scale factor	1.0	0.5	1.0	2.0
β_{min} $[k_B T]^{-1}$	1e-100	1e-100	1e-100	1e-100
β_{max} $[k_B T]^{-1}$	1e30	1e30	1e30	1e30
MC steps to change β	1	1	1	1
$\Delta\mu_{\text{ion}}$ $[k_B T]$	-4	-4	-4	-4
$\Delta\mu_{\text{substrate}}$ $[k_B T]$	2	2	2	2
$\Delta\mu_{\text{decoy}}$ $[k_B T]$	2	2	2	2
$\Delta\Delta G$ $[k_B T]$	1	1	1	1
Rate prefactor, k_0 $[\text{s}^{-1}]$	1e-3	1e-3	1e-3	1e-3
Energy function, E_{MC}	$-J_{\text{substrate}} \frac{\lvert J_{\text{substrate}}\rvert+\epsilon}{\lvert J_{\text{decoy}}\rvert+\epsilon}$	...	...	...
Numerical stability constant, ϵ	1e-15	1e-15	1e-15	1e-15

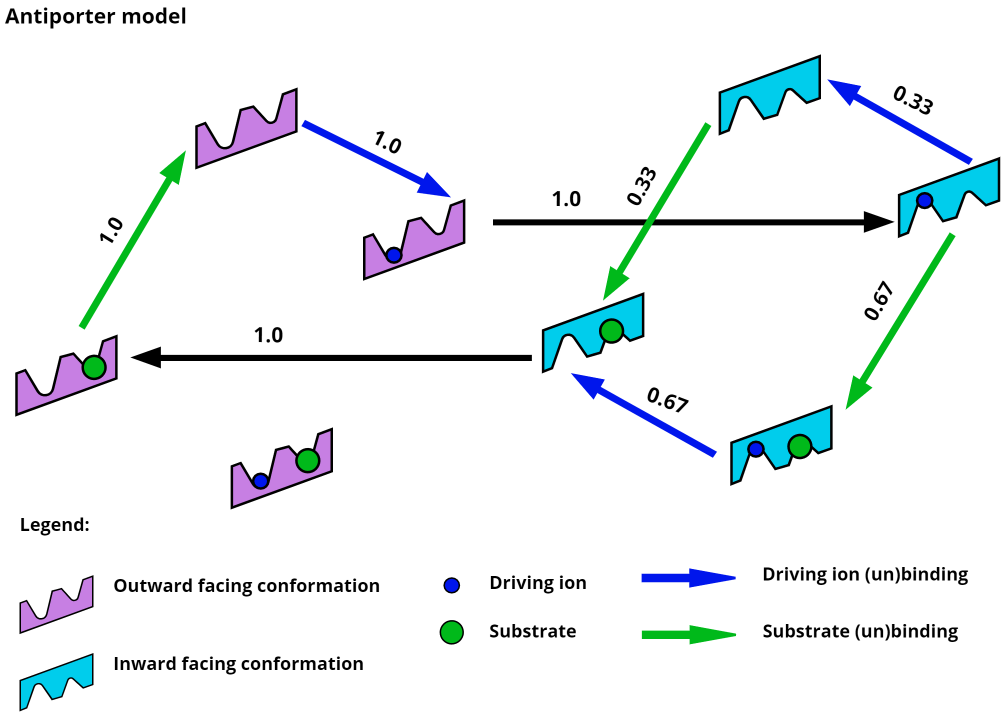
8.3 Symporter model pathway (without decoy substrate).

[[Note that the binding and dissociation events are shown implicity for visual clarity]]



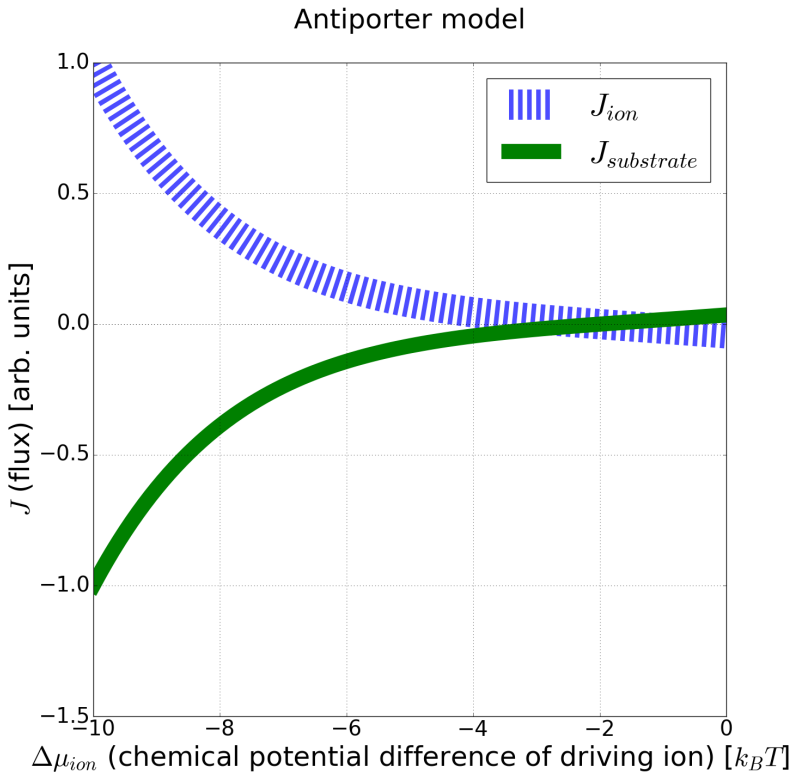
SI Fig 8.3 Pathway of a symporter model found at MC step = 1800. This model exhibits a combination of all four ideal symporter pathways which result in the intracellular transport of one substrate per ion. Note that the flows are scaled by the largest flow edge.

8.4 Antiporter model pathway (without decoy substrate).



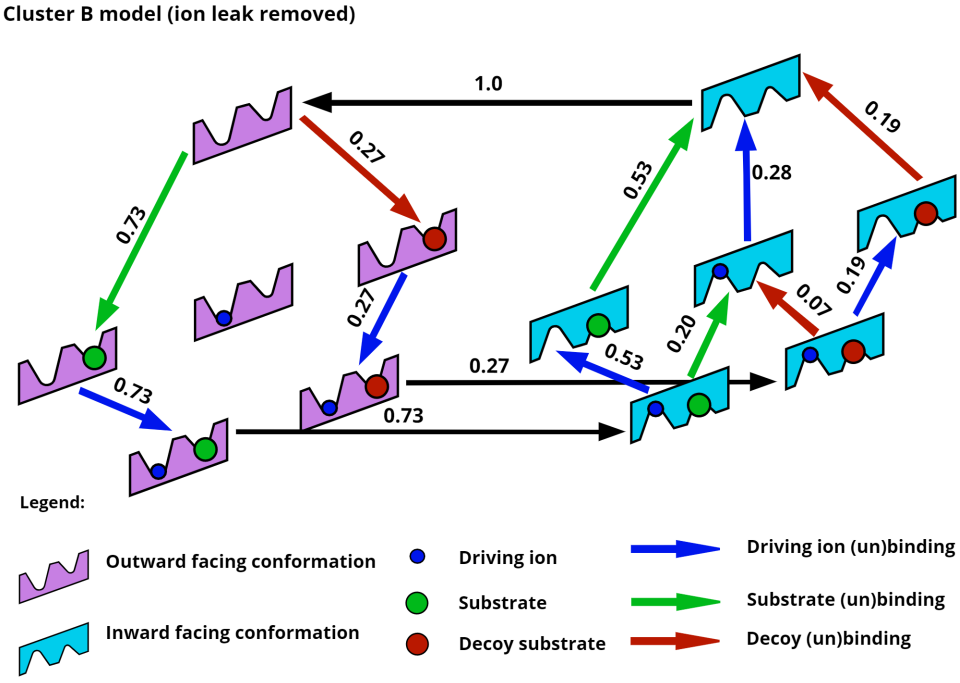
SI Fig 8.4 Pathway of an antiporter model found during the antiporter simulation run at MC step = 845000. This model exhibits a combination of two ideal antiporter pathways which result in the extracellular transport of one substrate per ion. Note that the flows are scaled by the largest flow edge.

8.5 Antiporter flux diagram (without decoy substrate).



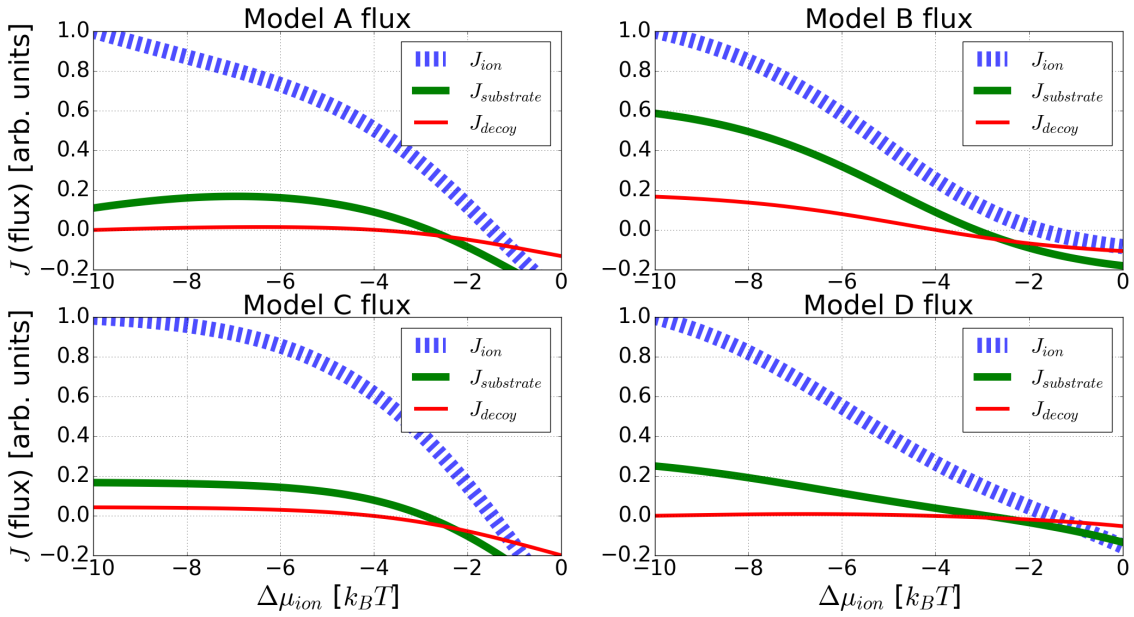
SI Fig 8.5 Flux of an antiporter model found during the antiporter simulation run at MC step = 845000, analyzed over a range of ion-chemical potential differences. This model exhibits a 1:1 ratio of ion influx to substrate efflux over a wide range of ion chemical potential differences. Note that the fluxes are scaled by the largest flux value.

8.6 Symporter model with ion leak removed (and with decoy substrate present)



SI Fig 8.6 Pathway of the model with enhanced selectivity representing cluster B, with the futile ion cycle removed. The energy barrier between the ion-only bound states in the inward and outward conformations were raised by 100 k_BT, effectively shutting off the ion leak. Note the two symmetrical pathways for substrate and decoy transport. The net flows have been scaled by the maximum flow edge.

8.7 Flux diagrams of the representative models for each cluster



SI Fig 8.7 Flux traces of the representative model for each cluster, scaled by the maximum flux, over a range of ion chemical potential differences. Each model has a narrow regime where the toxin flows down its gradient out of the cell, while the substrate is driven into the cell by the ion. Near the optimized conditions for the simulation, $\Delta\mu_{ion} = -4k_BT$, these models have negligible decoy flux, resulting in an unbounded substrate to decoy discrimination ratio.

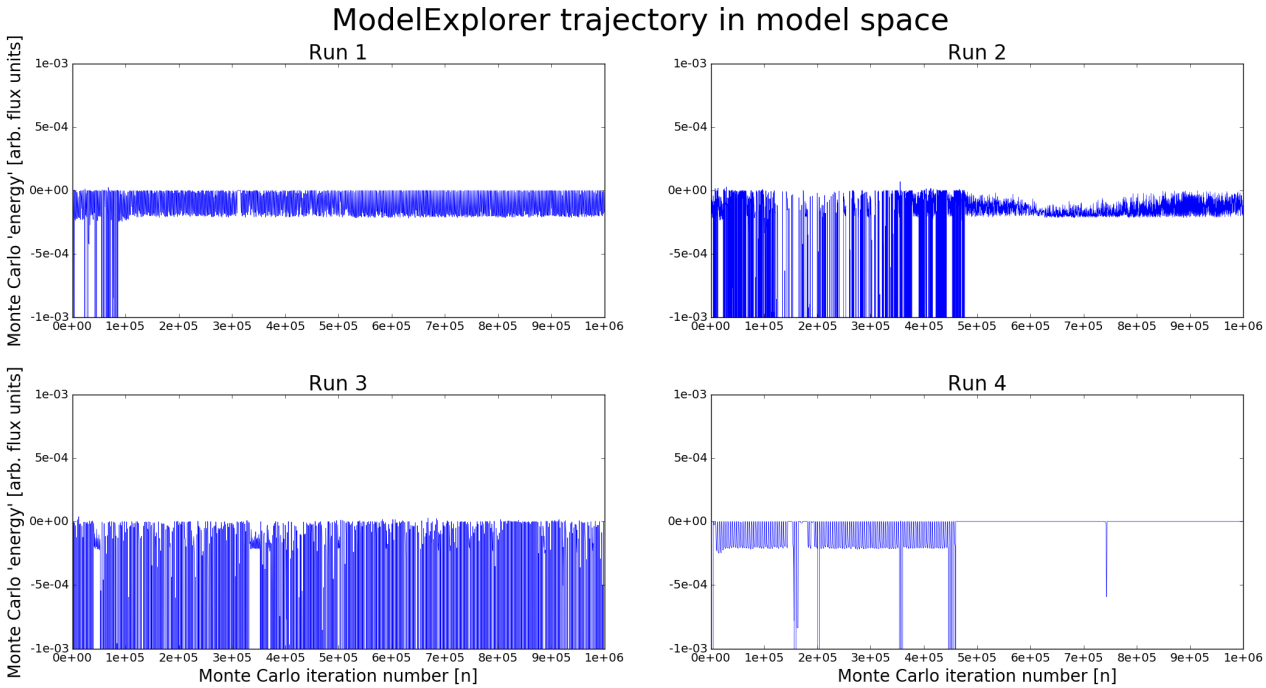
8.8 Model clustering and sampling

Simulations were run for transporters in a 'competitive' environment with a decoy. The resulting models were filtered based on a cost (ion to substrate flux ratio) below 10, and selectivity (substrate to decoy ratio) above $10e^{\Delta\Delta G=1}$. Clusters were determined using hierarchical clustering with complete linkage and the Euclidean distance between the scaled flows of each model. The threshold of 0.65 was determined empirically to produce qualitatively different kinetic pathways. This method produced four separate clusters. Representative models corresponding to each cluster were analyzed in the main text and SI.

The representative models used:

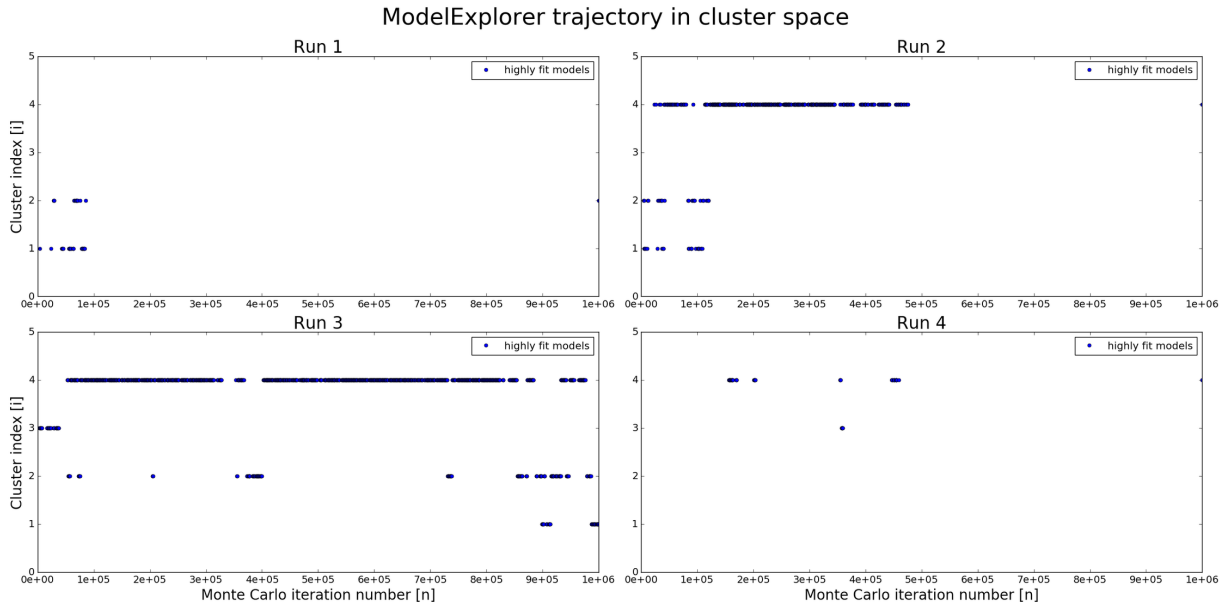
	Run	MC step number
Cluster A model	1	3500
Cluster B model	1	29000
Cluster C model	3	3000
Cluster D model	3	829000

8.9 Trajectory in model space



SI Fig 8.9 Trajectory in the model space of four different 1e6 Monte Carlo (MC) step simulations at different sampling settings. Simulations were run for transporters in a 'competitive' environment with a decoy using the MC energy function: $-J_{\text{substrate}} \frac{|J_{\text{substrate}}|+\epsilon}{|J_{\text{decoy}}|+\epsilon}$ where $J_{\text{substrate}}$ and J_{decoy} are the fluxes of the substrate and decoy respectively, and ϵ = 1e-15. Lower MC energy values denote models that are more fit, by convention. As shown in the figures, the tempering schedule aids in avoiding low-energy basins. Note that each point on the trajectory is a kinetic model. [[We note that each run finds a distribution of models, from highly fit (energy ≈1e-3) to moderately fit (energy ≈0.25e-4), to poorly fit (energy ~0).]]

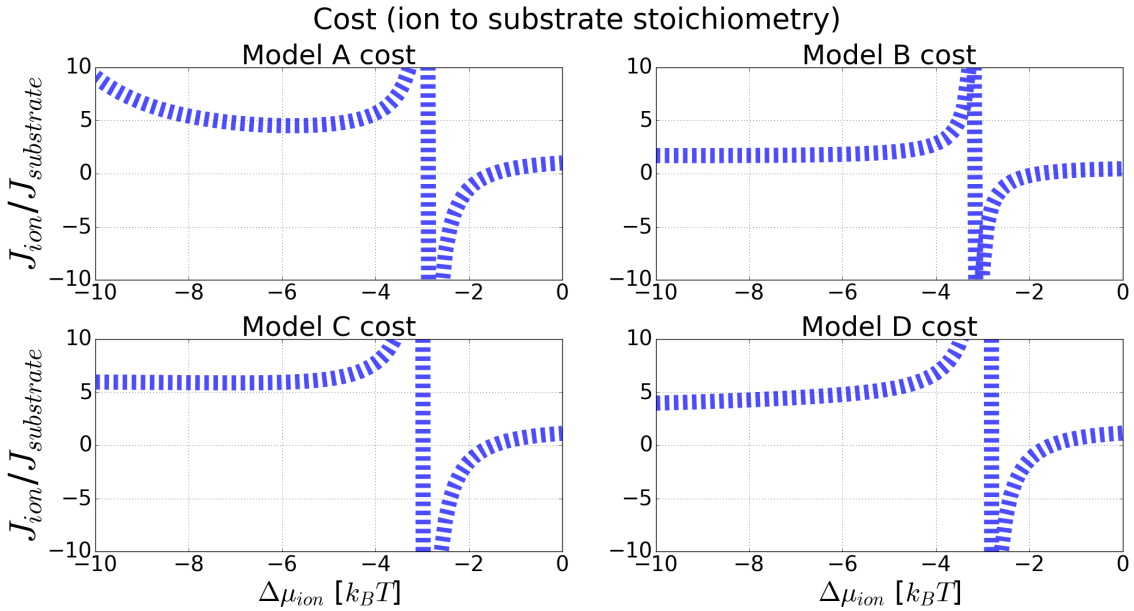
8.10 Trajectory in cluster space



SI Fig 8.10 Trajectory in the cluster space of four different 1e6 MC step simulations. Simulations were run for transporters in a 'competitive' environment with a decoy. Models were filtered based on a cost (ion to substrate flux ratio) below 10, and selectivity (substrate to decoy ratio) above $10e^{\Delta\Delta G=1}$. Clusters were determined using hierarchical clustering with complete linkage and the Euclidean distance between the scaled flows of each model. The threshold of 0.65 was determined empirically to produce qualitatively different kinetic pathways. These graphs indicate that each run only finds a few model classes during the simulation – implying the need for improved sampling methods. Note that in run 4, models meeting the selection criteria (i.e. cost and selectivity) were not found until approximately 1.5e5 MC steps. [[We note that certain runs 1, 2, and 4 fail to find models that are highly fit, likely due to poor sampling. It is likely that these sampling runs were 'trapped' in a low energy region and unable to escape to explore even lower energy regions. This could be further addressed in the future by better optimizing our tempering procedure, or with other sampling algorithms.]]

8.11 Cost of representative models

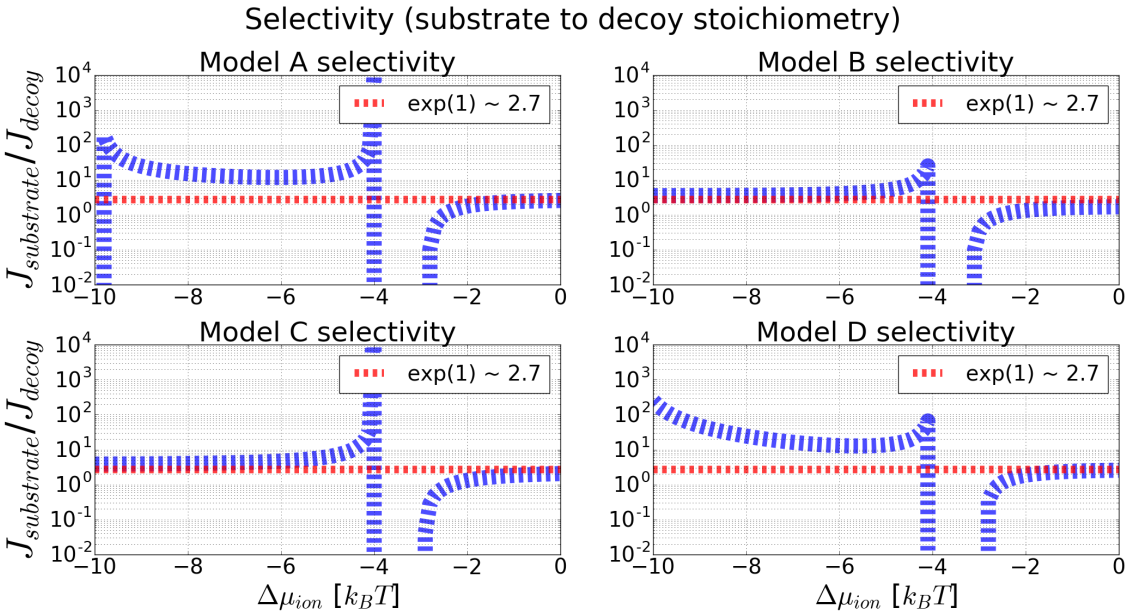
Here we examine the ion-to-substrate stoichiometry (cost) for each model cluster.



SI Fig 8.11 Cost of the representative model for each cluster, over a range of ion chemical potential differences. All of these models exhibit a cost above the ideal 1:1 stoichiometric ratio for a wide range of chemical potential differences. The extra ions transported relative to the substrate suggest a futile ion transport cycle - i.e. an ion leak. Note that the cost was not included as a constraint in the energy function.

8.12 Selectivity of representative models

Here we examine the substrate to decoy stoichiometry (selectivity) for each model cluster.



SI Fig 8.12 The stoichiometric ratio of the substrate to ion flux (selectivity), over a range of ion chemical potential differences. All the models demonstrate enhanced selectivity over a range of chemical potential differences, and unbounded selectivity at the optimized condition (at $\Delta\mu_{ion} = -4k_BT$). Models A and D exhibit enhanced discrimination over a wide range of conditions. The expected equilibrium value ($e^{\Delta\Delta G=1}$) is shown as a reference. Note that this stoichiometric ratio was used as a primary constraint in our energy function.

9 Chapter 3 Appendix

9.1 Transporter model definitions

We use a single reaction cycle of a 1:1 antiporter using alternating access, with the following coupled reactions:

Reaction 1: $$\mathrm{OF} + \mathrm{H_{out}} \underset{k_1^r}{\overset{k_1^f}{\rightleftharpoons}} \mathrm{OF\_Hb} \tag{28}$$

Reaction 2: $$\mathrm{OF\_Hb} \underset{k_2^r}{\overset{k_2^f}{\rightleftharpoons}} \mathrm{IF\_Hb} \tag{29}$$

Reaction 3: $$\mathrm{IF\_Hb} + \mathrm{S_{out}} \underset{k_3^r}{\overset{k_3^f}{\rightleftharpoons}} \mathrm{IF\_Hb\_Sb} \tag{30}$$

Reaction 4: $$\mathrm{IF\_Hb\_Sb} \underset{k_4^r}{\overset{k_4^f}{\rightleftharpoons}} \mathrm{IF\_Sb} + \mathrm{H_{in}} \tag{31}$$

Reaction 5: $$\mathrm{IF\_Sb} \underset{k_5^r}{\overset{k_5^f}{\rightleftharpoons}} \mathrm{OF\_Sb} \tag{32}$$

Reaction 6: $$\mathrm{OF\_Sb} \underset{k_6^r}{\overset{k_6^f}{\rightleftharpoons}} \mathrm{OF} + \mathrm{S_{out}} \tag{33}$$

with the following definitions for biochemical states and species:

OF	Outward-facing conformation
OF_Hb	Outward-facing conformation with ion bound
IF_Hb	Inward-facing conformation with ion bound
IF_Hb_Sb	Inward-facing conformation with ion and substrate bound
IF_Sb	Inward-facing conformation with substrate bound
OF_Sb	Outward-facing conformation with substrate bound
H_{out}	External (i.e. bath) ion
S_{out}	External (i.e. bath) substrate
H_{in}	Internal (i.e. inside lissome) ion
S_{in}	Internal (i.e. inside lissome) substrate

Note that we use molar concentrations unless specified otherwise, and have omitted the enclosing brackets for visual clarity.

9.2 Reaction rates and differential equations

Given the reactions above we have the following net reaction rates using mass action kinetics:

$$v_{rxn1} = k_1^f \cdot OF \cdot H_{out} - k_1^r \cdot OF\_Hb \tag{34}$$

$$v_{rxn2} = k_2^f \cdot OF\_Hb - k_2^r \cdot IF\_Hb \tag{35}$$

$$v_{rxn3} = k_3^f \cdot IF\_Hb \cdot S_{in} - k_3^r \cdot IF\_Hb\_Sb \tag{36}$$

$$v_{rxn4} = k_4^f \cdot IF\_Hb\_Sb - k_4^r \cdot IF\_Sb \cdot H_{in} \tag{37}$$

$$v_{rxn5} = k_5^f \cdot IF\_Sb - k_5^r \cdot OF\_Sb \tag{38}$$

$$v_{rxn6} = k_6^f \cdot OF\_Sb - k_6^r \cdot OF \cdot S_{out} \tag{39}$$

$$\tag{40}$$

Which yields the following ordinary differential equations:

$$\frac{dH_{in}}{dt} = v_{rxn4} \tag{41}$$

$$\frac{dS_{in}}{dt} = -v_{rxn3} \tag{42}$$

$$\frac{dOF}{dt} = v_{rxn6} - v_{rxn1} \tag{43}$$

$$\frac{dOF\_Hb}{dt} = v_{rxn1} - v_{rxn2} \tag{44}$$

$$\frac{dIF\_Hb}{dt} = v_{rxn2} - v_{rxn3} \tag{45}$$

$$\frac{dIF\_Hb\_Sb}{dt} = v_{rxn3} - v_{rxn4} \tag{46}$$

$$\frac{dIF\_Sb}{dt} = v_{rxn4} - v_{rxn5} \tag{47}$$

$$\frac{dOF\_Sb}{dt} = v_{rxn5} - v_{rxn6} \tag{48}$$

$$\tag{49}$$

where the external ion and substrate solutions are constant for a given SSME assay condition (i.e. $\frac{dH_{out}}{dt} = 0$ and $\frac{dS_{out}}{dt} = 0$ for each perturbation stage).

9.3 Constraint from detailed balance

At equilibrium, the forward flow must equal the reverse flow for any given reaction due to microscopic reversibility. In a cycle, this means that the product of the forward reaction rate constants must equal the product of the reverse reaction rate constants [101]. As a result, one of the reaction rate constants is not independent, and is defined by the remaining reaction rate constants. This equilibrium relationship holds even when the system is not at equilibrium:

$$\prod \frac{k_i^f}{k_i^r} = 1 \tag{50}$$

For our model we set k_6^r as constant:

$$k_6^r = \frac{k_1^f k_2^f k_3^f k_4^f k_5^f k_6^f}{k_1^r k_2^r k_3^r k_4^r k_5^r} \tag{51}$$

9.4 Synthetic SSME assay conditions

Below are the conditions used for the synthetic SSME assays, rounded for visual clarity:

N liposomes total	1e11
N transporters total	5e12
Total external volume	70 uL
Total internal volume	0.07 uL
Total membrane volume	0.02 uL
Liposome membrane capacitance	2.4e-16 F

Summary of parameters

OF	0.44 mM
OF_Hb	0 mM
IF_Hb	0 mM
IF_Hb_Sb	0 mM
IF_Sb	0 mM
OF_Sb	0 mM

Initial transporter state concentrations (t=0 s)

H_{out} (t=0 s)	1e-4 mM
S_{out} (t=0 s)	1 mM
H_{out} (t=1 s)	0.5e-4 mM
S_{out} (t=1 s)	1 mM
H_{out} (t=2 s)	1e-4 mM
S_{out} (t=2 s)	1 mM

Assay protocol for external bath solution perturbation

We assume uniformity of the liposomes, such that the concentrations are the same between each liposome, and single liposome volumes are equal to the total volume divided by the number of liposomes. Similarly, the voltage across the membrane of a single liposome is assumed to be equal to the voltage across the membrane from all the liposomes. We simulate a single liposome (with multiple transporters), and then convert that current into an aggregate net current (see manuscript).

As described in the manuscript, we only explicitly model the driving ion and substrate that are directly coupled to the transport process studied under SSME-like conditions. Therefore we do not explicitly model water, anions, or other chemical species used in the bath solution. We assume that these species concentrations are controlled such that they have a negligible effect on transport during an SSME assay. More specifically, we base our approach on related SSME experiments [240] for the Gdx protein, which holds the chloride (Cl^-) anion concentrations fixed throughout the bath

solution via a controlled buffer, removing potential confounding effects from varied anion concentrations. Future work will examine the role of anions and the aqueous external solution in greater detail.

The differential equations are integrated with the CVODES integrator with an absolute tolerance of 1e-15, relative tolerance of 1e-12, and with 200 points stored per assay stage (600 points total). A separate integration run is used for each of the stages (i.e. t=0-1s, t=1-2s, t=2-3s) to account for the discrete-time events. For the purposes of calculating the log-likelihood, we remove the data from the equilibration stage (t=0 to t=1 s), and for the remaining assay stages, we remove the first data point and the last 101 points, which leaves a total of 2 x 198 data points.

9.5 Reference values used for transporter model parameters

Rate constant	Nominal value at V=0	Reaction type
k_1^f	1e10 (1/(M*s))	Ion binding reaction
k_1^r	1e3 (1/s)	Ion unbinding reaction
k_2^f	1e2 (1/s)	Conformational change reaction
k_2^r	1e2 (1/s)	Conformational change reaction
k_3^f	1e7 (1/(M*s))	Substrate binding reaction
k_3^r	1e3 (1/s)	Substrate unbinding reaction
k_4^f	1e3 (1/s)	Ion unbinding reaction
k_4^r	1e10 (1/(M*s))	Ion binding reaction
k_5^f	1e2 (1/s)	Conformational change reaction
k_5^r	1e2 (1/s)	Conformational change reaction
k_6^f	1e3 (1/s)	Substrate unbinding reaction

Rate constants used for ground truth model. Note that k_6^r is defined from the cycle constraints previously described.

9.6 Log-likelihood function

As mentioned in the manuscript, we use a Normal log-likelihood distribution for Bayesian inference:

$$L(\theta, \sigma^2|D) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(D_i - D_{pred}(\theta)_i)^2 \tag{52}$$

For maximum likelihood estimation, the negative log-likelihood is minimized.

9.7 Rate constant priors

We use extremely broad uniform priors with a log10 transformation for the reaction rate constants:

9.8 Nuisance parameters and priors

We assume that the residual errors in the SSME data are from a Normal distribution with zero mean and unknown variance:

log10 rate constant	log10 nominal value at V=0	log10 prior range
log10 k_1^f	10	[6,12]
log10 k_1^r	3	[-1,5]
log10 k_2^f	2	[-2,4]
log10 k_2^r	2	[-2,4]
log10 k_3^f	7	[3,9]
log10 k_3^r	3	[-1,5]
log10 k_4^f	3	[-1,5]
log10 k_4^r	10	[6,12]
log10 k_5^f	2	[-2,4]
log10 k_5^r	2	[-2,4]
log10 k_6^f	3	[-1,5]

Priors used for rate constants, using wide uniform priors and a log10 transform

$$D_{obs} = D_{true} + \delta \tag{53}$$

where $\delta \sim \text{Normal}(0, \sigma^2)$

For the 16D transporter model, we introduce additional sources of uncertainty by using scaling factors in the initial transporter concentrations, the external bath perturbations concentrations, and the observed data.

$$OF(0)_{obs} = f_{transporter} \cdot OF(0)_{true} \tag{54}$$

$$H_{out_{obs}} = f_{H_{out}} \cdot H_{out_{true}} \tag{55}$$

$$S_{out_{obs}} = f_{S_{out}} \cdot S_{out_{true}} \tag{56}$$

$$D_{obs} = f_{bias} \cdot D_{true} + \delta \tag{57}$$

$$\tag{58}$$

We use uniform priors for the nuisance parameters, with a log10 scale for the standard deviation.

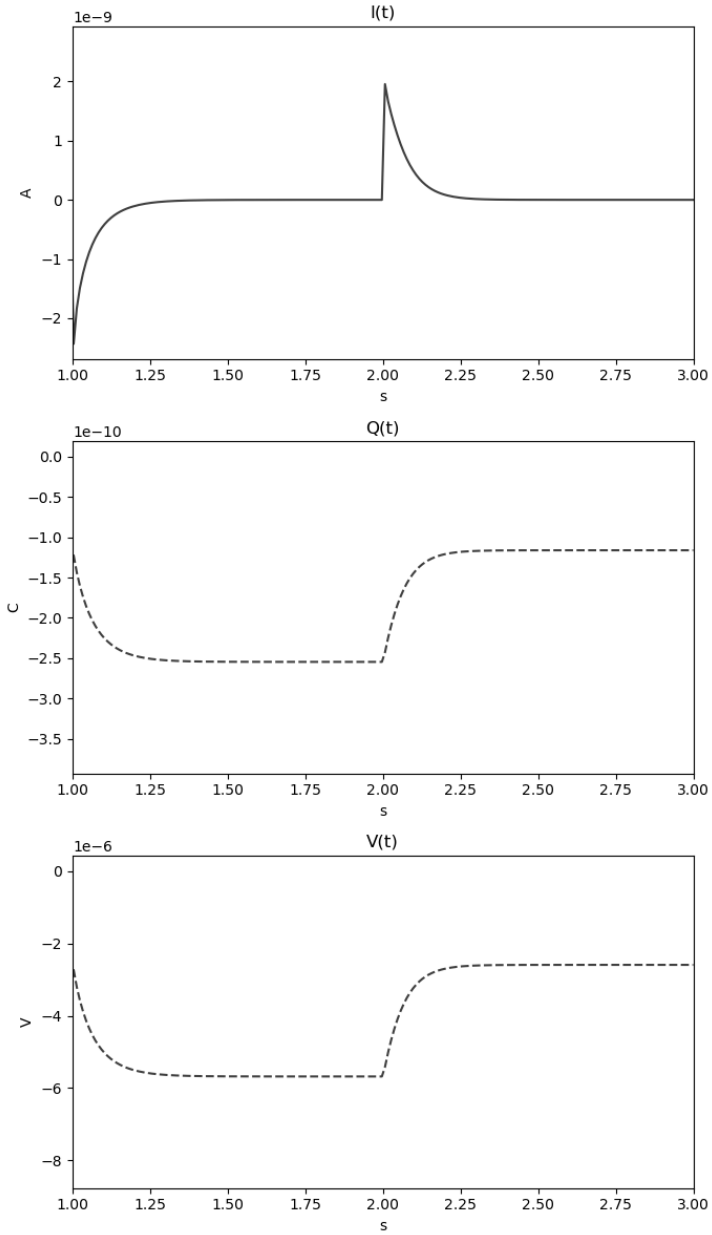
nuisance parameter	nominal value	prior range
log10 σ	-10.5	[-11,-9]
$f_{\text{transporter}}$	1	[0.8,1.2]
$f_{H_{out}}$	1	[0.8,1.2]
$f_{S_{out}}$	1	[0.8,1.2]
f_{bias}	1	[0.8,1.2]

Priors used for nuisance parameters

Electrogenic characterization of simulated SSME assay

9.9 Current, charge, and voltage

We examine the relationship between membrane voltage, charge, and current for our transporter model and simulated SSME assays: $V_m(t) = Q(t)/C_m = \int I(t)dt/C_m$



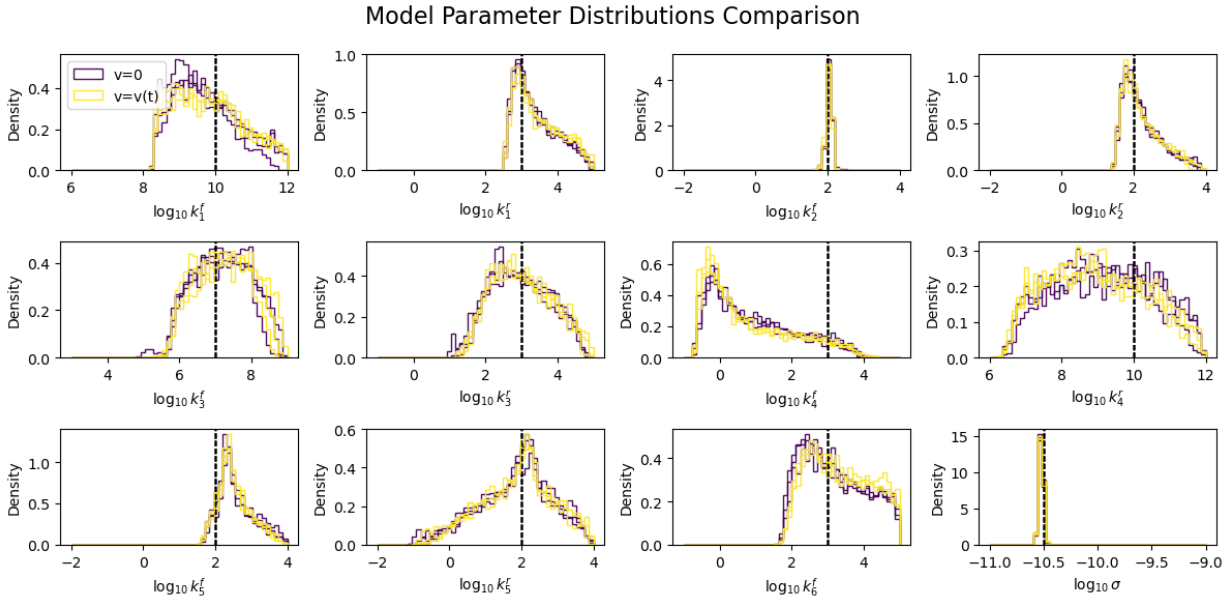
Characterizing current, charge, and membrane voltage The total current, transported charge, and membrane voltage for a simulated SSME assay for t=1-3 seconds. The equilibration phase creates a buildup of positive charge on the membrane's extracellular side, giving a negative charge Q(t) by convention.

9.10 Effects of membrane potential on rate constants

We find that the membrane voltage has a negligible effect on the reaction rate constants and the resulting current, for the conditions studied. Consider the modified rate constant form: $k(v) = k_0 * \exp(-\epsilon * \frac{VF}{RT})$. While exact values will vary depending on the transporter and experimental setup, the peak of the net currents is often in the nA range and transport a total charge in the nC range (see Bazzone et al, and Henzler-Wildman et al). This results in a per liposome membrane total charge of 1e-23 C, which when divided by a typical liposome membrane capacitance of 1e-16 F results in a voltage in the μV range. If we assume a maximum voltage of 1e-5 V, room temperature of 298K, and $\epsilon = 1$, this yields: $k(v) \approx k_0 * 0.9996$. This suggests that the membrane voltage adjusts the rate constants by a negligible amount for the conditions studied. We note that under physiological conditions, the resting membrane voltage is typically maintained in the 50-100mV magnitude range. Under these conditions, the membrane voltage will significantly adjust the rate constants, with $k(v) \approx k_0 * 0.15$ for 50 mV.

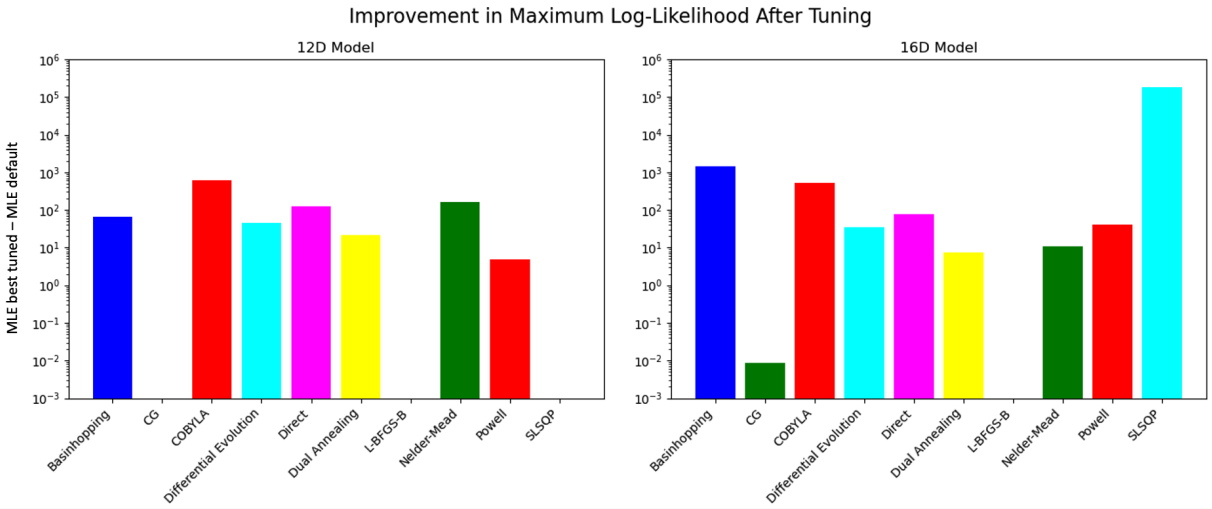
9.11 Comparing Bayesian inference results with and without dynamic voltage

We compare the same 12D transporter model using a dynamic voltage, $k(v) = k_0 * \exp(-\epsilon * \frac{VF}{RT})$, and under a zero voltage condition, $k(v) = k_0$. As expected from our model described above, we find a negligible difference between the two voltage formulations. Further work will investigate the membrane potential in greater detail.

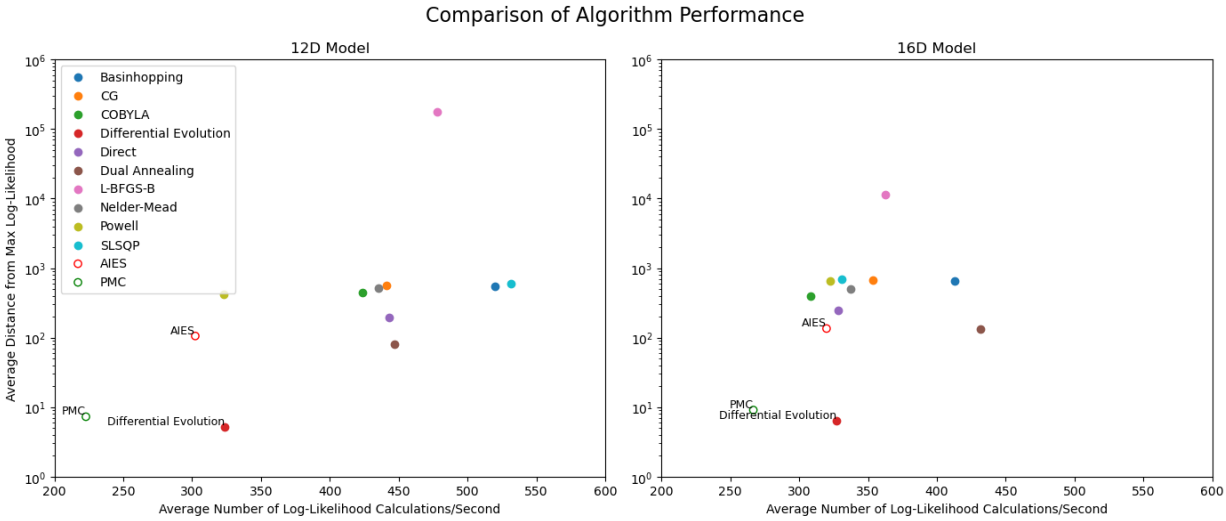


Parameter distributions The 1D parameter distributions with and without dynamic voltage. Both voltage formulations result in nearly identical parameter distributions, with three replicas for each formulation shown. The ground truth values are shown in vertical dashed lines for reference.

9.12 Additional Figures for MLE Tuning and Algorithm Performance



Improvement in MLE algorithms after tuning The difference between the maximum likelihood before and after randomized hyperparameter tuning. Future work will explore more robust tuning methods.



Comparison of Algorithm Performance The average distance from the maximum log-likelihood value vs. the number of log-likelihood calculations per second. We note that only comparing the averages may bias against Bayesian methods that generate distributions as opposed to point estimates found during maximum likelihood estimation.

10 Chapter 4 Appendix

10.1 Detailed Methods

The same simulation values in our prior work are used in this study unless otherwise noted.

10.1.1 Initial conditions

Model 2: Species Initial Concentrations

Species Label	Initial Concentration
conc_OF_Hb_Sb	0.0004369941792375325
conc_OF_Hb	0
conc_IF_Hb	0
conc_IF_Hb_Sb	0
conc_IF_Sb	0
conc_OF_Sb	0
conc_H_in	1.e-7
conc_S_in	1.e-3

Model 3: Species Initial Concentrations

Species Label	Initial Concentration
conc_OF	0.0004369941792375325
conc_OF_Hb	0
conc_IF_Hb	0
conc_IF	0
conc_IF_Sb	0
conc_OF_Sb	0
conc_H_in	1.e-7
conc_S_in	1.e-3

Model 4: Species Initial Concentrations

Species Label	Initial Concentration
conc_OF_Hb_Sb	0.0004369941792375325
conc_OF_Hb	0
conc_IF_Hb	0
conc_IF	0
conc_IF_Sb	0
conc_OF_Sb	0
conc_H_in	1.e-7
conc_S_in	1.e-3

10.1.2 Reactions and differential equations

Model 1 uses the same reactions, rates, and differential equations as described in our previous work and in the main manuscript.

Model 2 reactions and rates:

rxn1: $\text{OF\_Hb\_Sb} \rightarrow \text{OF\_Hb} + S_{\text{out}}$;

$(k_{1\text{f}} \times \text{OF\_Hb\_Sb} - k_{1\text{r}} \times \text{OF\_Hb} \times S_{\text{out}})$;

rxn2: $\text{OF\_Hb} \rightarrow \text{IF\_Hb}$;

$(k_{2\text{f}} \times \text{OF\_Hb} - k_{2\text{r}} \times \text{IF\_Hb})$;

rxn3: $\text{IF\_Hb} + S_{\text{in}} \rightarrow \text{IF\_Hb\_Sb}$;

$(k_{3\text{f}} \times \text{IF\_Hb} \times S_{\text{in}} - k_{3\text{r}} \times \text{IF\_Hb\_Sb})$;

rxn4: $\text{IF\_Hb\_Sb} \rightarrow \text{IF\_Sb} + \text{H\_in}$;

$(k_{4\text{f}} \times \text{IF\_Hb\_Sb} - k_{4\text{r}} \times \text{IF\_Sb} \times \text{H\_in})$;

rxn5: $\text{IF\_Sb} \rightarrow \text{OF\_Sb}$;

$(k_{5\text{f}} \times \text{IF\_Sb} - k_{5\text{r}} \times \text{OF\_Sb})$;

rxn6: $\text{OF\_Sb} + \text{H\_out} \rightarrow \text{OF\_Hb\_Sb}$;

$(k_{6\text{f}} \times \text{OF\_Sb} \times \text{H\_out} - k_{6\text{r}} \times \text{OF\_Hb\_Sb})$;

Model 2 differential equations:

$$\frac{d[H_{\text{in}}]}{dt} = k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}]$$
$$\frac{d[S_{\text{in}}]}{dt} = -(k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb])$$
$$\frac{d[OF\_Hb\_Sb]}{dt} = -(k_1^f[OF\_Hb\_Sb] - k_1^r[OF\_Hb][S_{\text{out}}]) + k_6^f[OF\_Sb][H_{\text{out}}] - k_6^r[OF\_Hb\_Sb]$$
$$\frac{d[OF\_Hb]}{dt} = k_1^f[OF\_Hb\_Sb] - k_1^r[OF\_Hb][S_{\text{out}}] - (k_2^f[OF\_Hb] - k_2^r[IF\_Hb])$$
$$\frac{d[IF\_Hb]}{dt} = k_2^f[OF\_Hb] - k_2^r[IF\_Hb] - (k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb])$$
$$\frac{d[IF\_Hb\_Sb]}{dt} = k_3^f[IF\_Hb][S_{\text{in}}] - k_3^r[IF\_Hb\_Sb] - (k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}])$$
$$\frac{d[IF\_Sb]}{dt} = k_4^f[IF\_Hb\_Sb] - k_4^r[IF\_Sb][H_{\text{in}}] - (k_5^f[IF\_Sb] - k_5^r[OF\_Sb])$$
$$\frac{d[OF\_Sb]}{dt} = k_5^f[IF\_Sb] - k_5^r[OF\_Sb] - (k_6^f[OF\_Sb][H_{\text{out}}] - k_6^r[OF\_Hb\_Sb])$$

Model 3 reactions and rates:

rxn1: $\text{OF} + H_{\text{out}} \rightarrow \text{OF\_Hb};$

$(k_{1\text{f}} \times \text{OF} \times H_{\text{out}} - k_{1\text{r}} \times \text{OF\_Hb});$

rxn2: $\text{OF\_Hb} \rightarrow \text{IF\_Hb};$

$(k_{2\text{f}} \times \text{OF\_Hb} - k_{2\text{r}} \times \text{IF\_Hb});$

rxn3: $\text{IF\_Hb} \rightarrow \text{IF} + H_{\text{in}};$

$(k_{3\text{f}} \times \text{IF\_Hb} - k_{3\text{r}} \times \text{IF} \times H_{\text{in}});$

rxn4: $\text{IF} + S_{\text{in}} \rightarrow \text{IF\_Sb};$

$(k_{4\text{f}} \times \text{IF} \times S_{\text{in}} - k_{4\text{r}} \times \text{IF\_Sb});$

rxn5: $\text{IF\_Sb} \rightarrow \text{OF\_Sb};$

$(k_{5\text{f}} \times \text{IF\_Sb} - k_{5\text{r}} \times \text{OF\_Sb});$

rxn6: $\text{OF\_Sb} \rightarrow \text{OF} + S_{\text{out}};$

$(k_{6\text{f}} \times \text{OF\_Sb} - k_{6\text{r}} \times \text{OF} \times S_{\text{out}});$

Model 3 differential equations:

$$\frac{d[H_{\text{in}}]}{dt} = k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}]$$
$$\frac{d[S_{\text{in}}]}{dt} = -(k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb])$$
$$\frac{d[OF]}{dt} = -(k_1^f[OF][H_{\text{out}}] - k_1^r[OF\_Hb]) + k_6^f[OF\_Sb] - k_6^r[OF][S_{\text{out}}]$$
$$\frac{d[OF\_Hb]}{dt} = k_1^f[OF][H_{\text{out}}] - k_1^r[OF\_Hb] - (k_2^f[OF\_Hb] - k_2^r[IF\_Hb])$$
$$\frac{d[IF\_Hb]}{dt} = k_2^f[OF\_Hb] - k_2^r[IF\_Hb] - (k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}])$$
$$\frac{d[IF]}{dt} = k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}] - (k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb])$$
$$\frac{d[IF\_Sb]}{dt} = k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb] - (k_5^f[IF\_Sb] - k_5^r[OF\_Sb])$$
$$\frac{d[OF\_Sb]}{dt} = k_5^f[IF\_Sb] - k_5^r[OF\_Sb] - (k_6^f[OF\_Sb] - k_6^r[OF][S_{\text{out}}])$$

Model 4 reactions and rates:

rxn1: $\text{OF\_Hb\_Sb} \rightarrow \text{OF\_Hb} + S_{\text{out}}$;

$(k_{1\text{f}} \times \text{OF\_Hb\_Sb} - k_{1\text{r}} \times \text{OF\_Hb} \times S_{\text{out}})$;

rxn2: $\text{OF\_Hb} \rightarrow \text{IF\_Hb}$;

$(k_{2\text{f}} \times \text{OF\_Hb} - k_{2\text{r}} \times \text{IF\_Hb})$;

rxn3: $\text{IF\_Hb} \rightarrow \text{IF} + H_{\text{in}}$;

$(k_{3\text{f}} \times \text{IF\_Hb} - k_{3\text{r}} \times \text{IF} \times H_{\text{in}})$;

rxn4: $\text{IF} + S_{\text{in}} \rightarrow \text{IF\_Sb}$;

$(k_{4\text{f}} \times \text{IF} \times S_{\text{in}} - k_{4\text{r}} \times \text{IF\_Sb})$;

rxn5: $\text{IF\_Sb} \rightarrow \text{OF\_Sb}$;

$(k_{5\text{f}} \times \text{IF\_Sb} - k_{5\text{r}} \times \text{OF\_Sb})$;

rxn6: $\text{OF\_Sb} + H_{\text{out}} \rightarrow \text{OF\_Hb\_Sb}$;

$(k_{6\text{f}} \times \text{OF\_Sb} \times H_{\text{out}} - k_{6\text{r}} \times \text{OF\_Hb\_Sb})$;

Model 4 differential equations:

$$\frac{d[H_{\text{in}}]}{dt} = k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}]$$
$$\frac{d[S_{\text{in}}]}{dt} = -(k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb])$$
$$\frac{d[OF\_Hb\_Sb]}{dt} = -(k_1^f[OF\_Hb\_Sb] - k_1^r[OF\_Hb][S_{\text{out}}]) + k_6^f[OF\_Sb][H_{\text{out}}] - k_6^r[OF\_Hb\_Sb]$$
$$\frac{d[OF\_Hb]}{dt} = k_1^f[OF\_Hb\_Sb] - k_1^r[OF\_Hb][S_{\text{out}}] - (k_2^f[OF\_Hb] - k_2^r[IF\_Hb])$$
$$\frac{d[IF\_Hb]}{dt} = k_2^f[OF\_Hb] - k_2^r[IF\_Hb] - (k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}])$$
$$\frac{d[IF]}{dt} = k_3^f[IF\_Hb] - k_3^r[IF][H_{\text{in}}] - (k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb])$$
$$\frac{d[IF\_Sb]}{dt} = k_4^f[IF][S_{\text{in}}] - k_4^r[IF\_Sb] - (k_5^f[IF\_Sb] - k_5^r[OF\_Sb])$$
$$\frac{d[OF\_Sb]}{dt} = k_5^f[IF\_Sb] - k_5^r[OF\_Sb] - (k_6^f[OF\_Sb][H_{\text{out}}] - k_6^r[OF\_Hb\_Sb])$$

10.1.3 Prior distributions

Model 2: Priors

Name	Bounds	Nominal
log10_k1_f	[-1, 5]	3
log10_k1_r	[3, 9]	7
log10_k2_f	[-2, 4]	2
log10_k2_r	[-2, 4]	2
log10_k3_f	[3, 9]	7
log10_k3_r	[-1, 5]	3
log10_k4_f	[-1, 5]	3
log10_k4_r	[6, 12]	10
log10_k5_f	[-2, 4]	2
log10_k5_r	[-2, 4]	2
log10_k6_f	[6, 12]	10

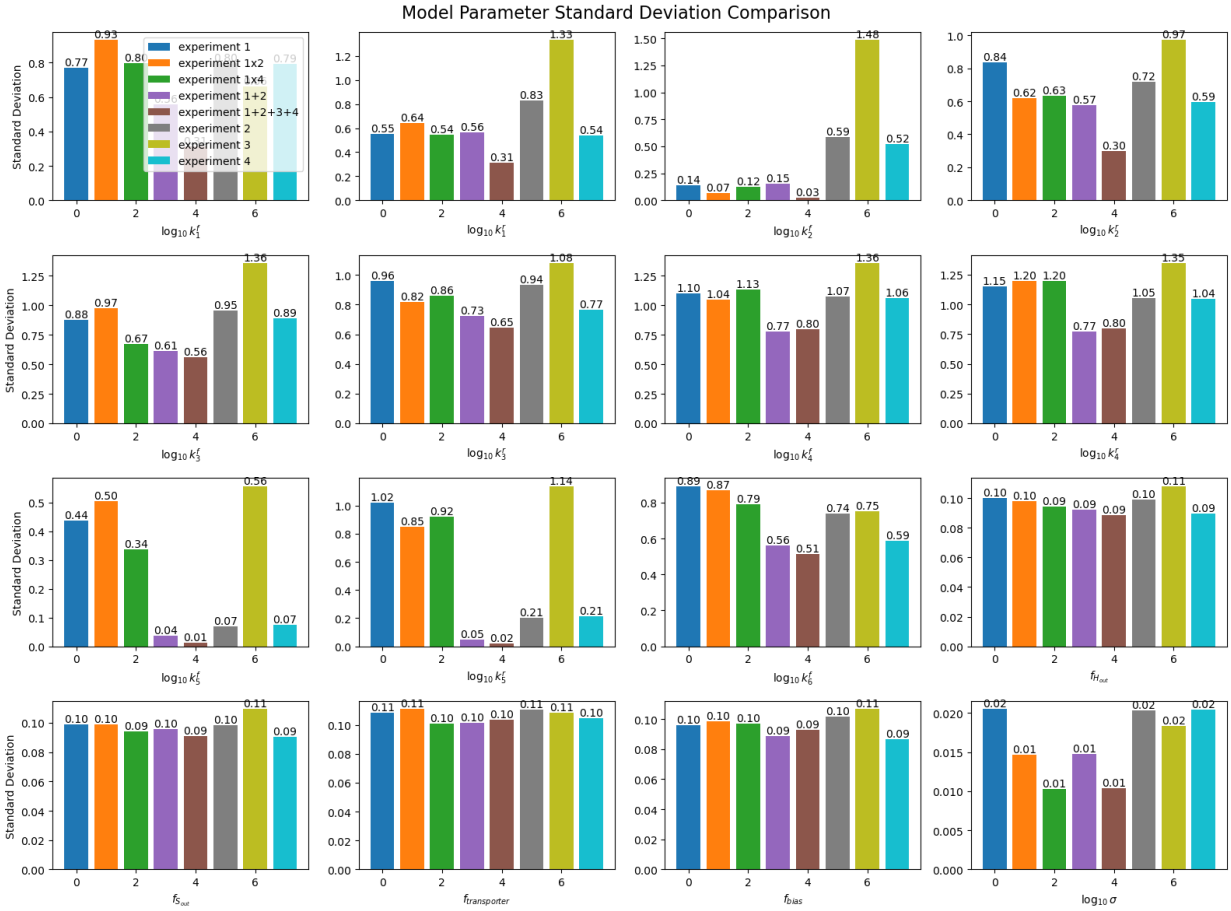
Model 3: Priors

Name	Bounds	Nominal
log10_k1_f	[6, 12]	10
log10_k1_r	[-1, 5]	3
log10_k2_f	[-2, 4]	2
log10_k2_r	[-2, 4]	2
log10_k3_f	[-1, 5]	3
log10_k3_r	[6, 12]	10
log10_k4_f	[3, 9]	7
log10_k4_r	[-1, 5]	3
log10_k5_f	[-2, 4]	2
log10_k5_r	[-2, 4]	2
log10_k6_f	[-1, 5]	3

Model 4: Priors

Name	Bounds	Nominal
log10_k1_f	[-1, 5]	3
log10_k1_r	[3, 9]	7
log10_k2_f	[-2, 4]	2
log10_k2_r	[-2, 4]	2
log10_k3_f	[-1, 5]	3
log10_k3_r	[6, 12]	10
log10_k4_f	[3, 9]	7
log10_k4_r	[-1, 5]	3
log10_k5_f	[-2, 4]	2
log10_k5_r	[-2, 4]	2
log10_k6_f	[6, 12]	10

10.2 Marginal Standard Deviations



Marginal standard deviations across each data set Note the order of data sets.

10.3 Model Evidence and Bayes Factor

Sequential Monte Carlo methods (like the pocoMC implementation of preconditioned Monte Carlo) can estimate the normalization constant in Bayes' Theorem, P(D).

This can be used to compute the Bayes factor between models:

$\text{BF}_{12} = \frac{P(D|M_1)}{P(D|M_2)}$

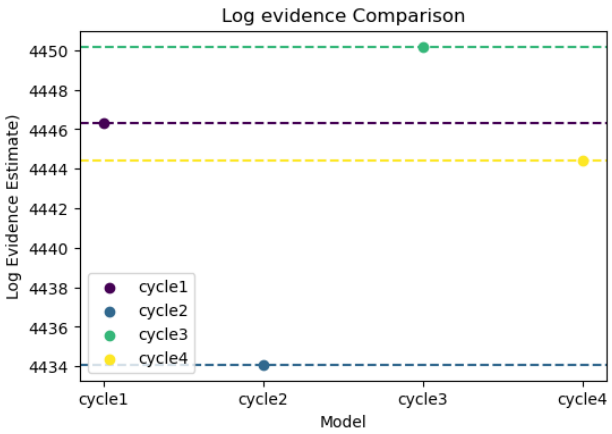
where $P(D|M1)$ is the marginal likelihood (model evidence) of the data under Model 1,

$P(D|M2)$ is the marginal likelihood (model evidence) of the data under Model 2,

and $P(D|M) = \int P(D|\theta, M) \cdot P(\theta|M)\, d\theta$.

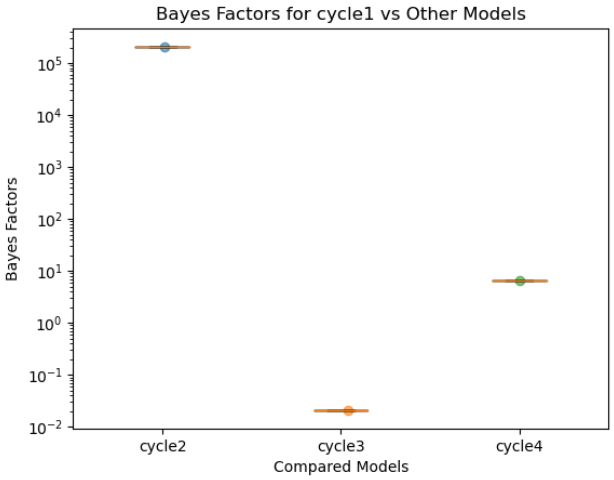
The Bayes factor calculates the odds of one model vs. another and is helpful for model selection.

For a single experiment (protocol 1, experiment 1), the model evidence and Bayes' factors were calculated against all four models, as shown below.

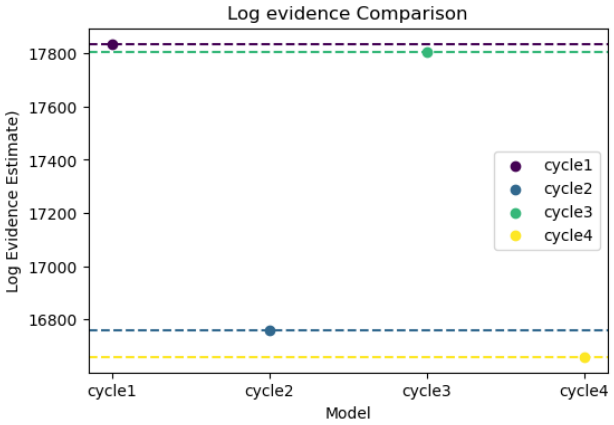


Model evidence using a single experiment

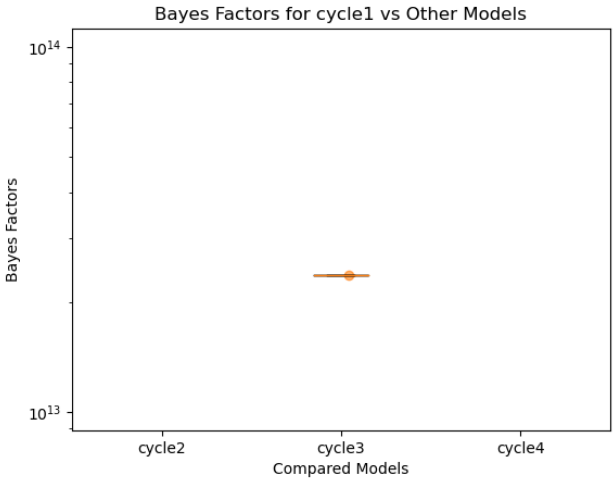
For a combined experiment (protocol 8, experiments 1,2,3, and 4 aggregated), the model evidence and Bayes' factors were calculated against all four models, as shown below.



Bayes factors using a single experiment



Model evidence using multiple experiments



Bayes factors using multiple experiments Note the large difference in model evidence between models 1, 2, and 4 results in Bayes factors between those models that are extremely large as compared to models 1 and 3, and so are omitted from the plot.

10.4 Gaussian Mixture Model Validation

Gaussian mixture models require a careful choice of the number of Gaussians to mix, K.

We use and compare both the Bayesian information criterion (BIC) and Akaike Information Criterion (AIC) and find they both give similar results.

BIC is given by:

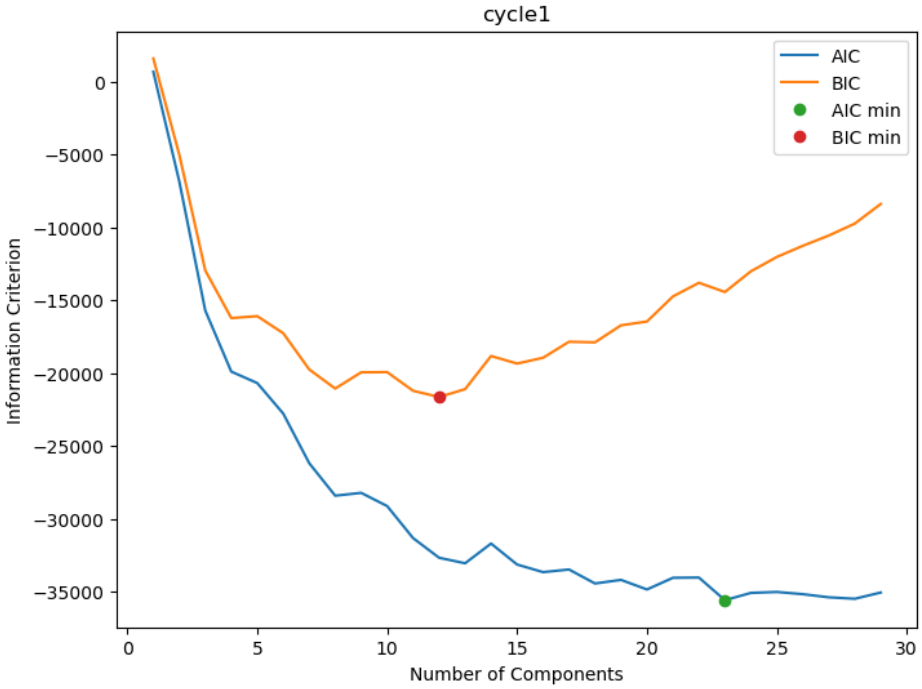
$$BIC = \ln(n)k - 2\ln(\hat{L}) \tag{59}$$

where n is the number of observations, $\hat{L}$ is the computed likelihood and k is the number of free parameters in the model.

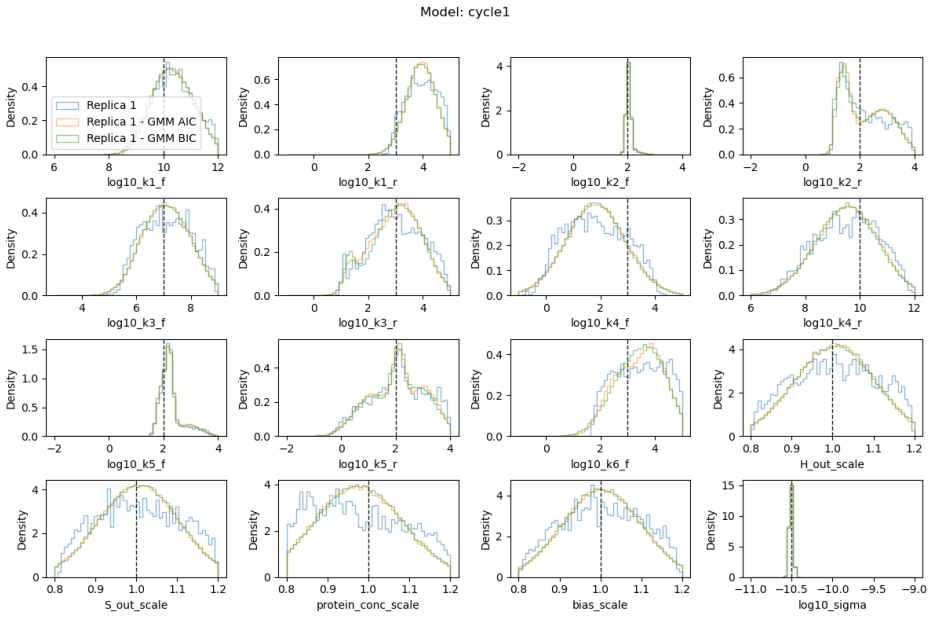
Where AIC is given by:

$$BIC = 2k - 2\ln(\hat{L}) \tag{60}$$

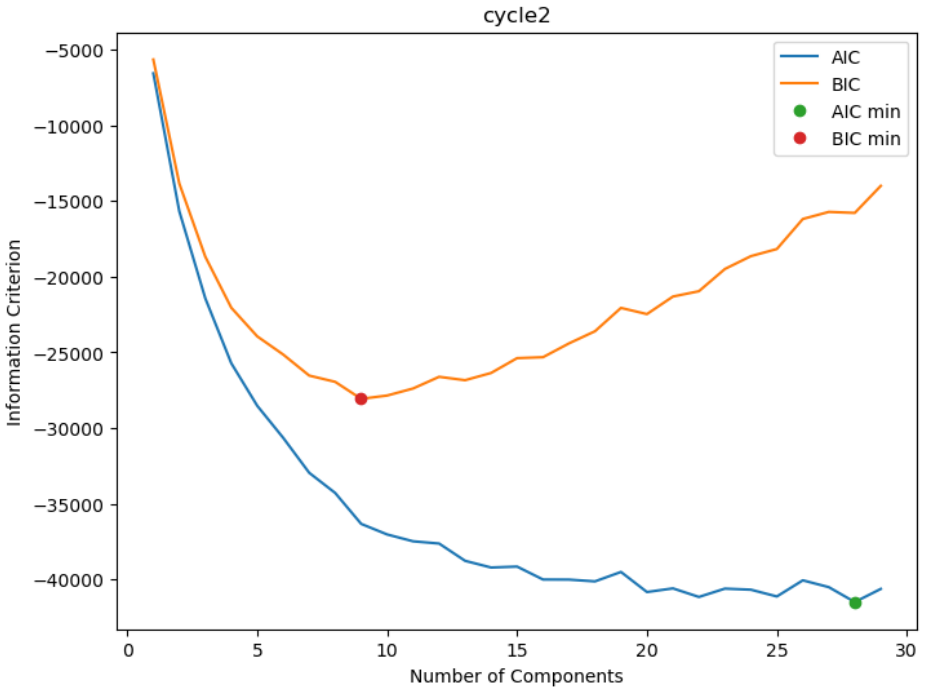
Below are the hyperparameter optimization and validation plots for all four models and the experiment 1 data set for protocol 1, as well as the hyperparameter optimization and validation plots for all four models and the combined experiments 1, 2, 3, and 4 data set for protocol 8:



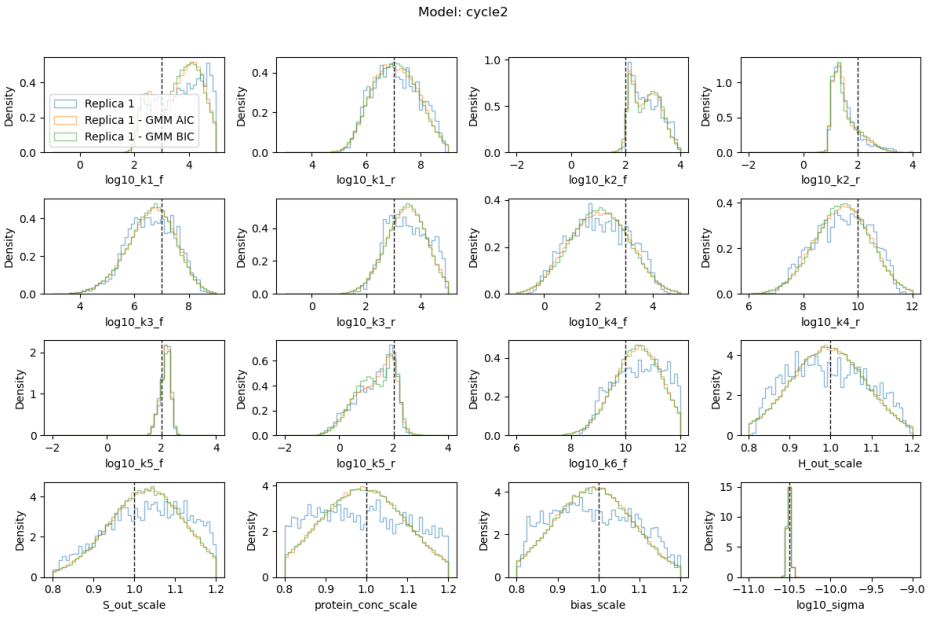
Information Criterion Minimization for Cycle 1 and Protocol 1



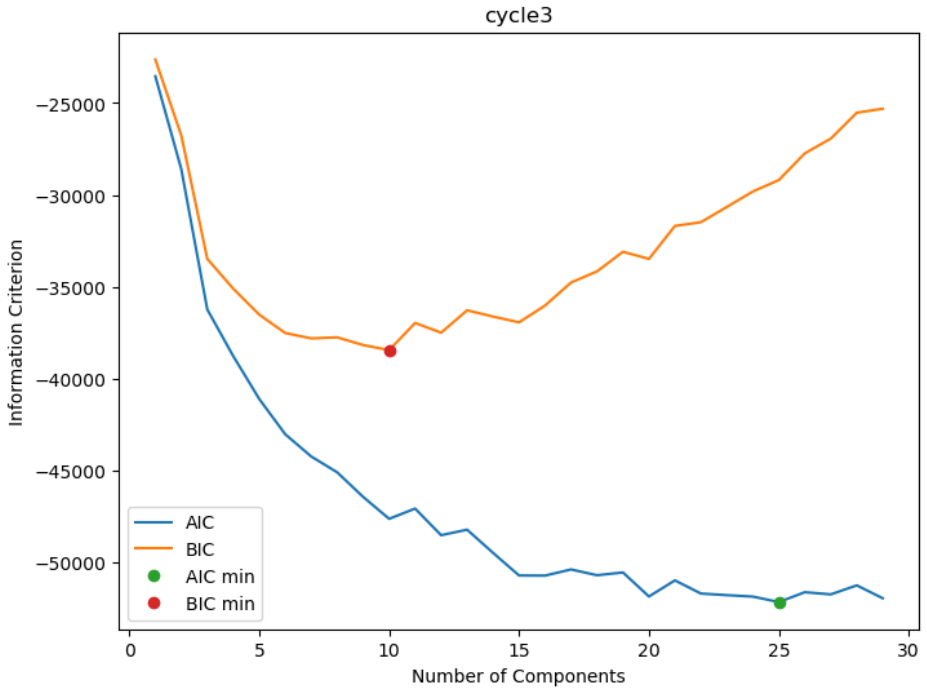
1D marginal with GMM using min AIC and BIC for Cycle 1 and Protocol 1



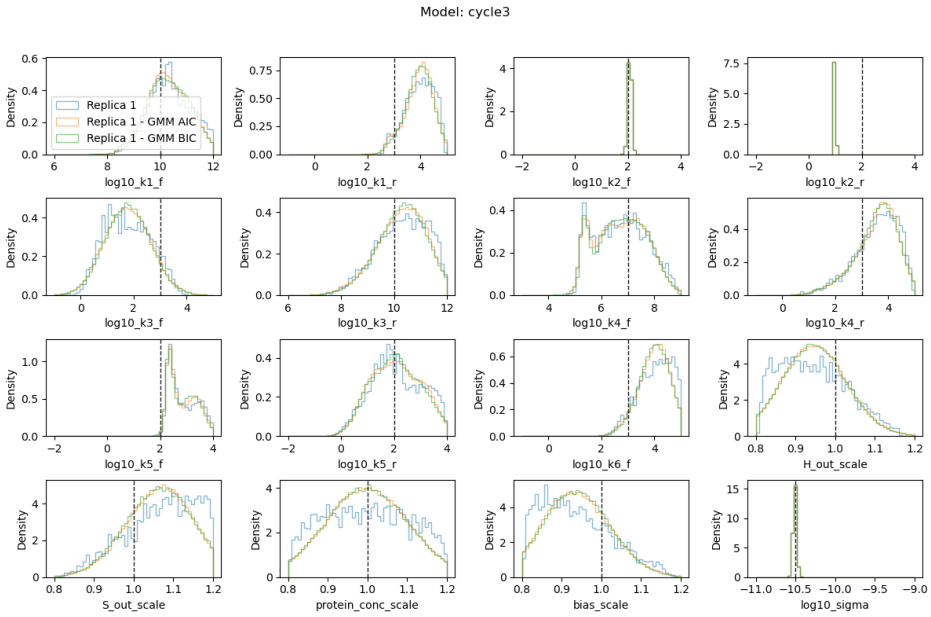
Information Criterion Minimization for Cycle 2 and Protocol 1



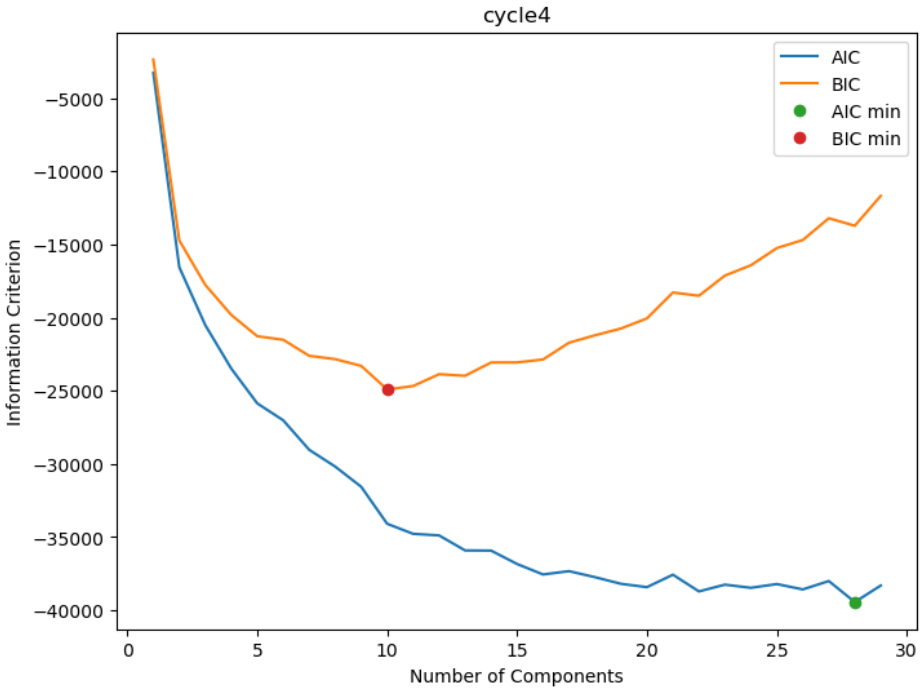
1D marginal with GMM using min AIC and BIC for Cycle 2 and Protocol 1



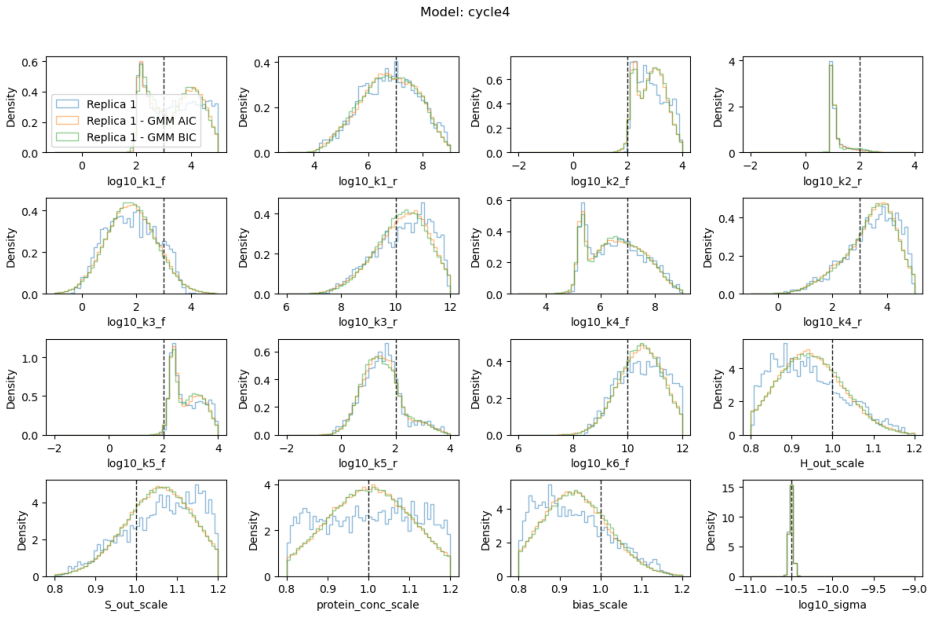
Information Criterion Minimization for Cycle 3 and Protocol 1



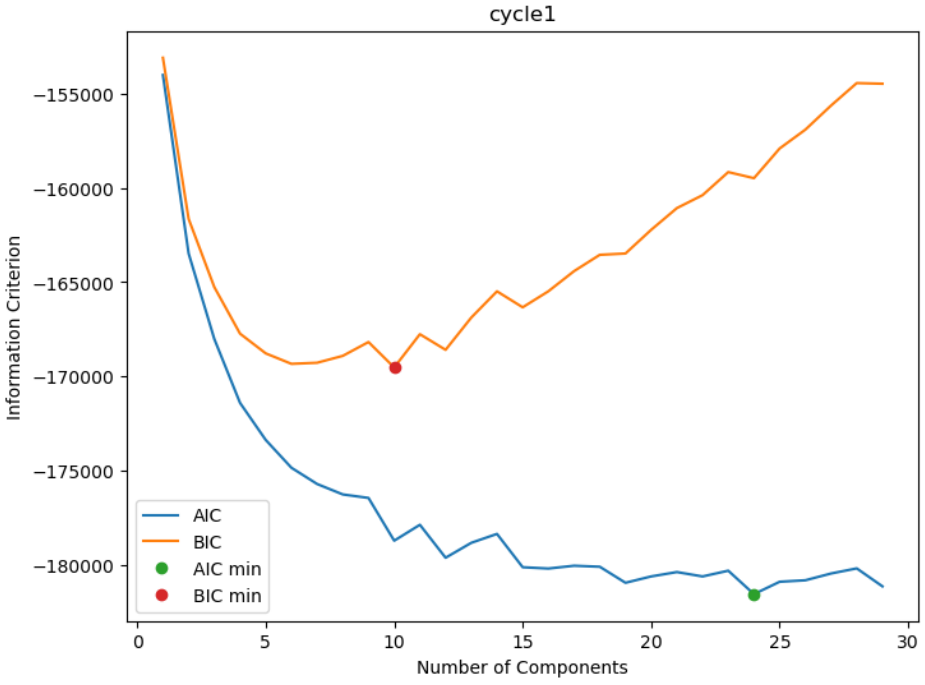
1D marginal with GMM using min AIC and BIC for Cycle 3 and Protocol 1



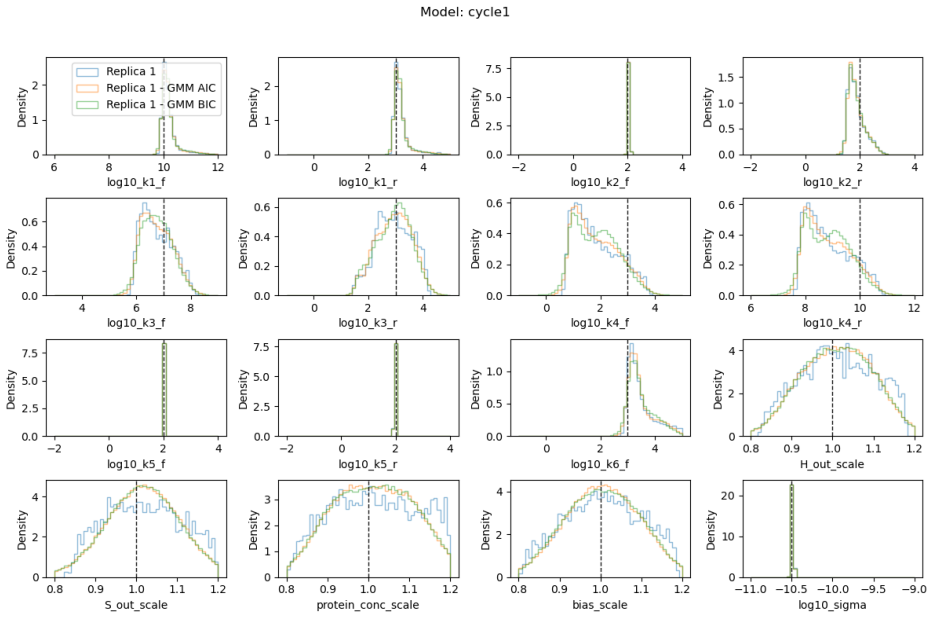
Information Criterion Minimization for Cycle 4 and Protocol 1



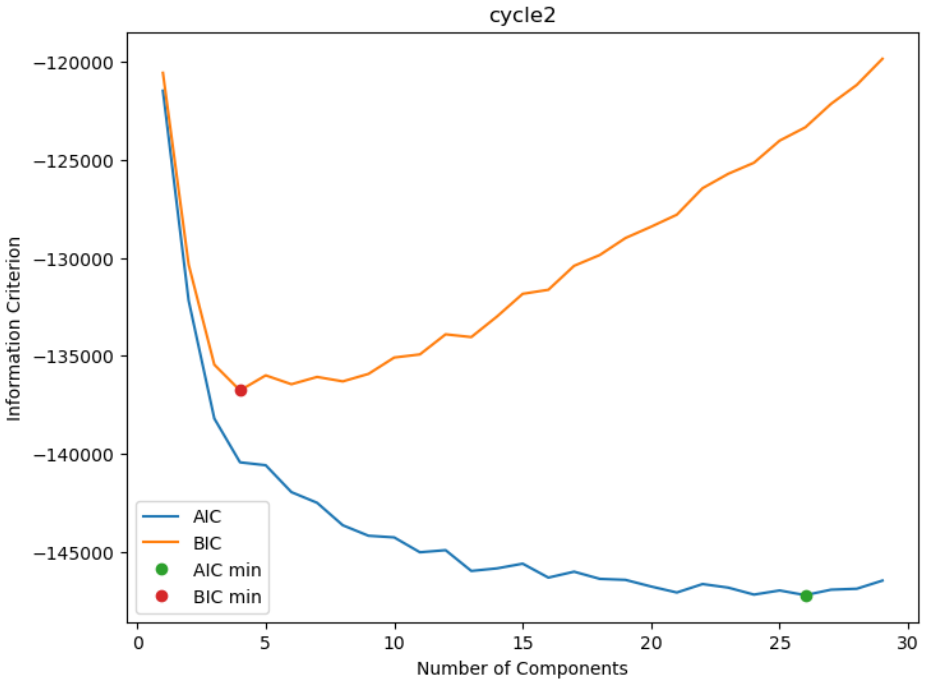
1D marginal with GMM using min AIC and BIC for Cycle 4 and Protocol 1



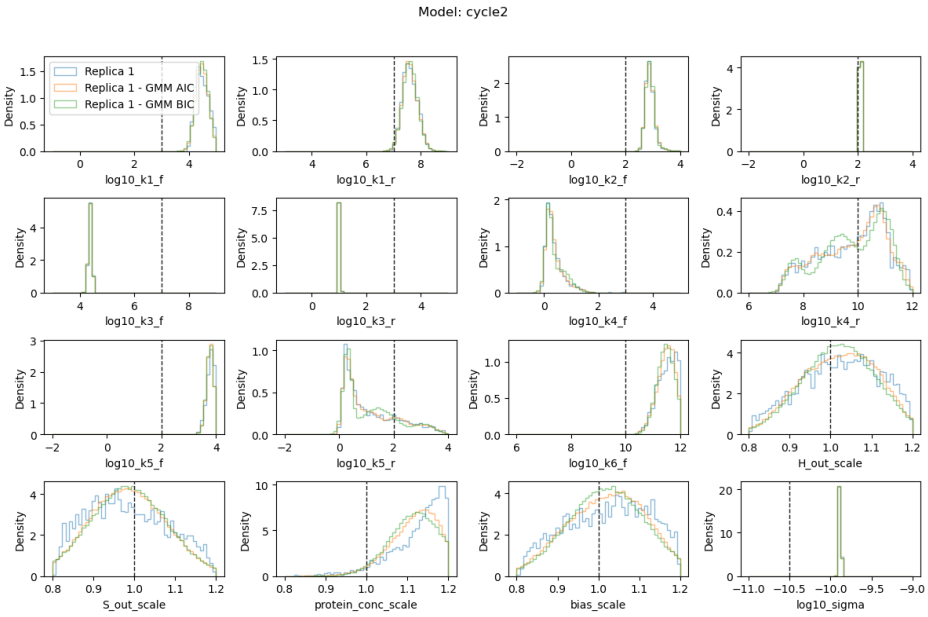
Information Criterion Minimization for Cycle 1 and Protocol 8



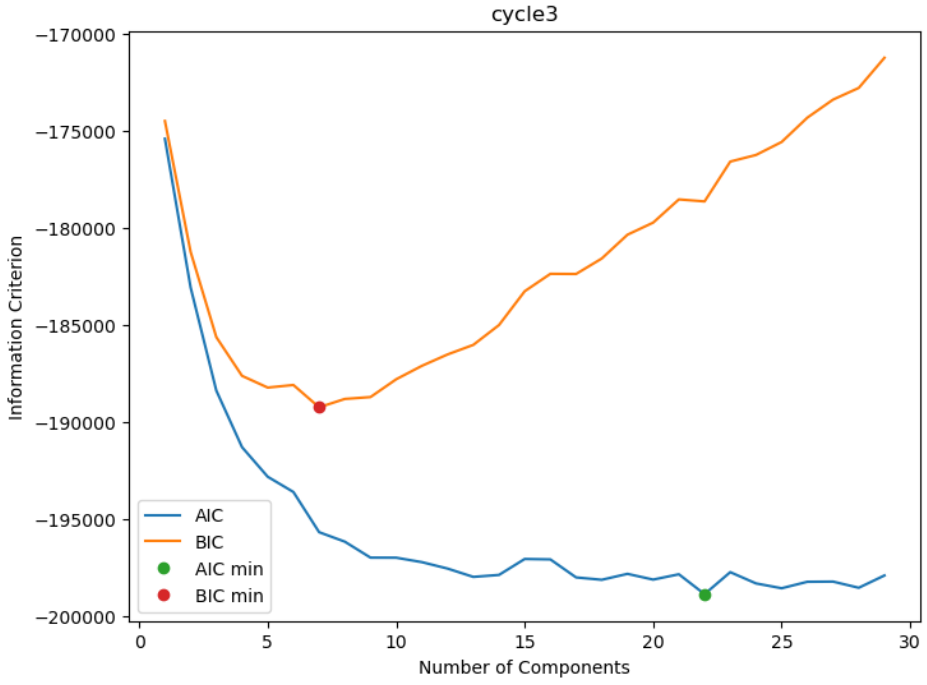
1D marginal with GMM using min AIC and BIC for Cycle 1 and Protocol 8



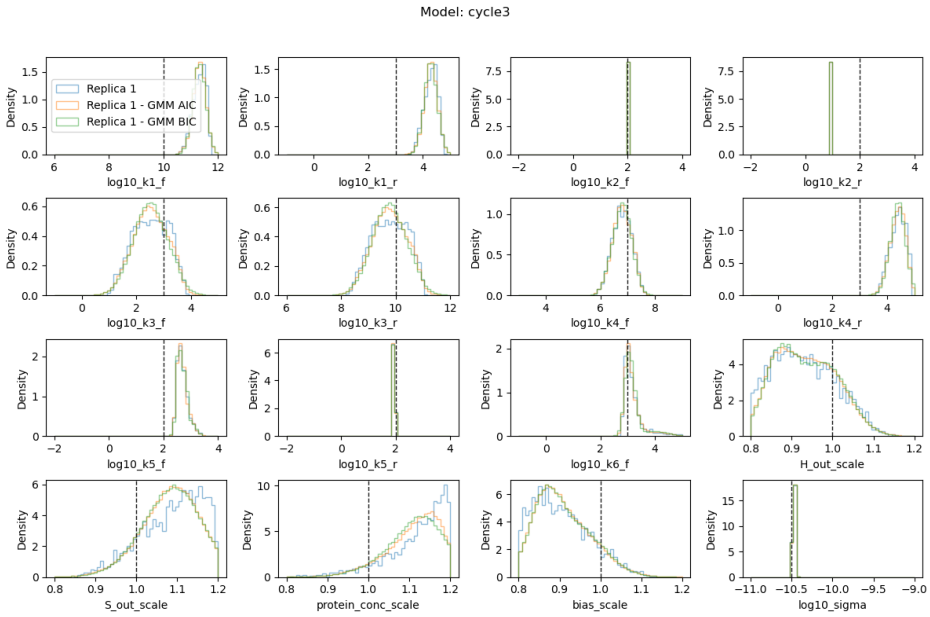
Information Criterion Minimization for Cycle 2 and Protocol 8



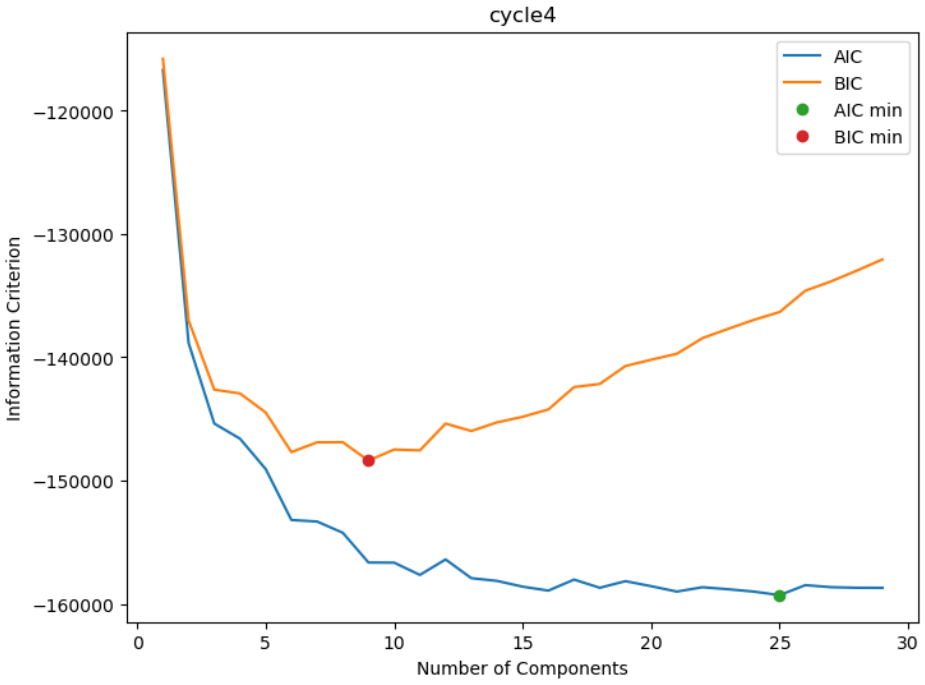
1D marginal with GMM using min AIC and BIC for Cycle 2 and Protocol 8



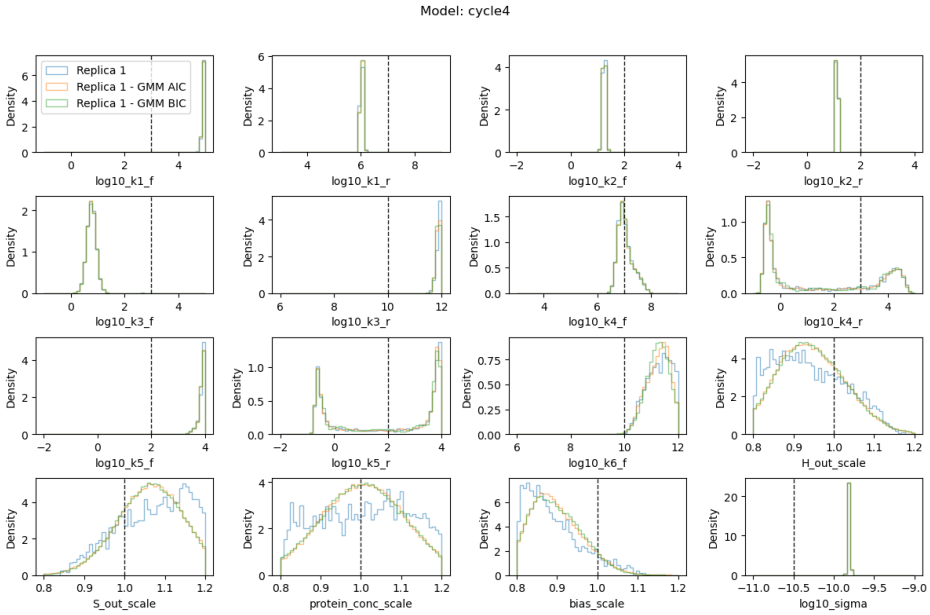
Information Criterion Minimization for Cycle 3 and Protocol 8



1D marginal with GMM using min AIC and BIC for Cycle 3 and Protocol 8



Information Criterion Minimization for Cycle 4 and Protocol 8



1D marginal with GMM using min AIC and BIC for Cycle 4 and Protocol 8