FROM PHYSICS TO PHENOTYPE: NOVEL APPROACHES FOR THE STUDY OF ALLOSTERIC MECHANISMS

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Michael V. LeVine

May 2016

FROM PHYSICS TO PHENOTYPE: NOVEL APPROACHES TO THE STUDY OF ALLOSTERIC MECHANISMS

Michael V. LeVine, Ph.D.

Cornell University 2016

It is well documented that molecular processes can be thermodynamically coupled such that the shift in the equilibrium of one process (e.g. ligand binding) can modify the kinetics and/or equilibrium of another process (e.g. receptor activation). This form of thermodynamic coupling is known as allostery and is believed to be a ubiquitous mechanism of function throughout cell, especially in the function of membrane proteins such as G protein-coupled receptors and transporters. In addition, the existence of ligand-specific allosteric modulation in both transporters and GPCRs emphasizes the importance of understanding how allostery works in these systems in terms of atomic-level physical mechanisms. Towards that goal, the work described in this dissertation will focus on two specific aims: i) the development of theoretical models that provide insight into the structural and dynamic features required for systems to be allosteric, and ii) the development of computational methods that can identify these features in specific systems of interest. First, we present a new theoretical model of allostery, the Allosteric Ising Model, which leads to several analytical conclusions regarding the structural and energetic requirements for long-distance allostery. Next, we present N-body Information Theory (NbIT) analysis, which improves on existing methods for identifying the structural components that act as allosteric channels. We illustrate the power of NbIT by identifying the allosteric channel underlying allosteric modulation of intracellular domain motions by substrate

in LeuT. Then we present a random forest-based method for identifying class-specific behavior from ensembles of the same protein bound to different ligands. This method is able to identify interactions that respond in a hallucinogen-specific manner in the serotonin receptor 5-HT_{2A}R. Finally, we present a generalized form of the two-state allosteric efficacy that can be applied to discrete and continuous variables. This description of allosteric coupling suggests that mutual information, a common measure of allostery, is fundamentally related to allostery but in itself is not a good quantification of it. The new quantification of allosteric coupling is then used to identify allosteric couplings in the simplest allosteric system, alanine dipeptide.

BIOGRAPHICAL SKETCH

Michael V. LeVine is a fifth year Ph.D. candidate in the Physiology, Biophysics, & Systems Biology program at Weill Cornell Medical College of Cornell University. He attended Coyle and Cassidy High School in Taunton, MA and received his Bachelor of Arts with high honors from Wesleyan University in Middletown, CT (triple major in Chemistry, Molecular Biology & Biochemistry, and Neuroscience & Behavior, certificate in Molecular Biophysics). His current research in the Weinstein lab focuses on using information theory and statistical mechanics to develop theoretical and computational methods for identifying and understanding the molecular-level mechanisms underlying allosteric phenomenon in signaling proteins such as GPCRs and transporters.

ACKNOWLEDGEMENTS

Foremost, I would like to gratefully acknowledge my thesis advisor, Harel Weinstein, without whom this dissertation would be impossible. I am forever grateful for the guidance that he has provided throughout my tenure in his group, in terms of research direction, research philosophy, and my career as a researcher. I would also like to thankfully acknowledge my thesis committee members, John Chodera and Scott Blanchard for their ongoing support throughout my time at Weill Cornell, and my thesis chair, David Christini. In addition, I appreciatively acknowledge the contributions of the many current and past members of the Weinstein lab, as well as our collaborators, who have made this thesis possible. In particular, I would like to thank George Khelashvili, Michel Cuendet, Jose Perez Aguilar, Jaime Medina, Michelle Sahai, Sebastian Stoltzenberg, Lei Shi, Jonathan Javitch, Mattias Quick, Daniel Terry, Scott Blanchard, Nathaniel Stanley, and Gianni De Fabritiis, who made significant contributions to the work presented in this dissertation and who have provided many invaluable discussions which have helped drive the conceptual basis of the work presented here. The NIH is gratefully acknowledged for its support of my training through both a T32 Training Grant in Pharmacological Sciences 5T32GM073546-07 (2012-2013) and a Ruth L. Kirschstein National Research Service Award F31 DA035533 (2013-2017). The NIH is also gratefully acknowledged for supporting the research described in this thesis through P01DA012408, R01DA17293, K05DA022413, U54GM087519, R01DA015170, and R01MH054137. The following computational resources used throughout the work are gratefully acknowledged: the Acellera supercomputer cluster at Barcelona Biomedical Research Park; an XSEDE allocation at the Texas Advanced Computing Center at the University of Texas at Austin (Stampede supercomputer, Projects TG-MCB090132 and TG-MCB120008); resources of the Oak Ridge Leadership Computing Facility (ALCC allocation BIP109)

at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725; an allocation at the National Energy Research Scientific Computing Center (NERSC, repository m1710) supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; an allocation on the Anton supercomputer (Grant No. PSCA14026P); and the computational resources of the David A. Cofrin Center for Biomedical Information in the HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine at Weill Cornell Medical College.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF SCHEMES

1. Introduction

Note: much of the text in this chapter has been adapted from two previously published manuscripts[1,2], with permission from the publisher.

1.1. Proteins as Molecular Machines

British science fiction writer Arthur C. Clark famously stated, "any sufficiently advanced technology is indistinguishable from magic"[3]. Technology has played a crucial role in the evolution of human society, and while Clark's quote is often used in reference to potential alien or futuristic technologies, the technology of today is often indistinguishable from the magic of yesterday. Since the earliest days of civilization, we have invented and constructed tools and machines to aid us in performing the many difficult tasks that are require for human survival and flourishing. In particular, much of the history of mankind was forged on the back of mechanical engineering – on the back of clocks, steam engines, pumps, and power plants. But while to the modern, educated eye, man-made machines are entirely distinguishable from magic, as we have the tools from mathematics, physics, and engineering to build them and understand how they work, there is a whole hidden world of machines that were not built by man. These machines were built over millions of years by the process of biological evolution. They are the machines that the human species is made of – the proteins, nucleic acids, and other biomolecules that make up human cells and the cells of every other living organism on the planet. However, because man didn't design these machines, because they arose out of random mutation and natural selection, even the most well-studied modern scientist does not have a complete understanding as to how the machines work, and up to this point, there has been limited success in building our own synthetic biomolecular machines de novo. This lack of

understanding may explain to some extent, why for much of history, humanity largely subscribed to a vitalistic view of biology; it was believed that the matter composing living organisms was somehow fundamentally different from that of non-living matter. Vitalism[4], which mistook molecular biology for magic, did eventually fall out of favor, in part due to Friedrich Wohler's 1828 discovery[5] that a biological substance, urea, could be synthesized without the use of biological material. Only within the last century, in 1931, did the physiologist John Scott Haldane declare "biologists have almost unanimously abandoned vitalism as an acknowledged belief"[6]. With the replacement of vitalism with mechanistic and reductionist scientific philosophies, biological entities such as protein began to be described in the language of man-made machines, merging the life sciences with the physical sciences and giving birth to the fields of biochemistry and biophysics.

Proteins have been conceptualized as molecular machines since as early as 1950[7], when the term was used to describe the oxygen transport protein hemoglobin. Just as mechanical machines are not random assemblies of parts, but instead are able to perform their functions due to the purposeful arrangement of those parts by engineers, proteins are not simply linear biopolymers of amino acids, and have evolved to perform their many functions by folding into 3-dimensional structures that are composed of a hierarchy of secondary and tertiary structural elements. These 3-dimensional structures can now be determined using techniques like x-ray crystallography[8], nuclear magnetic resonance (NMR)[9], and cryo-electron microscopy (cryoEM)[10], and the growing accessibility of protein structural data has led to a conserved effort in the theme of structure-function relationships[11–13], in which we would like to deduce the mechanism of a protein's function through the structures the protein takes on.

However, a protein's function cannot be fully understood through only the structures or states in which it can be found in. In order to describe the physical mechanisms involved in the function of a protein it is essential to i) define the states involved in the functional process in terms of molecular structure, ii) define the relations among those states in a kinetic model dependent on rate and equilibrium constants, and iii) express the protein function in terms of the kinetic model. The thermodynamics and kinetics of protein function have been studied by biochemists and biophysicists for some time, and detailed experiment and analysis has revealed that essential rate and equilibrium constants that describe the state and conformational changes of the protein are often modulated by its environment (especially in the case of membrane proteins) and by the ions, substrates, and ligands involved in the functional process. This crucial modulation of the kinetics and thermodynamics of the protein by outside actors is known as allostery. Despite the apparently fall of vitalism in biology, allostery is still largely invoked as a magic-like biophysical phenomenon with little mechanistic understanding. Allostery is often used as the answer for questions of "how", such as "how does a ligand activate a receptor", but this answer is not satisfying as it provides very little new understanding of the nature of the receptor itself or the nature of the specific receptor-ligand interaction. It is likely that more important are the questions such as "how does the ligand activate the receptor allosterically", which is a question that is often difficult to answer due to the fact that allostery has been largely phenomenological. Developing a physics-based, mechanistic description of the phenomenon of allostery will allow for us to answer questions that forward out understanding of biomolecular machines, and has been the focus of my doctoral research and this dissertation.

1.1.1. Allostery

1.1.1.1. History

It is often stated that the first recorded observation of allostery was by Christian Bohr, who found that oxygen displayed an unusual shaped equilibrium binding curve[14]. In the simple case of a ligand L binding a protein P, the equilibrium is written as:

$$L + P \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} LP \tag{1.1}$$

where k_{on} and k_{off} are the rate constants for binding and unbinding, respectively. At equilibrium, this defines the equilibrium binding dissociation constant,

$$\frac{[LP]}{[L][P]} = \frac{k_{off}}{k_{on}} = K_D \tag{1.2}$$

It can easily be shown that the fraction of ligand:protein complexes as a function of the concentration of ligand is simply the Michaelis-Henri equation:

$$\frac{[LP]}{[P]+[LP]} = \frac{[L]}{K_D + [L]} \tag{1.3}$$

If one plots this fraction as a function of the concentration of ligand, the resulting curve is known as a saturation binding curve (see **Figure 1**).

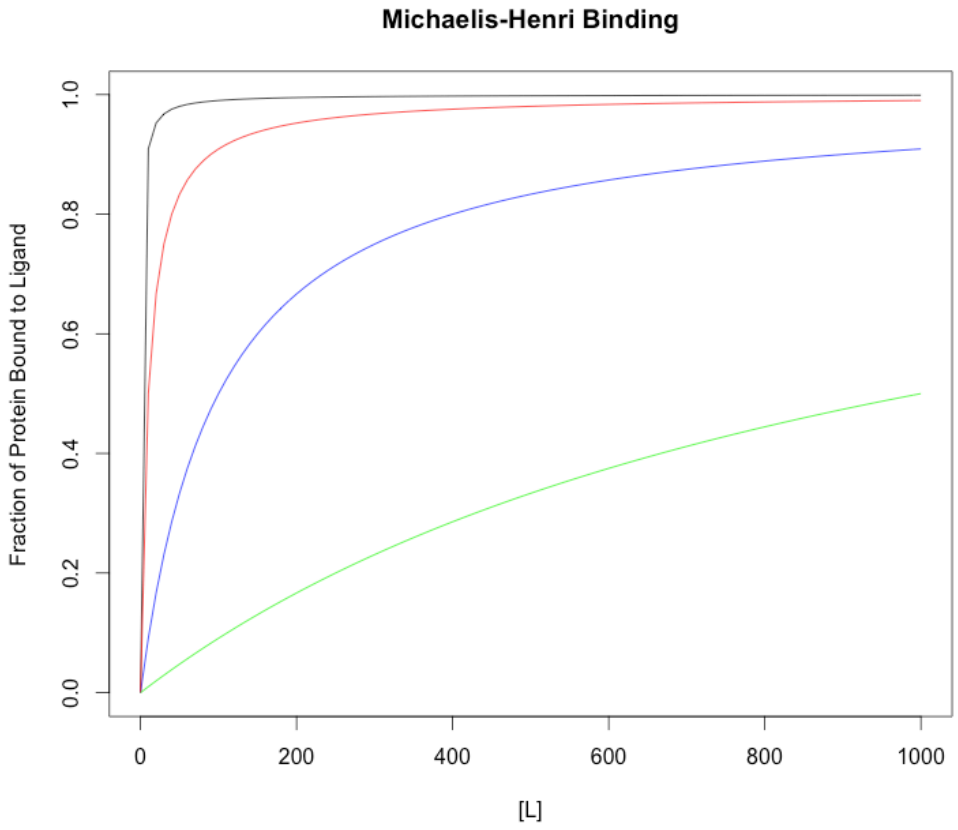


Figure 1. Expected saturation binding curve for Michaelis-Henri binding.
The fraction of protein bound to ligand is shown as a function of ligand concentration (in arbitrary unites) for K_D of 1 (black), 10 (red), 100 (blue), and 1000 (green).

As hemoglobin is a four subunit protein and can binding one oxygen in each subunit, if the binding were independent in each subunit, (1.3) would be sufficient and binding would display the expected saturation behavior. However, Bohr found that hemoglobin did not generate display the expected saturation binding curve, but rather a sigmoidal binding curve that was dependent on the partial pressure of carbon dioxide (see **Figure 2**).

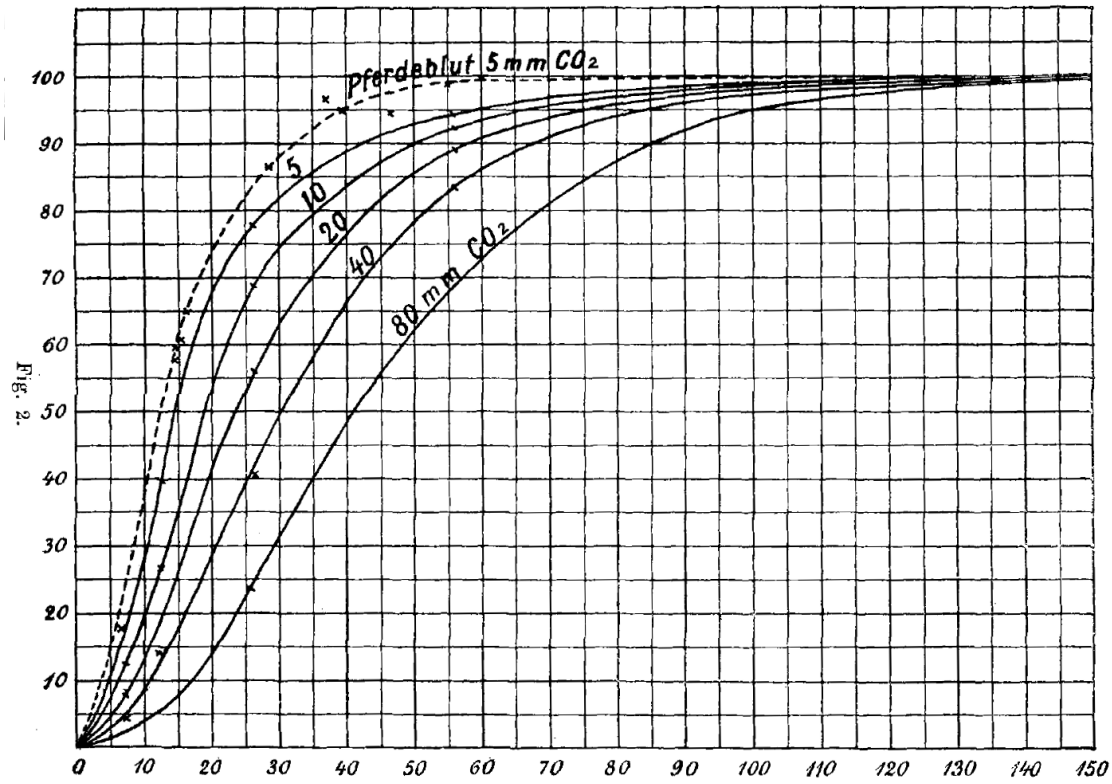


Figure 2. Sigmoidal oxygen binding in hemoglobin.

The original plot by Bohr[14] indicating both cooperative binding of oxygen and competitive binding by carbon dioxide, reproduced with permission. The x-axis is the oxygen partial pressure in mmHg, and the y-axis is the percentage oxy-hemoglobin. Reproduced with permission.

This led to the interpretation that binding in one subunit of hemoglobin was changing the binding affinity of oxygen binding in the other subunits. This phenomena and others like it are generally referred to as cooperativity.

Despite the role of hemoglobin as a prototypical allosteric system, there were many examples of allostery in a non-oligomeric system by the mid 1900s. In 1963, Monod, Changeux, and Jacob noted that "[it] would appear, in other words, that certain proteins, acting at critical metabolic steps, are electively endowed with specific

functions of regulation and coordination; through the agency of these proteins, a given biochemical reaction is eventually controlled by a metabolite acting apparently as a physiological "signal" rather than as a chemically necessary component of the reaction itself"[15]. Here, they referred to the common process of end-product inhibition, in which enzymes are often inhibits by downstream metabolites in a non-competitive manner. While this observation can be seen at the spark that drove Monod, Changeux, and Jacob to being their work on a theory of allostery, which will be described in the following section (Section 1.1.1.2. Theoretical Background), many instances of non-oligomeric allostery have been observed since[16], and it is now believed that nearly all protein may be allosteric[17]. Of specific interest to pharmacology and medicine has been the allostery involved in the activation of receptor proteins such as the G protein coupled receptors (GPCRs). In addition, many other membrane proteins have been noted to be allosteric, such as the secondary active symporters. These two systems will be the focus of study in this dissertation, and will be described in significant detail in Section 1.1.2. Allostery in Membrane Proteins.

1.1.1.2. Theoretical Background

Due to the ubiquitous nature of allostery, many have sought to define it using quantitative, theoretical models. In order to interface with experiments, many of these models are thermodynamic in nature. The first model is known as the Monod-Wyman-Changeux (MWC) model[18] and was constructed to describe the cooperativity among several ligands binding to the same protein. In the following description of the model, we will differ from the original notation used by MWC in order to be consistent in model notation throughout the dissertation. Rate constants will be written as $k^{state}_{process}$, where the subscript denotes the transformation process and the superscript denotes relevant characteristics about the state of the system on which the process is acting.

Similarly, an equilibrium constant will be written as $K^{state}_{process}$, where the state and process correspond to those of the forward reaction. In the cases presented throughout the dissertation, these characteristics will include, but are not limited to, being *activated* or *bound* to ligands; the unbound, inactive state will be taken as default and not indicated as a superscript. In the MWC model, one imagines an oligomer of identical protomers. Each protomer has two states, R and T, which refer to be "relaxed" or "tense", respectively. Additionally, the protomers within an oligomer are forced to be in identical states such that the whole oligomer has only two states, R and T. We will refer to the protomer as P, and will note the state with subscript, and their equilibrium constant will be denoted as Ktense.

$$\frac{[P_T]}{[P_R]} = K_{tense} \tag{1.4}$$

Ligand can bind sequentially to either state such that

$$\frac{[P_R L]}{[P_R][L]} = K^{R}_{bind,L} \tag{1.5}$$

and

$$\frac{[P_T L]}{[P_T][L]} = K^{T}_{bind,L} \tag{1.6}$$

We will define the ratio of these binding affinities as

$$\alpha = \frac{K^{R}_{bind,L}}{K^{T}_{bind,L}} \tag{1.7}$$

Taking into account the probabilities for n identical binding sites, the following equilibrium equations are written:

$$
\begin{aligned}
\left[P_R L\right] &= n\frac{\left[P_R\right]\left[L\right]}{K^R_{bind,L}} & \left[P_T L\right] &= n\frac{\left[P_T\right]\left[L\right]}{K^T_{bind,L}} \\
\left[P_R (L)_2\right] &= \frac{n-1}{2}\frac{\left[P_R L\right]\left[L\right]}{K^R_{bind,L}} & \left[P_T (L)_2\right] &= \frac{n-1}{2}\frac{\left[P_T L\right]\left[L\right]}{K^T_{bind,L}} \\
&\cdots & &\cdots \\
\left[P_R (L)_n\right] &= \frac{1}{n}\frac{\left[R(L)_{n-1}\right]\left[L\right]}{K^R_{bind,L}} & \left[P_T (L)_n\right] &= \frac{1}{n}\frac{\left[P_T (L)_{n-1}\right]\left[L\right]}{K^T_{bind,L}}
\end{aligned}
\tag{1.8}
$$

The fraction of protein bound to the ligand as a function of ligand concentration, Y_L is then:

$$
\bar{Y}_L = \frac{K_{tense}\alpha\frac{\left[L\right]}{K^R_{bind,L}}\left(1+\alpha\frac{\left[L\right]}{K^R_{bind,L}}\right)^{n-1} + \frac{\left[L\right]}{K^R_{bind,L}}\left(1+\frac{\left[L\right]}{K^R_{bind,L}}\right)^{n-1}}{K_{tense}\left(1+\alpha\frac{\left[L\right]}{K^R_{bind,L}}\right)^{n} + \left(1+\frac{\left[L\right]}{K^R_{bind,L}}\right)^{n}}
\tag{1.9}
$$

In the MWC model, the parameters that control the observed coopertivity are K_{tense} and α. K_{tense}, the intrinsic conformational preference of the protein, controls how slowly the sigmoid saturates. As K_{tense} is increased, the saturation occurs more slowly. α, which describes the degree of ligand preference for binding the R state over the T state, controls how sharp the sigmoidal function is (e.g. the slope around the inflexion point). While this model can qualitatively predict many of the features of cooperative proteins, several assumptions are made that limit the use. First, it assumes that the protomers must all be in the same state. This implies that there must be an additional parameter that describes the coupling between protomers, and that the parameter is at the limit of maximal coupling. Additionally, once a ligand binds, there is no equilibrium established between bound T and R states, which can only be assumed if there is a separation of time scales between the equilibriums between the T and R states while bound to ligand and all other equilibriums. While extended approaches of the MWC model have been developed to remove these assumptions, they most relate

to modeling cooperativity in oligomeric systems, which is not the specific focus of this dissertation.

Even within these generalizations, the MWC model fails to describe allostery within monomers, such as allosteric modulation of an enzyme's activity by a non-substrate ligand, or activation of a receptor by the receptor's ligand. To describe allostery within the context of a ligand activating a single protomer, the two-state allosteric model (TSAM) was developed. As an example, one can imagine a description of protein function in terms of its two distinct states, one of which is "active" and the other is "inactive". We can represent the equilibrium of such a protein transitioning between an "inactive" state (P) and an "active" state (P*).

$$P \underset{k_{inactivate}}{\overset{k_{activate}}{\rightleftarrows}} P^{*} \tag{1.10}$$

and

$$K_{activate} = \frac{k_{activate}}{k_{inactivate}} \tag{1.11}$$

The binding of ligand (L) to P

$$L + P \underset{k_{unbind,L}}{\overset{k_{bind,L}}{\rightleftarrows}} LP \tag{1.12}$$

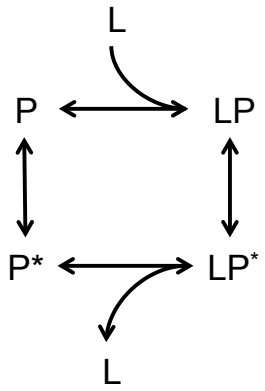
can modify the equilibrium between in active and active states in some way:

$$LP \underset{k^{L}_{inactivate}}{\overset{k^{L}_{activate}}{\rightleftarrows}} LP^{*} \tag{1.13}$$

so that a new equilibrium constant is achieved

$$K^{L}_{activate} = \frac{k^{L}_{activate}}{k^{L}_{inactivate}} \tag{1.14}$$

This corresponds to the following thermodynamic cycle connecting the activation and the ligand binding processes:



Scheme 1. The thermodynamic cycle for protein activation and ligand binding.

For this thermodynamic cycle, it is possible to quantify the allosteric efficacy. The ***allosteric efficacy*** is a measure of the allosteric coupling between two equilibrium processes[18–20], and can be used to characterize the allostery in the thermodynamic cycle presented in Scheme 1. The allosteric efficacy, α, with which this particular ligand binding process modifies the activation equilibrium, is expressed as the ratio of the equilibrium constants:

$$\alpha_{activate}^{bind,L} = \frac{K_{activate}^{L}}{K_{activate}} \tag{1.15}$$

It should be noted that the α of the TSAM turns out to be equivalent to the α used in the MWC model, which may suggest that it is a fundamental characteristic of models of allostery in general. In an equilibrium regime, the binding equilibrium constants of L to P and P*, will also change proportionately, such that the allosteric efficacy can equivalently be defined as

$$\alpha_{\text{activate}}^{\text{bind,L}} = \frac{K_{\text{bind,L}}^{*}}{K_{\text{bind,L}}} \tag{1.16}$$

Recalling that an equilibrium constant is a function of the difference in free energy of the two states:

$$K = e^{-\beta \Delta G} \tag{1.17}$$

then

$$-\frac{1}{\beta}\log\left(\alpha_{\text{activate}}^{\text{bind,L}}\right) = G\left(LP^{*}\right) + G(L+P) - G(LP) - G\left(L+P^{*}\right) \tag{1.18}$$

When $\alpha = 1$, the ligand binding is not coupled to the state of the protein, whereas an $\alpha > 1$ denotes positive coupling (i.e., the binding of ligand increases the probability of the active state) and $\alpha < 1$ denotes a negative coupling (i.e., the binding decreases the probability of the inactive state). This type of allostery in which the ligand modulates an equilibrium constant, is known as ***K-type allostery*** and is recognizable in a great variety of systems[16,21]. One of the most notable examples is the activation of receptors (e.g., GPCRs) by ligands, which will be discussed in the following section.

Notably, however, in addition to K-type allosteric modulation of equilibrium constants, the experimental evidence pointing to the modulation of maximum velocity of enzymatic reaction, v_{max} , by allosteric ligands indicates that there is a second type of allosteric modulation possible. In Michaelis-Menten kinetics[22], one uses a two-step, irreversible kinetic model:

$$E + S \underset{k_{\text{unbind,S}}}{\overset{k_{\text{bind,S}}}{\rightleftarrows}} ES \xrightarrow{k_{\text{cat}}} E + P \tag{1.19}$$

The rate of product formation is then:

$$\frac{d[P]}{dt} = \frac{v_{max}[S]}{K_D + [S]} \tag{1.20}$$

where

$$v_{max} = k_{cat}([E] + [ES]) \tag{1.21}$$

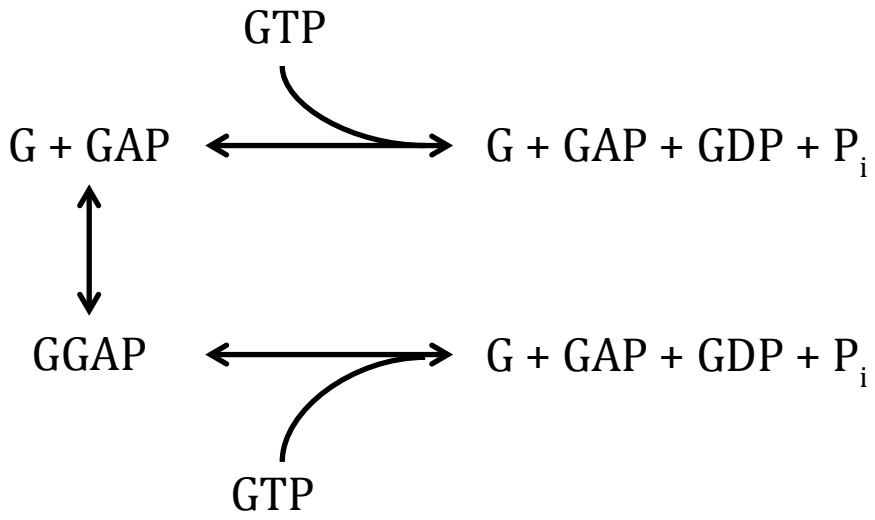
The second type of allosteric modulation describes the case in which v_{max} is modulated by changing the rate constants, and is known as **V-type allostery**[18,21]. V-type allostery has been identified in several enzymes, although it is currently thought to be rare, accounting for less than 1% of allosteric mechanisms. Notably, some G proteins have been shown to exhibit V-type allosteric regulation. In particular, GTPase activating proteins (GAPs) bind to G proteins and increase the k_{cat} of GTP hydrolysis. For p21ras, for example, k_{cat} was shown to increased by over four orders of magnitude[23]. This type of allostery couples a binding reaction,

$$GAP + G \underset{k_{unbind,GAP}}{\overset{k_{bind,GAP}}{\rightleftharpoons}} GAPG \tag{1.22}$$

to GTP hydrolysis,

$$G + GTP \xrightarrow{k_{hydrolysis}} G + GDP + P_I \tag{1.23}$$

as is shown in Scheme 2:



Scheme 2. The kinetic scheme for receptor GAP binding coupled to GTP hydrolysis

To illustrate this type of allostery we can quantify the modulation effects using the V-type allosteric efficacy, β:

$$\beta_{bind,GAP}^{hydrolysis} = \frac{k_{hydrolysis}^{GAPG}}{k_{hydrolysis}} \tag{1.24}$$

Using transition state theory (TST)[24] and the Eyring-Polanyi equation[25] with the rate constant k expressed as :

$$k = \kappa \frac{k_B T}{h} e^{-\frac{\Delta G^{\dagger}}{RT}} \tag{1.25}$$

where k_B is Boltzmann's constant, h is Planck's constant, κ is the transmission coefficient, and $\Delta G^{\dagger}$ is the activation free energy, the V-type allosteric efficacy β can be written as a function of the change in the energy of the transition state upon GAP binding, (assuming κ is a constant) as:

$$\beta_{bind,GAP}^{hydrolysis} = e^{-\frac{\Delta\Delta G\dagger}{RT}} \tag{1.26}$$

While the MWC and TSAM models have a long history of illustrating their power in the interpretation and analysis of experiments, they both fail to provide a mechanistic understanding of allostery. While these models can be extended to systems with multiple ligand binding sites and/or allosterically regulated sites (for a detailed review of extension of TSAM, see [19]), this clearly provides only a phenomenological explanation of allostery. According to this description, often considered "the thermodynamic" perspective, allostery occurs because of the differences in free energy of the respective states. However, this conclusion appears to be a definition, i.e. that allostery is the phenomena in which that the stability of the *on* state relative to the *off* state is greater when the ligand is bound, and lesser when the ligand is unbound. From a "structural" perspective, one needs to consider the differences in free energy as emerging from some feature of the underlying network of interacting structural components, and it is this feature that makes the system allosteric. This unresolved problem will be the focus in this dissertation.

1.1.2. Allostery in Membrane Proteins

While allostery is thought to be a ubiquitous process[17], it has been frequently claimed to be involved in the mechanisms of membrane proteins. In particular, allostery has been invoked in both membrane transporters and membrane receptors, where there is allosteric coupling observed between the intracellular and extracellular domains of the proteins. The communication of information regarding the external environment to intracellular machinery that can initial the appropriate adaptive response is crucial to cellular survival, and long-distance allostery through transmembrane (TM) domains is an intuitive physical mechanism by which this information can be transmitted. In this section, thermodynamic and kinetic models of transporter and receptor function will be described in the context of what is known about these systems structurally,

thermodynamically, and kinetically, in order to motivate the need for theoretical and computational methods that can determine the physical mechanisms of these allosteric systems.

1.1.2.2. G Protein-Coupled Receptors

G protein-coupled receptors (GPCRs) are 7 TM receptor proteins that act as mediators of information flow between the extracellular space and the intracellular signaling machinery, playing an essential role in cell-cell signaling across many cell and tissue types. Given this crucial positioning in cellular physiology, GPCRs are the targets of a large fraction of the pharmacopeia, and there are numerous mutations across the many GPCR subtypes that are implicated with disease[26].

While all GPCRs have a conserved 7TM topology, when clustered by sequence similarity, they can be broken down into several classes. The classes include the rhodopsin family (Class A), secretin and adhesion family (Class B), the glutamate family (Class C), and the frizzled/TAS2 family. The Class A GPCRs, for which there is the largest and most diverse set of crystal structures, will be the focus of study in this dissertation. At the neuronal synapse, Class A GPCRs play a large role in neurotransmission and signal transduction by responding to the presence neurotransmitters such as dopamine, norephinephrine, and serotonin[27]. Most hallucinogens, painkillers, and anti-psychotics are thought to act by competitively binding to these biogenic amine GPCRs and influencing their function[28,29,30] .

The experimental investigation of these receptor systems points to allosteric mechanisms as the central mode of molecular function for intracellular signal transduction in response to extracellular ligand binding. The fraction of the

concentration of receptors in the active state, f_{P*}, can be written as a function of the ligand concentration [L] and total concentration of receptor [P_0]:

$$f_{P^*}\left([L],[P_0]\right)=\frac{[P^*]+[P^*L]}{[P_0]}=\frac{[P^*]+[P^*L]}{[P^*]+[P^*L]+[P]+[PL]} \tag{1.27}$$

Assuming that activation of downstream signaling is a linear function of the number of active receptors, simplification of (1.27) using the equilibriums described in Scheme 1 results in:

$$\text{activation}\left([L],[P_0]\right)\approx\frac{K_{activate}+\alpha_{activate}^{bind,L}K_{bind,L}K_{activate}[L]}{1+K_{activate}+K_{bind,L}[L]+\alpha_{activate}^{bind,L}K_{bind,L}K_{activate}[L]} \tag{1.28}$$

Equation (1.28) is analogous to (1.9) of the MWC model. An allosteric mechanism in which ligand binding shifts the equilibrium between the receptor's active and inactive states explains the action of agonists, which activate the receptor ($\alpha > 1$), neutral antagonists, which block activation by agonists without activating the receptor ($\alpha = 1$), and most importantly the inverse agonists that inactivate the receptor ($\alpha < 1$). These types of allosteric behavior have traditionally been observed through dose response curves (see **Figure 3**), in which some downstream signaling marker is observed as a function of the ligand. Assuming the linear model in (1.28), these experiments can directly monitor the allosteric efficacy of the ligand without any direct information regarding the structure of the GPCR or the relative probability of its active and inactive states.

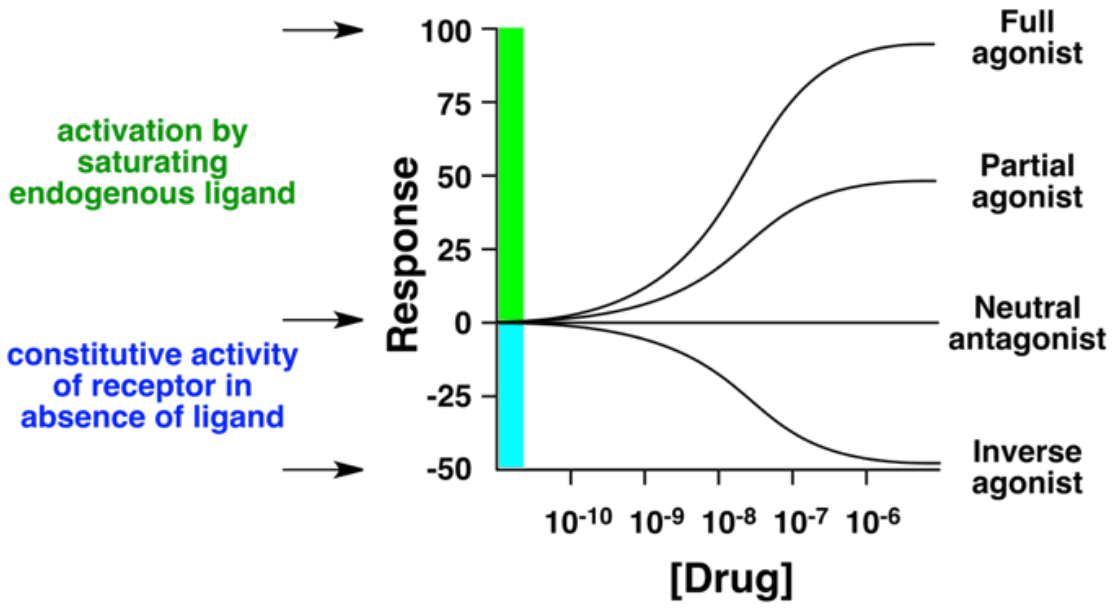


Figure 3. Ideal behavior of different pharmacological classes.

The dose-response curve for an idealized full agonist, partial agonist, neutral antagonist, and inverse agonist. Adapted from Wikipedia[31].

The allosteric mechanism described above predicts the existence of an active and inactive state for each class A GPCR, and that these states should be able to be crystalized through the use of strong agonists and inverse agonists, or even antagonists (given the receptor has low basal activity). While the model predates crystal structures of class A GPCRs, a large number of x-ray structures have been solved since 2007. These structures have largely supported the existence of active and inactive states that are associated with ligand binding, although it is now known that it is unlikely that there are singular active and inactive states associated with any given GPCR (this will be discussion further in Section 1.1.3.1. Ligand-specific Allosteric Modulation in GPCRs).

The first crystal structure of a class A GPCR came from bovine rhodopsin, which was crystallized in what was assigned to be an inactive state[32]. Rhodopsin, unlike other

class A GPCRs, contains a covalently bound ligand, 11-cis retinal, and is activated when the ligand undergoes a light-induced *cis* to *trans* transition. While this mechanism differs from the above-described mechanism in detail, the state of the covalently bound ligand can be seen as equivalent to the binding state (bound versus unbound) of a non-covalently bound ligand. Thus, the structure of rhodopsin bound to 11-cis retinal was assigned as the inactive state (also known as the "dark state"), whereas a hypothetical structural of rhodopsin bound to 11-trans retinal would be expected to be in the active state. Later, retinal-free rhodopsin (known as opsin) was crystalized in a state in which opsin bound to a synthetic Gα carboxy terminus (GαCT) peptide[33], and thus the structure was expected to be active. The structure featured prominent conformational changes, including a 6-7 A tilt of TM6 (see **Figure 4**) and reorganization of both a salt bridge composing the conserved E(D)RY motif and an aromatic stacking interaction in the conserved NPxxY.

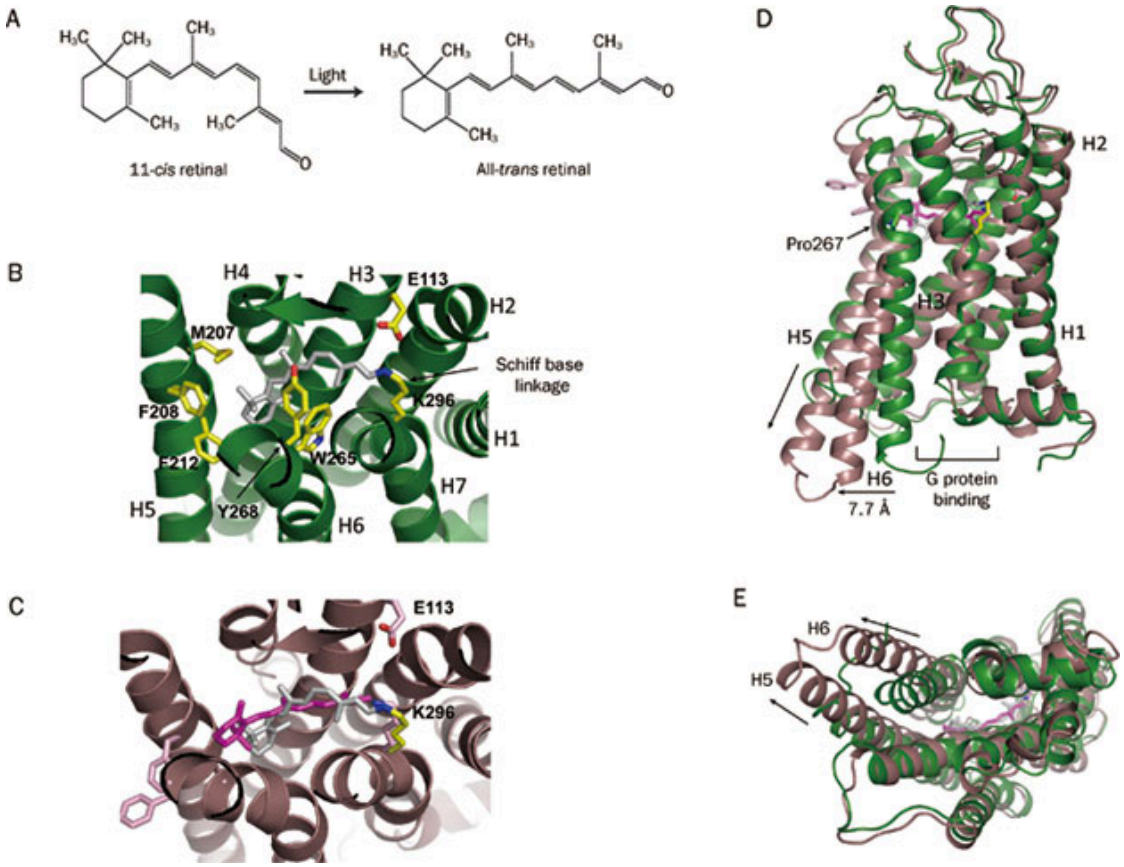


Figure 4. Ligand binding and conformational changes in rhodopsin.

(A) Chemical structures of 11-*cis*- and all-*trans*-retinal. (B) 11-*cis*-retinal (in gray) in the ligand binding pocket (green, PDB: 1F88). (C) Conformational changes in retinal and the binding pocket of rhodopsin upon photoactivation. The photoactivated all-*trans*-retinal (PDB: 3PQR) is magenta and the ground-state 11-*cis*-retinal (PDB: 1F88, gray) is superposed on the activated all-*trans*-retinal for comparison. The activated protein (PDB: 3PQR) is dark brown. (D) The outward tilting of the cytoplasmic end of helix 6 (indicated by the horizontal arrow) and the elongation of the cytoplasmic end of helix 5 (indicated by the vertical arrow). Green shows the inactive conformation (PDB: 1F88), and brown shows the activate conformation (PDB: 3PQR). (E) Bottom view of panel D. Figure and legend reproduced from [34] with permission.

While the opsin state requires a more complicated model than the simple two-state model (as the ligand has effectively three states, unbound, bound/cis, and bound/trans), one would expect that the GαCT-bound receptor would be in the active state, if it is assumed that the active state is described phenomenologically as "active" because it has higher affinity for the down-stream signaling G protein (i.e. the activation is also allosterically coupled to G protein binding via a K-type mechanism). However, light states were eventually crystalized in various states that were considered intermediates to the fully active state, such as bathorhodopsin[35], metarhodopsin I[36], and then the fully activated, deprotonated metarhodopsin II[37]. However, due to the covalent nature of the ligand, the quantum nature of light-induced isomerization of retinal, and the several intermediates, rhodopsin may not be the best model for the mechanism of activation of class A GPCRs by non-covalent agonists.

X-ray structures of the β2 adrenergic receptor and A_2A adenosine receptor in inverse agonist, antagonist, and agonist bound states have made it possible to directly address the structures involved in non-covalent ligand-induced allosteric modulation of function. The first crystal structure of β2AR was solved in complex with the inverse agonist carazolol. While the structure was incredibly similar to that of dark rhodopsin, with a TM RMSD of 1.56 A, there was a notable difference in TM3 and TM6 local to the so-called "ionic lock" composed by $E^{6.30}$ and $R^{3.50}$ of the E(D)RY motif, where the ionic lock mimicked the state seen in light-activated rhodopsin[32]. This difference was seen in the subsequent structures of a β2AR/T4L chimera bound the inverse agonist timolol[38], avian β1AR bound to the antagonist cyanopindolol[39], and human A_2A bound to the antagonist ZM241385[40]. However, the active state structure of β2AR, bound to the high affinity BI-167107 agonist and Nb80[41], a nanobody that acts as a G protein mimetic, revealed similar conformations changes as seen in the active state

opsin and rhodopsin structures, with large outward displacement of TM6. Outward displacements of TM6 are consonant with the predictions of early biophysical experiments[42,43], which suggested a TM6 conformational change being characteristic of ligand-induced activation.

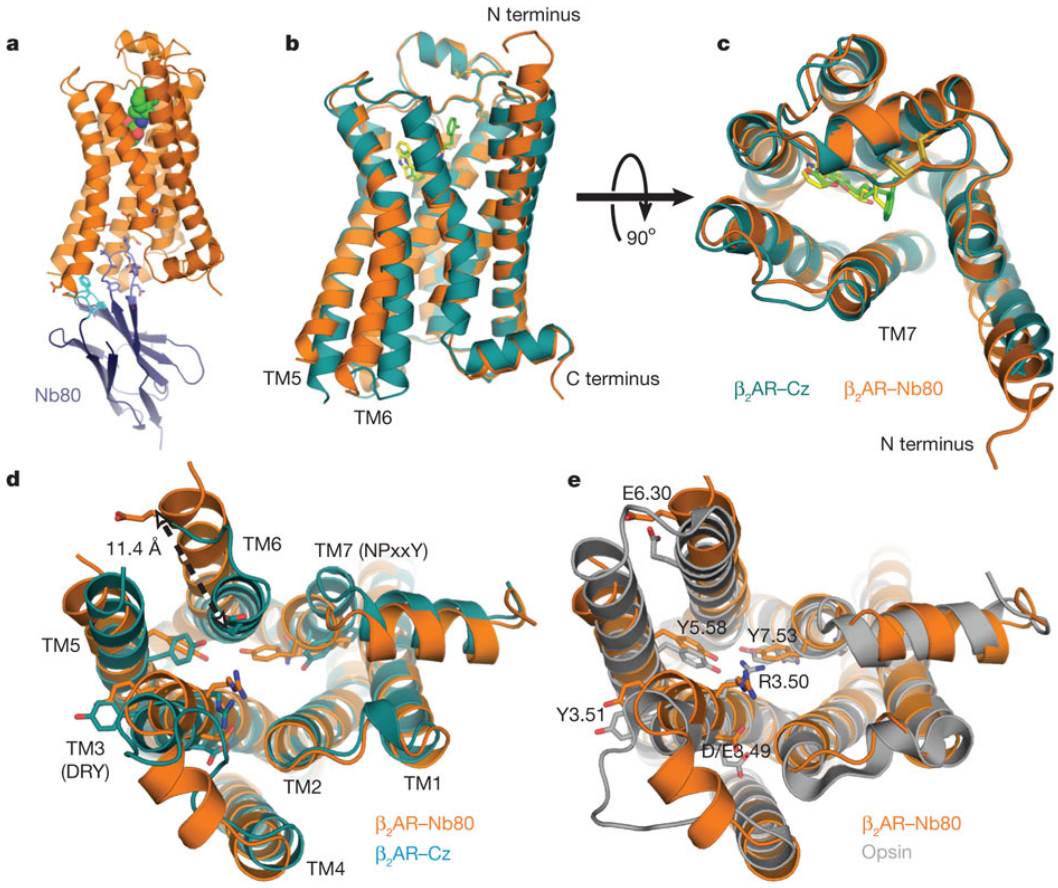


Figure 5. The agonist-Nb80 stabilized crystal structures of β_2AR.

The structure of β_2AR-T4L bound to the inverse agonist carazolol (β_2AR-Cz) is shown with β_2AR-T4Lin blue and carazolol in yellow. The structure of BI-167107 agonist bound and Nb80 stabilized β_2AR-T4L (β_2AR-Nb80) is shown with β_2AR-T4L in orange and BI-167107 in green. **(a)** The β_2AR-Nb80 complex with β_2AR in orange and CDRs of Nb80 in light blue (CDR1) and blue (CDR3). (**b**) Superposition of β_2AR-Cz and β_2AR-Nb80. **(c)** Extracellular view of the superposition of β_2AR-Cz and β_2AR-Nb80. Extracellular view of the superposition of β_2AR-Cz and β_2AR-Nb80. **(d)** Intracellular view of the superposition of β_2AR-Cz and β_2AR-Nb80. **(e)** Superposition of β_2AR-Nb80 with the structure of opsin crystallized with the C-terminal peptide of G_t (transducin). Adapted from [41] with permission.

However, a later structure of β_2AR in complex with a heterotrimeric G_S complex[44] indicated that the conformational change required to accommodate the G protein was significantly larger than that expected from the structures of agonist-bound receptor (see **Figure 6**).

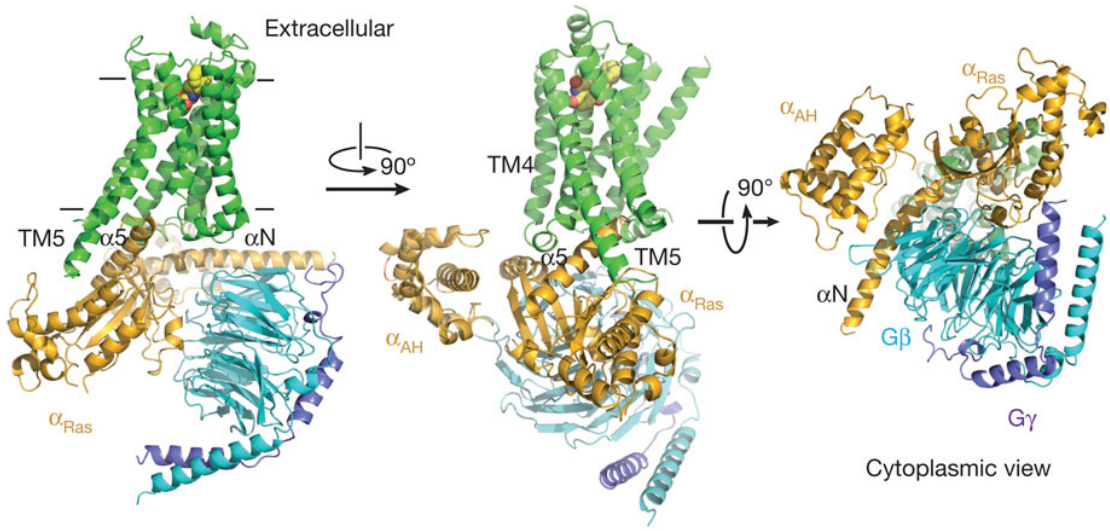
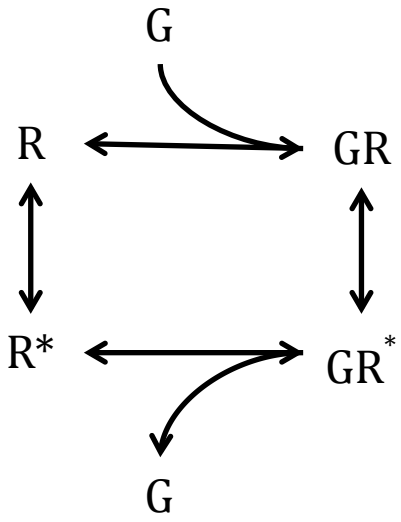


Figure 6. The structure of β2AR bound to a heterotrimeric G_s protein complex. The overall structure of the β_2AR (green) bound to an agonist (yellow spheres) and the heterotrimeric Gs composed of Gαs (orange), Gβ (cyan) and Gγ (purple). Reproduced with permission from [44].

It should be noted that while GPCR allostery is generally discussed in the context of the ligand's ability to stabilize the activate state of the GPCR, the function of the GPCR comes from its own ability to activate a G protein. One mechanism of activation of a G protein by a GPCR that utilizes allostery is a K-type mechanism in which either state of the GPCR can activate the G protein, but the active state of the

GPCR has higher affinity for the G protein, and thus the coupling of G protein binding to receptor activation must be considered:

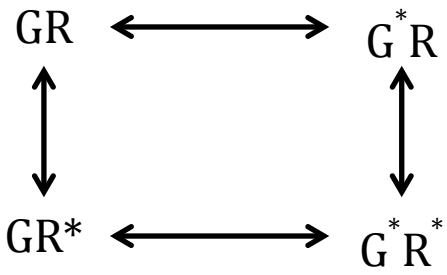


Scheme 3. The thermodynamic cycle for the coupling of G protein binding to receptor activation.

And thus there is a K-type coupling between the activation state of the receptor and the binding of the G protein:

$$\alpha_{\text{activate}}^{\text{bind,G}} = \frac{K_{\text{activate}}^{\text{G}}}{K_{\text{activate}}} = \frac{K_{\text{D,G}}^{\text{inactive}}}{K_{\text{D,G}}^{\text{active}}} \tag{1.29}$$

However, there is additional evidence that G proteins may be pre-coupled to their receptors. If this is the case, the inactive state of the G protein may have high affinity for the inactive form of the receptor, and the active state of the G protein may have high affinity for the active form of the receptor. Thus, the allostery could involve the coupling of the active states (see Scheme 4).

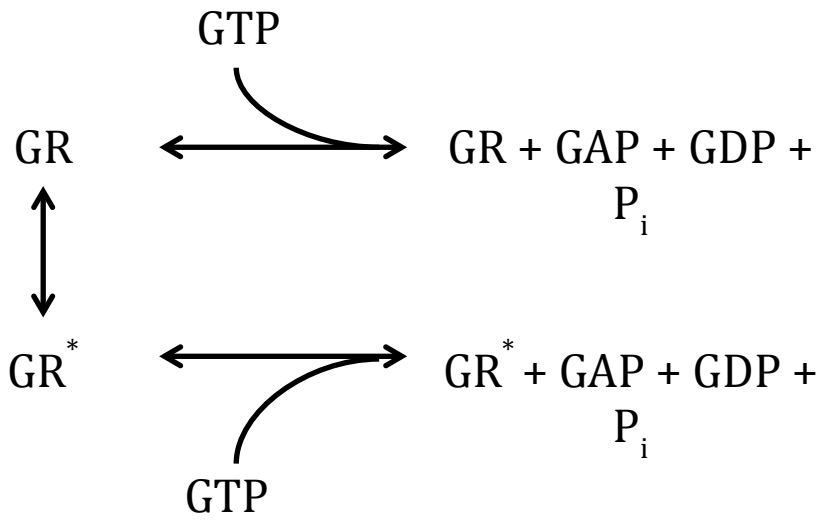


Scheme 4. The thermodynamic cycle for receptor activation coupled to G protein activation.

Thus,

$$\alpha_{activate,R}^{activate,G} = \frac{K_{activate,R}^{G_{active}}}{K_{activate,R}^{G_{inactive}}} = \frac{K_{active,G}^{inactive}}{K_{active,G}^{active}} \tag{1.30}$$

Lastly, as described in Section 1.1.1.2. Theoretical Background, the GPCR may act as a GAP and increase the rate of GTP hydrolysis by the G protein.



Scheme 5. The kinetic scheme for receptor GAP binding coupled to GTP hydrolysis

Thus, the complex process of inducing intracellular G protein signaling through activation of a GPCR by an extracellular ligand has the potential to involve several

different allosteric mechanisms that likely will involve very different underlying physical mechanisms.

1.1.2.2. Membrane Transporters

The transport of solutes across cell membranes is an essential process in the life of each cell. Membrane transport maintains homeostasis and connects the cell to its environment by establishing and keeping ion gradients[45], and absorbing essential substrates such as sugar[46,47] and amino acids[48]. In multicellular organisms most physiological processes utilize solute transport to enable the specific function of tissues and organs, from concentrating the urine in the kidney[49], to reuptake of released neurotransmitter that enables neurotransmission in the brain[50]. Not surprisingly, transport malfunction has been implicated in many disease states[51], and the molecular machines involved in transport are the targets of both medications and various drugs of abuse[52].

Three main mechanisms of transport across the cell membrane have been identified, and each can be described in the framework of thermodynamics and chemical kinetics. The simplest mechanism is *passive transport*[45], in which solutes diffuse across the membrane without the assistance of other molecules. The concentration gradient and the membrane potential determine the net direction of the diffusion. With the equation for a transport of a solute S from the extracellular space to the intracellular space written as:

$$S_{out} \rightarrow S_{in} \tag{1.31}$$

the free energy change, ΔG, associated with this transport process is simply

$$\Delta G_{passive} = -RT\log\left(\frac{[S_{out}]}{[S_{in}]}\right) - zFE_m \tag{1.32}$$

where $[S_{out}]$ and $[S_{in}]$ are the solute concentrations on the outside and inside of the cell, R is the gas constant, T is the temperature, z is the charge of the solute, F is the Faraday constant, and E_m is the membrane potential. The solute equilibrium can be written as

$$S_{out} \underset{k_{-p}}{\overset{k_p}{\rightleftarrows}} S_{in} \tag{1.33}$$

where k_p and k_{-p} are the rate constants for the forward and backward passive transport, respectively. These rate constants implicitly include the effect of the membrane potential. At equilibrium,

$$\Delta G_{passive} = -RT\log\left(\frac{k_p}{k_{-p}}\right) = 0 \tag{1.34}$$

and assuming (for simplicity) the membrane potential is 0, the new diffusion rates in either direction become equal, and the equilibrium substrate concentrations are such that

$$[S_{out}]_{eq} = [S_{in}]_{eq} \tag{1.35}$$

While small, uncharged solutes can equilibrate via passive diffusion at a reasonable rate, the membrane is not permeable to larger, charged molecules. These move across the membrane through *facilitated diffusion*[45]. Facilitated diffusion is mediated by a class of transport proteins known as uniporters, which can be either channels or carrier proteins. Channels act as regulated pores that open in response to a stimulus and allow the free flow of specific solutes. Carrier proteins bind one molecule at a time and

transport it across the membrane with the solute concentration gradient. In both cases, uniporters act by increasing the effective membrane permeability for the solute and thus the rate of equilibration of concentrations on either side of the membrane. Denoting the transporter protein as T, the facilitated diffusion equilibrium is

$$S_{out} + T \underset{k_{-f}}{\overset{k_f}{\rightleftharpoons}} S_{in} + T \tag{1.36}$$

where k_f and k_{-f} are the rates of forward and reverse facilitated transport, respectively. Given the fact that the free energy of facilitated transport is the same as $\Delta G_{passive}$, which depends only on the intra- and extracellular substrate concentrations and the membrane potential,

$$\frac{k_p}{k_{-p}} = \frac{k_f}{k_{-f}} \tag{1.37}$$

However, the free energy barrier for transport is lowered, such that

$$k_f > k_p, k_{-f} > k_{-p} \tag{1.38}$$

Both passive transport and facilitated transport are determined by the direction of the gradient or the electric field. However, much of the transport required for cell physiological processes do not occur in the direction of the solute's concentration gradient. When transport of a solute alone is not thermodynamically spontaneous, a third mechanism of transport, *active transport*[45], is required. To achieve transport against a concentration gradient, the thermodynamically unfavorable transport of the solute is coupled to an energy source. The nature of the energy source classifies active transport into "primary" and "secondary".

In *primary active transport*[45], the thermodynamically unfavorable transport process is coupled to a chemical reaction that proceeds in a thermodynamically favorable direction. This can be written as:

$$S_{out} + R + T \underset{k_{-a}}{\overset{k_a}{\rightleftarrows}} S_{in} + P + T \tag{1.39}$$

where R and P are the reactant and product of the chemical reaction, respectively, and k_a and k_{-a} are the rate constants for forward and reverse transport. Here, the relation to passive transport rates is

$$\frac{k_a}{k_{-a}} = \Delta G_{active} < \Delta G_{passive} \tag{1.40}$$

Although $\Delta G_{passive} > 0$, the energy released by the chemical reaction allows reversal of the equilibrium, such that $\Delta G_{active} < 0$ when the system is out of equilibrium and inward transport becomes effective. A common energy source for primary active transport is ATP hydrolysis[45], which is used by the family of transporters known as transmembrane ATPases[53]. For this family, (1.39) becomes

$$S_{out} + ATP + T \underset{k_{-a}}{\overset{k_a}{\rightleftarrows}} S_{in} + ADP + P_i + T \tag{1.41}$$

The manner in which the energy from ATP hydrolysis is used to enable solute transport against its gradient by the primary active transporter is a key consideration in understanding the molecular mechanism of these membrane proteins. While the binding of ATP to a primary active transporter has been suspected for a long time to be separate from that of a solute[54], it is now made clear from the known molecular structures of such transporters that the binding sites are well separated spatially[55–57]. By virtue of this spatial relationship, it is reasonable to consider the coupling between ATP hydrolysis and solute transport to involve an allosteric mechanism.

Indeed, potential allosteric mechanisms for the coupling of ATP hydrolysis to solute transport against its gradient have been proposed since as early as the 1960s. In 1966, Jardetzky proposed a "simple allosteric model" for phosphorylation-driven transport of an ion[58], which corresponds to the following formulation:

$$S_{out} + ATP + T_{out} \underset{k_{-s}}{\overset{k_s}{\rightleftarrows}} ATP + ST_{out} \xrightarrow{k_{release}} S_{in} + ADP + pT_{in} \qquad (1.42)$$

In this mechanism, the solute binding to the transporter occurs from the outside (i.e., the extracellular environment) when the transporter adopts an "outward-facing" conformation, denoted T_{out}. The substrate binding and unbinding rate constants are denoted k_s and k_{-s}, respectively. In a second step ATP-dependent phosphorylation of the transporter drives it into an inward-facing conformation so that the phosphorylated form is pT_{in}, which has low affinity for the solute and releases it into the cell. Because of the large amount of energy released by ATP hydrolysis, the second step is assumed to be irreversible and the transport reaction leading to intracellular substrate release is described by the forward rate constant $k_{release}$. This mechanism has often been referred to as a "rocker switch" for the transition between an outward-facing and an inward-facing conformation of the transporter[59], and is generally cited as the origin of what is now known as the "alternating access" model or mechanism (the authors note that while we were unable to locate the first instance of the use of the name "alternating access", it has been in use since as early as 1977 in [60]).

The alternating access model describes transport as the following process:

$$S_{out} + T_{OF} \rightleftarrows ST_{OF} \rightleftarrows ST_{OCC} \rightleftarrows ST_{IF} \rightleftarrows S_{in} + T_{IF} \qquad (1.43)$$

In this mechanism, the transporter has three states: outward-facing (T_{OF}), occluded (T_{OCC}), and inward-facing (T_{IF}). While Jardetzky's simple representation of the

allosteric model (see (1.42)) for the ATP dependent transporters did not include an occluded state as in (1.43), it still fits within the alternating access framework, and there is a substantial body of work clarifying how ATP hydrolysis is allosterically coupled to solute transport in an alternating access manner. Indeed, for some systems such as the sodium/potassium exchanger Na^+/K^+ ATPase, the mechanism of Na^+ import is strikingly similar to the original model[61], but for the important class of transporters identified as the ATP Binding Cassette (ABC) transporters, it appears that the mechanism can be fundamentally different within subfamilies, as illustrated below.

The ABC transporter superfamily is the largest transporter gene family and is responsible for the transport of many different types of substrates, both as exporters and importers[62]. Based on an abundance of structural, biochemical, and biophysical data, an "ATP switch" model has been proposed for ABC transporters in which the translocation step is driven by ATP binding and not hydrolysis[63,64]. According to this mechanism (presented for ABC exporters), substrate binding to the inward-facing transporters causes the transition to an occluded state of the transporter, and increases the transporter's affinity for ATP. ATP binding then stabilizes the outward-facing conformation, which has low affinity for substrate, and substrate is released:

$$T_{IF} + S_{in} + ATP \rightleftarrows T_{OCC}S + ATP \rightleftarrows T_{OF}ATP + S_{out} \tag{1.44}$$

Finally, the hydrolysis of ATP is then proposed to drive the resetting of the empty transporter to the inward-facing state in preparation for the next cycle of substrate transport:

$$T_{OF}ATP \rightarrow T_{OCC}ADP + P_i \rightleftarrows T_{IF} + ADP + P_i \tag{1.45}$$

In this component of the mechanism, ATP hydrolysis with the transporter in the outward-facing state, leads to occlusion of the binding site from the extracellular environment. Dissociation of the low affinity ADP then produces an apo state of the transporter, which favors the inward-facing conformation. This combined mechanism enables primary active transport of the solute against its gradient, but appears considerably more complicated than Jardetzky's simple allosteric model in (1.42). Despite the apparent increased complexity, this mechanism has significant empirical support (for reviews, see [63,64] and references therein).

In contrast to the primary active transporters, in which the overall thermodynamically unfavorable substrate transport process is coupled to a favorable chemical reaction, the secondary active transporters couple the unfavorable transport process to the favorable transport process of one or several ions, (denoted as I). The ions are either transported in the same direction of the substrate (symport) or in the opposite direction (antiport). In inward symport, the equilibrium becomes:

$$S_{out} + I_{out} + T \underset{k_{-s}}{\overset{k_s}{\rightleftharpoons}} S_{in} + I_{in} + T \tag{1.46}$$

The net free energy change of the coupled process can be written as $\Delta G_{symport} = \Delta G_S + \Delta G_I$, with the transport free energies for solute and ions defined as in (2). This yields:

$$\Delta G_{symport} = -RT\log\left(\frac{[S_{out}][I_{out}]}{[S_{in}][I_{in}]}\right) - (z_S + z_I)FE_M \tag{1.47}$$

Here, z_S and z_I indicate the charges of the solute and ion, respectively. When $DG_{symport}$ is negative, unfavorable substrate import can be driven against the concentration gradient. In particular, sodium uptake is used in many of the known secondary active

transports, with the large difference in extracellular and intracellular sodium concentrations in most tissues and environments providing a significant source of electrochemical energy.

The sodium-coupled symporters are a large family of great interest because the solutes transported span a vast array of chemical compositions and they are essential components of many physiological functions. For example, sodium-coupled symporters play an essential role in neurotransmission, where they mediate the reuptake of neurotransmitter into the presynaptic and glial cells, and thus enable the transduction of information. Consequently, these transporters are also efficient targets for psychoactive therapeutics. The reasons for the attention accorded here to sodium-coupled symporters go beyond the importance due to their sheer abundance in biological systems, and their diversity. It includes as well the central role of allostery in the mechanisms of transport that have been proposed for their various subfamilies, and their diversity.

While the sodium-coupled symporters are also believed to use an alternating access mechanism as shown in (1.43), the molecular and thermodynamics details of the mechanism are not obvious. For sodium-coupled symporters, neither of the simple allosteric model or the model described above for the ABC transporters are supported by experimental or computational evidence.

1.1.2.1.1 LeuT- a prototype for secondary transporters

The small amino acid transporter LeuT, originally identified as a bacterial leucine transporter, has proven to be an extremely useful tool in understanding the allosteric mechanisms involved in secondary transport. LeuT was originally identified as a 12-transmembrane segment (TM) homologue of the Na^+/Cl^- -dependent transporters[65],

but with the rapid growth in structural and functional information about other members of the SLC6 gene family, the current view is that the pair of pseudo-symmetric 5 TM bundles in LeuT (TMs 1-5/6-10) represent a more general protein fold motif termed the "LeuT-fold", shared by many membrane proteins performing a variety of transport functions[66]. For these reasons, the structure and dynamics of LeuT and its functional mechanisms became the focus of intense investigations and it continues to serve as a prototype for the large class of mammalian monoamine transporters (MATs) that are Na^+/Cl^--dependent neurotransmitter symporters (NSS) and carry out the symport of Na^+ and a biogenic amine, together with Cl^- antiport (see below).

The first step towards understanding the transport functions of LeuT-fold proteins from the context of a 3D molecular structure was the determination of a high resolution x-ray crystal structure of LeuT bound to two sodium ions and to leucine [65], which is a slowly transported as a substrate (v_{max} = 334 pmol/min/mg). The 1.7 Å resolution leucine-bound structure (PDB 2A65) revealed the topology of the twelve TM domains, ten of which form the two TM bundles arranged in the pseudo-symmetry characteristic to the LeuT-fold family (see Fig. 1A). The substrate binding site, termed S1, was found at the midpoint of the two pseudo-symmetric domains, local to TM3, TM8, and the unstructured regions at the centers of TM1 and TM6 helices. Adjacent to the substrate are two distinct sodium sites, termed Na1 and Na2. The sodium in the Na1 site is directly coordinated by the leucine's carbonyl oxygen (see Fig. 1B), as well as by residues A22 and N27 of TM1, T254 of TM6, and N286 of TM7. In the Na2 site, the Na^+ ion does not interact directly with the substrate, but is coordinated by residues G20 and V23 of TM1, and A351, T354, and S355 of TM8. Na+ titration experiments showed that no substrate will bind in the absence of Na+,

and subsequent studies indicated that the binding of Na+ and leucine was cooperative[67], as might be expected from the interaction of substrate and the Na1 sodium identifiable in the crystal structure.

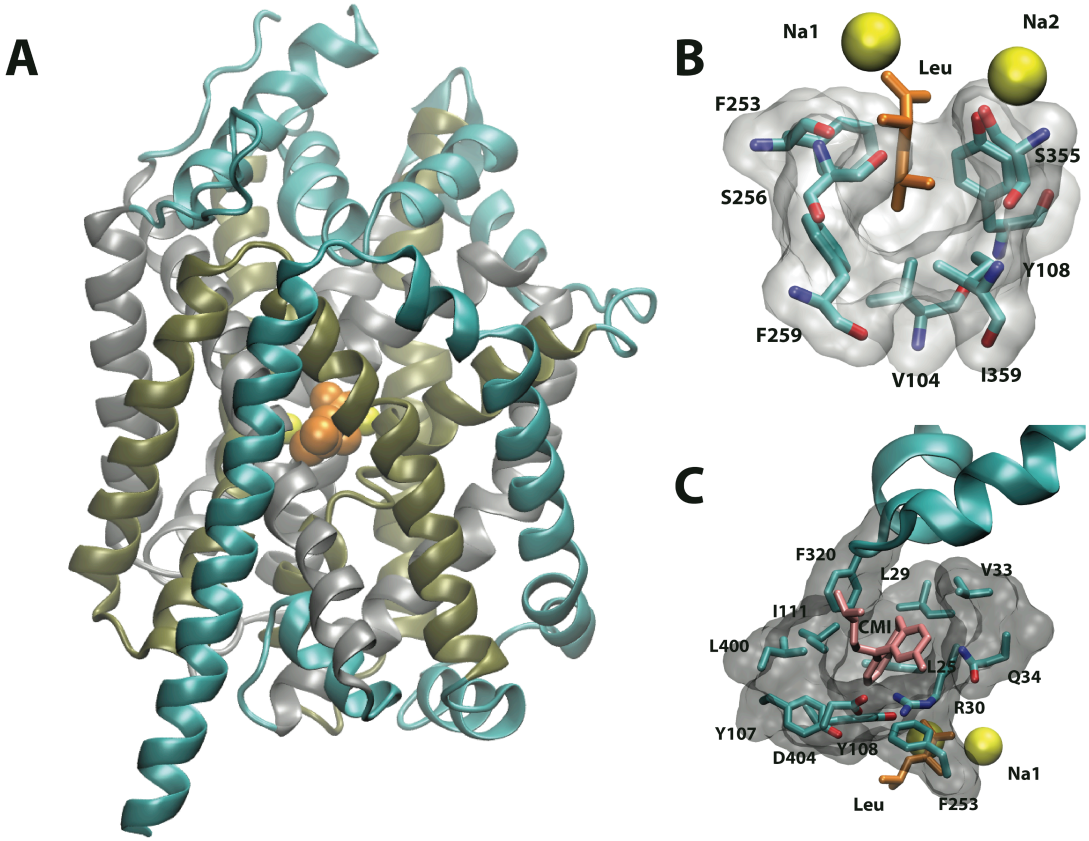
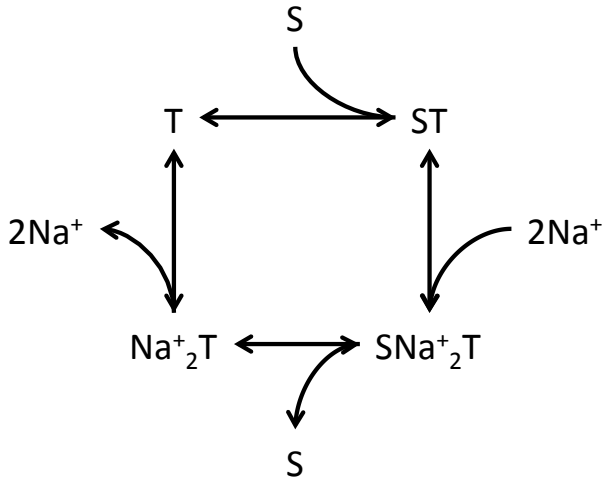


Figure 7. LeuT fold and binding sites.

(A) Crystal structure of LeuT bound to two Na^+ and a leucine (2A65). The pseudosymmetric TM repeats are represented in silver ribbons (TMs 1–5) and gold illustration (TMs 6–10). TMs 11 and 12 and loops are shown in cyan ribbons. Leucine and Na^+ are represented as orange and yellow van der Waals spheres, respectively. (B) Primary substrate binding site S1. Leucine is shown in orange, and the binding residues are colored by element (cyan for carbon, blue for nitrogen, and red for oxygen). Na^+ ions are shown as yellow van der Waals spheres and labeled according to their binding site. (C) Clomipramine binding site in S2. Clomipramine is shown in pink, EL4 is shown as cyan ribbon, and the remaining residues and substrates are colored as in B, except that the backbone is omitted for clarity.

The detailed interactions between the Na^+ ions and the substrate can be considered as a basic form of allostery, in which two structural components influence each other's binding equilibria (hence K-type allostery) through direct interaction, as described by (1.15) and (1.16) in Section 1.1.1.2. Theoretical Background. But in MD simulations, however, the Na+ binding led to opening of the extracellular vestibule that would allow the substrate to bind[68], thus pointing to a more intricate network of allosteric interactions. Indeed, later studies with electron paramagnetic resonance (EPR) confirmed Na^+ -induced outward opening[69,70], suggesting that Na^+ binding may allosterically modulate the extracellular vestibule and the S1 site to increase substrate binding through a mechanism other than just direct interactions.

Together, these results of the functional and structural analysis of LeuT functions can be considered to *represent the first steps in an allosteric transport mechanism in which the transporter can only bind its substrate when the energy source to which it is coupled is already bound*. This allosteric coupling is represented by the following thermodynamic cycle:



Scheme 6. The thermodynamic cycle for binding of Na^+ and substrate.

The typical substrate binding equilibrium can be written as:

$$\mathrm{Leu} + \mathrm{T} \underset{k_{unbind,Leu}}{\overset{k_{bind,Leu}}{\rightleftharpoons}} \mathrm{LeuT} \tag{1.48}$$

and the dissociation constant for leucine binding to the apo transporter will be denoted as:

$$K_{D,Leu} = \frac{k_{unbind,Leu}}{k_{bind,Leu}} \tag{1.49}$$

In the presence of Na^+, one can write the modified equilibrium as:

$$\mathrm{Leu} + \mathrm{Na_2^+T} \underset{k_{unbind,Leu}^{Na_2^+}}{\overset{k_{bind,Leu}^{Na_2^+}}{\rightleftharpoons}} \mathrm{LeuNa_2^+T} \tag{1.50}$$

with the new dissociation constant

$$K_{D,Leu}^{Na_2^+} = \frac{k_{unbind,Leu}^{Na_2^+}}{k_{bind,Leu}^{Na_2^+}} \tag{1.51}$$

so that the allosteric coupling between leucine and Na^+ binding can be quantified as a function of the dissociation or association constants:

$$\alpha_{bind,Leu}^{bind,2Na^+} = \frac{K_{D,Leu}}{K_{D,Leu}^{Na_2^+}} \tag{1.52}$$

Thus, $\alpha > 1$ for cooperative binding. Assuming that the concentration of Na^+ is saturating, and thus LeuT is in the sodium-bound state before binding substrate, the transport process can be represented as:

$$\mathrm{Leu_{out}} + \mathrm{Na_2^+T} \underset{k_{unbind,Leu}^{Na_2^+}}{\overset{k_{bind,Leu}^{Na_2^+}}{\rightleftharpoons}} \mathrm{LeuNa_2^+T} \xrightarrow{k_{cat}} \mathrm{Leu_{in}} + 2\mathrm{Na_{in}^+} + \mathrm{T} \tag{1.53}$$

For this process, we can write a MM equation that accounts for the K-type allosteric modulation of transport due to the coupling of leucine and sodium binding:

$$\frac{d[Leu_{in}]}{dt} = \frac{k_{cat}[T_0][Leu_{out}]}{\left(\alpha_{bind,Leu}^{bind,2Na^+}\right)^{-1} K_{D,Leu} + [Leu_{out}]} \tag{1.54}$$

Equation (1.54) demonstrates the first allosteric component in LeuT's secondary active transport mechanism, with the following implications: *The stronger the allosteric coupling between the binding of leucine and of sodium, the higher the transport rate will be, because more transporters will be in the fully bound state and prepared for release.* Notably, transporters could still perform symport if these two binding events were independent, as long as both binding events are thermodynamically favorable. However, *their allosteric coupling makes transport significantly faster when release is the rate-limiting step.*

Moreover, by favoring the fully bound state, the allosteric coupling makes the substrate-only bound state less populated, and thus less Na^+-independent substrate export would be expected than when compared to binding that is not allosterically coupled.

While the leucine-bound structure of LeuT described above enabled the appreciation of an allosteric coupling can be achieved between sodium binding and substrate binding, structures alone cannot show how domains of the transporter are involved in outward and inward opening, nor suggest the presence of any intrinsic allosteric coupling between these domains. These elements of allostery in the function of the transporter emerge from the analysis in the context of previous studies of residues believed to be involved in gating in mammalian homologous[71], from steered molecular dynamics simulations that simulated the translocation of leucine[68],

additional structures of the sodium-bound outward-open state and a mutant apo inward-open state[72]. All these studies helped reveal the domains involved in gating, their potential open and closed conformations, and the networks of interactions that stabilized these conformations. Despite the pseudo-symmetry of the TM domains, it became clear from these simulations and the various structures that significant differences between the intracellular and extracellular domains are likely to be important for the functional mechanism. This is illustrated by the specific details of the extracellular gate (EG, see **Figure 8**) and the intracellular gate (IG, see **Figure 9**).

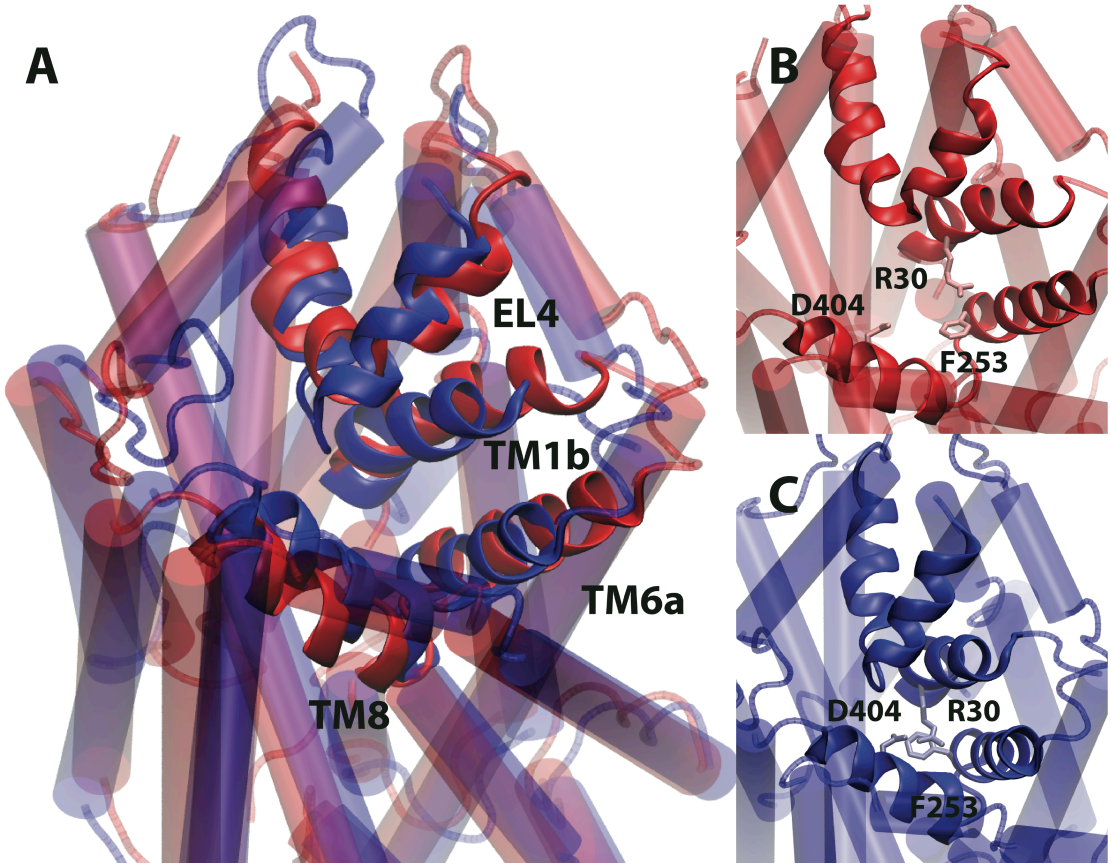


Figure 8. The extracellular gate of LeuT in the open and closed states.

A. The structure of the extracellular gate in the outward-open (red) and outward-closed (blue) conformations. Domains involved in the conformational change (EL4, TM1b, TM6a, and TM8) are shown as ribbons, while the rest of the protein is shown as transparent cylinders. B and C. A closer view of the open and closed extracellular gates, respectively. Residues involved in stabilizing the closed conformations are shown in pink and light blue.

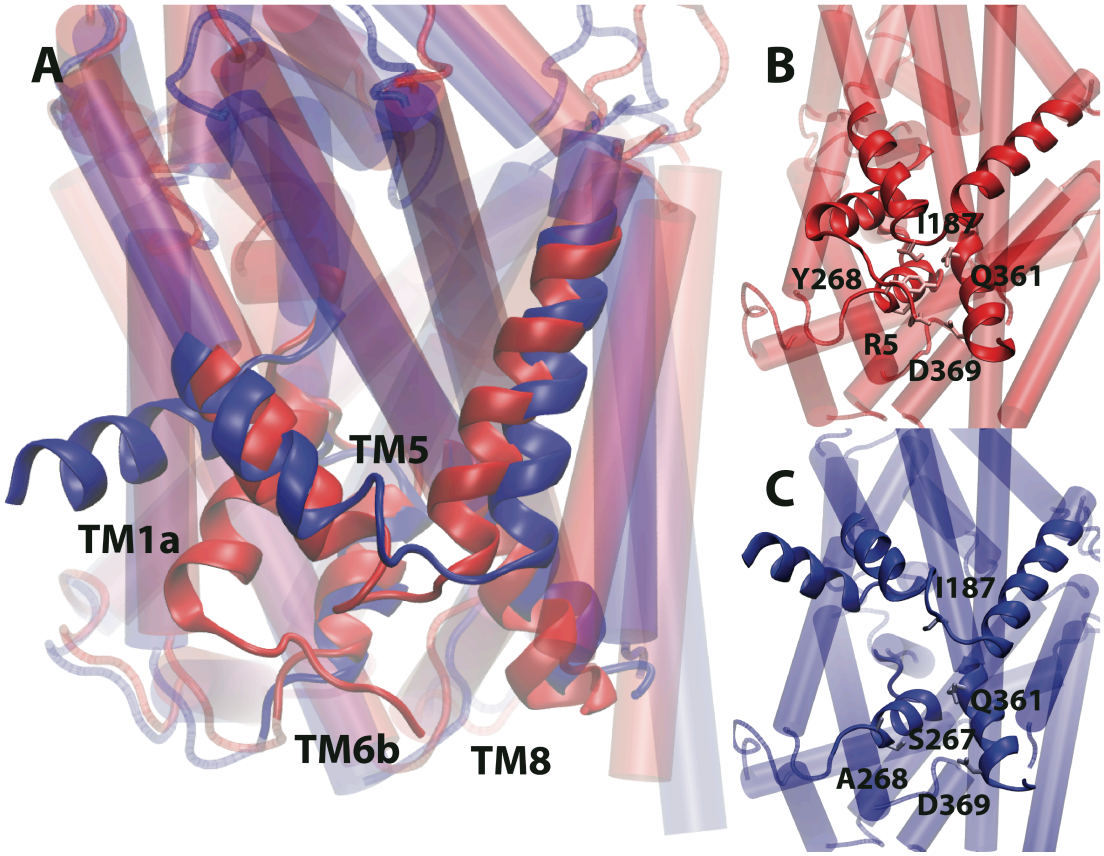


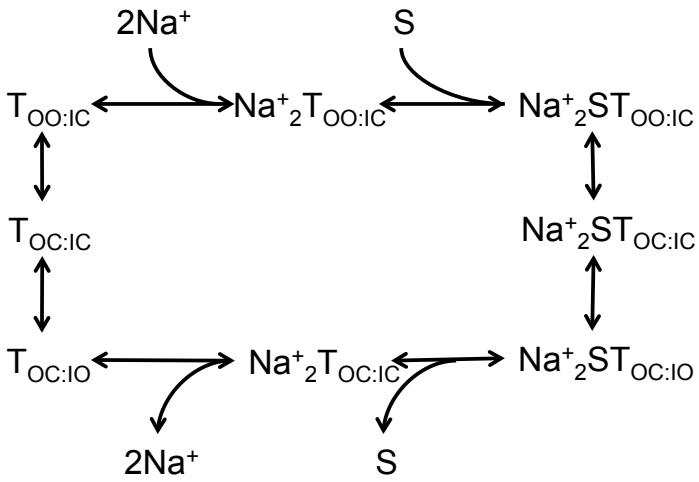
Figure 9. The intracellular gate of LeuT in the open and closed states.

A. The structure of the intracellular gate in the inward-closed (red) and inward-open (blue) conformations. Domains involved in the conformational change (TM5, TM8, TM6b, TM1a) are shown as ribbons, while the rest of the protein is shown as transparent cylinders. B and C. A closer view of the open and closed intracellular gates, respectively. Residues involved in stabilizing the closed conformations are shown in pink and light blue.

The EG is formed by F253, which occludes the substrate from extracellular water, a salt bridge between R30 and D404, and it is capped by the extracellular loop 4 v-helix (see **Figure 8**). Outward opening corresponds to the disruption of the R30/D404 salt bridge, isomerization of F253, a reorientation of the EL4, and the outward motion of TM1b, TM6a, and TM8. Conversely, the intracellular gate (IG) is composed of a more extensive interaction network of residues from TM1a (R5), TM6b (S267 and Y268), TM8 (D369), and TM2 (I187), and these interactions are all disrupted in the inward-open mutant, leading to a large displacement of TM1a, smaller displacements of TM6b and TM8, and an unwinding of the TM5 kink (see **Figure 9**). These crystal structures, and the experimentally determined existence of intracellular and extracellular gates with open and closed states, supported the proposal of a *gated pore alternating access mechanism* for transport, in which (i)-LeuT binds an extracellular substrate while in an outward open/inward closed state (OO:IC), (ii)-transitions to a doubly-occluded state (OC:IC), and then into (iii)-an outward closed/inward open state (OC:IC), from which the substrate can be released into the intracellular space. Forward transport is often written with only three states, as:

$$2Na^{+}_{out} + Leu_{out} + T_{OO:IC} \rightleftarrows LeuNa^{+}_{2}T_{OC:IC} \rightarrow 2Na^{+}_{in} + Leu_{in} + T_{OC:IO} \qquad (1.55)$$

and the mechanism is represented in the following thermodynamic cycle:



Scheme 7. The thermodynamic cycle for substrate binding and gating.

However, the complete mechanism actually implies a number of additional states, including a doubly open state and all possible combinations of gate states bound to substrate.

Notably, the simplified three-state model implied in (1.55) would be approximately accurate in the regime of high K-type allostery between the EG and IG (e.g. the limit as α goes to infinity). While an allosteric coupling between the gates appears reasonable given the alternating access model and the available crystal structures, this inference cannot be drawn from the crystal structures, each of which represents only one conformation in an ensemble of microstates. Because these structures do not provide evidence for an allosteric coupling between the gates, nor can they suggest the role that such an allosteric coupling would play in the molecular mechanism of transport, as is necessary to evaluate the relative free energies of all states in Scheme 7, it is impossible to determine the allosteric efficacy between gate conformational changes. *Thus, in order to understand the properties and mechanism of LeuT as a transporter, the determination of structures representing several conformational states along the transport cycle needs to be complemented by the characterization of*

the many kinetic and thermodynamic parameters that are required to form a kinetic model of transport composed of these states.

Soon after the initial occluded structure, several additional structures of LeuT were solved in which the transporter was crystallized in complex with bound antidepressants [73,74] that are known to act as inhibitors of mammalian NSS transporters, and also act as inhibitors of LeuT. In the structures of these complexes with LeuT, the antidepressants were bound in an extracellular vestibule containing the extracellular gate, and were positioned above the substrate binding site (S1) observed in the original leucine-bound crystal structure. This binding site is now referred to as the secondary substrate site S2, and its properties and proposed functions will be detailed further below.

Because the inhibitors bound in S2 occludes access to S1, which contained a bound leucine, (see **Figure 7**B), it was inferred that inhibitor binding has to occur after ligand binding, and it was further suggested that the inhibitors impeded substrate transport in a non-competitive manner by locking the extracellular gate in a state that is incompatible with inward-opening or substrate release. In terms of the MM representation typically used to analyze transport, this implies that in order to achieve inhibition by this mechanism, inward opening must be the rate-limiting step. In fact, kinetic analysis of alanine transport under saturating sodium and inhibitor concentrations found[73] that the inhibitors decreased the v_{max}, (from 1890 ± 90 without inhibitor to 770 ± 40 pmol/min/mg) while K_D was unchanged (450 ± 70 versus 480 ± 80 nM). This indicates a V-type allosteric mechanism for antidepressant inhibition. Indeed, binding experiments found that inhibitors caused the transporter to retain the bound alanine substrate, indicating that they greatly reduced the rate constant for intracellular release.

With the assumption that substrate binds first, followed by the binding of inhibitor is supported by the crystal structure, so that the mechanism of transport in the presence of an inhibitor, Inh, can be written as:

$$\mathrm{Inh_{out} + LeuNa_2^+T \underset{k_{unbind,Inh}^{Na_2^+}}{\overset{k_{bind,Inh}^{Na_2^+}}{\rightleftharpoons}} InhLeuNa_2^+T \xrightarrow{k_{cat}^{Inh}} Leu_{in} + 2Na_{in}^+ + InhT} \quad (1.56)$$

The V-type allosteric efficacy is

$$\beta_{cat}^{bind,Inh} = \frac{k_{cat}^{Inh}}{k_{cat}} \quad (1.57)$$

so that the rate of transport under saturating sodium and inhibitor concentrations becomes

$$\frac{d[Leu_{in}]}{dt} = \frac{\beta_{cat}^{bind,Inh} k_{cat}[T_0][Leu_{out}]}{\left(\alpha_{bind,Leu}^{bind,2Na^+}\right)^{-1} K_{D,Leu} + [Leu_{out}]} \quad (1.58)$$

This analysis provided significant, quantitative evidence for allosteric modulation of transport that is achieved by modulating the conformation of the extracellular gate. However, at the time it was carried out[73] there was no structure of an inward open state, and the structures of substrate-bound transporters were all inward closed. Consequently, there was no structural evidence for allosteric modulation of inward opening, and such an inference could be made from the existing inhibitor-bound structures only by assuming a model of the transporter as an allosteric gated pore. But, the mechanism implied by the structures differed from that of a typical gated pore, because the inhibitor-bound structures did not exhibit any more outward opening than the original leucine-bound structure, which had been described as “doubly occluded”, and thus was unlikely to represent an outward open state. However, a conformation nearly identical to the original leucine-bound structure, in which a β-octoglycoside (β-

OG) detergent bound in S2 was crystalized later[75], and EPR experiments revealed that the β-OG bound conformation corresponded to an outward open state[69], which led to the suggestion that the states may be at least partially inward open.

Consequently, the inhibition mechanism can be considered consistent with an allosteric gated pore model, but without the clear open/closed two-state behavior described for the extracellular gate. *We note, moreover, that in itself, the finding of a V-type allosteric mechanism of inhibition by the synthetic tricyclic antidepressants does not necessarily implicate V-type allostery as a required component of the physiological mechanism of transport.*

A detailed assessment of the function role of the allosteric coupling between the intracellular and extracellular gates in the physiological mechanism of transport has emerged from a combination of computational and biochemical experiments[68]. The results of these studies led to the conclusion that by modulating inward opening, this allosteric coupling could serve both to induce, and to inhibit. A particular role in this modulation was suggested for the functional secondary binding site (S2) that identified from steered molecular dynamics simulations (SMD)[68] of substrate translocation. As leucine was pulled through the transporter in order to identify the conformational changes required for transport the presence of a relatively stable binding site in the extracellular vestibule emerged, which overlapped significantly with the inhibitor binding site observed in the crystal structures. As mentioned above, this site was deemed to be the secondary substrate site, S2. The first experimental evidence for this site was provided by careful evaluations of binding stoichiometry under various conditions using scintillation proximity assays. These experimental results revealed a 2:1 substrate:transporter binding stoichiometry, which was inconsistent with the 1:1 binding stoichiometry seen in the leucine-bound structure. Interestingly, in long

dissociation experiments in the presence of Na+, this measured binding stoichiometry decreased to 1:1, with the remaining leucine trapped in the S1 site. However, addition of non-radiolabelled leucine led to rapid dissociation of the trapped radiolabelled leucine, suggesting that a second leucine was acting on the transporter in a way that induced release of the trapped one. Importantly, mutations of residues identified from the computation to be within the S2 site were shown to reduce the experimentally determined binding stoichiometry from 2:1 to 1:1, and prevented substrate-induced dissociation of trapped substrate[68].

That substrate-induced substrate release was related to transport was subsequently demonstrated by repeating the experiments with LeuT reconstituted in proteoliposomes[68]. These displayed both S1 substrate trapping and S2-dependent substrate-induced substrate release. Importantly, the S2 mutants essentially completely abolished transport in the proteoliposomes, indicating an essential role in the transport mechanism.

The corresponding mechanism of transport dependent on substrate binding in both sites (assuming only the S1 substrate is released) in the MM representation can be described as:

$$\mathrm{Leu^{S2}Leu^{S1}Na_2^+T} \xrightarrow{\beta_{cat}^{bind,Leu^{S2}} k_{cat}} \mathrm{Leu_{in} + 2Na_{in}^+ + Leu^{S2}T} \qquad (1.59)$$

and thus the rate of transport at initial high, saturating concentrations of extracellular leucine would be

$$\frac{d[\mathrm{Leu_{in}}]}{dt} = \beta_{cat}^{bind,Leu^{S2}} k_{cat} [\mathrm{T_0}] \qquad (1.60)$$

This representation of the results suggests that the S2 site could allosterically modulate substrate release in both positively and negatively, i.e., the binding of a second substrate molecule in S2 will induce the release of the S1 substrate ($\beta > 1$), whereas inhibitor binding in S2 will inhibit the release of the substrate in S1 ($\beta < 1$).

It is interesting to note that the mechanistic model implies that by increasing the rate of transport when substrate is bound, the relative degree of substrate-independent Na+ import (which we will refer to as "leak", but note to the reader that we refer here to ion transport, rather than a conductive, channel-like process) could be minimized is the transporter has low, substrate-free inward opening.

The key elements that were revealed by the SMD simulations regarding the molecular process required for leucine to be released intracellularly included (i)-the evidence for the S2 site, (ii)-conformational changes that resulted in outward opening movements of intracellular domains TM1a and TM6b, and (iii)-the increased accessibly of water from the intracellular side that reached all the way to the substrate and ion binding sites. The binding of substrate in the S2 site has not yet been confirmed by crystallographic evidence, and it became clear from a well-documented controversy[75–80] that the experimental conditions have much to do with the availability of this site for ligand binding. However, the specific structural rearrangements suggested by the simulation results were confirmed by the subsequently determined x-ray structure of a LeuT mutant in the *apo* inward-open state[72]. Moreover, MD simulations suggested that binding of alanine in S2 was strongest when the extracellular gate was closed and the intracellular gate was in the process of opening[81]. This is consistent with the gated pore allosteric mechanism, according to which binding in S2 can facilitate the opening of the intracellular gate.

Together, the various results from computational modeling suggested that intracellular conformational changes at the intracellular end of the transporter involve TM1a and TM6b, and that these conformational changes can be induced by binding of substrate in S2. It should be noted, however, that the potential conformational transitions between gate opening states were proposed on the basis of SMD simulations[68], which like other trajectories calculated subsequently using MD and path finding algorithms[81–83], provide important information regarding the processes by which the states exchange, but do not on their own suggest anything about an allosteric coupling mechanism that modulates the equilibrium between those states. In order to extract any suggestions about allosteric mechanisms themselves from these simulations, accurate free energy differences and barriers would need to be calculated. Thus, notwithstanding the reasonable and compelling mechanistic models inferred from both computation and experiment, direct evidence that intracellular gate opening is rate limiting, or modulated by the inhibitors, remained elusive.

The direct measurement of molecular dynamics of LeuT in experiments utilizing single molecule Förster resonance energy transfer (smFRET)[84,85] and electron paramagnetic resonance (EPR)[69,70], have provided important support and validation for the conformational changes observed in the SMD simulations and from the comparisons of various x-ray structures. Moreover, these results also provided quantitative measures of the equilibrium populations of states visited by the LeuT protein in the corresponding experimental conditions, as well as the rates of transition between. This type of information is essential for building a full model of transport, and as detailed below it supports the proposed allosteric modulation of the conformational changes by substrates and inhibitors.

The smFRET experiments on LeuT[84,85], the first such investigations of a membrane transporter, revealed intrinsic gating dynamics, as well as allosteric modulation of those gating dynamics by substrates and inhibitors. Specifically, intracellular gating dynamics were measured in this set of experiments from the interactions of fluorophore labels at the positions of His7 in the N-terminus and Arg86 in intracellular loop 1 (IL1). The dynamics of the extracellular gate were assessed with labels at the position of Lys239 in extracellular loop 3 (EL3) and His480 in EL6. The results showed that on the intracellular side, the *apo* transporter dynamically exchanged between two kinetically distinct FRET states (referred to as the high and low FRET states), but preferred the low FRET state. Based on their positions and response to ligands, these states were presumed to correspond to inward closed (IC) and inward open (IO) states of the transporter, respectively. The equilibrium between these states can be written as

$$T_{IC} \underset{k_{close}}{\overset{k_{open}}{\rightleftharpoons}} T_{IO} \tag{1.61}$$

where

$$K_{open} = \frac{k_{open}}{k_{close}} < 1 \tag{1.62}$$

Addition of saturating Na+, leucine, or the inhibitor CMI further stabilized the high FRET, inward closed state. Thus,

$$\alpha_{open}^{bind,Leu} = \frac{K_{open}^{LeuNa_2^+}}{K_{open}} < 1 \tag{1.63}$$

Additionally, under conditions of saturating Na^+, or leucine in the presence of non-saturating Na^+, or CMI, decreased the rate of transitions between states (7-fold, 3.5-fold, and unreported, respectively)[85]. Thus,

$$\beta_{open}^{bind,Leu} = \frac{k_{open}^{LeuNa_2^+}}{k_{open}} < 1 \quad (1.64)$$

In fact, transition state theory (TST) calculations suggested that the combination of Na+ and leucine led to an increase in the transition state energy of approximately 3 kJ/mol, which, corresponds to a β value of ~ 0.3.

The mechanistic inferences from the smFRET study were aided by the investigation of mutants with previously determined, well-known functional properties[86]. Thus, constructs with mutations in the intracellular gate, R5A and Y268A, were found to exhibit a stabilized low FRET state, reinforcing the expectation that the low FRET state indeed corresponded to a state in which the intracellular gate was open, in agreement with the SMD simulations[68] and inward-open crystal structure[72]; conversely, the high FRET state corresponded to a state in which the intracellular gate was closed, as had been seen in the leucine-bound[65] and inhibitor-bound[73,74] crystal structures. Notably, the result that the tricyclic antidepressant CMI both closed the intracellular gate and decreased the transitions between the open and closed state, supports a mechanism in which it blocks transport via V-type allosteric modulation of the inward opening.

The smFRET studies also showed that on the extracellular side, the *apo* transporter displayed a unimodal FRET distribution. Importantly, this distribution was found to be sensitive to the R5A and Y268A mutations at the distal, intracellular end of the transporter: they shift the FRET distribution to higher values, indicating that the

extracellular domain may become more closed as the intracellular gate opens. *Taken together, the evidence for induction of intracellular closing by CMI and for extracellular closing induced by the intracellular gate mutants supports the model of an allosteric gated pore mechanism for transport by LeuT*. But it also suggests that the behavior at the extracellular gate may not be well described by a two-state model.

The main inferences about the allosteric interconnection of the intracellular and extracellular gates of LeuT obtained from the smFRET studies were supported by results from extensive mapping of LeuT's structural ensemble obtained using site-directed spin labeling and double electron-electron resonance (DEER) EPR experiments[70] under varying conditions of ion and substrate concentrations, and mutations. The results from these measurements were interpreted in a structural context[70] with the application of restrained ensemble MD (REMD) calculations[87]. Measurements of extracellular pairs revealed distinct open and closed populations (unlike what had been seen in smFRET) that displayed Na+-induced opening and Na^+- and leucine-induced closing[69]. On the intracellular side, EPR measurements of the H7C/R86C pair reproduced the smFRET results[84,85] with the finding of two well-separated peaks in the distance distribution that were modulated by ions and substrate to favor a closed state. In addition, the EPR measurements revealed modulation of several other distances on the intracellular side, primarily involving the TM6b segment and TM7[70]. However, the measurements in this study did not support the large conformational change in TM1a suggested by the inward-open crystal structure[72]. Instead, these measurements suggested that the conformational change observed by monitoring H7C/R86C pair in both smFRET and EPR corresponded to the movement of the N-terminus rather than a substantial movement of TM1a. Notably, both the Y268A mutation, which had been used to stabilize the inward-open

state captured in the crystal structure[72], and the R5A mutation yielded large displacements of TM1a in the EPR measurements. This led to the suggestion that the large displacement of TM1a seen in the crystal structure could be an artifact of the mutated construct used for crystallization, in which an intrinsically available motion of this TM1a segment was exacerbated when the mutations loosened constraints on that region of the transporter structure.

Despite their disagreement with the x-ray structure, the combined results from smFRET[84,85] and EPR[69,70] demonstrate definitively the allosteric modulation produced by the substrate and ions. *Even if the exact atomic details of the conformational changes are still not entirely clear, the biophysical evidence from these and other experiments, amplified by, and interpreted in the context of results from simulations of various states of the transporter, implicate the binding processes at a distal part of the transporter in the allosteric modulation of the conformational ensemble and dynamics of the intracellular gate.*

The mammalian monoamine transporters (MATs) in the subclass of the LeuT-fold transporters are Na^+/Cl^--dependent neurotransmitter:symporters, which carry out the symport of Na^+ and a biogenic amine, together with Cl^- antiport. The three major types of plasma membrane synaptic MATs (sMATs) include the dopamine transporters (DAT), serotonin transporters (SERT), and norepinephrine transporters (NET). The three classes share high homology and are believed to be both structurally and functionally similar to each other, and to a lesser extent, to LeuT[71]. It is reasonable to question if and how the allosteric mechanism described above for LeuT translates to these transporters, as they are of great importance in basic neurobiology (for their role in the fundamental mechanisms composing neurotransmission) and medicine (for their role as validated targets for a large variety of drugs).

Given the great interest in mechanistic insights about the sMATs, they were subjected to kinetic analyses long before the molecular structures of NSS family members became available from x-ray crystallography. Kinetic analysis of dopamine uptake in rat striatal synapatosomes[88] revealed that the sodium dependence of DAT function was evident both in a change in maximum velocity due to the sodium gradient, and in a change in the affinity of dopamine for DAT. The reported allosteric efficacy of ~2 for DA binding was observed as well for SERT[89]. These results are consistent with the findings for LeuT, and suggest that K-type allosteric coupling between Na^+ and substrate binding discussed in the previous Sections is a fundamental component of the family's transport mechanism. Interestingly, the more recent x-ray structures[90–92] and homology models[93,94] of sMATs found that the Na^+ ions do not interact directly with the substrate as is seen in LeuT[65], indicating that the coupling can be accomplished via more than one structural mechanism.

The presence of an allosteric coupling between ligand binding and the proposed extracellular and intracellular gates of the sMATs had been inferred from structure-function analysis with a variety of approaches by monitoring ligand-induced conformational changes with the substituted cysteine accessibility method (SCAM)[95–98]. In DAT, such experiments in which the sensitivity of extracellular and intracellular cysteine point mutants to reaction with methanethiosulfonate was monitored, revealed that the binding of cocaine, a transport inhibitor, increased the accessibility of extracellular residues, while decreasing the accessibility of intracellular residues[95]. The binding site of cocaine was unknown at the time, and these results suggested that cocaine locked DAT in a conformational state with outward-open/inward-closed characteristics. To various degrees, combinations of such conformational changes were identified for other substrates and inhibitors as well, clearly demonstrating a

coupling between accessible conformations and ligand binding[97]. The finding that homologous intracellular residues in SERT were also shown to be less reactive after binding serotonin or cocaine[99,100] reinforced the consideration of common coupling mechanisms in the sMAT family. The structural context for such coupling mechanisms was offered recently by x-ray structures of drosophila DAT (dDAT) bound to various inhibitors[90–92], including cocaine, as well as dopamine, which were found in outward open/inward closed configurations. Furthermore, the intracellular Cys342 in DAT was shown to become more reactive during Na^+-dependent, inward transport of the substrate m-tyramine, and this was interpreted as an indication that an opening of this intracellular region was required for transport[98].

Other structure-function studies also pointed to residues in the intracellular domain that were involved in determining the intracellular conformation equilibrium. Specifically, an endogenous Zn^{2+} binding site in DAT was used to show that while Zn^{2+} binding usually inhibits transport[101], the Y355A transport-inactivating mutation reversed this effect of Zn^{2+} binding, so that Zn^{2+} activated transport in the inactive mutant[102]. The Y355A mutant also decreased the affinity of cocaine-like inhibitors, suggesting that this tyrosine was somehow required for the stabilization of an inward-closed conformation usually associated with cocaine binding.

Computational analysis of the conformational changes associated with the Y355A mutation in models of hDAT constructed by homology to LeuT attributed the effects to the role of Y355 as part of the conserved intracellular network described in [86]. Indeed, mutations of other residues in this network, R60A and D436A, were shown to lead as well to transport inhibition, and to activation of transport by Zn^{2+} binding[86]. The proposed mechanism for the effects of these mutations was the stabilization of an inward open state, and this was confirmed when the homologous mutations in LeuT

(Y265A and R5A) were examined with smFRET[84], EPR[66], and x-ray crystallography[72] and found to stabilize the inward open state (for details see above).

A LeuT-based hDAT homology model was further used to investigate the conformational changes associated with substrate release, as well as the existence and mechanistic role of the secondary substrate binding site S2 in the extracellular vestibule that had been previously identified in LeuT[93]. By using an SMD procedure based on the one used to study LeuT[68], a homologous S2 site was identified in hDAT, and so were many of the same conformational changes observed in LeuT during the transition to an inward-open conformation.

1.1.3. Ligand-Specific Allosteric Modulation

While allostery is well documented in both GPCRs and transporters, the simple models of allostery described in Section 1.1.1.2. Theoretical Background are not sufficient to describe many new functional observations. In particular, growing evidence suggests that the functions of these two classes of membrane proteins (GPCRs and transporters) are not binary. GPCRs are not simply active or inactive, and transporters do not simply transport whatever substrate can bind to their substrate binding site. Instead, the allosteric modulation of functional equilibrium and kinetics function appears to be ligand-specific. In the section below, we refer to this new behavior as ligand-specific allosteric modulation.

1.1.3.1. Ligand-specific Allosteric Modulation in GPCRs

The traditional allosteric receptor activation model assumes the receptor is phenomenologically a two-stat system (active and inactive, in respect to its ability to activate a G protein) and also assumes that ligand binding modifies the relative energy

of the two states without modulating the distribution (or ensemble) of conformations within those states. However, this model predicts that a GPCR only has a single mode of activation. This prediction has been long contested by the observation that GPCRs can activate many different G proteins (e.g., G_i, G_s, and G_q) to different extents[103], in addition to signaling through of arrestin[104,105]. However, if one invokes a phenomenon known as "biased signaling", the two-state model's parsimony is preserved. It has been observed that different receptors can have their own inherent efficacy for activating different pathways, due to differences in either sequence or structure, and thus it would be expected that there exists receptor-specific active and inactive states, and each receptor can still be a two-state system. While biased signaling can explain multi-model signaling with different efficacies profiles, the two-state model has recently been refuted by the observation that ligands of differing structures that target the same receptors appear to activate these multi-model signaling profiles in ligand specific manners (for examples, see [106]). The preferential activation of specific signaling or downstream phenotypes is referred to as **"functional selectivity"** or "biased agonism"[107,108,109]. The existence of functional selectivity requires a more complex model for GPCR activation. This new model must either i) allow for more "active" states, i.e. at least one state per potential signaling modality, ii) allow for ligand-specific active states, i.e. that the character of the active state is modulated by the ligand itself, or iii) some combination of the two, i.e. within the active state there are substates that are involved in activating different signaling modalities, and thus the overall active state is stabilized by all agonists, but the specific distribution of functional substates are modulated differently depending on the agonist. However, multiple activate states present a major problem for crystallography, which has generally relied on the assumption that ligands stabilize a single crystallographically resolved structure given their two-state pharmacology. If many different active states

exist, which of these states have been captured in past structures? Can these states be observed without the downstream effector bound? The second model, in which the active state is specific to the ligand, also presents a problem crystallographically, as it would be difficult to infer the activation mechanism of a GPCR by one ligand based on previously solved structures of that GPCR with other ligands, unless their activation profiles are significantly similar. However, mass spectroscopy studies of β_2AR support the second model. The accessibility of nine endogenous cysteine and lysine residues to reaction with N-ethylmaleimide and succinic anhydride reagents was quantified, and compared for nine ligands of differing pharmacology. While two residues, Cys77$^{2.48}$ and Cys327$^{7.54}$, were found to respond in accordance with the pharmacological class of the ligand (agonist versus antagonist/inverse agonist), many other residues were found to respond in a ligand-specific manner without any correlation with the functional output of the ligand (see **Figure 10**). These data suggested that while GPCRs may have somewhat discrete active and inactive states, these active states display significant ligand-specific character, which is currently not understood through any physical mechanism.

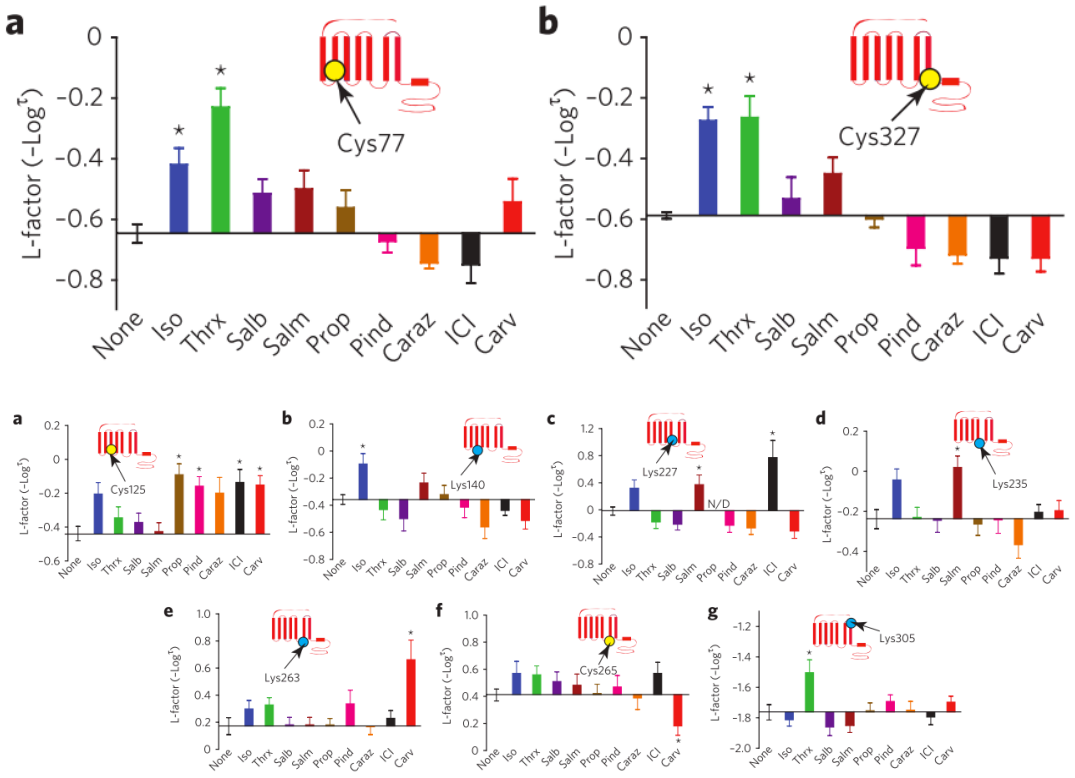


Figure 10. Ligand-specific allosteric modulation of β_2AR.

Top: effects of nine β_2AR ligands on neM reactivity at cys77 (a) and at cys327 (b). Bottom: (a–g) The effects of various ligands on the changes in the l-factors of seven different sites of the β_2AR, expressed relative to the receptor without ligand: cys125 (a), lys140 (b), lys227 (c), lys235 (d), lys263 (e), cys265 (f) and lys305 (g). Data correspond to the means ± standard errors from at least three independent experiments. Asterisks indicate statistical significance (*$P < 0.05$) compared to control receptor alone by one-way ANOVA.

1.1.3.2. Ligand-Specific Allosteric Modulation in Transporters

While LeuT does transport leucine, it does so at an incredibly slow rate (v_{max} = 334 pmol/min/mg)[110]. In contrast, LeuT transports alanine much more efficiently (v_{max} = 1730 pmol/min/mg)[110], and thus alanine has become a popular substrate for functional

and dynamics experiments. Interestingly, smFRET experiments[85] reveal that unlike leucine, which closes the intracellular gate and reduces dynamics, alanine increases the dynamics at the intracellular end of the transporter without altering the relative populations of each state with a β value of ~ 5. The differential effects of the substrates and their consequences for the transport function led to the proposition[85,111] that the increased dynamics are responsible for the increased v_{max} measured experimentally. This proposal is consistent with the theoretical result encoded in (1.60), namely that the velocity of transport can be increased by a greater rate of inward opening, and not necessarily by an increase in the equilibrium population of that inward open state.

However, while this striking observation of ligand-specific allosteric can be linked to transport on a theoretical level, the physical basis for this allosteric effect was unknown at the time. While several x-ray structures of LeuT bound to various substrates in S1 have been solved[110], nearly all of them were identical in terms of Cα RMSD, with only very minor differences in the binding pocket. Furthermore, these structures were for the most part the same state as the original leucine-bound structure – only tryptophan crystallized in a new state, which was outward open with tryptophan bound in both S1 and S2. With these structures alone, it is not possible to make any strong hypotheses regarding the mechanism of allosteric modulation of intracellular gating, nor is it clear as to how the various substrates of different transport efficacies were engaging that mechanism differently. These open questions will be addressed in this dissertation.

1.2. Dissertation Overview

The existence of ligand-specific allosteric modulation in both transporters and GPCRs emphasizes the importance of understanding how allostery works in these systems in term of atomic-level physical mechanism. In order to understand how ligand-specific allostery occurs, a general model of allostery that allows for ligand-specific allosteric modulation is required, and thus it must be a physical model rather than a phenomenological model. The overarching hypotheses driving the work described in this dissertation is that proteins such as transporters and GPCRs are intrinsically allosteric (i.e. structural components are allosterically coupled in the absence of external perturbations) and that ligand-specific allosteric modulation is due to differential engagement of structural components involved in the protein's intrinsic allosteric behavior (i.e. the response of the receptor to ligands with different pharmacological profiles is the results of ligand-specific engagement of the structural components involved in activation). Towards that goal, the work described in this dissertation will focus on two specific aims: i) the development of theoretical models that provide insight into the structural and dynamic features required for systems to be allosteric, and ii) the development of computational methods that can identify these features in specific systems of interest. The utility of these aims will be demonstrated through application to the membrane transporters LeuT and DAT, and the serotonin receptor 5-HT_{2A}R. In specific, these advances will be used to i) generate new understanding regarding the ligand-specific allosteric effects that have been observed in these systems, and ii) generate novel hypotheses that can be addressed experimentally.

2. Theoretical Models and Computational Methods

The following section will describe theoretical models and computational methods that were developed towards the goals of the dissertation. The section is divided into i) Allosteric Ising Models, a theoretical model for allostery that can be related analytically to the allosteric efficacy for simple systems, ii) N-body Information Theory Analysis, a computational method for identifying allosteric channels in proteins, and iii) a random forest-based method for differentiating class-specific allosteric modulation from ligand-specific allosteric modulation.

2.1. Allosteric Ising Models

Note: much of the text in this chapter has been adapted from a previously published manuscript[2] with permission from the publisher.

2.1.1. Motivation for Model

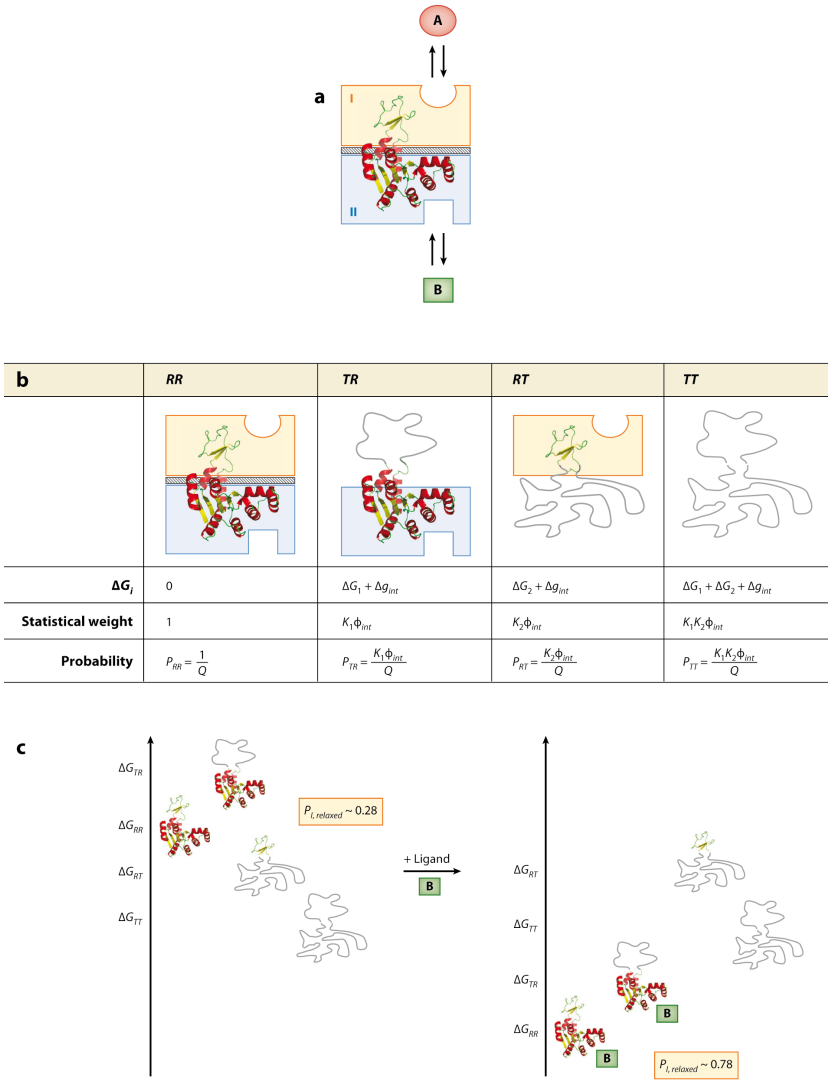
As described in Section 1.1.1.2. Theoretical Background, the allosteric efficacy is a powerful tool for quantifying the allosteric coupling between molecular processes. However, it provides only a phenomenological explanation of allostery. According to the description, often considered "the thermodynamic" perspective, allostery occurs because of the differences in free energy of the respective states. However, this conclusion appears to be a definition, i.e. that allostery is the phenomena in which that the stability of the *on* state relative to the *off* state is greater when the ligand is bound, and lesser when the ligand is unbound. From a "structural" perspective, one needs to consider the differences in free energy as emerging from some feature of the underlying network of interacting structural components, and it is this feature that makes the system allosteric.

To understand allostery at a level that explains how allosteric biomolecular systems work in a structural context requires a quantitative theoretical description that bridges the features of the structural components and their interactions, to the thermodynamic allosteric parameters.

2.1.1.1. Previous Statistical Mechanical Models of Allostery

While the structural features of proteins are often invoked when attempting to describe the physical mechanism underlying a particular allosteric response in a particular system, little has been done to construct a theory which can describe how the energetics of individual structural components and networks of their interactions between them leads to the emergence of allostery. At the thermodynamic level, allostery can be described through the free energy of specific states, and thus a theory of allostery must be able to calculate probabilities of specific states system and express those probabilities in terms of the system's potential energy function. One model that comes close to providing this level of insight in the framework of statistical mechanics is the *ensemble allosteric model* (EAM)[112], which describes multi-component cooperative systems. For the sake of brevity, we will describe the model in terms of a two-component cooperative system, although the model is not limited in size in theory. In the EAM, each component of the system has two states, R or T, but the system is not presumed to be a homo-oligomer. The reference state is taken to be RR. The free energy difference between each other possible state (TR, RT, TT) from RR is expressed as a change in free energy due to the conformational change, and then a change free energy due to the interaction between the domains. There is an identical change in free energy due to the interaction between the domains for all non-RR states. From here, the partition function can be calculated (see **Figure 11**). To account of the effect of ligand binding, it is assumed that ligands can bind their respective

domains only when the domain is in the R state. Given that assumption, one can calculate how the binding of ligand B binding in domain 2 modulates the binding of ligand A in domain 1 by calculating the change in probability of state RR when the free energy of states RR and TR is increased or decreased by a given ligand binding energy.



Hilser VJ, et al. 2012.
Annu. Rev. Biophys. 41:585–609

Figure 11. The two-domain ensemble allosteric model.

(a) A hypothetical two-domain protein (blue and orange boxes) contains effector and active sites. (b) Each state of the two-domain protein (RR, TR, RT, and TT), where the T state is denoted as a grey random loop. The free energy differences from the RR state, the statistical Boltzmann weight, and the probability are shown for each. (c) The relative free energy of the states before and after ligand binding. The values used for this example are $\Delta G_1 = -0.7$ kcal mol^{-1}, $\Delta G_2 = -2.3$ kcal mol^{-1}, $\Delta g_{int} = +1.6$ kcal mol^{-1}, $\Delta g_{Lig,B} = -3.0$ kcal mol^{-1}. Reproduced from [112].

To quantify the allostery in these systems, the allosteric efficacy is not calculated, although it is possible. Instead, the coupling response (CR) is calculated:

$$CR = \frac{\left| p_B(RR) + p_B(RT) - p(RR) - p(RT) \right|}{\log(Z_B) - 1\log(Z)} \tag{1.65}$$

The CR describes the change in probability relative to the amount of free energy introduced into the system due to the binding of ligand B.

The EAM suffers from some severe limitations that narrow the applicability of its predictions. First, while it is not required, the model assumes that a ligand can only bind to the R state of its corresponding domains, which is referred to as the "high affinity state". This is assumption is non-physical, and implicitly assumes that the allosteric efficacy for the "activation" of any binding domain by its ligand is INF, which is contradicted by the known existence of ligands that bind at the same site of the same protein with similar affinities and yet have different allosteric effects in terms of agonism/antagonism/inverse agonism. In the EAM, the allosteric effect is encoded in the coupling between the domains, and thus does not allow for different ligands to differentially modulate the same system. Even if the "high affinity state" is switched from R to T to model an antagonist, partial agonism and antagonism is not possible. Thus, it is necessary to add the ability of the ligand to bind both states is required. Additionally, the use of CR rather than allosteric efficacy as a quantification of allostery leads to apparent insights that are actually just artifacts of the CR function. For example, as the CR uses raw changes in probabilities, changing the free energy of the R and T states by the same amount will always result in a larger CR if the R state was initially low probability. Thus, the prediction that systems that begin in the T state are more allosteric is simply due to the construction of the CR measure. Furthermore,

while it has been claimed that the EAM model proves that a structural pathway for allosteric coupling is unnecessary, this apparent insight is due to the use of an unreasonable interaction term. There is no reason for the interaction energies between domains for the RT, TR, and TT states to be equivalent, and they likely are not. Instead, each state may have a potentially unique interaction energy term, which would result in a minimum of two more parameters in the model. While some of these parameters can be assumed to be merged into the conformational free energy changes, the EAM can only be recovered when $\Delta G_{int}(RT) + \Delta G_{int}(TR) = \Delta G_{int}(TT)$, which is an added constraint that has no justification. In fact, the value of four potentially unique interaction energies is likely to relate to the specific type of interactions that can be formed between the differing conformations of the domains, i.e. their values report on something about how the two domains are interacting physically.

In the following sections, we will present a statistical mechanical model for allostery that is constructed to i) include the allosteric efficacy of a ligand for shifting the conformation of its binding site, ii) naturally express the allosteric efficacy of a ligand for shifting the conformation of an allosteric site in terms of the model parameters, and iii) include the presence of allosteric modulation through indirect coupling through intermediate structural elements.

2.1.2. Derivation and Results

2.1.2.1. The allosteric efficacy as a function of local interactions

We approach the problem of formulating a theory of "how allostery works" by studying the statistical mechanics of a system of interacting structural components. These structural components may be any subset of a biomolecular system that can be treated as a unit when described at some level of coarse-graining (i.e, a helix, a β

strand, a helical bundle, a binding site, etc). The approach we will pursue is conceptually similar to the EAM[112], but with the goal of introducing a structural context that can be analyzed analytically. Defining an n-component system X where for a single configuration each component can be in one of m states, we write the potential energy function of a given configuration of X, U(X), as

$$U(X) = \sum_{i=1}^{n} U^{conf}(X_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{U^{int}(X_i, X_j)}{2} \tag{1.66}$$

The first term in (1.66) represents the conformational energy of each state of each component independent of other components, and the second term represents the pairwise interaction energy between components; all interaction terms when i = j are 0. We can write the probability of any conformation of the system according to the Boltzmann distribution as:

$$p(X) = \frac{e^{-\beta U(X)}}{Z} \tag{1.67}$$

β is $1/k_BT$, where k_B is the Boltzmann constant and T is the temperature in Kelvin. The numerator is known as the Boltzmann factor, and Z is the partition function, which sums over the Boltzmann factors of all states and normalizes the probability

$$Z = \sum e^{-\beta U(X)} \tag{1.68}$$

We can then define the specific case of ligand binding to a two-state receptor. This system can be defined as a two-component system in which each component is two-state: one component representing the receptor, R, with states *on* and *off*, and the second component representing the ligand, L, with states *bound* and *unbound*. It should be noted that for the ligand, the conformational energy term represents the

component of the binding energy that is independent of the state of the receptor. Using the explicit definition of the concentration:

$$\left[X\right]=\frac{N_X}{V} \tag{1.69}$$

where N_x is the number of molecules of X and V is the volume, we can rewrite (1.7) with the explicit definition of protein concentration,

$$K=\frac{\frac{Nf_{on}}{V}}{\frac{Nf_{off}}{V}}=\frac{f_{on}}{f_{off}} \tag{1.70}$$

where N is the total number of receptors and f_{on} and f_{off} are the fraction of receptors in the *on* and *off* states, respectively. Given that the system is ergodic, the frequency of a given state at steady state will converge to the ensemble probabilities. Rewriting (1.7) by substituting thermodynamic equilibrium constants with ratios of probabilities, we can define the allosteric efficacy as

$$\alpha\frac{p\left(L=\text{unbound},R=\text{on}\right)}{p\left(L=\text{unbound},R=\text{off}\right)}=\frac{p\left(L=\text{bound},R=\text{on}\right)}{p\left(L=\text{bound},R=\text{off}\right)} \tag{1.71}$$

Using (1.67) and (1.68), we can write (1.71) as

$$\alpha\frac{e^{-\beta\left[U^{conf}\left(L=\text{unbound}\right)+U^{conf}\left(R=\text{on}\right)+U^{int}\left(L=\text{unbound},R=\text{on}\right)\right]}}{e^{-\beta\left[U^{conf}\left(L=\text{unbound}\right)+U^{conf}\left(R=\text{off}\right)+U^{int}\left(L=\text{unbound},R=\text{off}\right)\right]}}=\frac{e^{-\beta\left[U^{conf}\left(L=\text{bound}\right)+U^{conf}\left(R=\text{on}\right)+U^{int}\left(L=\text{bound},R=\text{on}\right)\right]}}{e^{-\beta\left[U^{conf}\left(L=\text{bound}\right)+U^{conf}\left(R=\text{off}\right)+U^{int}\left(L=\text{bound},R=\text{off}\right)\right]}} \tag{1.72}$$

Equation (1.72) reduces to

$$\alpha=e^{-\beta\left[\left(U^{int}\left(L=\text{bound},R=\text{on}\right)-U^{int}\left(L=\text{bound},R=\text{off}\right)\right)+\left(U^{int}\left(L=\text{unbound},R=\text{off}\right)-U^{int}\left(L=\text{unbound},R=\text{on}\right)\right)\right]} \tag{1.73}$$

We then find the analogous expression of (1.18):

$$-\frac{1}{\beta}\log(\alpha)=\left(U^{int}\left(L=\text{bound},R=\text{on}\right)-U^{int}\left(L=\text{bound},R=\text{off}\right)\right)+\left(U^{int}\left(L=\text{unbound},R=\text{off}\right)-U^{int}\left(L=\text{unbound},R=\text{on}\right)\right) \quad (1.74)$$

As (1.74) indicates, the allosteric efficacy is a function the interaction energy between the states, and we have succeeded in expressing the thermodynamic allosteric efficacy as a function of local interactions in our simple two-component ligand/receptor system. However, this result is significantly more useful for considering multi-component systems if additional energetic symmetries are imposed by using an Ising model potential energy function. While these symmetries are not strictly realized in a biomolecular system, we will show that their application leads to concise analytic expressions that are qualitatively and quantitatively accurate as well for systems in which these symmetries are not present.

2.1.2.2. The Allosteric Ising Model (AIM) for multicomponent systems

The Ising model is a statistical mechanical model originally developed to describe phase behavior in ferromagnetic materials [113]. The Ising model, as well as Ising-like models, have since been applied to other complex systems with collective behavior [114,115], including cooperativity during folding [116–118] and in oligomeric assemblies [119,120]. In the Ising model, each particle has two states, corresponding to a spin state of up or down

$$s_X=\begin{cases}-1 & X=\downarrow\\ 1 & X=\uparrow\end{cases} \quad (1.75)$$

The potential energy function of an n-component Ising model is:

$$U(X)=-\sum_{i=1}^{n}h_i s_i-\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{j_{ij}}{2}s_i s_j \quad (1.76)$$

In the Ising model, h_i is the potential energy of particle i due to the magnetic field, and j_{ij} is the spin coupling between particles i and j, where j_{ii} is taken to be 0. If the field term is taken to be site-specific, one can see that the field term can be considered to correspond to the conformational energy, and the spin coupling term to the pairwise interaction energy. We can rewrite the potential function as:

$$U(X) = \sum_{i=1}^{n} u_i^{conf} s_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{u_{i,j}^{int}}{2} s_i s_j \tag{1.77}$$

where u_i^{conf} is the conformation energy of component i and $u_{i,j}^{int}$ is the interaction energy of components i and j. By using (1.77) for the potential energy function, we impose the following symmetries on the two-state components (with binary states represented by up and down arrows):

$$U^{conf}(X = \uparrow) = -U^{conf}(X = \downarrow)$$
$$U^{int}(X_i = \uparrow, X_j = \uparrow) = U^{int}(X_i = \downarrow, X_j = \downarrow) = -U^{int}(X_i = \uparrow, X_j = \downarrow) = -U^{int}(X_i = \downarrow, X_j = \uparrow)$$

(1.78)

For Ising models composed of several components and various interaction topologies, these symmetries allow for concise analytical expression for the allosteric efficacy and binding affinity. We will refer to these models as Allosteric Ising Models (AIMs).

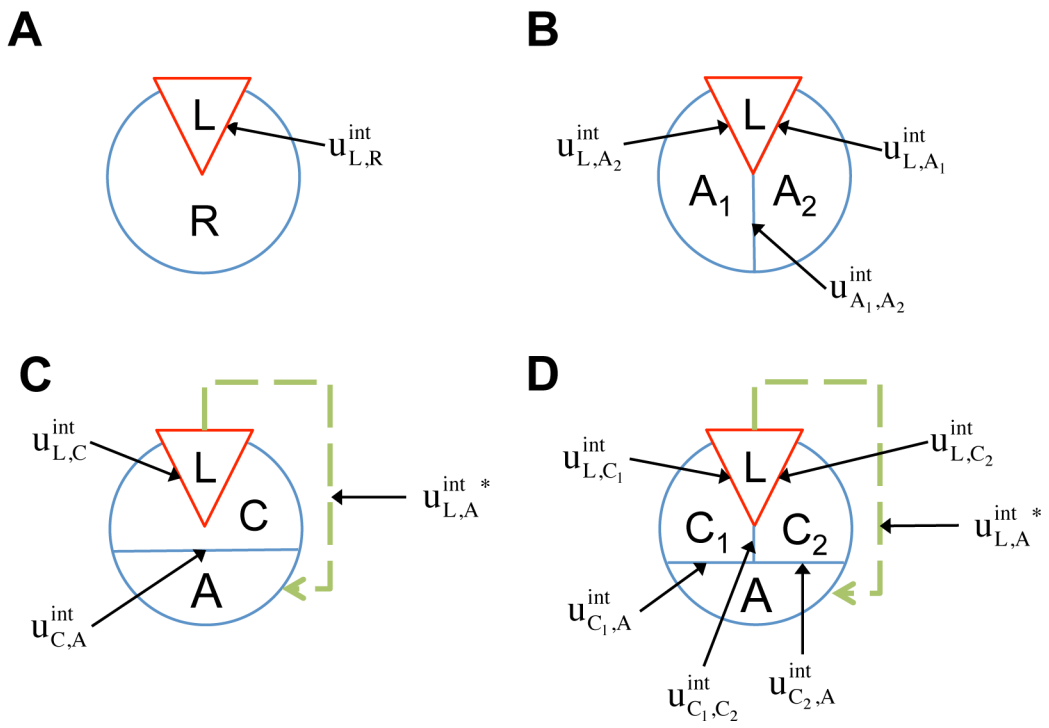


Figure 12. Schematic representations of allosteric Ising models (AIMs).

In the 4 AIMs analyzed here the ligand, L, is represented as a red triangle, and the protein is the blue circle subdivided into various constituent structural components. Lines separating ligand from protein or protein structural components from each other are labeled with the appropriate interaction energy term (as used in the text). Allosteric effective interactions are represented with green dotted lines. The schemes in **(A)** to **(D)** represent, respectively: **(A):** The simple two-component ligand/receptor system. **(B):** A three-component ligand/receptor system with two allosteric sites, A_1 and A_2. **(C):** A three-component ligand/receptor system with one channel, C, coupling the ligand and the allosteric site A. **(D):** A four-component ligand/receptor system with two channels, C1 and C2, coupling the ligand and the allosterically coupled site A.

Considering the analogy to the ligand(L)-receptor(R) systems and treating the *on*/*off* and *bound*/*unbound* states as *up*/*down* spins (see **Figure 12A**), the potential energy function according to (1.77) can be written as:

$$U\left(s_L, s_R\right) = u_L^{conf} s_L + u_R^{conf} s_R + u_{L,R}^{int} s_L s_R \tag{1.79}$$

As the interaction energy between the receptor and the ligand must be zero when the ligand is in the unbound state, we write an alternative non-Ising potential energy function where the interaction energy is 0 when the ligand is unbound:

$$U\left(s_L, s_R\right) = u_L^{conf} s_L + u_R^{conf} s_R + u_{L,R}^{int} \frac{s_L + 1}{2} s_R \tag{1.80}$$

This equation can be re-written as an Ising model potential energy function:

$$U\left(s_L, s_R\right) = u_L^{conf} s_L + \left(u_R^{conf} + \frac{u_{L,R}^{int}}{2} \right) s_R + \frac{u_{L,R}^{int}}{2} s_L s_R \tag{1.81}$$

Thus we will proceed with (1.79) despite the seemingly non-physical interaction, and later confirm that the relationships derived using this model accurately represent those of non-Ising systems. The allosteric efficacy using this potential energy function is:

$$\alpha \frac{p\left(L = \downarrow, R = \uparrow\right)}{p\left(L = \downarrow, R = \downarrow\right)} = \frac{p\left(L = \uparrow, R = \uparrow\right)}{p\left(L = \uparrow, R = \downarrow\right)} \tag{1.82}$$

and we can simplify (1.73) to:

$$\alpha = e^{-4\beta u_{L,R}^{int}} \tag{1.83}$$

Equation (1.83) indicates that in the Allosteric Ising Model for the ligand/receptor system ("ligand/receptor AIM"), the allosteric efficacy is simply a function of the

ligand-receptor interaction energy term. Positive allostery (agonism) is attributed to negative interaction energy; negative allostery (inverse agonism) is attributed to positive interaction energy. Note that as the interaction energy between the ligand and receptor is related to the allosteric efficacy by a log transformation, we will use here the allosteric efficacy and interaction energy interchangeably, and specifically use interaction energy for visual representations, where the log scale is required.

The two-component model assumes that the protein is entirely rigid, with two global states. However, it is possible for the ligand to allosterically modulate multiple distinct allosteric sites (see **Figure 12B**). It is well known that GPCRs can signal through multiple downstream signaling pathways through coupling to various G protein subtypes and β arrestin[121,122], and that different ligands can differentially activate these pathways[106,107]. This distinction is therefore necessary in the representation of receptor allostery. If we introduce two non-interacting allosteric sites, A_1 and A_2, we can write the potential energy function as:

$$U\left(L,A_1,A_2\right)=u_L^{conf}+u_{A_1}^{conf}+u_{A_2}^{conf}+u_{L,A_1}^{int}+u_{L,A_2}^{int} \tag{1.84}$$

Then the allosteric efficacy at a site as:

$$\alpha\frac{p\left(L=\downarrow,A_1=\uparrow\right)}{p\left(L=\downarrow,A_1=\downarrow\right)}=\frac{p\left(L=\uparrow,A_1=\uparrow\right)}{p\left(L=\uparrow,A_1=\downarrow\right)} \tag{1.85}$$

The probabilities of each state is the sum of the probability of two underlying states:

$$\alpha_{L,A_1}\frac{p\left(L=\downarrow,A_1=\uparrow,A_2=\uparrow\right)+p\left(L=\downarrow,A_1=\uparrow,A_2=\downarrow\right)}{p\left(L=\downarrow,A_1=\downarrow,A_2=\uparrow\right)+p\left(L=\downarrow,A_1=\downarrow,A_2=\downarrow\right)}=\frac{p\left(L=\uparrow,A_1=\uparrow,A_2=\uparrow\right)+p\left(L=\uparrow,A_1=\uparrow,A_2=\downarrow\right)}{p\left(L=\uparrow,A_1=\downarrow,A_2=\uparrow\right)+p\left(L=\uparrow,A_1=\downarrow,A_2=\downarrow\right)} \tag{1.86}$$

which is equal to:

$$\alpha_{L,A_1}\frac{e^{-\beta\left[-u_L^{conf}+u_{A_1}^{conf}+u_{A_2}^{conf}-u_{L,A_1}^{int}-u_{L,A_2}^{int}\right]}+e^{-\beta\left[-u_L^{conf}+u_{A_1}^{conf}-u_{A_2}^{conf}-u_{L,A_1}^{int}+u_{L,A_2}^{int}\right]}}{e^{-\beta\left[-u_L^{conf}-u_{A_1}^{conf}+u_{A_2}^{conf}+u_{L,A_1}^{int}-u_{L,A_2}^{int}\right]}+e^{-\beta\left[-u_L^{conf}-u_{A_1}^{conf}-u_{A_2}^{conf}+u_{L,A_1}^{int}+u_{L,A_2}^{int}\right]}}=\frac{e^{-\beta\left[u_L^{conf}+u_{A_1}^{conf}+u_{A_2}^{conf}+u_{L,A_1}^{int}+u_{L,A_2}^{int}\right]}+e^{-\beta\left[u_L^{conf}+u_{A_1}^{conf}-u_{A_2}^{conf}+u_{L,A_1}^{int}-u_{L,A_2}^{int}\right]}}{e^{-\beta\left[u_L^{conf}-u_{A_1}^{conf}+u_{A_2}^{conf}+u_{L,A_1}^{int}-u_{L,A_2}^{int}\right]}+e^{-\beta\left[u_L^{conf}-u_{A_1}^{conf}-u_{A_2}^{conf}-u_{L,A_1}^{int}-u_{L,A_2}^{int}\right]}} \quad (1.87)$$

This reduces to:

$$\alpha_{L,A_1}=e^{-4\beta u_{L,A_1}^{int}} \quad (1.88)$$

which indicates that the allosteric efficacy of a ligand for an allosteric site is independent of other allosteric sites it also modulates (provided the allosteric sites are not coupled through another interaction). In terms of receptor signaling, this analysis predicts that for ligands with absolute bias for only one signaling pathway to exist, the downstream effectors (i.e., G proteins, β arrestin) would need to interact with unique and independent allosteric sites.

In addition to the existence of multiple allosteric sites, allosteric conformational coupling can be propagated through specific regions within the protein, often called "paths" or "channels". Using the AIM approach described here, we can expand the treatment of allostery to proteins with multiple structural components, where some components are allosterically regulated, and some mediate the allosteric regulation. We begin with a three-component model, composed of the ligand L, a channel C, and an allosteric site A (see AIM represented in **Figure 12C**). The potential energy function is

$$U(L,C,A)=u_L^{conf}s_L+u_C^{conf}s_C+u_A^{conf}s_A+u_{L,C}^{int}s_Ls_C+u_{C,A}^{int}s_Cs_A+u_{L,A}^{int}s_Ls_A \quad (1.89)$$

The allosteric efficacy is then

$$\alpha_{L,A_1}\frac{e^{-\beta\left[-u_L^{conf}+u_C^{conf}+u_A^{conf}-u_{L,C}^{int}-u_{L,A}^{int}+u_{C,A}^{int}\right]}+e^{-\beta\left[-u_L^{conf}+u_C^{conf}-u_A^{conf}-u_{L,C}^{int}+u_{L,A}^{int}-u_{C,A}^{int}\right]}}{e^{-\beta\left[-u_L^{conf}-u_C^{conf}+u_A^{conf}+u_{L,C}^{int}-u_{L,A}^{int}-u_{C,A}^{int}\right]}+e^{-\beta\left[-u_L^{conf}-u_C^{conf}-u_A^{conf}+u_{L,C}^{int}+u_{L,A}^{int}+u_{C,A}^{int}\right]}}=\frac{e^{-\beta\left[u_L^{conf}+u_C^{conf}+u_A^{conf}+u_{L,C}^{int}+u_{L,A}^{int}+u_{C,A}^{int}\right]}+e^{-\beta\left[u_L^{conf}+u_C^{conf}-u_A^{conf}+u_{L,C}^{int}-u_{L,A}^{int}-u_{C,A}^{int}\right]}}{e^{-\beta\left[u_L^{conf}-u_C^{conf}+u_A^{conf}+u_{L,C}^{int}-u_{L,A}^{int}-u_{C,A}^{int}\right]}+e^{-\beta\left[u_L^{conf}-u_C^{conf}-u_A^{conf}+u_{L,C}^{int}-u_{L,A}^{int}+u_{C,A}^{int}\right]}} \quad (1.90)$$

Equation (1.90) simplifies to

$$\alpha_{L,A} = e^{-4\beta u_{L,A}^{int}} \frac{\cosh\left(2\beta\left(u_{L,C}^{int} + u_{C,A}^{int}\right)\right) + \cosh\left(2\beta u_{C}^{conf}\right)}{\cosh\left(2\beta\left(u_{L,C}^{int} - u_{C,A}^{int}\right)\right) + \cosh\left(2\beta u_{C}^{conf}\right)} \tag{1.91}$$

where cosh is the hyperbolic cosine function,

$$\cosh(x) = \frac{e^{x} + e^{-x}}{2} \tag{1.92}$$

It should be noted that the exponential term in (1.91) is the *conditional allosteric efficacy* (i.e. the allosteric efficacy contributed by the direct interaction between the two components). The conditional allosteric efficacy can be written as the sum of weighted allosteric efficacies, with each allosteric efficacy conditioned on a different state of the channel and then weighted by the corresponding probability of that state:

$$\alpha_{L,A|C} = p(C=\uparrow)\alpha_{L,A|C=\uparrow} + p(C=\downarrow)\alpha_{L,A|C=\downarrow} \tag{1.93}$$

where for a given state, s, of C,

$$\alpha_{L,A|C=s} = \frac{p(L=\uparrow, A=\uparrow, C=s)\,p(L=\downarrow, A=\downarrow, C=s)}{p(L=\uparrow, A=\downarrow, C=s)\,p(L=\downarrow, A=\uparrow, C=s)} \tag{1.94}$$

Equation (1.94) simplifies to

$$\alpha_{L,A|C} = e^{-4\beta u_{L,A}^{int}} \tag{1.95}$$

Comparing (1.95) with the allosteric efficacy of the two-component ligand/receptor system expressed in (1.83), it is clear that the conditional allosteric efficacies in the three-component system are simply the allosteric efficacies of the corresponding two-component systems.

We can then differentiate the allosteric efficacy contributed by the direct interaction of two components, the conditional allosteric efficacy, from the indirect contributions and write:

$$\alpha_{L,A} = \alpha_{L,A|C}\alpha_{L,A}^{indirect} \tag{1.96}$$

where the allosteric efficacy contributed by the indirect interaction is:

$$\alpha_{L,A}^{indirect,C} = \frac{\cosh\left(2\beta\left(u_{L,C}^{int} + u_{C,A}^{int}\right)\right) + \cosh\left(2\beta u_{C}^{conf}\right)}{\cosh\left(2\beta\left(u_{L,C}^{int} - u_{C,A}^{int}\right)\right) + \cosh\left(2\beta u_{C}^{conf}\right)} \tag{1.97}$$

Importantly, (1.97) provides a description of the allosteric efficacy as a function of the channel through which it is propagated. There are immediate inferences that can be drawn from this representation. First, the channel must have little preference for either one of its conformations, so that signaling through it can have a high intrinsic signal-to-noise ratio. Based on this inference, *mutations that further stabilize the intrinsically preferred conformation of a channel will decrease the allosteric efficacy of a ligand, whereas mutations that destabilize that conformation will increase the allosteric efficacy*. The existence of these two classes of mutations has immediate implications for the ability to test experimentally the role of specific domains in allosteric signaling. Second, because allosteric transmission through the channel depends on a balance between the channel's conformational energy and the interaction energy between the channel and ligand, and the channel and allosteric site, it follows that a low intrinsic signal-to-noise ratio can be overcome by an increased coupling of the ligand to the channel. Lastly, if the sign of the coupling of the ligand to the channel is opposite that of the channel to the allosteric site, the allosteric signal can be reversed. *Consequently, a binding site on a protein that has been evolved for positive allostery by endogenous ligands can be targeted as a site for negative allosteric modulation,*

and vice versa. It is well known that endogenous agonist-binding sites can be targeted by inverse-agonists, so this result is anchored in experimental evidence.

Comparison of (1.94) with (1.95) indicates that the allosteric efficacy can be written in terms of the conditional allosteric efficacies due to direct interactions:

$$\alpha_{L,A} = \alpha_{L,A|C} \frac{\cosh\left(\frac{1}{2}\log\left(\alpha_{L,C|A}\alpha_{C,A|L}\right)\right) + \cosh\left(2\beta u_C^{conf}\right)}{\cosh\left(\frac{1}{2}\log\left(\frac{\alpha_{L,C|A}}{\alpha_{C,A|L}}\right)\right) + \cosh\left(2\beta u_C^{conf}\right)} \tag{1.98}$$

In effect, the conditional allosteric efficacy is the signal-to-noise ratio for a single step in the signal propagation process, and the effective signal-to-noise ratio for the entire signal propagation system can be described by a non-linear function of all the constituent propagation steps.

Equation (1.98) can also be written as the effective interaction energy, $u_{L,A}^{int\ *}$

$$u_{L,A}^{int\ *} = u_{L,A}^{int} - \frac{1}{4\beta}\log\left(\frac{\cosh\left(2\beta\left(u_{L,C}^{int} + u_{C,A}^{int}\right)\right) + \cosh\left(2\beta u_C^{conf}\right)}{\cosh\left(2\beta\left(u_{L,C}^{int} - u_{C,A}^{int}\right)\right) + \cosh\left(2\beta u_C^{conf}\right)}\right) \tag{1.99}$$

and thus as the sum of the direct and indirect interactions

$$u_{L,A}^{int\ *} = u_{L,A}^{int} + u_{L,A}^{indirect,C} \tag{1.100}$$

It should be noted that the designation of channel versus allosteric site is purely an operational definition in which the site that performs the function of interest is referred to as the allosteric site. If both sites are functional, such as in the case of two independent allosteric sites described above, and if they interact, we can rewrite (1.98) as

$$\alpha_{L,A_1} = \alpha_{L,A_1|A_2} \frac{\cosh\left(\frac{1}{2}\log\left(\alpha_{L,A_2|A_1}\alpha_{A_1,A_2|L}\right)\right) + \cosh\left(2\beta u_{A_2}^{conf}\right)}{\cosh\left(\frac{1}{2}\log\left(\frac{\alpha_{L,A_2|A_1}}{\alpha_{A_1,A_2|L}}\right)\right) + \cosh\left(2\beta u_{A_2}^{conf}\right)} \tag{1.101}$$

The description of the allosteric efficacy as a function of the channel through which it is propagated, in (1.97), indicates that if the channel is a one-dimensional chain of interacting structural components, the allosteric efficacy is quickly diminished (it has been shown that the spin correlation function decays exponentially with distance in one-dimensional Ising models [113]). In **Figure 13**, the effective interaction energy between the first and last components of one-dimensional Ising chains with uniform conditional allosteric efficacies of 10, 100, 1000, 10000, and 100000 are shown as a function of chain length. For weakly interacting systems, channels formed by structural components interacting in series do not appear to be good mediators of allosteric efficacy. The prevalence of multi-segment transmembrane signaling complexes may indicate an evolutionary mechanism to overcome the limitations of serial channels.

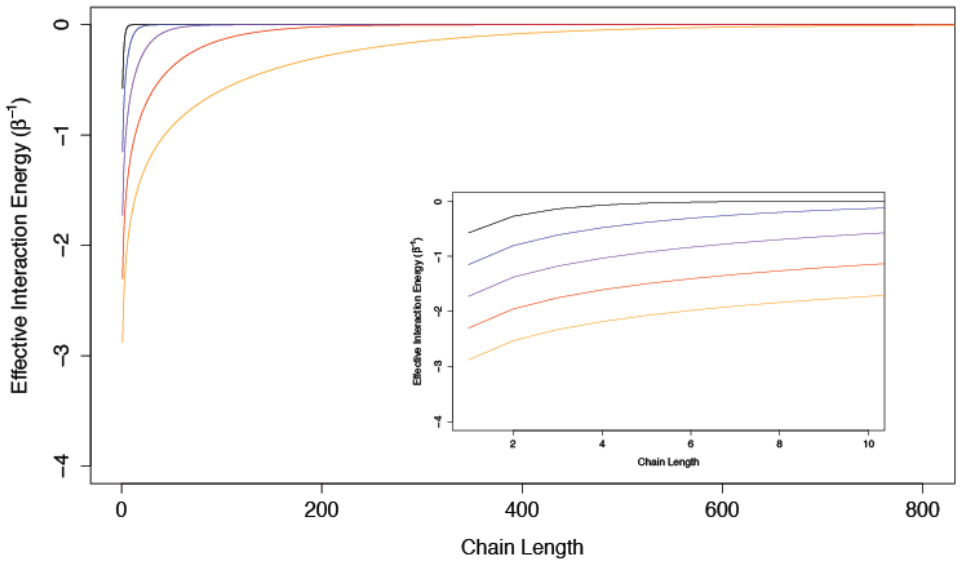


Figure 13. The effective interaction energy through serial channels.

Effective interaction energies of the first and last components of one-dimensional Ising chains are plotted as a function of chain length for direct allosteric efficacy values of 10 (black), 100 (blue), 1000 (purple) 10000 (red) and 100000 (orange). The inset shows detail for short chain lengths. The effective interaction energy is seen to decay exponentially with channel length.

As described above, this analysis is made possible through the energetic symmetries imposed by the Ising model. However, it is unlikely that these energetic symmetries exist in real allosteric proteins. Thus, it is important to consider how well the relationships derived from AIMs describe non-Ising two-state models, which are expected to be better representations of the types of interaction networks present in the biomolecular systems of interest.

To consider this problem, we sampled 100,000 non-Ising two-state allosteric systems with interaction energies and configurational energies sampled from normal distributions of mean 0 and standard deviation of β^{-1}, $3/\beta$, or $5/\beta$. The exact allosteric efficacies, calculated from the exact probabilities of each state, were then compared to the allosteric efficacies estimated from (1.98) using the direct allosteric efficacy terms. We should note that while direct allosteric efficacies can be calculated for non-Ising model, the calculation of the configuration energy term followed

$$2u_C^{conf} \approx U^{conf}\left(C=\uparrow\right) - U^{conf}\left(C=\downarrow\right) \tag{1.102}$$

As above, we addressed problems that may arise from the non-physical interaction energy between unbound ligand and the protein by setting to 0 all interaction energies with the unbound ligand. Results of these calculations are shown in **Figure 14**, where the corresponding effective interaction energies have been used for clarity. Our calculations indicate that (1.98) is a good estimate of the true allosteric efficacy in non-Ising systems in which the allosteric efficacy is high (see **Figure 14A**). As the standard deviation on the energy term distribution increases, and more systems have significant deviation from Ising-like behavior, two distinct groups of false positives (exact effective interaction energy is 0 but estimated interaction energy is non-zero) and true negatives appear (exact effective interaction energy is non-zero but estimated

interaction energy is 0), but the sign of the allosteric modulation is conserved (see **Figure 14B-C**). That the model maintains high accuracy for systems with high allosteric efficacy in spite of the two groups of inaccuracy, suggests that this model should reflect many of the qualitative and quantitative properties of actual allosteric systems.

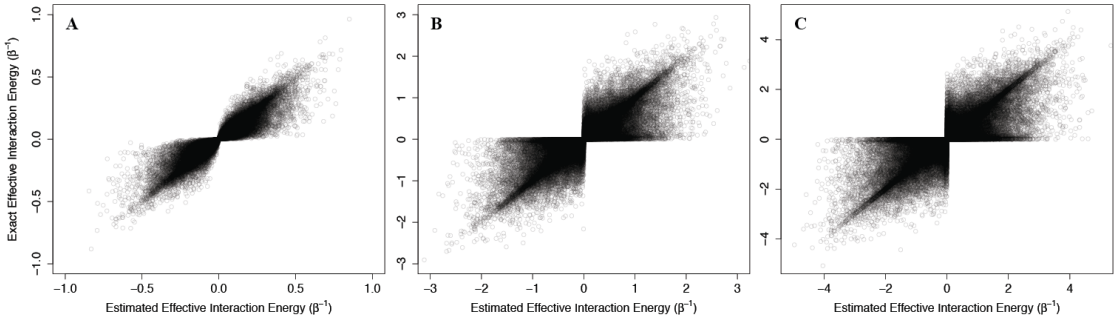


Figure 14. Using the Ising model to estimate effective interaction energies in non-Ising three-component/two-state systems.

The exact effective interaction energies of 100,000 three-component/two-state non-Ising systems are plotted against the effective interaction energy estimated using the equations derived for the three-component Ising model. The systems are generated using energy terms sampled from a normal distribution of mean 0 and standard deviation of 1/β (**A**), 3/β (**B**), and 5/β (**C**) and the points are plotted with 10% opacity.

Efforts to identify allosteric sites and channels in the structures of functional biomolecules have utilized estimates of correlation or mutual information between the structural dynamics of known allosteric sites and candidate modulation sites or channels, most often based on the analysis of molecular dynamics (MD) trajectories [111,123–125] or elastic network models (ENMs) [126,127]. Equation (1.99) indicates that

structural components that can act as channels will have high effective interaction energy with known allosteric sites (e.g., $u^{int}_{C,A}$).

It is not clear, however, how this relates to the mutual information that is evaluated from an MD simulation. As we and others have used mutual information successfully to interpret the structural dynamics and allostery from MD trajectories [111,125,128], it is interesting to test the use of mutual information as an identifier of allostery in the context of AIMs. To this end we calculated the *symmetric uncertainty* [129], a normalized variant of the mutual information, between each component in two-component Ising models and two-component non-Ising models, and compared the *symmetric uncertainty* to the absolute interaction energy. The *symmetric uncertainty* (SU) between components is

$$\mathrm{SU}\left(\mathrm{X_i},\mathrm{X_j}\right)=\frac{2\mathrm{I}\left(\mathrm{X_i},\mathrm{X_j}\right)}{\mathrm{H}\left(\mathrm{X_i}\right)+\mathrm{H}\left(\mathrm{X_j}\right)} \tag{1.103}$$

where I is the mutual information

$$\mathrm{I}\left(\mathrm{X_i},\mathrm{X_j}\right)=\mathrm{H}\left(\mathrm{X_i}\right)+\mathrm{H}\left(\mathrm{X_j}\right)-\mathrm{H}\left(\mathrm{X_i},\mathrm{X_j}\right) \tag{1.104}$$

and H is the Shannon entropy

$$\mathrm{H}\left(\mathrm{X}\right)=-\sum \mathrm{p}\left(\mathrm{X}\right)\log\left(\mathrm{p}\left(\mathrm{X}\right)\right) \tag{1.105}$$

We generated 100,000 two-component Ising systems and 100,000 two-component non-Ising systems with energy terms sampled from a normal distribution with mean 0 and standard deviation of 1, and calculated the symmetric uncertainty and allosteric efficacy of each. We find that the symmetric uncertainty enforces a lower limit on the allosteric efficacy, and allosteric efficacy increases with higher symmetric uncertainty

(see **Figure 15**). Thus, mutual information is a good predictor of allosteric activity in the two-state models explored here. The use of mutual information in systems that are not two-state will be discussed further below.

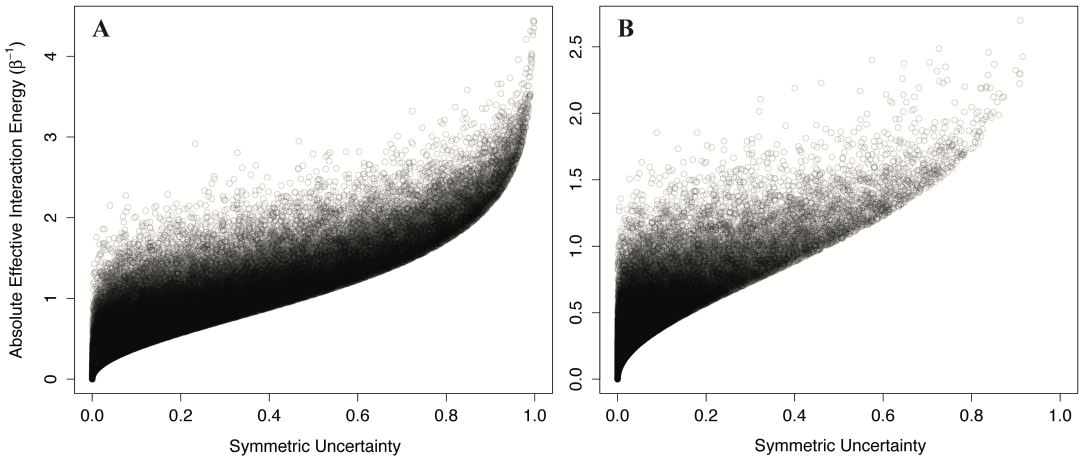


Figure 15. Calculated *mutual information* between the channel and allosteric sites sets a lower bound on the allosteric efficacy.

The symmetric uncertainty between the two components is plotted against the absolute effective interaction energy for 100,000 two-component/two-state non-Ising models (**A**), and two-component Ising models (**B**). The systems are generated using energy terms sampled from a normal distribution of mean 0 and standard deviation of 1/β, and the points are plotted with 10% opacity.

Many proteins have been suggested to have multiple allosteric channels [130]. Assuming that the channels are independent, careful algebra (not shown) reveals that to study the allosteric efficacy of a multi-channel system one can iteratively replace the direct interaction energy term with a direct interaction and indirect interaction of the same effective interaction energy. The effective interaction energy due to multiple independent channels is additive:

$$u_{L,A}^{int\ *} = u_{L,A}^{int} + \sum_{i=1}^{N} u_{L,A}^{indirect,C_N} \tag{1.106}$$

and the allosteric efficacy is then multiplicative

$$\alpha_{L,A} = \alpha_{L,A|\{C_1,\ldots,C_N\}} \prod_{i=1}^{N} \alpha_{L,A}^{indirect,C_N} \tag{1.107}$$

This formally obvious result reveals the advantage of multiple channels in an allosteric protein: perturbations such as mutations that disrupt the conformational stability of one channel will not abolish allosteric function completely. Many parallel weak channels introduce significant robustness when compared to the allosterically equivalent single strong channel built in series, because the latter is completely eliminated by disruption of even a single interaction between two of its structural components.

To test the ability of (1.107) to reflect accurately the behavior of non-Ising systems, we again constructed 100,000 two- and three-channel non-Ising allosteric systems using the methodology described for single channel systems, and compared the resulting allosteric efficacy to that calculated using (1.107) (see **Figure 16**). Again, we find good agreement between the estimates using (1.107) and the exact calculated efficacies, although the accuracy is slightly reduced as the number of channels increases from two to three.

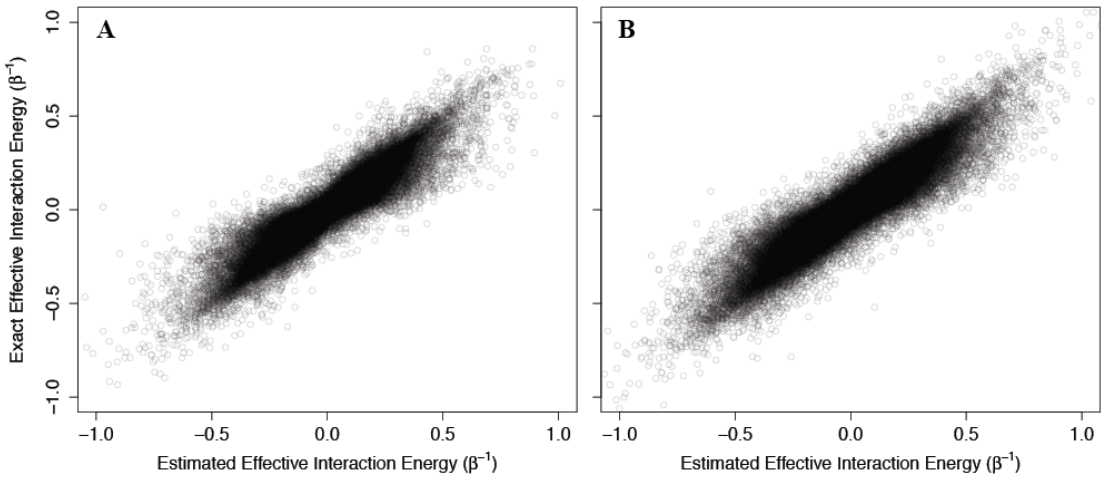


Figure 16. Relation of effective interaction energies in non-Ising two-state systems with multiple independent channels to estimates from the corresponding Ising model.

The exact effective interaction energies of 100,000 two-state non-Ising system is plotted against the effective interaction energy estimated using the equations derived for the n-channel Ising model for two (**A**), and three (**B**) independent channels. The systems are generated using energy terms sampled from a normal distribution of mean 0 and standard deviation of 1/β, and the points are plotted with 10% opacity.

Because it is unlikely that allosteric proteins consist of absolutely independent channels, we explored the effect of interaction between channels through the use of two AIMs: one two-channel system where both channels provide equal magnitude positive allosteric coupling, and one two-channel system where both channels are of equal magnitude but opposite direction. The allosteric efficacy was calculated for each system as a function of the interaction energy between the two channels of allostery for ligands that are coupled to one, or both channels.

As depicted in **Figure 17**, we found that when two channels mediating positive allosteric modulation have negative interaction energy, the allosteric efficacy of the ligand is increased, even if the ligand only interacts with one channel (**Figure 17A**). This is not unexpected; the second channel acts as an indirect channel from the first channel to the allosteric site and additionally multiplies the allosteric efficacy of the channel. However, if the ligand interacts with both channels, the allosteric efficacy is not the square of the allosteric efficacy of binding to one channel as would be for two identical, independent channels. This is because the interaction of the ligand with the first channel has already partially shifted the conformational distribution of the second channel, decreasing its channel efficacy by effectively increasing its intrinsic conformational preference.

For the second two-channel system, with channels providing allosteric coupling in opposite directions, we find that when the interaction energy between the channels is negative, there is decreased allosteric efficacy for the ligand in either channel, whereas positive interaction energy between the channels leads to increased allosteric efficacy (**Figure 17B**). From the perspective of the positive channel, if the channels are positively coupled the second (negative) channel is an indirect channel that flips the sign of the allosteric signal, which leads to reduced overall allosteric efficacy due to negation. However, if they are negatively coupled, the signal through the second channel is flipped twice and left unchanged, leading to increased allosteric efficacy. Interestingly, if the ligand interacts with both channels equally, the effective interaction energy from this pair of channels is 0, independent of the interactions between the channels. In a receptor with these characteristics, antagonists could interact with each channel without conformational preference for the channel, or interact with both channels with the same sign, leading to no allosteric signal.

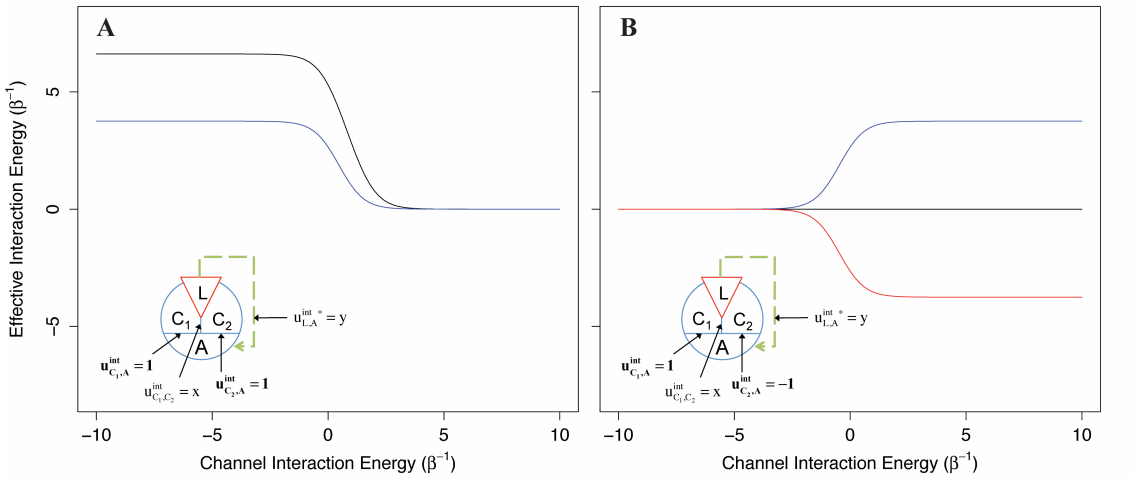


Figure 17. The effective interaction energy of a two-channel AIM as a function of the interaction energy between the channels.

(**A):** The two-channel system in which each channel contributes to positive allosteric modulation is shown for a ligand that interacts with one channel (blue) or both channels (black). (**B):** A two-channel system with one positive allosteric channel and one negative allosteric channel is shown for a ligand that interacts only with the positive channel (blue), only with the negative channel (red), or both channels (black). The effect of interactions between channels is seen to modify significantly the allosteric signal transduction.

2.2. N-body Information Theory Analysis

Note: much of the text in this chapter has been adapted from two previously published manuscripts[111,131], with permission from the publishers.

2.2.1. Motivation for Method

As mentioned previously, the specific process of allosteric signal propagation in a molecular system through intramolecular interactions between structural components has not yet been subjected to direct experimental measurements. This is somewhat surprising, because the allosteric effects can be observed experimentally from the apparent relation between distal parts of a macromolecule. Indeed, to date, there are no experimental methods capable of specifically and definitively defining the role of the intramolecular interactions involved in propagating allostery. Most proposed mechanisms are descriptions of series of local rearrangements that are at best presumed (but not demonstrated) to be causally sequential, but a specific, quantitative definition of the information flow does not exist. For example, a successful experimental method for determining residues that are coupled to ligand binding is the mutant cycle analysis[132]. While it is able to quantify thermodynamic coupling at a distance, the approach relies on these sequential descriptions to propose the underlying mechanism of propagation. In addition, the simple procedure of mutating a residue and measuring the allosteric efficacy in the mutant does not directly test hypotheses regarding the role of that residue in a specific mechanism under wild-type conditions. Viewing the system from the perspective of the AIM described above, one would like to be able to test if either the conformational preference of that residue, or its interaction with another residue, is involved in allostery. While the mutation of a given residue to a "benign" residue such as alanine, as is traditional, can modify interactions between residues involved in propagation, it also modifies the states that residue can adopt and their distribution. Thus, a mutation does not simply remove a residue from the AIM, but instead modifies its role in the AIM in a possibly unpredictable way. If a mutation is to be performed to test hypotheses regarding the role of a specific residue

in an allosteric mechanism, this must be done with specific care for the local structural ensemble both before and after mutation.

Due to this difficulty, many have developed computational methods through which structural information is used to derive an allosteric mechanism. Largely, these methods can be subdivided in two ways. In the first subbdivision, most methods can often be classified by the data they use: they are either structure-based or ensemble-based. In structure-based approaches, allosteric mechanisms are derived from one or more experimentally determined 3-dimensional structures, while in ensemble-based approaches, allosteric mechanisms are derived from the ensemble of conformations the system can adopt, generally estimated using a physics-based sampling method such as Molecular Dynamics. In the second, the methods can be classified by the observable that is used to indicate the presence of allostery: they either look for allosterically-induced conformational changes in terms of significant differences in the structure or ensemble between the apo and ligand-bound state, or for allosteric couplings between domains in terms of statistical or graph theoretical associations. Below, some of these methods will be briefly described.

2.2.2. Previous Methods

2.2.2.1. Structure-based Methods

The simplest method for proposing an allosteric mechanism is through the investigation of the differences between an apo and ligand-bound structure of a comparable construct. At the limits of very large allosteric efficacy and very low basal activity, one might expect that for a ligand-activatable protein, the apo structure is likely to be in an inactive state, whereas the ligand-bound state is likely to be in the active state. It is often, but not always, possible to trace conformational changes from

the ligand-binding site to distant domains. However, observations of an allosteric conformational change do not themselves generate hypotheses as to why that specific conformational change occurred. Rather, mechanistic hypotheses are generally constructed by visually investigating the structure and invoking biophysical intuition.

Returning to the statistical mechanical models of allostery, it should be possible to predict what kinds of allosteric behaviors are possible and the components crucial for those behaviors if one knows the components and their interactions, as well as their intrinsic conformational preferences. While a single structure does not inform about the conformational preferences of its structural components, it does contain the topology of interactions for one of the states. Most often, electrostatic or Lenard-Jones interactions are considered, and are inferred from certain geometric criteria, such as contact distances, or angles of relative orientation between atoms[133,134]. Many methods rely on the analysis of the interaction topology by using graph theoretical statistics[135–137], such as centrality, which quantifies the extent to which a given residue plays a role in the overall connectivity of the network. While these methods do not always directly reveal allosteric networks, as not all pathways identifiable from an interaction network will contribute strongly to the emergent allosteric efficacy, these methods have been able to find some motifs in protein structure, such as interaction hubs[136,138] of densely connected elements, and conclude that differences between these networks highlight interactions that may stabilize one conformational state over others[139].

2.2.2.1. Ensemble-based Methods

As allostery is fundamentally a statistical mechanical phenomenon that pertains to the relative probability of specific states of the system, methods that draw conclusions from the ensemble of conformational states available to the allosteric system of

interest are likely to generate more reasonable predictions regarding the underlying physical mechanism. The generation of these ensembles is not trivial however, as there is no existing experimental method for determining a multi-dimensional conformational ensemble, even at a coarse-grain scale. For example, while EPR can be used to determine many distance distributions simultaneously, it can only measure these distributions independently, and thus the full multivariate ensemble cannot be recovered from the experiment alone. In addition, while multi-color smFRET experiments could in theory measure a multivariate distribution, to our knowledge there are few examples of more than two distances being measured simultaneously, and the conformation of transmembrane domains is not accessible due to the bulk of the dyes required for smFRET. Due to these experimental limitations, computational conformational sampling approaches such as elastic network models (ENM) and Molecular Dynamics simulations (MD) have been extremely useful.

In ENMs[140], the system is represented with only the Calpha carbons, and each carbon is attached to each other carbon within a given cut-off distance by a spring. Here, we will detail only the Gaussian network, although other variants have been proposed[141]. The potential energy function is:

$$U_{ENM} = \frac{\gamma}{2}\left[\sum_{i}^{N}\sum_{j}^{N}\Delta\vec{r}_i\Gamma_{ij}\Delta\vec{r}_j\right] \tag{1.108}$$

where:

$$\Gamma_{ij} = \begin{cases} -1 & i \neq j, \quad r_{ij} < r_c \\ 0 & i \neq j, \quad r_{ij} > r_c \\ -\sum_{j,j\neq i}^{N}\Gamma_{ij} & i = j \end{cases} \tag{1.109}$$

The covariance between the fluctuations of any two atoms can be shown to be:

$$\left\langle \Delta\vec{r}_i \cdot \Delta\vec{r}_j \right\rangle = \frac{3k_B T}{\gamma}\left(\Gamma^{-1}\right)_{ij} \tag{1.110}$$

The covariance matrix, C, is then defined as:

$$C = \frac{3k_B T}{\gamma}\left(\Gamma^{-1}\right) \tag{1.111}$$

An eigenvalue decomposition of C can be performed,

$$\Gamma^{-1} = U\Lambda U^T = \sum_{k=1}^{N-1} \lambda_k^{-1}\left[u_k u_k{}^T\right] \tag{1.112}$$

which leads to independent normal modes whose low frequency (high eigenvalue) models contribute most to equilibrium fluctuations.

It should be noted that the covariance measure defined in (1.110) is not the typical covariance between variables, which will be discussed later. The dot product between two vectors is 0 whenever the vectors are orthogonal to each other. Thus, this covariance measure is dependent on the relative orientation of the fluctuation vectors, which is not desirable. We will refer to this measure as the vector covariance to differentiate it from other measures of covariance.

When using ENM approaches to study allostery, it is implicitly assumed that perturbations that have allosteric effects, such as ligand binding or mutations at key position in an interaction network, modulate the dominant eigenmodes. This assumption appears to be reasonable and has yielded important mechanistic insights about collective conformational changes, e.g., in response to different ligands binding to GPCRs[142], or to mutations at the intracellular gate of LeuT[86]. The successful investigation of such perturbations has been extended to even larger membrane protein

systems (as reviewed in [143]), such as the role of inter-domain or lipid–protein interactions in the conformational transition energy in GltPh[144,145]

While there are some successful applications of ENMs, modeling the free energy landscape as a single harmonic basic around an x-ray structure is unlikely to accurately represent the true ensemble of the system. In order to generate a more accurate estimate of the conformational ensemble, MD is used. In MD, the system is represented at the classical, atomic scale. The potential energy function, also known as the force field, defines the types of interactions between the atoms and the parameters of these interactions for a given set of atom types. A typical force field includes bonded and non-bonded potential energy terms:

$$U = U_{bonded} + U_{nonbonded} \tag{1.113}$$

The bonded terms are harmonic energy terms for the bond, angles, dihedrals, improper angles, Urey-Bradley, and correction map (CMAP) terms:

$$\begin{aligned}
U_{bonded} &= U_{bond} + U_{angle} + U_{dihedral} + U_{improper} + U_{UB} + U_{CMAP} \\
U_{bond} &= \sum_{bonds} k_r \left(r - r_0\right)^2 \\
U_{angle} &= \sum_{angles} k_\theta \left(\theta - \theta_0\right)^2 \\
U_{dihedral} &= \sum_{dihedrals} k_\varphi \left(\left(1 + \cos\left(n\varphi + \delta\right)\right)\right) \\
U_{improper} &= \sum_{angles} k_w \left(\omega - \omega_0\right)^2 \\
U_{UB} &= \sum_{Urey-Bradey} k_{ub} \left(r^{1-3} - r_0^{1-3}\right)^2 \\
U_{CMAP} &= \sum_{residues} u_{CMAP} \left(\Phi, \Psi\right)
\end{aligned} \tag{1.114}$$

In each equation, k represents a force constant. The Urey-Bradley terms impose an additional energetic constraint on each atom by imposing a pseudo-bond between the first and third atom, whereas the CMAP correction biases the backbone angles of each

residue to better resemble those sampled in quantum mechanical calculations. The non-bonded terms include the electrostatics and van der Waals interactions:

$$\begin{aligned} U_{non-bonded} &= U_{electrostatic} + U_{VDW} \\ U_{VDW} &= \sum_{non-bonded} \varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right] \\ U_{electrostatic} &= \sum_{non-bonded} \frac{q_i q_j}{\varepsilon r_{ij}} \end{aligned} \tag{1.115}$$

Here, the VDW term uses the standard 6-12 Leonard-Jones potential. Given this force field, a set of coordinates and velocities, and a function to numerically integrate Newton's equations of motion, the dynamics of the system can be simulated. Newton's equations define the force, F, exerted on a given particle as:

$$\vec{F} = m\vec{a} \tag{1.116}$$

where m is the particle's mass and a is its acceleration. Additionally, the force is equal to the negative of the potential energy gradient:

$$\vec{F} = -\nabla U \tag{1.117}$$

Combining the two leads to:

$$-\frac{dU}{dr} = m\frac{d^2r}{dt^2} \tag{1.118}$$

Typically, the equations of motion are numerically integrated using an algorithm such as Velocity Verlet[146]:

$$\begin{aligned} \vec{x}_{t+\Delta t} &= \vec{x}_t + \vec{v}_t \Delta t + \frac{\vec{a}_t}{2} \Delta t \\ \vec{v}_{t+\Delta t} &= \vec{v}_t + \frac{\vec{a}_t + \vec{a}_{t+\Delta t}}{2} \Delta t \end{aligned} \tag{1.119}$$

MD can be used to estimate the conformational ensemble of a system given that system is ergodic, i.e. the system is not periodic and there exists a number of time steps, n, in which the system can evolve from any state i to any other state j. If these conditions are met, the time average of some observable will approach the ensemble average of that observable:

$$\lim_{T\to\infty}\left[\int_0^T \frac{A(\vec{r}_t)}{T}dt\right] = \int A(\vec{r})p(\vec{r})d\vec{r} = \langle A(\vec{r})\rangle \tag{1.120}$$

Thus, the desired conformational ensemble can be estimated at the limit of very long simulation times:

$$\langle A(\vec{r})\rangle \approx \sum_{t=1}^{T} \frac{A(\vec{r}_t)}{T} \tag{1.121}$$

From an MD simulation of N atom for T time steps, the 3N x T time series X is produced:

$$X = \begin{bmatrix} r_{1,1} & r_{2,1} & \cdots & r_{3N,1} \\ r_{1,2} & r_{2,2} & \cdots & r_{3N,2} \\ \cdots & \cdots & \ddots & \cdots \\ r_{1,T} & r_{2,T} & \cdots & r_{3N,T} \end{bmatrix} \tag{1.122}$$

From this time series, several characteristics can be calculated. First, it is important to note that unlike in the ENM, each x, y, and z coordinate of each atom is implicitly represented in these time series. Thus, (1.122), is equivalent to:

$$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_N \end{bmatrix} \tag{1.123}$$

where

$$X_a = \begin{bmatrix} r_{a_x,1} & r_{a_y,1} & r_{a_z,1} \\ r_{a_x,2} & r_{a_y,2} & r_{a_z,2} \\ \cdots & \cdots & \cdots \\ r_{a_x,T} & r_{a_y,T} & r_{a_z,T} \end{bmatrix} \tag{1.124}$$

As is the case for the ENMs, the vector covariance can be calculated from these time series:

$$C_{a,b}^{vector} = \langle \vec{r}_a \cdot \vec{r}_b \rangle \approx \sum_{t=1}^{T} \frac{\vec{r}_{a,t} \cdot \vec{r}_{b,t}}{T} \tag{1.125}$$

While the vector covariance has been used as an end point of analysis in the past [59], a better measure of covariance can be calculated. The average value of each coordinate can be estimated,

$$\langle r_i \rangle \approx \sum_{t=1}^{T} \frac{r_{i,t}}{T} \tag{1.126}$$

and the 3N x 3N atomic fluctuation covariance matrix, C can be estimated:

$$C_{ij} = \left\langle \left(r_i - \langle r_i \rangle\right)\left(r_j - \langle r_j \rangle\right) \right\rangle \approx \sum_{t}^{T} \frac{\left(r_{i,t} - \langle r_i \rangle\right)\left(r_{j,t} - \langle r_j \rangle\right)}{T} \tag{1.127}$$

Unlike the vector covariance, the covariance terms in C correspond to true statistical covariance. However, the covariance for two atoms, a and b, is then described by a 3 x 3 non-symmetric cross-covariance matrix:

$$C_{a,b} = \begin{bmatrix} C_{a_x b_x} & C_{a_x b_y} & C_{a_x b_z} \\ C_{a_y b_x} & C_{a_y b_y} & C_{a_y b_z} \\ C_{a_z b_x} & C_{a_z b_y} & C_{a_z b_z} \end{bmatrix} \tag{1.128}$$

While this representation leads to difficulties in interpretation, as the covariance between each atom pair is represented by nine numbers rather than 1, eigenvalue

decomposition can also be performed on the full 3N x 3N covariance matrix. The eigenvalue decomposition of a covariance matrix is known as principal component analysis (PCA) or essential dynamics (within the MD community), and similar to normal mode analysis, PCA can be used to identify the system's highest variance (and potential largest spatial scale) motions. However, like normal mode analysis, PCA only identifies the largest linearly independent motions in a given trajectory, and these motions are not guaranteed to be functionally relevant in terms of responding to allosteric perturbations.

As PCA does not directly identify the motions that will be allosterically modulated by a given perturbation, methods have been developed to specifically search for networks of residues that are expected to mediate the long-distance allosteric couplings between specific residues or clusters of residues. We will refer to these methods as analyses of dynamical network models (DNMs)[123,147,148].

In the DNMs, network representations of the protein are built by treating each residue as a node and assigning weights to the edges between each node, e_{ab}, using the dot correlation, rho, between each atom:

$$e_{ab} = -\log\left(\left|\rho_{ab}^{dot}\right|\right) = -\log\left(\left|\frac{\langle \vec{r}_a \cdot \vec{r}_b \rangle}{\sqrt{\langle \vec{r}_a \cdot \vec{r}_a \rangle \langle \vec{r}_b \cdot \vec{r}_b \rangle}}\right|\right) \tag{1.129}$$

In this representation, maximally dot correlated residues have an edge weight of 0, whereas maximally uncorrelated residues have an edge weight of ∞. This representation allows for the calculation of allosteric pathways through the protein by using shortest pathway algorithms from graph theory and network theory. In this framework, an allosteric pathway is defined as a sequence of residues, each of which has high pairwise correlation with the residues before it and after it. In addition, this

framework has been extended to account for sub-optimal pathways[149], and allows for the quantification of statistics such as the centrality and the identification of structures such as communities[150]. This method has been illustrated in applications to tRNA:synthetase complexes[123], in which the identified allosteric pathways pinpointed interactions between conserved residues and specific nucleotides that were shown with mutagenesis experiments to affect synthetase kinetics. An alternative to the original dynamic network analysis in DNM, which is based on linear correlations, is to use the mutual information from information theory[151], which captures non-linear dependencies.

In information theory, the entropy is a functional over the probability distribution of a given variable, and is defined as the average value of the information content of that variable:

$$H\left[p(x)\right] = -\sum_{x \in X} p(x) \log\left(p(x)\right) \tag{1.130}$$

This measure of the entropy is analogous to the entropy of statistical mechanics, but was devised by Shannon due to its favorable characteristics: it is non-negative, is 0 when a variable can take on only one value and is thus non-informative to measure, and is additive for independent variables. While the entropy is additive for independent variables, it is not additive for dependent variables. When variables are dependent, the entropy of the joint distribution is less than the sum of the entropy of the marginal distributions, but no less than the minimum entropy of the two distributions:

$$H\left[p(x)\right] + H\left[p(y)\right] \geq H\left[p(x,y)\right] \geq \min\left(H\left[p(x)\right], H\left[p(y)\right]\right) \tag{1.131}$$

The mutual information is then defined as the entropy difference between the entropy

of the joint distribution and the sum of the entropy of the marginal distributions:

$$I_2\left[p(x,y)\right] = H\left[p(x)\right] + H\left[p(y)\right] - H\left[p(x,y)\right] \tag{1.132}$$

Equation (1.132) can be alternatively express as:

$$I_2\left[p(x,y)\right] = H\left[p(x)\right] + H\left[p(x|y)\right] = H\left[p(y)\right] + H\left[p(y|x)\right] \tag{1.133}$$

where H[p(x|y)] is the conditional entropy of x given y:

$$H\left[p(x|y)\right] = H\left[p(x,y)\right] - H\left[p(y)\right] \tag{1.134}$$

Many approaches have been developed based on the mutual information[152–154]. Because mutual information ranges from 0 to ∞, it is desirable to normalize it in some way. Normalization is not trivial, however. The mutual information is bounded:

$$I_2\left[p(x,y)\right] \leq \min\left(H\left[p(x)\right], H\left[p(y)\right]\right) \tag{1.135}$$

Thus, in the discrete case, several intuitive normalizations are available. The symmetric redundancy, R, is particularly useful:

$$R\left[p(x,y)\right] = \frac{2I_2\left[p(x,y)\right]}{H\left[p(x)\right] + H\left[p(y)\right]} \tag{1.136}$$

However, entropy is often estimated from continuous estimates of the distribution. This is done using the differential entropy, which is the continuous counterpart to the discrete entropy:

$$H_{\text{diff}}\left[p(x)\right] = -\int p(x)\log\left(p(x)\right)dx \tag{1.137}$$

Unlike the discrete entropy, the differential entropy can be negative. It is well known that the differential entropy is infinitely shifted from the true entropy[155], but the

mutual information is unaffected by this property. However, because the true entropy in infinite, the mutual information is unbounded. To bound it, the generalized correlation coefficient[156], which uses the relationship between the mutual information among two normal distributions and their correlation coefficient, r:

$$I_2\left[p(x,y)\right] = \frac{1}{2}\log\left(1 - r_{xy}{}^2\right) \tag{1.138}$$

where:

$$r_{xy} = \frac{\int p\left(r_i, r_j\right)\left(r_i - \langle r_i \rangle\right)\left(r_j - \langle r_j \rangle\right) dr_i\, dr_j}{\int p\left(r_i\right)\left(r_i - \langle r_i \rangle\right)^2 dr_i \int p\left(r_j\right)\left(r_j - \langle r_j \rangle\right)^2 dr_i} = \lim_{T\to\infty}\left[\frac{\sum_t^T \left(r_{i,t} - \langle r_i \rangle\right)\left(r_{j,t} - \langle r_j \rangle\right)}{T\sum_t^T \left(r_{i,t} - \langle r_i \rangle\right)^2 \sum_t^T \left(r_{j,t} - \langle r_j \rangle\right)^2}\right] \tag{1.139}$$

Assuming that the maximum mutual information scales with dimension, the generalized correlation coefficient between two d-dimensional distributions is then defined as:

$$r_{GC}\left[p(x,y)\right] = \sqrt{1 - e^{-\frac{2}{d} I_2\left[p(x,y)\right]}} \tag{1.140}$$

This normalization has several problems, which will be discussed later. Despite these problems, using the generalized correlation coefficient with the DNM has been used in the analyses of allosteric networks in thrombin[128] and imidazole glycerol phosphate synthase[157] to some success, although with minimal experimental validation. However, while DNM and mutual information methods can identify pathways with high consecutive pairwise correlation or mutual information, these formulations do not guarantee that all components of the pathway are correlated. We will illustrate with the use of a simple two-dimensional three-body chain (see **Figure 18**) that if the axes of covariance are not aligned, the pairwise correlations between bodies that are consecutive in the chain may be high, but the system cannot transmit information.

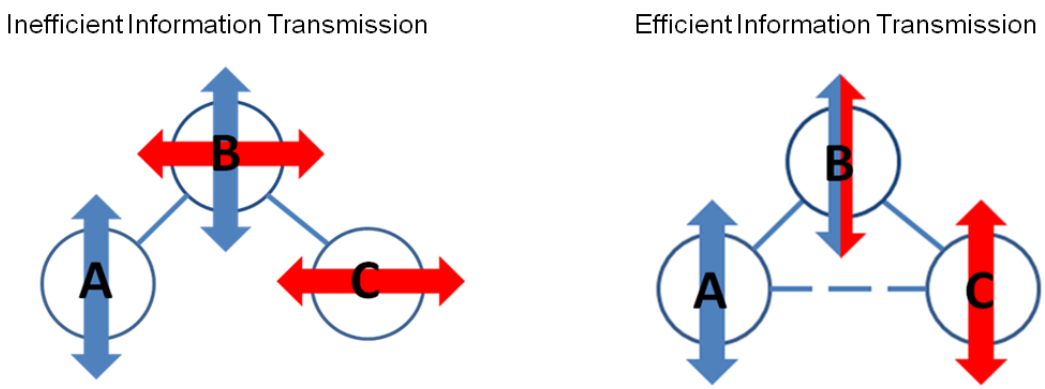


Figure 18. Efficient information transmission by a 3-body system.
Information transmission through a 3-body system (solid blue lines represent direct interactions), moving in 2D, is inefficient if the axes of covariance of each pair (thick arrow) are not aligned (left). Information transmission it is efficient if the axes of covariance are aligned (right); the dotted blue lines represent indirect allosteric interaction as a result of information sharing.

If A and B co-vary on the blue axis, information about the position of A on the blue axis is present in the position of B on the blue axis, and if B co-varies with C on the red axis, information about the position of B on the red axis is present in the position of C on the red axis. When A and C co-vary with B on different axes, no information about the position of B on the blue axis is present in the position of C on the red axis, and thus no information about the position of A on the blue axis is present in the position of C on the red axis, i.e., there is no allosteric information transmission (see **Figure 18**, left).

However, when the blue and red axes are aligned, A, B and C all co-vary on the same axis and the 3-body correlation leads to information about the position of A on the blue axis being present in the position of C on the red axis (see **Figure 18**, right). This

model illustrates a weakness in the use of network theoretical methods that do not maximize the higher n-body correlations: while shortest path analysis maximizes the pairwise correlations, one would expect that many pathways found using such network theoretical methods may not actually be efficient information channels.

Indeed, higher-order correlations between multiple residues in the network, which can be described using higher-order mutual information[158]*, are required for a system to transmit information through the network. In the following sections, we will present a method that uses these high-order mutual information terms to identify allosteric information channels in proteins.*

2.2.3. N-body Information Theory (NbIT) Analysis

The new NbIT analysis method presented here utilizes a generalization of the concept of n-body mutual information, also known as co-information or interaction information[158–161], an information theoretical measure which enables a description of the possible contribution that a variable makes to the mutual information shared between two other variables. The N-body information is calculated recursively as:

$$I_N\left[p\left(x_1,x_2,\ldots,x_N\right)\right]=I_{N-1}\left[p\left(x_1,x_2,\ldots,x_{N-1}\right)\right]-I_{N-1}\left[p\left(x_1,x_2,\ldots,x_{N-1}\middle|x_N\right)\right] \qquad (1.141)$$

where

$$I_{N-1}\left[p\left(x_1,x_2,\ldots,x_{N-1}\middle|x_N\right)\right]=I_{N-2}\left[p\left(x_1,x_2,\ldots,x_{N-2}\middle|x_N\right)\right]-I_{N-2}\left[p\left(x_1,x_2,\ldots,x_{N-2}\middle|x_N,x_{N-1}\right)\right] \qquad (1.142)$$

Recursion will eventually lead to the 3-body information terms, which are the central information measurements in the analysis. The 3-body information is:

$$I_3\left[p(x,y,z)\right] = I_2\left[p(x,y)\right] - I_2\left[p(x,y|z)\right] \tag{1.143}$$

where

$$I_2\left[p(x,y|z)\right] = H\left[p(x|z)\right] + H\left[p(y|z)\right] - H\left[p(x,y|z)\right] \tag{1.144}$$

3-body information can be visualized easily using an information Venn diagram (see **Figure 19**). While several representations of this information are found in the literature with varying signs, we have chosen to use the sign convention described by [159,161]. Using this convention, when 3-body information is positive, the third body may increase the information transmission between the two others, whereas when it is negative, the third body diminishes it.

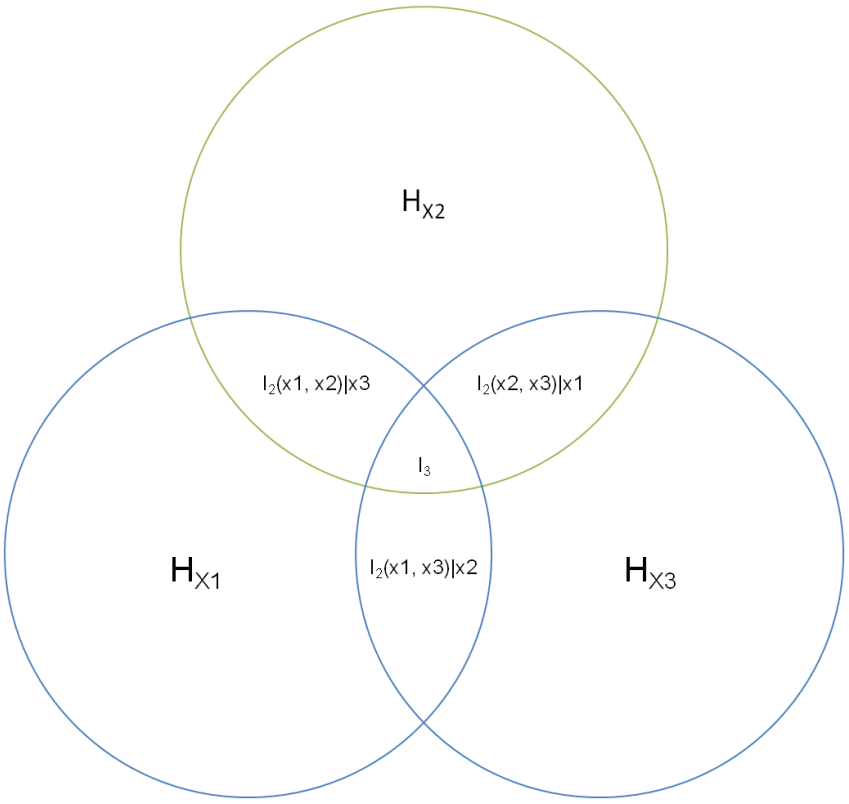


Figure 19. The 3-body information Venn diagram.

In a 3-body system, the co-information between three variables is the 3-way intersect, denoted as $I_3(X_1,X_2,X_3)$. Blue circles denote the transmitter and receiver, whereas the green circle denotes the channel.

It is important to discuss the interpretation of negative 3-body information. If X_1 and X_2 are positively correlated by direct interaction, but X_2 is positively correlated to X_3 while X_1 is negatively correlated to X_3 (both by direct interaction), the information shared by X_1 and X_2 is diminished due to their interaction with X_3 for certain parameters (for example, when the correlation between X_1 and X_2 is 0.1, the correlation between X_1 and X_3 is -0.7, and the correlation between X_2 and X_3 is 0.7). While this can occur in allosteric biomolecular systems, we have found it to be rare in our applications. In fact, it appears to occur when data is limited (data not shown), which may indicate that it is an indication of poor sampling.

When calculating the 3-body information in order to analyze whether some body acts as a channel for the information transmission between two others, we will refer to the 3-body information as *co-information*. In order to compare co-body information and quantify the potential fraction contribution of a channel variable to information transmission between two other residues, we calculate the normalized co-body information, defined as:

$$\overline{I_3\left[p(x,y,z)\right]} = \frac{I_3\left[p(x,y,z)\right]}{I_2\left[p(x,y)\right]} * 100\% \tag{1.145}$$

In this normalized form, the third variable (in this case z) is specifically taken to be a potential channel for information transmission between the first two variables (x and y). This normalization is useful as it allows for a distinction between potential channel topologies. If x and y are conditionally independent given z, then:

$$I_2\left[p(x,y|z)\right] = 0 \tag{1.146}$$

and

$$\overline{I_3\left[p(x,y,z)\right]} = 100\% \tag{1.147}$$

However, unless all three variables are maximally dependent, x and z will not be conditionally independent given y, and y and z will not be conditionally independent given x. Thus,

$$\begin{aligned} \overline{I_3\left[p(x,y,z)\right]} &> \overline{I_3\left[p(x,z,y)\right]} \\ \overline{I_3\left[p(x,y,z)\right]} &> \overline{I_3\left[p(y,z,x)\right]} \end{aligned} \tag{1.148}$$

The property in (1.148) is desirable, as the co-information itself is invariant to the order of the variables and thus cannot on its own reveal anything about the topology. In order to illustrate the power of using the normalized co-information to identify allosteric channels, we designed symmetric $K_{1,4}$ networks of coupled univariate normal distributions (see **Figure 20**, top left). Each of the three illustrative systems - weak, moderate, and strong - is defined by a covariance matrix with the diagonal elements (the variances) equal to 1. In the weak system, the covariance between directly coupled distributions, r_{direct}, is 0.25 and the covariance between indirectly coupled distributions, $r_{indirect}$, is 0.0625. In the moderate system, $r_{direct} = 0.5$ and $r_{indirect} = 0.25$, and in the strong one, $r_{direct} = 0.75$ and $r_{indirect} = 0.2625$. For each system, inverting the covariance matrix produces 0 in all the elements corresponding to interactions between indirectly coupled distributions, which indicates that they are conditionally independent as intended.

We sampled the complete multivariate distribution of the symmetric $K_{1,4}$ networks with observations ranging from 200 to 10000 (in multiples of 200), each with ten realizations, using the mvrnorm function within the R package MASS. We then tested how well one could differentiate node 1, which is the true channel between nodes 2 and 3, from node 4, which is a false channel, by using the normalized co-information.

The results are summarized in **Figure 20**, showing that even for the weak system, where the true indirect correlations are lower than would traditionally be considered for investigation, the normalized co-information can be used to determine the true channels from the false channels if one has over 8000 observations of the systems. For the moderate system, channels can be identified with fewer than 2000 observations.

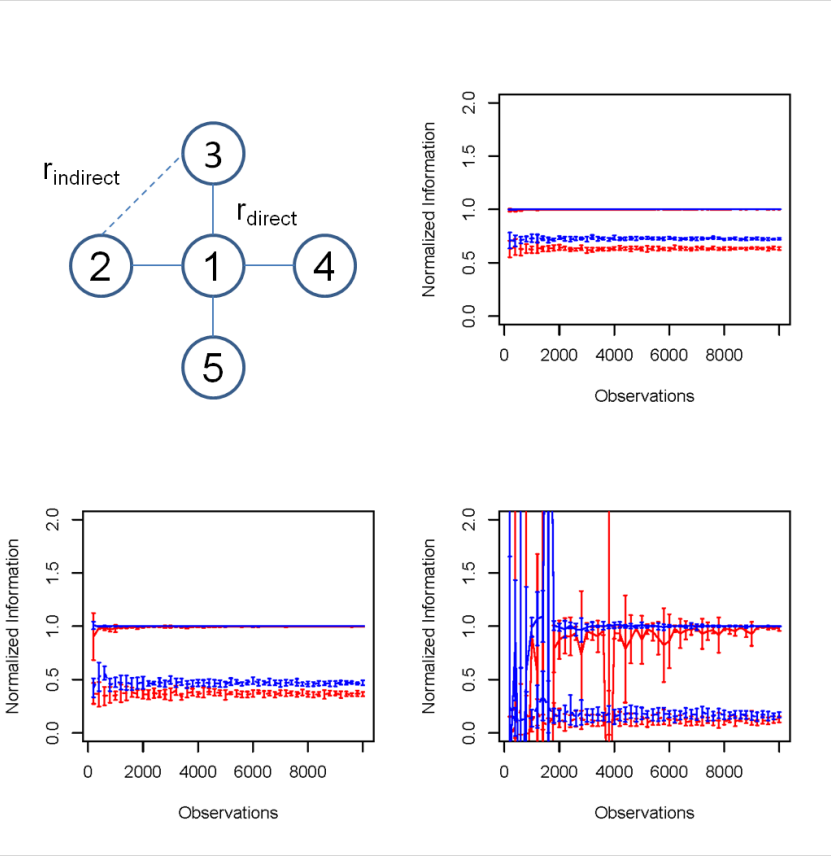


Figure 20. Co-information and Mutual Coordination Information can identify channels in $K_{1,4}$.

Top left: The $K_{1,4}$ network that serves to illustrate the ability of these measures to discriminate between true and false channels. Each circle is a node and the connecting lines are the edges. Edges represent direct interactions with covariance r_{direct}, and all nodes that are not connected by an edge display indirect interactions with covariance $r_{indirect}$. Top right: Figure shows the results from separate calculations with different numbers of observations, of normalized co-information (red) and normalized mutual coordination information (blue) in the strong $K_{1,4}$ network. True channels are shown in solid lines and false channels are shown in dashed lines, and bars represent the standard deviation of 10 realizations. Bottom left: Same as top right, for the moderate $K_{1,4}$ network. Bottom right: Same as left panel, for the weak $K_{1,4}$ network.

However, more complex information transmission networks may be imagined than the simple 3-body system for which co-information applies. For example, more than two domains may display coupled motions, and these coupled motions may be due to a central channel. While the N-body information can be used to quantify how much information in shared by all members in a set, there may be many different collective motions existing involving different subsets of the full set of N. To quantify overall amount of information shared by N bodies, including all possible n-body coupled motions ranging from n=2 to n=N, we can calculation the total correlation, also known as the multi-information:

$$\mathrm{TC}\left[p(x_1,x_2,\ldots,x_N)\right]=\sum_{i=1}^{N}\mathrm{H}\left[p(x_i)\right]-\mathrm{H}\left[p(x_1,x_2,\ldots,x_N)\right] \tag{1.149}$$

We can generalize the co-information to describe how much information that is shared by a set of variables of arbitrary size is also shared with another variable. This is calculated as the difference between the TC and the conditional TC, which we will call the *coordination information, (CI)*:

$$\mathrm{CI}\left[p(x_1,x_2,\ldots,x_N),p(y)\right]=\mathrm{TC}\left[p(x_1,x_2,\ldots,x_N)\right]-\mathrm{TC}\left[p(x_1,x_2,\ldots,x_N|y)\right] \tag{1.150}$$

This contribution describes the amount of total correlation in a set of variables (the "coordinated set") that is shared with a variable (or multivariate distribution) that is not included in the coordinated set ("the coordinator"). When calculated in this manner, CI describes the contribution of a site to all possible n-body correlations within another site. We can define the *normalized coordination information* (NCI), analogous to the normalized co-information, in which the coordination information is normalized to the total correlation within the coordinated site:

$$\overline{\mathrm{CI}\left[p(x_1,x_2,\ldots,x_N),p(y)\right]}=\frac{\mathrm{CI}\left[p(x_1,x_2,\ldots,x_N),p(y)\right]}{\mathrm{TC}\left[p(x_1,x_2,\ldots,x_N)\right]} \quad (1.151)$$

It should be noted that *coordinators* are not all *coordination channels*. *Coordinators* can be coupled to *coordination channels*, and thus perturbation to the *coordinator* leads to a perturbation in the coordinated set. In order to define channels that mediate coordination information, we calculate the amount of coordination information that is shared between two residues and the same set, which we call *mutual coordination information, (MCI)*:

$$\mathrm{MCI}\left[p(x_1,x_2,\ldots,x_N),p(y),p(z)\right]=\mathrm{CI}\left[p(x_1,x_2,\ldots,x_N),p(y)\right]-\mathrm{CI}\left[p(x_1,x_2,\ldots,x_N),p(y)|p(z)\right] \quad (1.152)$$

The mutual coordination information can also be normalized:

$$\overline{\mathrm{MCI}\left[p(x_1,x_2,\ldots,x_N),p(y),p(z)\right]}=\frac{\mathrm{MCI}\left[p(x_1,x_2,\ldots,x_N),p(y),p(z)\right]}{\mathrm{CI}\left[p(x_1,x_2,\ldots,x_N),p(y)\right]} \quad (1.153)$$

Using the $K_{1,4}$ network, we demonstrate how well one could differentiate node 1 - the true coordination channel for the coordination of nodes 2 and 3 by node 5 - from node 4, a false coordination channel, using *mutual coordination information*. We find results similar to those of the *co-information*, indicating that the mutual coordination information is also a good tool for identify allosteric channels.

Of additional interest is the study of rigid bodies and rigid-body-like behavior. Because much of the dynamics in proteins is often considered qualitatively in terms of "rigid body motions", it is of interest to study such behavior, and rigid-body-like behavior, in the context of information. In general, a rigid body is considered to be a solid body in which internal deformations can be neglected. But formal description of what constitutes a rigid body in a molecular system, which can be independent of the particular physics of the system of interest, is lacking. Nevertheless, the qualitative

description implies specific constraints on the conformational entropy and N-body information of the system. We can consider as an example a sphere of densely packed atoms. The coordinate of any atom in this sphere after a translation can be determined by measuring only one atom. Thus, for any atom i:

$$H\left[p_{\text{translation}}\left(x_1,x_2,\ldots,x_N\right)\right]=H\left[p_{\text{translation}}\left(x_i\right)\right] \tag{1.154}$$

As all atoms have the same translational entropy:

$$H\left[p_{\text{translation}}\left(x_1,x_2,\ldots,x_N\right)\right]=H\left[p_{\text{translation}}\left(x_i\right)\right]=I_N\left[p_{\text{translation}}\left(x_1,x_2,\ldots,x_N\right)\right] \tag{1.155}$$

Thus, a system is maximally rigid in regard to translations if (1.155) is true. In the same sense, as the constituents atoms can be considered points in space, the coordinate of any atoms after a rotation can be determined, given the axis of rotation, by measuring only one atom. Thus, ignoring atoms that lie perfectly on the axis of rotation:

$$H\left[p_{\text{rotation}}\left(x_1,x_2,\ldots,x_N\right)\right]=H\left[p_{\text{rotation}}\left(x_i\right)\right]=I_N\left[p_{\text{rotation}}\left(x_1,x_2,\ldots,x_N\right)\right] \tag{1.156}$$

If the rotations and translations are taken to be independent, the total entropy is then:

$$H\left[p\left(x_1,x_2,\ldots,x_N\right)\right]=I_N\left[p\left(x_1,x_2,\ldots,x_N\right)\right]=I_N\left[p_{\text{rotation}}\left(x_1,x_2,\ldots,x_N\right)\right]+I_N\left[p_{\text{translation}}\left(x_1,x_2,\ldots,x_N\right)\right] \tag{1.157}$$

Thus, an intuitive quantification of the rigid-body dynamics of a system, R, can be written as:

$$R\left[p\left(x_1,x_2,\ldots,x_N\right)\right]=\frac{I_N\left[p\left(x_1,x_2,\ldots,x_N\right)\right]}{H\left[p\left(x_1,x_2,\ldots,x_N\right)\right]} \tag{1.158}$$

However, R is not useful for a continuous distribution, in which the entropy is infinite for all non-Dirac delta distributions[162], and the commonly used differential entropy

can be negative. Instead, it is useful to normalize to a strictly positive quantity that is also strictly greater than the N-body information and equal to the N-body information when the system is maximally rigid. One such quantity is the average 2-body information, and thus we can calculate the rigid-body fraction, RBF, as:

$$\mathrm{RBF}\left[p(x_1,x_2,\ldots,x_N)\right]=\frac{I_N\left[p(x_1,x_2,\ldots,x_N)\right]}{\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}\frac{N^2-N}{2}I_2\left[p(x_i,x_j)\right]} \tag{1.159}$$

However, the RBF can be problematic. Given a system composed of one rigid body, adding a single particle to the system that is independent of the existing rigid body results in a RBF of 0. In general, it would be useful to have a measure that could quantify rigid-body behavior in a system of multiple, possibly coupled, rigid bodies. To do this, one can consider the mutual information expansion of the entropy:

$$H\left[p(x_1,x_2,\ldots,x_N)\right]=\sum_{i=1}^{N}H\left[p(x_i)\right]+\left(\sum_{j=2}^{N}-1^{j-1}\sum\frac{I_j}{\binom{N}{j}}\right) \tag{1.160}$$

where the sum over I_j is the sum of all j-body information terms. Given that:

$$I_N\left[p(x_1,x_2,\ldots,x_N)\right]\leq\min\left(I_{N-1}\right) \tag{1.161}$$

then

$$I_N\left[p(x_1,x_2,\ldots,x_N)\right]\leq\sum\frac{I_{N-1}}{\binom{N}{N-1}}\leq\sum\frac{I_{N-2}}{\binom{N}{N-2}}\leq\ldots\leq\sum\frac{I_2}{\binom{N}{2}} \tag{1.162}$$

Thus, the curve representing the average n-body information as a function of n is strictly decreasing. In fact, for a model system consisting of a finite one-dimensional lattice of one-dimensional normal distributions with unit variance and uniform

covariance between neighbors, we find that an approximately exponential decay of the average n-body information is expected for a range of covariances (see **Figure 21**). Additionally, adding heterogeneity by modifying some of the covariances did not change the decay (see **Figure 21**).

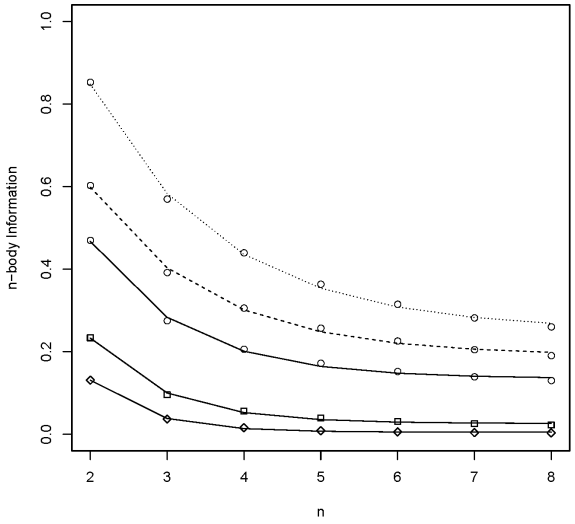


Figure 21. Approximately exponential decay of the average n-body information in model 1-dimensional lattices of coupled 1-dimensional normal distributions. Covariance between neighbors are 0.7 (diamond), 0.8 (square), and 0.9 (circle, solid line). The dashed and dotted lines correspond to systems where the fourth and fifth distributions have greater covariance with their neighbors (0.95 and 0.99, respectively). All lines are parameterized as exponential decays using Eq. (1.163).

Thus, in order to describe the average n-body information term as a function of n from 2 to N, we can parameterize a function with the following exponential form:

$$\langle I_n \rangle = Ae^{\frac{-(n-2)}{CO}} + B \tag{1.163}$$

By parameterizing the exponential function, we calculate the *correlation order*, CO, which describes the rate of decay of n-body correlations in the system. A CO of 1 would indicate that the average nat (the unit of information when the natural logarithm is used) of n-body information is contributed by an n + 1 body correlation. In the model system described above, if the correlations between neighboring distributions are low, the exponential will decay quickly and have a low CO, and if the correlations between neighbors are high, the exponential decays slowly due to the emergence of higher correlations and have a high CO.

Finally, we must address an issue regarding the normalization of the mutual information between multivariate continuous distributions through the lens of two rigid bodies. The generalized correlation coefficient, shown in (1.140), normalizes the mutual information between d-dimensional distributions by dividing the information by d. However, the mutual information is not expected to scale linearly with the additional of new dimensions. Consider two rigid bodies, each composed of some number of atoms. If they behave as perfectly rigid bodies, their translational entropy does not increase as new atoms are added to the rigid bodies, while their dimensionality clearly will. Thus, as the number of constituent atoms in the rigid bodies approaches infinity, the generalized correlation coefficient describing their translations will approach 0, even in the case where their translations are perfectly coupled. To remedy this, we developed a new quantification of the information shared between two multivariate distributions, which can be normalized similar to the generalized correlation coefficient. We begin with two multivariate distributions, X and Y.

$$\begin{aligned} X &= \left\{ X_1, X_2, \ldots, X_{d_x} \right\} \\ Y &= \left\{ Y_1, Y_2, \ldots, Y_{d_x} \right\} \end{aligned} \tag{1.164}$$

We start by calculation the joint total correlation:

$$\mathrm{TC}\left[\mathrm{p}\left(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{\mathrm{d}_x},\mathrm{y}_1,\mathrm{y}_2,\ldots,\mathrm{y}_{\mathrm{d}_x}\right)\right]=\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{H}\left[\mathrm{p}\left(\mathrm{x}_\mathrm{i}\right)\right]+\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{H}\left[\mathrm{p}\left(\mathrm{y}_\mathrm{i}\right)\right]-\mathrm{H}\left[\mathrm{p}\left(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{\mathrm{d}_x},\mathrm{y}_1,\mathrm{y}_2,\ldots,\mathrm{y}_{\mathrm{d}_x}\right)\right] \quad (1.165)$$

However, this measure of information also includes information shared with X and within Y that is not shared between X and Y. Thus, we subtract out that excess information to result at the total intercorrelation:

$$\mathrm{TC}\left[\mathrm{p}(\mathrm{x},\mathrm{y})\right]=\mathrm{TC}\left[\mathrm{p}\left(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{\mathrm{d}_x},\mathrm{y}_1,\mathrm{y}_2,\ldots,\mathrm{y}_{\mathrm{d}_x}\right)\right]-\mathrm{TC}\left[\mathrm{p}\left(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{\mathrm{d}_x}\middle|\mathrm{y}_1,\mathrm{y}_2,\ldots,\mathrm{y}_{\mathrm{d}_x}\right)\right]-\mathrm{TC}\left[\mathrm{p}\left(\mathrm{y}_1,\mathrm{y}_2,\ldots,\mathrm{y}_{\mathrm{d}_x}\middle|\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{\mathrm{d}_x}\right)\right]$$

(1.166)

The *total intercorrelation* describes the total amount of information shared between two multivariate distributions through any n-body correlation that contains at least one dimension from both distributions, and equals the 2-body information between X and Y in the univariate case. *Total intercorrelation* is distinctly different from the (d_x + d_y)-body *information* as it counts the n-body information between dimensions of A and B that is not shared by all dimensions of both A and B.

For illustration, we discuss a system of two atoms where the marginal entropy of each dimension of each atom has been standardized to H. If all dimensions share maximum information, then:

$$\max\left(\mathrm{TC}_{\mathrm{INTER}}\left[\mathrm{p}(\mathrm{x},\mathrm{y})\right]\right)=\max\left(\mathrm{TC}\left[\mathrm{p}\left(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{\mathrm{d}_x},\mathrm{y}_1,\mathrm{y}_2,\ldots,\mathrm{y}_{\mathrm{d}_x}\right)\right]\right)=\left(\mathrm{d}_x+\mathrm{d}_y-1\right)\mathrm{H} \quad (1.167)$$

Thus, the maximum *total intercorrelation* can be easily scaled to account for dimensionality. Since in the one-dimensional case, *total intercorrelation* is equivalent to 2-body information, we can write an intercorrelation coefficient that is analogous to the generalized correlation coefficient:

$$r_{inter}\left[p(x,y)\right] = \sqrt{1 - e^{-\frac{2}{(d_x + d_y - 1)} TC_{INTER}\left[p(x,y)\right]}} \tag{1.168}$$

However, it should be noted that the total intercorrelation behaves differently from the mutual information in important ways. The total intercorrelation maximizes when X and Y are perfectly coupled rigid bodies, whereas the mutual information maximizes when X and Y have no rigid-body like behavior but are perfectly coupled to each other.

Lastly, the domains involved in allostery, such as ligand binding sites, are generally composed of many residues. Thus, it is desirable to be able to differentiate between residues that are involved in information transmission with other sites (we will refer to these residues as "communicators" versus those that may be essential to the internal dynamics of its constituent site (we will refer to these residues as "stabilizers").

To identify residues that contribute significantly to information measures, we calculated the contribution of a variable x to an arbitrary information metric describing a set of variables X that contains x, which we will denote as I[f(X)], as:

$$\text{Contribution}\left(I\left[f(X)\right], x\right) = \frac{I\left[f(X)\right] - I\left[f(X|x)\right]}{I\left[f(X)\right]} \tag{1.169}$$

Lastly, we described a method that can be used to identify residues that play a significant role as a channel. To identify the residues with high *co-information* (or mutual coordination information), we take advantage of an empirically observed relationship between the *co-information* and the *co-information rank*. We find that when plotting *co-information* against the *co-information rank* for an arbitrary pair of sites, the midsection contains a large linear region, surrounded with a large co-information extreme on the left and a small co-information extreme on the right. To

identify a cut-off for the high *co-information* extreme region, we calculate a linear fit to the middle residues (the exact number determined on a case-by-base basis) and calculate the root mean squared residual (RMSR) for the fit. We then project the fit across all residues, and define the residues with high *co-information* as those that have a residual of greater than 1 RMSR. An example distribution and the corresponding fit and cut-off are shown in **Figure 22**. This type of method has been used previously [163] to identify residues that are important for function from multiple sequence alignments.

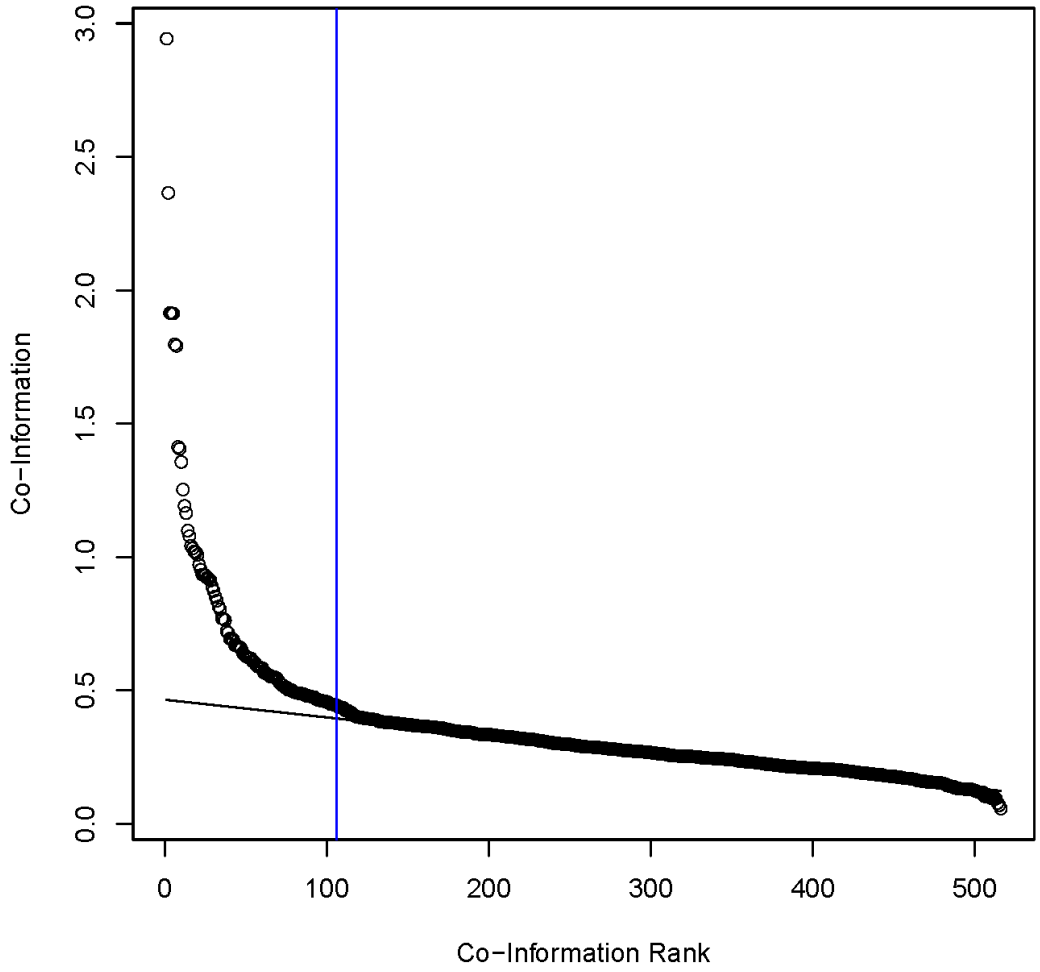


Figure 22. The typical co-information plot.

The co-information for a given residue (the channel) with the INI (receiver) and S1 (transmitter) is plotted against the co-information rank of that residue (black circles). The black line is the linear fit to the middle 200 residues and the blue line is the cut-off for high co-information residues.

2.3. Random Forest-Based Identification of Ligand-Specific Allosteric Modulation

2.3.1. Motivation for Method

While residues involved in function are often known from detailed biochemical and pharmacological analysis, their responses to ligand-specific allosteric modulation are hard to characterize experimentally outside of the use of structural method such as x-ray crystallography. However, both the crystallization of specific ligand-protein complexes and the determination of high resolutions structures from these crystals are non-trivial undertakings. Even when structures are available, modulation that affects either the distribution or dynamics of the protein, such as the local flexibility of loops or the frequency of interactions, are difficult or impossible to quantify from single structures. To identify these differences, estimation of the ensemble through MD is commonly employed.

2.3.2. Previous Work

Many methods have been proposed to differentiate ensembles under different conditions. These methods sometimes utilize measures from probability theory and information theory, such as the Kullback-Leibler divergence[164,165] and the related Jensen-Shannon divergence[166,167], or methods from machine learning such as support vector machines[168]. While these methods are theoretically rigorous, it is beneficial to utilize a method that specifically analyzes characteristics of the protein that can be modulated experimentally through mutagenesis, such as pairwise interactions. The dimensionality of the pairwise interaction space can be very high and cannot be estimated using quasi-harmonic approximations of the distributions. However, it has been shown that rather simple statistical analysis of the differences in pairwise

interaction frequencies between simulations with and without allosteric modulators or mutations that have allosteric effects can reveal interactions that are likely to compose the allosteric interaction networks (AINs) that propagate long-distance conformation change[134]. Still, these statistical differences are not guaranteed to be statistical differences related to allosteric modulation; they may be statistical differences that would also appear if two MD simulations of the same system were run and neither reached the true equilibrium distribution, which is expected to be common even as we reach the micro- and millisecond times scales. Since the number of simulation is finite even when multiple simulations are performed, the proper reweighting procedure is required to ensure that an outlier simulations trapped in a low probability region of conformational space does not bias the entire analysis.

This problem is further amplified when the goal is to identify class-specific statistical differences rather than ligand-specific differences. The naïve approach would be to combine the simulations of ligands of the same classes, and compare these agglomerated simulations. However, due to this implicit averaging, the naive approach is sensitive to mistaking ligand-specific interactions for class-specific interactions. For example, imagine that one wishes to compare the allosteric modulation of interaction frequencies by three ligands of class A to that of three ligand of class B. If an interaction is formed in two of the ligands in class A and none of class B, pooling the data for class A and class B ligands will result in a significant difference in the interaction frequency between class A and class B. However, as one ligand in class A does not allosterically induce that interaction, it cannot actually be a class-specific interaction. *Thus, the problem must be divided into two steps: first, ligand-specific interaction frequencies within the classes must be identified and removed, and then the*

remaining interaction frequencies can be compared between classes. To solve this two-step problem, we developed a two-step method utilizing random forests.

2.3.3. 2-Step Random Forest Identification of Class-Specific Features

We treat the problem of determining which interaction frequencies display class- or ligand-specific modulation as a classification problem, in which we would like determine which interactions are most useful in classifying which ligand or class of ligand is likely to be bound to a given protein structure. As interaction frequencies deal with binary variables, *decision trees* become a powerful tool for this classification problem.

A *decision tree* is a tree-like graph in which each internal node denotes a splitting process that divides the given attribute by categories: each branch represents the categorical output of the preceding node, and each terminal leaf represents a class label. In order to classify a data point using a decision tree, one begins at the first internal node (called the root) and traverses the decision tree until a leaf is reached, which outputs the class prediction for that data point. While there are many methods for training decisions trees, most methods are known to be sensitive to over-fitting the training data. To overcome this over-fitting problem, random forest methods build an ensemble of decisions trees using different subsets of the training data, and then outputting a class prediction using the mode classification across all trees. Random forests have been shown to be less sensitive to over-fitting and outperform single decision trees.

However, we are not primarily interested in classification per se. Instead, we would life to first identify interactions that are most important for the classification of structures by ligands of the same class. To do so, we perform random forest

classification and then calculate a normalized variant of the variable importance for each interaction. The variable importance describes the contribution of a given variable to a random forest, and is most often described by the mean decrease in accuracy across all trees in the random forest when a predictor variable in randomly permuted and the accuracy is re-calculated. The accuracy, A, is calculated as:

$$A(d) = \frac{n_{correct}}{n_{total}} \tag{1.170}$$

where d is a set of data points, $n_{correct}$ is the number of correctly classified data points, and n_{total} is the total number of data points.

Thus, the mean decrease in accuracy for a variable x, <dA>x, is calculated as:

$$\langle \Delta A \rangle_x = \frac{\sum_{i=1}^{N} A_i(d) - A_i(d_x)}{N} \tag{1.171}$$

where A_i is the accuracy of a given decision tree, N is the number of decision trees in the random forest, and dx is the set of data points with variable x permuted. Another measure, the mean decrease in node impurity, is also used but is difficult to normalize across variable of differing number of categories. For this reason, we chose to use the mean decrease in accuracy to quantify the role of an interaction in predicting the ligand or class of ligand bound to a given structure.

To normalize the decrease in accuracy, we use Cohen's kappa, k.

$$\kappa = \frac{A - A_{exp}}{1 - A_{exp}} \tag{1.172}$$

where A_{exp} is the expected accuracy if data points were classified by random given the frequency of that class in the data set. The value of κ is 1 if classification is perfect,

and 0 if classification is no better then random assignment. Thus, we write a normalized variable importance as:

$$\Delta\kappa_x = \frac{\langle \Delta A \rangle_x}{1 - A_{exp}} \tag{1.173}$$

To solve the first step of the problem, we first pool the trajectory data by class, and then we then build for each class a random forest to classify the structures by the ligand, rank the interactions by Dk, and remove all interactions that exceed some threshold, which we will call k. After doing this for each class, we pool the classes together, remove their ligand labels, and classify the structures by class with the remaining interactions. All interactions with Dk > k are then deemed to be important for the classification of class but not ligand, and thus are likely to be signatures of hallucinogen-specific allosteric modulation.

One difficulty is then choosing the cut-off for importance, k. The goal of the analysis is to find the interactions that best classify the ligand class and not the individual ligands. Thus, it seems reasonable that enough interactions should be removed such that, given the reduced set of interaction data, a random forest for predicting class outperforms random forests for predicting the ligands within class. To identify when class prediction outperforms ligand prediction, we calculate the ratio of the Cohen's kappa for classification by ligand class over the average k for classification of ligands within each class:

$$\alpha = \frac{\kappa_{class}}{\langle \kappa_{ligand} \rangle} \tag{1.174}$$

While ideally α should exceed 1, it is not necessarily the case that the classification of structures by ligand class will be highly accurate. One could imagine that there exist

two active states of the receptor, one that results in hallucinogenic activity and one that mediates all other activities, hallucinogens may increase the probability of the hallucinogenic state more than the non-hallucinogenic ligands, but the resulting ensembles will still be mixtures of the two states and thus significantly overlapping. Thus, we test many values of k and choose the cut-off that maximized α.

3. Application to Membrane Protein Systems

While the models and methods described above may be of interest to theoreticians, the merit of their development is determined by their ability to help us describe and understand real systems. Below, the use of several of these methods will be illustrated in the context of membrane transporters and GPCRs.

3.1. Allostery in the Transport Mechanisms of LeuT

Much of the content in this section has been adapted with permission from [110].

The prototypical member of the family of neurotransmitter:sodium symporters (NSS), the bacterial transporter LeuT has been particularly well studied, and the results from many experimental and computational investigations suggest that transport is driven by a complex allosteric mechanism spanning the entire length of the transporter. From single molecule FRET (smFRET) experiments carried out on LeuT, a number of transport-related structural transitions were identified in the intracellular gate region that occludes the substrate from the cytoplasm[169], and these were shown to be modulated by binding events at the extracellular end [84,170]. Crystallographic studies have also revealed that a second binding site in the extracellular vestibule (termed S2) is the target of several transport inhibitors (including many of the psycho-active drugs acting on the cognate NSS neurotransmitter transporters) [73,171], and biochemical and computational evidence suggests that the release of substrate is allosterically connected to the binding of a second substrate in this site[68,81,93]. These results bring to light the cross talk between several allosterically coupled domains in the transport mechanism of NSS transporters, and suggest that modulation of these domains can both facilitate and hinder function. However, the smFRET and crystallographic data did not provide a basis for understanding which domains and interactions were crucial

for the observed substrate-modulated dynamics. Thus, we proposed that the NbIT analysis method would be able to identify tentative channels and crucial interactions between the channel and substrates that mediate the allosteric modulation.

Problematically, simulating the complete equilibrium distribution of LeuT, in any substrate or ion condition, is still unfeasible due to computational restrictions on the time scales of MD simulations. We reasoned that if the metastable states of two domains are coupled (e.g., if the population of the open and closed state of the intracellular gate, and/or the transitions between them, are coupled to the occupancy state of the substrate sites), their microstates were likely to also exhibit coupling (e.g., the fluctuations within the closed state of the intracellular gate would be coupled to the fluctuations within the bound state of the substrate site). It should be noted that while there is no theoretical basis that requires this to be true, it has been previosuly found that fluctuations around a single state can be indicative of dynamics that are relevant to function[172].

3.1.1. NbIT Identifies Allosteric Channels and Functional Residues in LeuT

The application of the NbIT analysis to LeuT will be presented below. The section will be subdivided into two subsections. First, the results of the analysis performed in [111] will be presented and discussed (text and figures have been adapted with permission). In the next subsection, unpublished results that are currently in preparation for submission will be presented, including follow-up computational work and experimental validation.

3.1.1.1. Methods

3.1.1.1.1. Simulations

Two separate trajectories of the same LeuT structure were analyzed with the NbIT method, denoted as $LeuT_{POPE/POPG}$ and $LeuT_{MNG-3}$. The $LeuT_{POPE/POPG}$ trajectory is a simulation of the occluded LeuT structure[75] (PDB ID 3GJD) bound to the two sodium ions and leucine, but with the octyl-glucoside (OG) detergent molecule removed, which has been described previously[173]. The $LeuT_{MNG-3}$ trajectory is for the same LeuT structure simulated in lauryl maltose-neopentyl glycol (MNG-3), a detergent known for its excellent stabilization of transmembrane proteins, including LeuT, in micellar environments[174,175]. Both simulations were run at in an NPT ensemble under semi-isotropic pressure coupling conditions and at 310 K temperature using the CHARMM27 force field with CMAP corrections for proteins[176] and CHARMM36 lipid force field[177] in NAMD 2.7[178] using the Nose-Hoover Langevin piston algorithm and PME for electrostatic interactions. The trajectories used for the analysis are from the production phase and only include the segment of the simulations after the Cα RMSD had converged. The total lengths of the equilibrated trajectories were 148 ns for $LeuT_{POPE/POPG}$ and 146 ns for $LeuT_{MNG-3}$.

3.1.1.1.2. Definition of functional residue clusters

Mechanistic and structure-function studies of LeuT as a prototypical NSS transporter have identified specific residues and structural microdomains that have significant roles in functional mechanisms. These include the binding sites for substrate and ions identified in the crystal structures [65,73,75], as well as the intracellular gate and surrounding interaction network, which has been shown to be involved in the transport mechanism [169]. We used these findings to define functional residue clusters (*frc*-s).

Specifically, we defined the S1-frc to include the substrate, leucine, and residues L25, G26, V105, Y108, F253, T254, S256, F259, S355, and I359. The NA1-frc includes the bound ion, leucine, and residues A22, N27, T254, and N286 of the Na1 binding site. The NA2-frc is composed of the second ion bound, and residues G20, V23, A351, T354, and S355 of the Na2 binding site. We defined the S2-frc as composed of L29, R30, Y107, I111, W114, F253, A319, F320, F324, L400, and D404, and the intracellular gate region as an "intracellular network of interactions", INI-frc, composed of R5, I187, S267, Y268, Q361, and D369. The locations of these sites in the LeuT structure are presented in **Figure 23**.

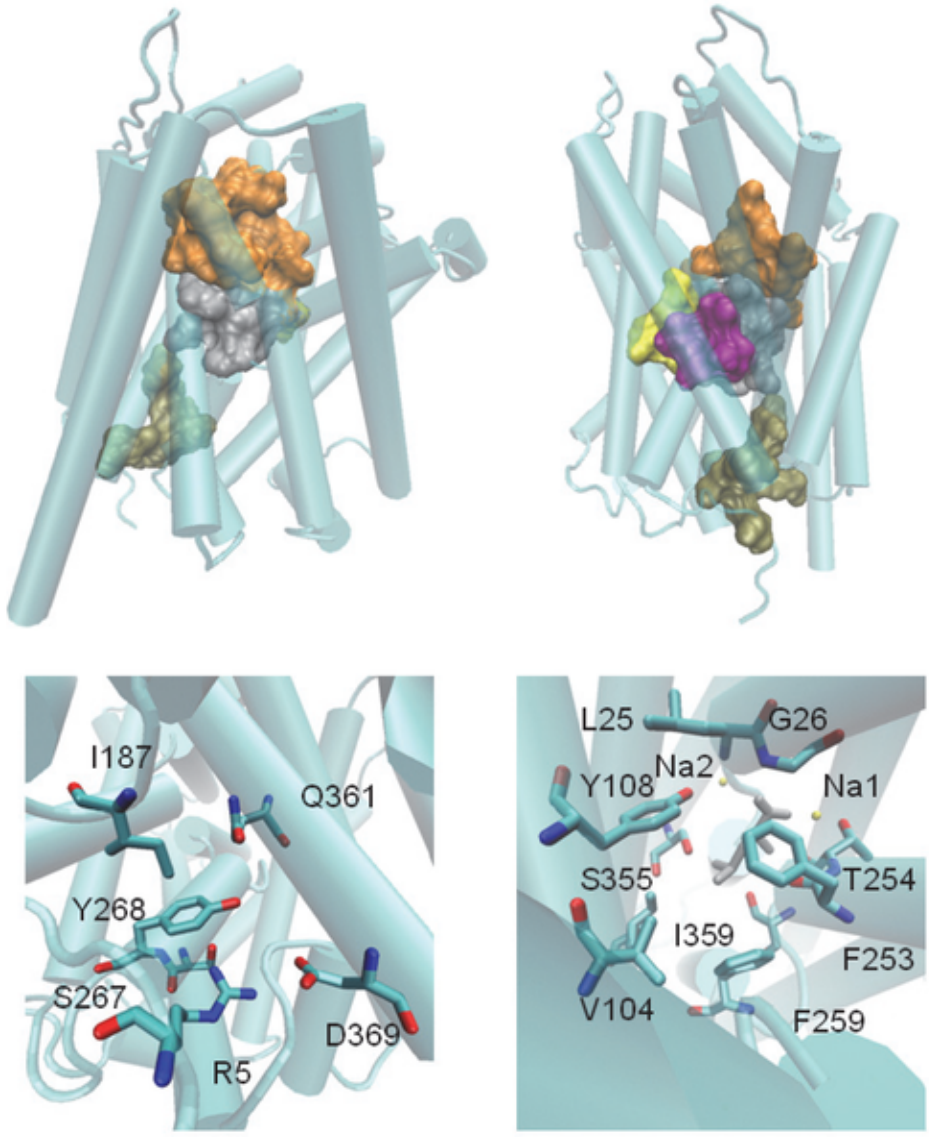


Figure 23. The structure of LeuT.

Top panels: The 3GJD crystal structure of LeuT from two perspectives. TMs are displayed as cyan cylinders connected by loops. Each frc-site is represented by an outer surface: S1 (grey), S2 (orange), INI (tan), Na1 (yellow) and Na2 (purple). Bottom left: The INI-frc; numbers refer to the residue identity. Bottom right: The S1-frc (the leucine substrate is in grey, Na2 is added for reference).

3.1.1.1.3. Correcting for symmetric side chain conformations

Some post-processing was required for the analysis by the NbIT method. In order to estimate entropy from MD simulations, the coordinate of each atom is tracked throughout the trajectory to create a distribution of Cartesian coordinates. For side chains that display symmetry (Phe, Tyr, the carboyxlate groups of unprotonated Glu and Asp), simple tracking of atoms based on their numbering in the structure file can make symmetric states appear non-symmetric. To account for this, we used a clustering algorithm to group states by dihedral angles, and then divide the states by symmetry. For Phe and Tyr, we defined the state of the ring by the dihedral angle formed by the Cα, Cβ, the benzyl carbon bound to Cβ, and a benzyl carbon para to that carbon. For Glu and Asp, the state of the carboxylate was defined as the dihedral angle formed by N, Cα, the carbonyl carbon, and a carboxylate oxygen. For each residue, the *sin* and *cos* of each angle was calculated in order to project the angles onto the unit circle. Finally, the projections were collected into two clusters using the k-means clustering algorithm (implemented in R using the kmeans function in the stats package). If the angle between the centers of the two clusters was > 90°, the position of the fourth atom was rotated by 180° relative to the plane formed by the first three atoms (as listed above) in frames from the second cluster.

3.1.1.1.4. Clustering

The MD trajectories analyzed with NbIT for this illustration of the method include only the long segments in which the interaction between R5, D369, and S267, which is observed crystallographically, is maintained. From analysis of a large number of LeuT simulations in our lab, we became aware of long-lived rearrangements in the conformation of the INI. We determined first if there were distinct substates of the

INI, by using k-means clustering on the minimum distances between side chains in the INI. Indeed, this revealed the transition between two long-lived states in the two simulations used for the NbIT analysis. Specifically, in $LeuT_{POPE/POPG}$, the system transitioned after ~118 ns from the crystal structure configuration in which R5 interacts with D369 and S267 in the INI, to a new configuration where R5 interacts with the surrounding water. In $LeuT_{MNG-3}$, the equilibrated portion of the simulation begins with R5 interacting with the D369 and S267, but after ~25 ns there is a transient rearrangement event, leading to a state in which R5 breaks away from D369, followed by a return of the INI to its original state after ~20 ns. In order to isolate these states, MD simulation trajectories were clustered by the minimum distance between non-hydrogen side chain atoms of residues within the *frc*-s using the k-means clustering algorithm. Distance time series were smoothed over 1 ns windows to minimize thermal noise, and the best clustering was taken from 100 k-means runs. We performed the same clustering analysis using each *frc* individually, and found that not only did the INI have the most conformational variability (nearly an order of magnitude greater sum of square distance between frames in comparison to the other frc-s), but clustering into two states accounted for most of the variability (see Table S2). Furthermore, we determined the similarity between results of clustering by the conformation of a specific *frc* versus all *frc*-s, by calculating the overlap as:

$$\text{overlap} = \frac{\text{occluded}_{\text{frc}} \bigcap \text{occluded}_{\text{all}}}{\text{occluded}_{\text{frc}} \bigcup \text{occluded}_{\text{all}}} \tag{1.175}$$

where $\text{occluded}_{\text{frc}}$ corresponds to the set of frames in the occluded state when clustered by a given frc, whereas $\text{occluded}_{\text{all}}$ corresponds to the set when clustered by all *frc-s*. We find that clustering by all residues in the *frc*-s of interest provided a near identical result to clustering specifically by the INI. These results indicate that the INI

rearrangement is the only significant rearrangement of a structural motif that takes place in the simulation trajectories. As the interaction between R5, D369, and S267 is observed crystallographically, we focused the study herein on comparing only this state from both simulations, in trajectories of over 100 ns from each simulation. While it might be interesting eventually to study as well the minor states of the INI not observed crystallographically, in which the gate is broken, these were not sampled sufficiently in either trajectory and thus are not yet adequate for rigorous analysis.

3.1.1.1.5. Entropy estimations

In order to estimate the configurational entropy from the MD simulations, we first approximated the probability distributions of the atomic coordinates as a 3N-dimensional multivariate normal distribution. The probability density function of a multivariate normal distribution is:

$$p(\vec{x}) = \frac{e^{-\frac{1}{2}(\vec{x}-\langle\vec{x}\rangle)^T C^{-1}(\vec{x}-\langle\vec{x}\rangle)}}{\sqrt{(2\pi)^k |C|}} \tag{1.176}$$

where k is the rank of the covariance matrix. Covariance matrices were calculated using **carma**[179]. The entropy of the continuous multivariate normal distribution can be calculated analytically through the differential entropy:

$$H\left[p(\vec{x})\right] = \frac{1}{2}\log\left(2\pi e|C|\right) \tag{1.177}$$

It should be noted that while the multivariate normal distribution is the maximum entropy distribution given constraints on the mean and covariance matrix, it can be a very poor estimator of the entropy if the distribution is multi-modal, and while the multivariate normal approximation of the mutual information between two distributions has been claimed to define the lower bound on the true mutual

information, it has been shown to be untrue. Thus, care must be taken when using the multivariate approximation. However, calculating the configuration entropy through more rigorous means is computationally intensive and not reasonable in the case of calculating the entropy of different subsets of the space many times as is done when performing NbIT analysis.

3.1.1.2. Results

3.1.1.2.1. The Pairwise Mutual Information

The analysis of pairwise *mutual information* for each of the functional residue clusters (*frc*-s) is summarized in Table 1. The calculated values show that the component residues in each of the *frc*-s exhibit coupled motions within the leucine-bound state studied here, as indicated by the mutual information that is greater than zero. Note, however, that it is difficult to compare the strength of coupling between two different sets of *frc*-s, because mutual information cannot be easily normalized from differential entropies calculated from multivariate normal distributions (see Section 2.2.3. N-body Information Theory (NbIT) Analysis). Therefore, we will not discuss further below the coupling strength between sites until we discuss other measures of information that can be normalized.

Table 1. Mutual information between known function sites in $LeuT_{POPE/POPG}$.

	S1	**S2**	**Na1**	**Na2**	**Na1, Na2**	**Na1, Na2, S1**	**Na1, Na2, S1, S2**	**INI**
S1	***-328.1 (0.5)***	23.4 (0.6)	9.3 (0.1)	7.1 (0.1)	13.2 (0.2)	X	X	12.9 (0.3)
S2	X	***-356.3 (0.7)***	14.9 (0.3)	7.5 (0.2)	21.6 (0.6)	33.0 (1.0)	X	15.1 (0.4)
Na1	X	X	***-141.2 (0.1)***	8.3 (0.1)	X	X	X	4.8 (0.1)
Na2	X	X	X	***-112.9*** **(0.1)**	X	X	X	4.0 (0.1)
Na1, Na2	X	X	X	X	***-262.4 (0.12***	X	X	8.4 (0.3)
Na1, Na2, S1	X	X	X	X	X	***-519.2 (0.9)***	X	18.1 (0.6)
Na1, Na2, S1, S2	X	X	X	X	X	X	***-869.6 (2.5)***	31.7 (1.2)
INI	X	X	X	X	X	X	X	***-136.8 (1.4)***

Off-diagonal elements correspond to the mutual information between two given *frc-s*, where as the diagonal elements correspond to the entropy of a given *frc*. Units are in nats.

3.1.1.2.2. The Communication Channel Coupling the S1-frc to the INI-frc utilizes TM6

A central mechanistic question regarding the functional dynamics of transporters is how the binding of substrate can trigger the conformational reorganization leading to the intracellular-open state from which the substrate is eventually released. Because studies have shown that just the binding of Na^+ and substrate cause measurable dynamic effects at the intracellular end of the LeuT molecule, even in the absence of transport [84,170], we sought to determine the information channel enabling this allosteric behavior. To this end, we performed *co-information analysis* to evaluate which residues played the role of channel in the information exchange between the substrate sites and the INI. Applying co-information analysis reveals that S1 and the INI are coupled through a set of residues consisting largely of residues from TM6b, TM8, and TM2 (see **Figure 24**).

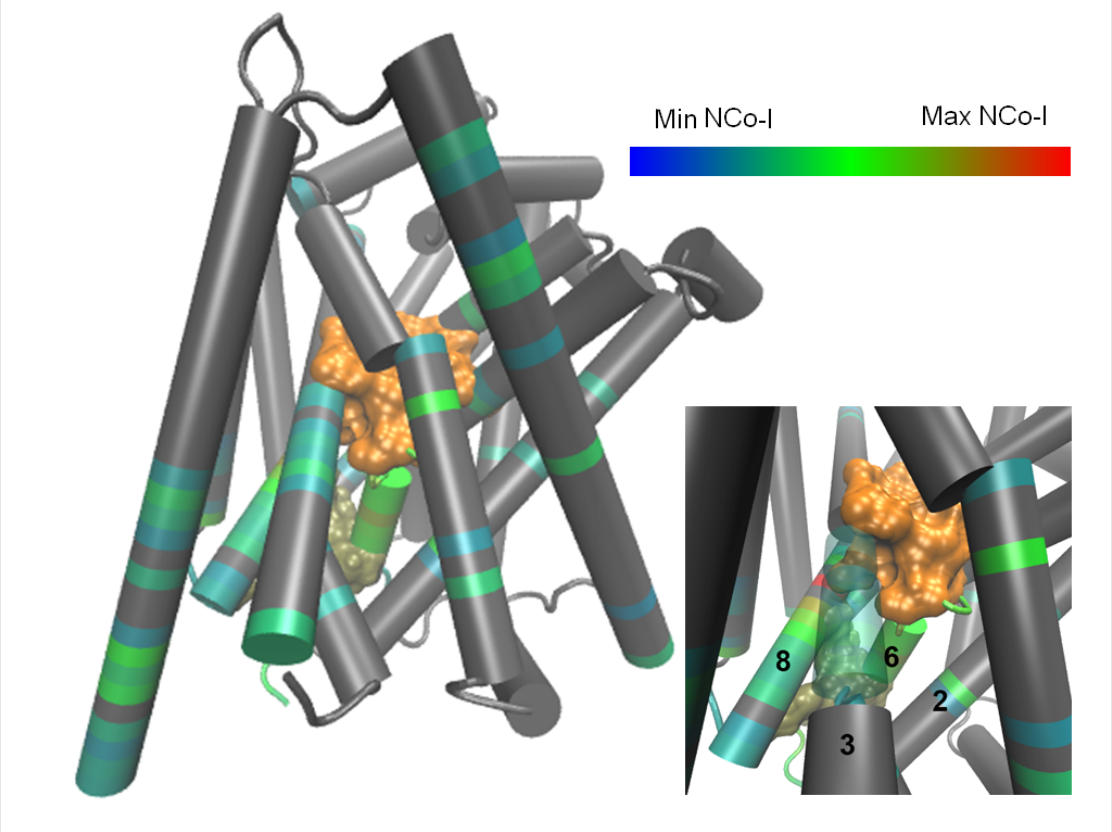


Figure 24. TMs 2, 6b, and 8 form a co-information channel between S1 and the INI.

Main: Residues found to have high *co-information* with S1 and the INI are colored by their calculated *normalized co-information* (NCo-I) values using the scale at the top right, where the Min and Max NCo-I refer to the minimum and maximum values among all possible residues. All other residues are represented in grey. Bottom right: A close up of the TM6b and TM8 interface.

Co-information analysis also reveals a channel between S2 and the INI, which is similarly composed of residues from TM6b and TM8, in addition to residues from S1 in the unstructured region between TM6a and TM6b (see **Figure 25**).

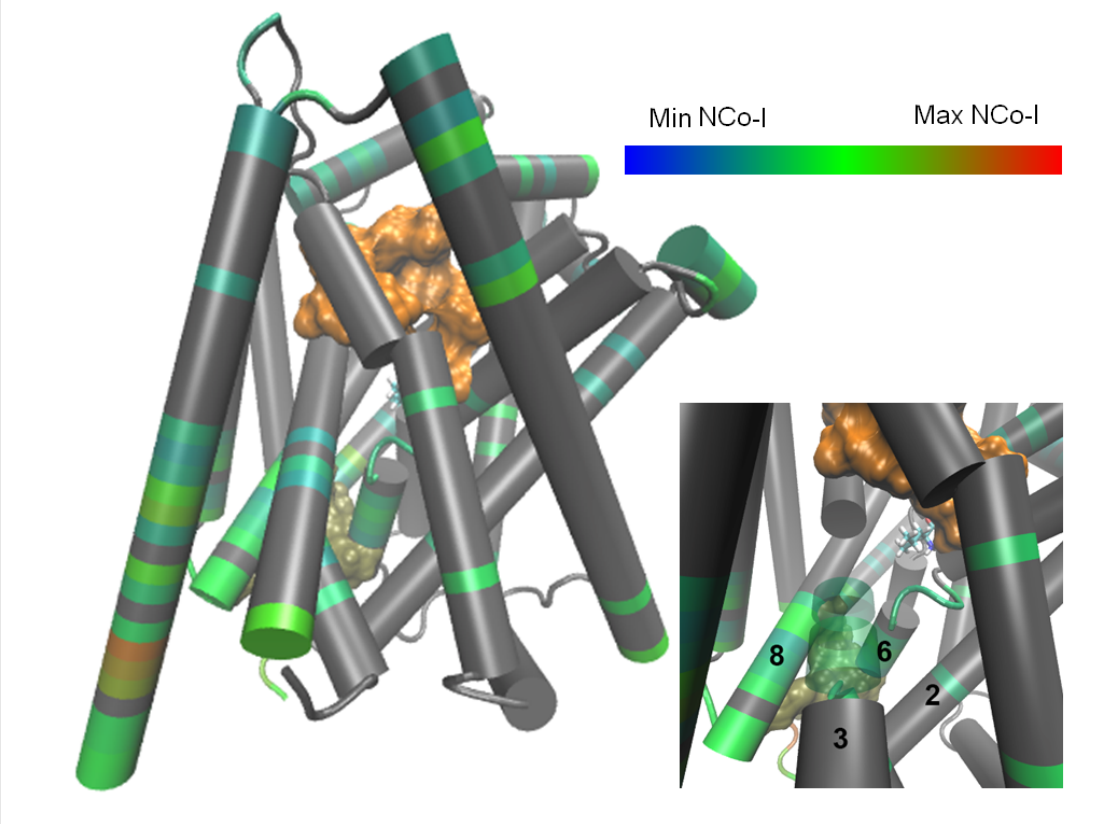


Figure 25. TMs 2, 6b, and 8 form a co-information channel between S2 and the INI.

Main: Residues found to have high *co-information* with S2 and the INI are colored by their calculated *normalized co-information* (NCo-I) values using the scale at the top right, where the Min and Max NCo-I refer to the minimum and maximum values among all possible residues. All other residues are represented in grey. Bottom right: A close up of the TM6b and TM8 interface.

Not all the residues in a particular *frc* contribute equally to the allosteric communication. In order to identify which residues within the substrate sites and the INI are essential for allosteric communication we identified the residues within these sites that made large contributions to the mutual information. Such residues contribute by coupling the sites directly to the channel, and by distributing the information

throughout the rest of their respective site. It is essential to note that the total sum of contribution from all residues does not necessarily sum to 100%. This occurs because just as the residues share information, they can also share their contribution to the mutual information, so the sum of the contribution will exceed 100%. This is also the case for other contribution measures, as described further below.

We found that for the coupling between the S1-frc and the INI, it is residues I359, F259, F253 in the S1-frc that make the largest contributions (21.2% 18.8%, and 12.5% respectively), and in the INI the largest contribution is from residues Q361, R5, and Y268 (28.3%, 21.6%, and 21.3% respectively). These very specific identifications underscore the validity of the calculated communication channel, as they are consistent with results from previous work in which mutations of I359 and F259 were shown to modulate transport efficacy [180]. Interestingly, we find that for the coupling between the S2-frc and the INI, residues R30, F324, and W114 make the largest contributions in S2 (20.1%, 12.9%, and 12.5%), and in the INI residues R5, I187, and Y268 make the largest contributions (27.1%, 23.3%, and 9.5% respectively). Because R30 is considered to form an extracellular gate with D404, the significant role we find for it here in the coupling of S2 and the INI underscores the strong relationship between the extracellular and intracellular gates. These results are summarized in Table 2**A** and Table 2**B**.

Table 2A. Specific residues highly contribute to mutual information between S1 and the INI in $LeuT_{POPE/POPG}$.

S1	**Leu**	**L25**	**G26**	**V104**	**Y108**	**F253**
	10.5%	9.9%	6.4%	8.4%	11.8%	**12.5%**
	(0.1%)	(0.0%)	(0.0%)	(0.1%)	(0.1%)	**(0.1%)**
	T254	**S256**	**F259**	**S355**	**I359**	**Na1**
	8.8%	9.3%	**18.8%**	7.7%	**21.2%**	3.0%
	(0.1%)	(0.1%)	**(0.1%)**	(0.1%)	**(0.2%)**	(0.0%)
INI	**R5**	**I187**	**S267**	**Y268**	**Q361**	**D369**
	21.6%	19.7%	14.6%	**21.3%**	**28.3%**	15.6%
	(0.3%)	(0.4%)	(0.1%)	**(0.1%)**	**(0.3%)**	(0.1%)

The contribution of specific residues in S1 (top) and the INI (bottom) to the communication between S1 and the INI (top 3 in each site are bold).

Table 2B. Specific residues highly contribute to mutual information between S2and the INI in $LeuT_{POPE/POPG}$.

S2	**L29**	**R30**	**Y107**	**I111**	**W114**	**F253**
	8.8%	**20.1%**	9.9%	7.5%	**12.5%**	10.6%
	(0.6%)	**(0.0%)**	(0.1%)	(0.1%)	**(0.1%)**	(0.1%)
	A319	**F320**	**F324**	**L400**	**D404**	
	6.1%	10.2%	**12.9%**	9.1%	8.6%	
	(0.1%)	(0.1%)	**(0.1%)**	(0.1%)	(0.1%)	

INI	R5	I187	S267	Y268	Q361	D369
	27.1%	**23.3%**	14.6%	**19.5%**	17.3%	18.2%
	(0.3%)	**(0.5%)**	(0.2%)	**(0.1%)**	(0.2%)	(0.2%)

The contribution of specific residues in S2 (top) and the INI (bottom) to the communication between S2 and the INI (top 3 in each site are bold).

3.1.1.2.3. The Coordination within frc-s is Performed by Known Functional Residues

We hypothesized that that the proper fold and specific local function of a given *frc*, such as substrate binding, are maintained through short-distance allosteric couplings underlying collective behavior among the residues in the clusters. We probed this by calculating the *total correlation* (TC) for each *frc* to obtain a measure of the total amount of information shared by a set of size *N* through any type of correlation from 2 to *N*-body. We then calculated the contribution of a given residue in the *frc* to this TC.

With this approach, we find that in the INI, the three largest contributors are Y268 (60.7%), S267 (59.0%) and R5 (42.7%). This is consistent with their central location in the INI topology and with previous reports that mutation of the highly conserved Y268 and R5 to alanine has a strong effect on the structure and dynamics of the intracellular gate [84,169]. In the S1-frc, the largest contributions to the TC were calculated to come from T254 (40.3%), the leucine substrate (38.9%), and F253 (38.9%). The bound Leu is expected to contribute strongly, as seen here, because it interacts with all other residues in S1. Furthermore, as mutation of F253 has been shown to greatly reduce binding in S1[69,170], it is possible that its role is not only to stabilize Leu binding through direct interaction, but also to stabilize the site as a whole by coordinating the rest of the S1 residues.

In the other *frc*-s we also found a small number of specific high contributions. Thus, in the Na1 site the largest contributions to the total correlation are made by the Na1 sodium ion (61.7%), T254 (60.1%), and by leucine (58.4%). Interestingly, in the Na2 site, T354 and S355 contribute significantly more (70.9% and 66.4%, respectively) than the Na^+ ion (52.1%). Finally, in S2, residues F320, A319, and R30 are found to make the largest contributions of 39.6%, 33.0%, and 31.1%, respectively. These results are summarized in Table 3.

Table 3. The contribution of specific residues to the total correlation of their sites in $LeuT_{POPE/POPG}$.

S1	**Leu**	**L25**	**G26**	**V104**	**Y108**	**F253**
	38.9%	36.2%	32.3%	13.2%	23.3%	**38.9%**
	(0.2%)	(0.2%)	(0.3%)	(0.1%)	(0.1%)	**(0.2%)**
	T254	**S256**	**F259**	**S355**	**I359**	**Na1**
	40.3%	29.1%	20.1%	13.6%	12.1%	20.2%
	(0.3%)	(0.2%)	(0.2%)	(0.2%)	(0.1%)	(0.2%)
S2	**L29**	**R30**	**Y107**	**I111**	**W114**	**F253**
	25.6%	**31.1%**	17.4%	17.4%	18.5%	10.9%
	(0.1%)	**(0.2%)**	(0.1%)	(0.1%)	(0.1%)	(0.0%)
	A319	**F320**	**F324**	**L400**	**D404**	
	33.0%	**39.6%**	20.9%	14.0%	15.0%	
	(0.3%)	**(0.3%)**	(0.1%)	(0.1%)	(0.1%)	
Na1	**Na1**	**A22**	**N27**	**T254**	**N286**	**Leu**

	61.7%	49.5%	50.0%	**60.1%**	36.3%	**58.4%**
	(0.3%)	(0.2%)	(0.2%)	**(0.2%)**	(0.2%)	**(0.2%)**
Na2	**Na2**	**G20**	**V23**	**A351**	**T354**	**S355**
	52.1%	37.6%	40.1%	38.6%	**70.9%**	**66.4%**
	(0.2%)	(0.2%)	(0.2%)	(0.2%)	**(0.1%)**	**(0.1%)**
INI	**R5**	**I187**	**S267**	**Y268**	**Q361**	**D369**
	42.7%	34.8%	**59.0%**	**60.7%**	23.8%	28.8%
	(0.4%)	(0.8%)	**(0.6%)**	**(0.4%)**	(0.4%)	(0.4%)

For each *frc*, the contribution of each residue to the total correlation is presented. The top 3 residues in each site are shown in bold.

3.1.1.2.4. Both the S1-frc and the S2-frc Coordinate Multi-Body Collective Motions in the INI

Key findings from smFRET experiments investigating the allosteric modulation of intracellular gating in LeuT [84] were that conformational changes in the intracellular gates require collective motions resulting in large spatial displacements, and that these motions are modulated (in some undetermined way) by the state of the substrate binding sites, S1 and S2 [170]. In order to investigate the role of these substrate binding sites in the collective dynamics within the INI-frc, we calculated how much each of the two binding sites contributed to the total correlation of INI. This contribution, termed here *coordination information* (CI), describes the amount of total correlation in a set of variables (the "coordinated set", here the INI-frc) that is shared with a variable (or multivariate distribution) that is not included in the coordinated set ("the

coordinator", here the S1 or S2 frc-s). When calculated in this manner, CI describes the contribution of a site to all possible n-body correlations within another. Here we used as the descriptor the *normalized coordination information* (NCI), in which the coordination information is normalized to the total correlation within the coordinated site. It should be noted that *coordinators* are not all *coordination channels*. *Coordinators* can be coupled to *coordination channels*, and thus perturbation to the *coordinator* leads to a perturbation in the coordinated set.

As summarized in Table 4, the NCI calculated for S1 and S2 show that they both coordinate the INI, with values of 19.1% for S1, and 21.2% for S2. The Na1 and Na2 sites coordinate the INI only weakly (NCI = 9.0% and 6.9%, respectively), and their combined NCI in coordinating the INI is 11.1%. The coordination of INI by the combination of S1, S2, and the Na1 and Na2 frc-s is 27.1%, indicating that just under a third of all the correlated motions in the INI are related to these sites. The coordination exerted by INI on the binding sites was also calculated, because coordination information is not symmetric. We find that while S1 and S2 coordinate the INI strongly, the INI coordinates the two only moderately (NCI = 12.0% and 7.4%, respectively). Interestingly, in the MD trajectory we analyzed, the coordination by INI of the Na1 (NCI = 14.2%) and Na2 (NCI = 10.5%) sites is stronger than in the opposite direction. These results, along with results for all comparisons of sites, are summarized in Table 4.

Table 4. Normalized Coordination Information between sites in $LeuT_{POPE/POPG}$.

	S1	**S2**	**Na1**	**Na2**	**Na1, Na2**	**Na1, Na2, S1**	**Na1, Na2, S1, S2**	**INI**
S1	***30.6 (0.2)***	23.8% (0.5%)	27.5% (0.4%)	17.3% (0.3%)	31.0% (0.5%)	X	X	12.0% (0.4%)
S2	14.2% (0.4%)	***33.1*** **(0.4)**	8.2% (0.2%)	4.5% (0.2%)	8.9% (0.3%)	15.0% (0.5%)	X	7.4% (0.3%)
Na1	51.5% (0.5%)	44.2% (0.5%)	***9.2*** **(0.1)**	39.1% (0.3%)	X	X	X	14.2% (0.5%)
Na2	40.1% (0.4%)	16.7% (0.3%)	32.4% (0.3%)	***12.04*** **(0.1)**	X	X	X	10.5% (0.3%)
Na1, Na2	32.1% (0.3%)	23.3% (0.4%)	X	X	***29.8*** **(0.2)**	X	X	10.1% (0.3%)
Na1, Na2, S1	X	16.3% (0.5%)	X	X	X	***67.2*** **(0.6)**	X	8.8% (0.3%)
Na1, Na2, S1, S2	X	X	X	X	X	X	***132.9*** **(2.0)**	6.2% (0.4%)
INI	19.1% (0.6%)	21.2% (0.7%)	9.0% (0.3%)	6.9% (0.3%)	11.1% (0.4%)	20.5% (0.7%)	27.1% (1.2%)	***14.3*** **(0.1)**

For each pair of *frc-s*, the normalized coordination information is presented, with residues on the top (columns) acting as the coordinator and residues on the left (rows) being coordinated. On the diagonal, the total correlation of the site is shown in bold.

To estimate the importance of these coordination values for the allosteric mechanism, we performed control calculations of the normalized coordination information for S1 and S2, with several other intracellular sites not known for their functional roles, including specific helices, loops, and interfaces between them. In all cases, S1 and S2 coordination of any of these control sites was half (or much less) that of the INI (see Table 5).

Table 5. Coordination of control regions by S1 and S2 in $LeuT_{POPE/POPG}$ and $LeuT_{MNG-3}$.

A. $LeuT_{POPE/POPG}$						
S1	**1. 506-511**	**2. 65-70**	**3A. 77-84**	**3B. 87-96**	**3C. 77-96**	**3D. 78, 81-82, 85-87, 90, 91, 94**
	5.4% (0.2%)	5.0% (0.2%)	4.0% (0.1%)	4.8% (0.1%)	3.3% (0.2%)	4.0% (0.2%)
	3E. 78, 81-82, 90, 91, 94	**4A. 437-445**	**4B. 269-274**	**4C. 437-445, 269-274**	**4D. 438-440, 272-274**	**5. 11-22**
	5.9% (0.1%)	2.6% (0.1%)	4.7% (0.2%)	3.0% (0.2%)	4.2% (0.1%)	10.7% (0.2%)
S2	**1. 506-511**	**2. 65-70**	**3A. 77-84**	**3B. 87-96**	**3C. 77-96**	**3D. 78, 81-82, 85-87, 90, 91, 94**
	10.7% (0.3%)	4.4% (01%)	7.1% (0.2%)	6.8% (0.1%)	5.4% (0.3%)	7.0% (0.2%)
	3E. 78, 81-82, 90, 91, 94	**4A. 437-445**	**4B. 269-274**	**4C. 437-445, 269-274**	**4D. 438-440, 272-274**	**5. 11-22**
	9.3% (0.2%)	3.7% (0.1%)	4.8% (0.2%)	3.6% (0.2%)	4.6% (0.2%)	6.2% (0.2%)
B. $LeuT_{MNG-3}$						
S1	**1. 506-511**	**2. 65-70**	**3A. 77-84**	**3B. 87-96**	**3C. 77-96**	**3D. 78, 81-82, 85-87, 90, 91, 94**
	5.9% (0.2%)	6.5% (0.3%)	6.3% (0.2%)	7.9% (0.2%)	5.9% (0.3%)	8.3% (0.2%)
	3E. 78, 81-82, 90, 91, 94	**4A. 437-445**	**4B. 269-274**	**4C. 437-445, 269-274**	**4D. 438-440, 272-274**	**5. 11-22**
	10.2% (0.3%)	7.3% (0.3%)	7.7% (0.2%)	7.2% (0.3%)	8.8% (0.3%)	13.9% (0.3%)
S2	**1. 506-511**	**2. 65-70**	**3A. 77-84**	**3B. 87-96**	**3C. 77-96**	**3D. 78, 81-82, 85-87, 90, 91, 94**
	7.6% (0.2%)	6.1% (0.2%)	6.8% (0.2%)	8.2% (0.2%)	6.3% (0.3%)	9.5% (0.2%)
	3E. 78, 81-82, 90, 91, 94	**4A. 437-445**	**4B. 269-274**	**4C. 437-445, 269-274**	**4D. 438-440, 272-274**	**5. 11-22**
	10.9% (0.2%)	7.7% (0.3%)	8.5% (0.2%)	7.8% (0.3%)	9.3% (0.3%)	8.8% (0.2%)

The coordination of each control region by S1 and S2 are presented with the standard error of the mean in parenthesis, estimated using moving block bootstrapping with 50 realizations.

Given the importance of the INI in the function of the transporter, we also determined which individual residues make the largest contributions to coordination of the INI. For each residue in the S1-frc and S2-frc residue we calculated the contribution of the residue to the particular frc coordination of the INI, as well as the contribution of INI

residues to receiving that coordination, using Equation (S5). Results summarized in Table 6 show that for coordination of the INI-frc by S1, the top 3 *coordinators* are F259 (contribution = 69.6%), S256 (contribution = 34.9%), and I359 (contribution = 34.6%), and the top 3 *receivers* are R5 (contribution = 67.8%), I187 (contribution = 63.8%), and S267 (contribution = 59.9%). For coordination by S2, the top 3 *coordinators* are R30 (contribution = 54.7%), F253 (contribution = 28.7%), and F324 (contribution = 24.0%), and the top 3 *receivers* are R5 (contribution = 80.8%), I187 (contribution = 71.0%), and D369 (contribution = 58.1%). This underscores the important role of INI residues R5, I187, and S267 in the coordination of the INI-frc by the known allosteric substrate sites.

Table 6A. Specific residues highly contribute to coordination of the INI by S1 in $LeuT_{POPE/POPG}$.

S1	**Leu**	**L25**	**G26**	**V104**	**Y108**	**F253**
	24.5% (0.1%)	22.8% (0.2%)	21.9% (0.2%)	18.8% (0.2%)	13.9% (0.1%)	31.0% (0.2%)
	T254	**S256**	**F259**	**S355**	**I359**	**Na1**
	27.3% (0.2%)	**33.6%** (0.3%)	**67.6%** (0.2%)	13.3% (0.2%)	**33.2%** (0.4%)	16.6% (0.1%)
INI	**R5**	**I187**	**S267**	**Y268**	**Q361**	**D369**
	66.1% (0.3%)	**63.1%** (0.2%)	**58.7%** (0.1%)	57.3% (0.2%)	57.2% (0.4%)	48.1% (0.3%)

The contribution of specific residues in the S1-frc (top) and the INI-frc (bottom) to the coordination of the INI-frc by the S1-frc (top 3 in each site are bold).

Table 6B. Specific residues highly contribute to coordination of the INI by S2 in $LeuT_{POPE/POPG}$.

S2	**L29**	**R30**	**Y107**	**I111**	**W114**	**F253**
	20.1%	**53.8%**	10.6%	9.0%	14.6%	**28.0%**
	(0.2%)	**(0.5%)**	(0.2%)	(0.3%)	(0.2%)	**(0.2%)**
	A319	**F320**	**F324**	**L400**	**D404**	
	10.7%	11.4%	**23.2%**	16.2%	18.8%	
	(0.0%)	(0.1%)	**(0.1%)**	(0.1%)	(0.1%)	
INI	**R5**	**I187**	**S267**	**Y268**	**Q361**	**D369**
	78.3%	**69.0%**	48.5%	42.5%	40.0%	**57.6%**
	(0.2%)	**(0.2%)**	(0.2%)	(0.3%)	(0.3%)	**(0.4%)**

The contribution of specific residues in the S2-frc (top) and the INI-frc (bottom) to the coordination of the INI-frc by the S2-frc (top 3 in each site are bold).

3.1.1.2.5. The Coordination Channel Mediating the INI-frc Coordination by the Substrate frc-s is Through TM6b

Because TM6b emerged as the major channel for communication between S1 and the INI, we investigated whether it was also the major channel for the CI between the substrate sites and the INI. We calculated the *mutual coordination information* (MCI), which described how much of the coordination information is shared between two coordinators that are coordinating the same set, and then normalized to the coordination information of the coordinator of interest (NMCI). Using this analysis, we identified residues in the high NMCI region using the same criteria described for

co-information. The results identify a *coordination channel* that is nearly identical to the channel revealed by the *co-information* analysis, with a significantly larger signal in TM6b than that calculated with co-information analysis (see **Figure 26**).

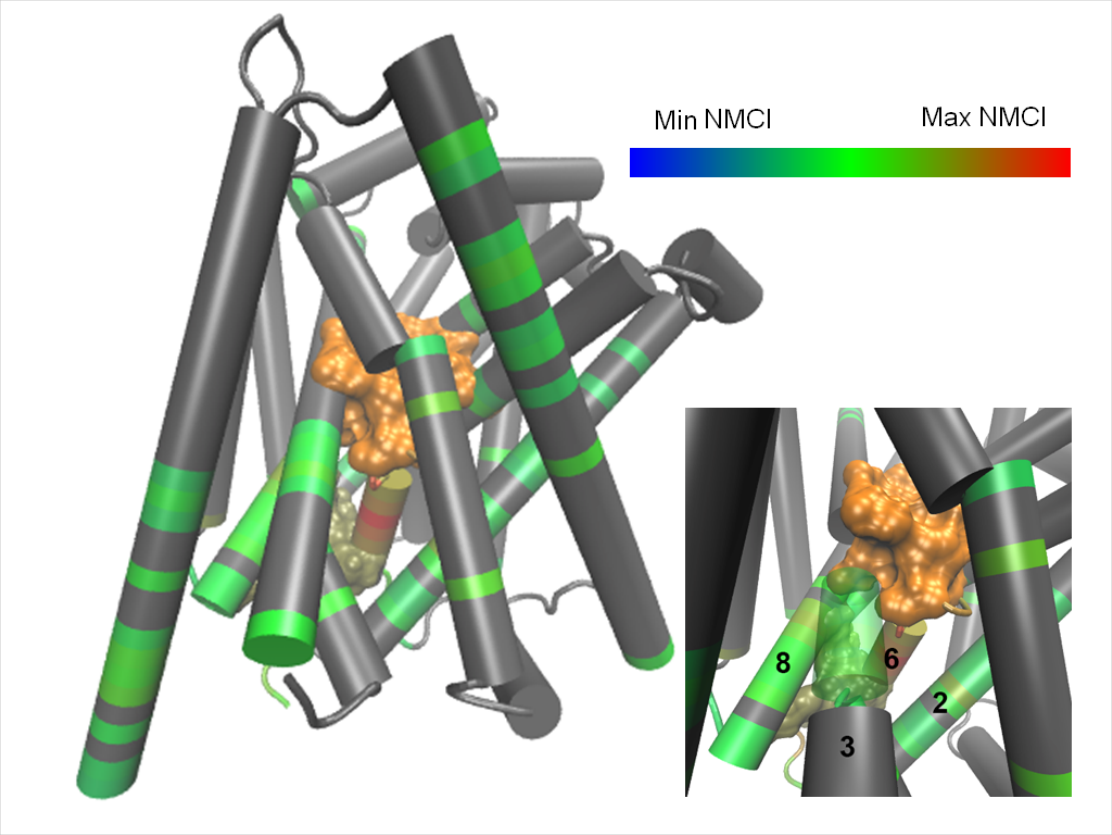


Figure 26. TMs 2, 6b, and 8 form a coordination channel between S1 and the INI. **Main:** Residues found to have high *mutual coordination information* with S1 and the INI are colored by their calculated *normalized mutual coordination information* (NMCI) values using the scale at the top right, where the Min and Max NMCI refer to the minimum and maximum values among all possible residues. All other residues are represented in grey. **Bottom right**: A close up of the TM6b and TM8 interface. interface.

We are able to identify a similar coordination channel for S2 (see **Figure 27**). These results indicate that TM6b is the major channel for the coordination of the INI by S1 and S2.

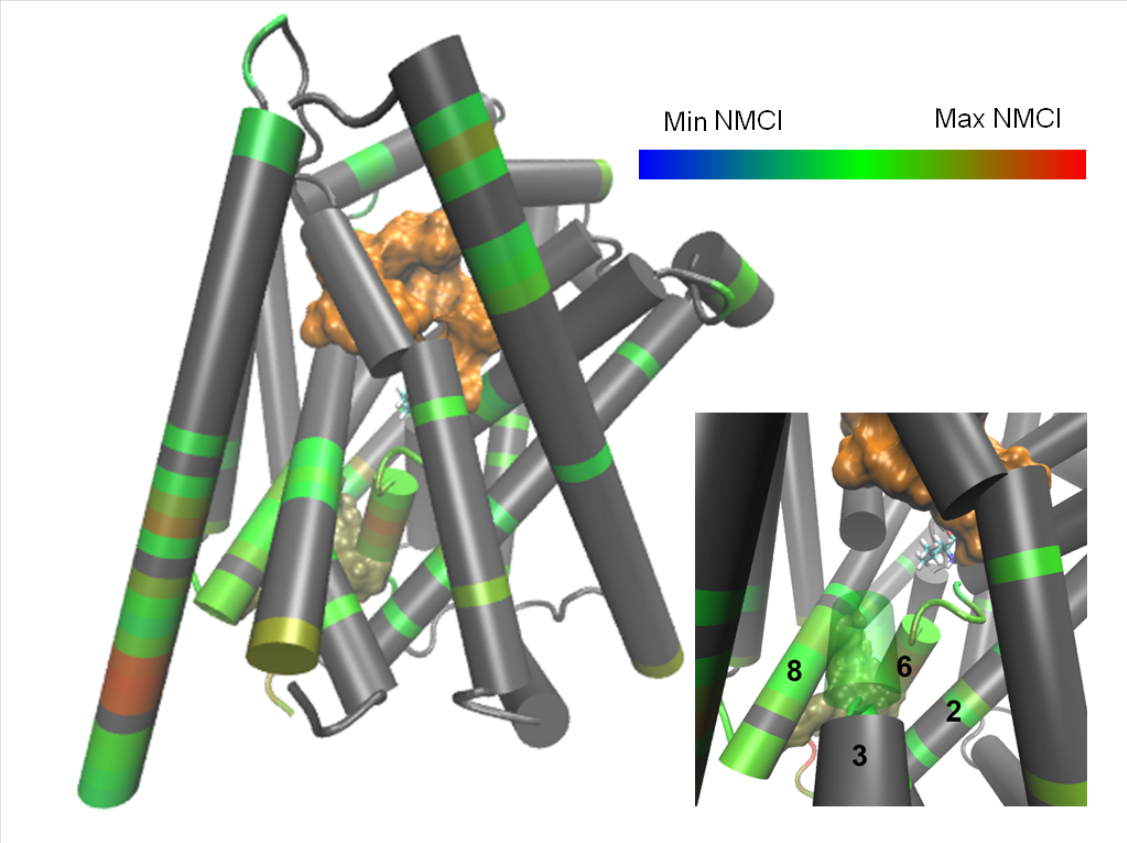


Figure 27. TMs 2, 6b, and 8 form a coordination channel between S2 and the INI in LeuT$_{POPE/POPG}$.

Main: Residues found to have high *mutual coordination information* with S2 and the INI are colored by their calculated *normalized mutual coordination information* (NCMI) values using the scale at the top right, where the Min and Max NCMI refer to the minimum and maximum values among all possible residues. All other residues are represented in grey. **Bottom right**: A close up of the TM6b and TM8 interface.

3.1.1.3. Discussion

Taking advantage of the information about specific functional motifs for the allosteric transporter LeuT, the illustration of the new NbIT analysis method brings to light how it identifies the details of allosteric couplings, and can quantify them at a previously unattained level of detail. Moreover, the choice of LeuT for this illustration of NbIT allowed us not only to start from well-defined *frc*-s, but also to compare the results and the inferences from NbIT analysis to known mechanistic elements in the allosteric process underlying LeuT function. Indeed, the allosteric pathway between the known ligand (ions, substrate) binding sites and previously proposed functional elements such as the intracellular gate (in INI), were identified by the NbIT analysis as the channels that propagate these couplings. This agreement with previous mechanistic insights is important because computational approaches, and in particular the type of MD simulations utilized here as well, have been used successfully to study the dynamics of transporter molecules and to infer on residues and motifs that play essential roles in the allosteric mechanisms[81,181–183], By taking advantage of this kind of data, the novel NbIT analysis provides the first rigorous method for the identification of specific channels by which information is transmitted between functional sites of an allosteric molecular system. Key observations from the present application of NbIT analysis are discussed below to stress the specific molecular detail of the results, and to indicate the predictive power that this new method can bring to the many other allosteric protein systems for which the type of information available for LeuT is currently lacking.

3.1.1.3.1. Allosteric Coordination of the INI by S1 and S2

The CI calculations were essential in revealing that the S1 and S2 sites coordinate the internal dynamics of the INI (see Table 4). The allosteric modulation of the intracellular gate considered on the single molecule macro scale has been noted previously in the dynamic changes revealed by smFRET experiments with LeuT in detergent; this study showed how the allosteric connection enabling modulation at the micro scale is effectuated. *Coordination information* as calculated here connects the collective coordination of the INI domain to the individual components (specific residues) and interactions (within, and outside the *frc* to which they belong) that underlie it. This provides insight at unprecedented detail about the elaborate coordination in the allosteric mechanism underlying ligand-induced opening of the gate. An intriguing observation in view of the ongoing controversy surrounding the role of the S2 binding site [68,75–78,81,93,184] is that the S2-frc coordinates the INI through a channel that includes the S1 site. The coordination found here, of the INI by the *apo* S2 site (the MD trajectories analyzed here did not include substrate bound in S2) may explain why mutations to the S2 site have been shown to affect intracellular gating dynamics [84]. Although they demonstrate the ability of the S2-frc to coordinate the intracellular gate, the present results cannot inform about the role of substrate binding in S2 in the transport process, since this was not covered in the MD simulation.

3.1.1.3.2. Propagation of Information between S1, S2, and the INI Requires TM6b

The channel that propagates the coordination of the INI by S1 and S2 was found here to consist largely of residues in TM6b (see **Figure 26** and **Figure 27**). Indeed, several residues in the S1 site and the INI are part of the highly conserved TM6, and its intracellular end, TM6b, was shown to undergo a large rotation of 17° in a recent

crystal structure of a LeuT mutant stabilized in what is believed to be an *apo* intracellular-open state [185]; TM1a and TM8 also contain many residues from S1 and the INI.

Notably, while this work was originally in preparation, a set of LeuT mutants have been described that were constructed to resemble the human serotonin transporter [186], and all constructs containing a mutation of the TM6b residue Y265 to F, were found to lack transport activity despite retaining high affinity inhibitor binding. This indicates a possible role of TM6b in function, and we interpret the observed rotation of TM6b and the effect of the Y265F mutation as support for their role in propagating information from the substrate site to the intracellular gate during the transition between LeuT states. The fact that the role of TM6b became evident from the NbIT analysis of the S1-occupied occluded state supports its role as an information conduit from the substrate sites to the intracellular gate.

3.1.1.3.3. The Intramolecular Allosteric Mechanism Involves a Subset of Residues Known to Have Functional Roles

With NbIT analysis, we identified specific residues that play a role in allosteric connections related to function, and were able to discern different contributions (i.e., "stabilizers" and "communicators"). In the S1-frc we find that while the bound leucine substrate, F253, and T254 coordinate the binding site's internal correlations (hence acting as stabilizers), residues F259, S256, and Q359 contribute to the coupling between S1 and the INI (Table 5A) and belong to "communicators", which are involved in between-site allosteric communication. We know of no previous computational method that offered such functionally specific discrimination.

The identification of functional roles for specific residues in the allosteric communication revealed further details of their mechanistic involvement:

3.1.1.3.3.1. F259

Our analysis predicted that F259 interactions may have a significant effect on transport. Earlier crystallographic studies had indicates that F259 may be involved in the diversity of transport phenotypes produced by various LeuT substrates [110]. Three basic modes of interaction have been observed: (i)-in crystal structures of LeuT in complex with leucine, methionine, or p-flurophenylalanine, the hydrophobic side chains interact with F259; (ii)-in LeuT structures with alanine or glycine, this interaction is lost, leading to a 30° rotation of the F259 side chain; (iii)-in the structure bound to tryptophan, the indole ring makes a ring-ring contact with the F259 side chain. The three distinct modes of interaction observed for F259 correlate with distinct transport phenotypes. Thus, although the overall binding modes could appear nearly identical, the transport efficiencies differ, with alanine being transported with highest efficiency (k_{cat}/K_m); leucine, methionine, and p-flurophenylalanine displaying low efficiency, and tryptophan acting as an inhibitor. While the efficiency for glycine is even lower than for the low efficiency amino acids mentioned above, the difference may in fact be due to the very low affinity of Gly for LeuT which may not allow it to remain bound to the transporter long enough to initiate transport (no k_{on} or k_{off} values have been reported). Together, these structure/function relations suggest that substrate interactions with F259 may lead to different effects on transport. Our analysis predicted a specific participation in the allosteric mechanism. We suggest that because alanine does not interact with F259 and induces a change in the rotameric state of F259 relative to that observed for the less efficiently transported substrates, F259 plays an inhibitory role by allosterically blocking transport. Clarification of the specific role

that this type of allosteric modulation plays in the transport cycle with the NbIT method must await a complete trajectory of the transition among the different states, but the insights gained in this study offer an intriguing avenue for future experimentation.

3.1.1.3.3.2. Y268, S267, R5, and I187 – stabilizers and communicators in the INI

We find that Y268 R5, and S267 all play the role of both strong stabilizers and communicators in the INI. Both R5 and Y268 are known to be involved in function, with mutation of either residue to alanine resulting in disruption of the intracellular gate [84,169], characterized by an increased "open" (intracellular gate) population observed in smFRET experiments of the intracellular gate. However, the R5A mutation has also been shown to cause increased transitions between the "open" and "closed" (intracellular gate) state in the presence of leucine [84]. Considered together, these experimental findings indicate that mutation of R5 can affect the allosterically modulated gating dynamics; in agreement, R5 is predicted to be the strongest *coordinator* within the INI. The result that Y268, S267, and R5 all play the role of both *coordinator* and *stabilizer* is especially noteworthy because one would expect that residues that are essential to the stability of the gate would need to be modulated in order to initiate large collective conformational changes, such as the opening of the gate. That such residues are also communicators substantiates the allosteric modulation of the conformational change that opens the gate. Indeed, these residues are highly conserved in NSS transporters[71], and our finding leads to the prediction that disruption of interactions between S267 and its surrounding network will strongly affect transport. Future experiments should be able to better define the role of S267 in the transport function based on this testable hypothesis. In addition, we find that while I187 has a minor stabilizer role in the INI, it plays a significant role as a

communicator. This leads to the mechanistic prediction that mutation of I187 may lead to disruption of allosteric modulation without disrupting the structure of the intracellular gate.

3.1.1.2. Related Work

Another quantitative computational approach was used to investigate allosteric couplings in LeuT from MD simulation trajectories, utilizing a comparative analysis of the results from a large set of simulations of the transporter and mutant constructs (Y268A/R5A/D369A) in complexes with various combinations of ions (Na^+/Li^+) and substrates (no substrate/leucine/alanine)[187]. The MD simulation of the Na^+-bound, substrate free state of LeuT was used as a reference relative to which the various trajectories for different ion binding states and mutations were considered as perturbations. To follow the manner in which the "perturbations" affected the allosteric coupling in a detailed structural context, the comparative analysis was formulated in terms of the interaction frequencies between residue pairs observed in the compared trajectories. These interaction frequencies were used to build a network, termed "allosteric interaction network" (AIN), that contains the conformational changes produced by each of the "perturbations".

A key finding of this analysis of the interactions involved in the conformational changes is the consistency of the AIN in the various constructs. Thus, the perturbations - whether induced by ion, substrate, or mutation - led to changes in a core interaction network. This network, the AIN, surrounds S1 substrate binding site and spreads out to the intracellular and extracellular domains. The large changes in this core interaction network were observed in the unwound region of TM6 and the central region of TM10. Notably, the analysis predicted that the Y268A mutation at

the intracellular end of the LeuT transporter would perturb the Na^+ binding sites, through a propagation of changes involving TM6b, TM8, and F259 in particular. The excellent agreement of these results with the findings from the NbIT analysis[111] supports the involvement of these structural elements in the coupling between binding sites and the intracellular gate region. Indeed, the associated experiments reported in this study[187] found that the Y268A mutation disrupted Na^+ binding in the distal binding site, and that due to the cooperative binding of Na^+ and substrate, it also disrupted substrate transport in a clear coupling of distal structural motifs.

3.1.2. Additional Computational and Experimental Studies

We hypothesized that the role of F259 in the substrate-specific allosteric modulation of intracellular gating dynamics could be investigated by a systematic study of how the interaction between F259 and substrate affected F259 dynamics (using MD simulations) and the corresponding allosteric modulation of intracellular gating (using smFRET). Because it is difficult to modify the interaction between F259 and the substrate from the side of F259 in a semi-continuous fashion by modifying F259, we modify the interaction by changing the substrate instrad. We chose to study the substrates leucine, valine, alanine, and glycine, which maintains the interaction as a van der Waals interaction but is expected to reduce the interaction strength as the side chain is shortened and eventually removed completely. The results of this study will be detailed below.

3.1.2.1. Methods

3.1.2.1.1. Computational Methods

We constructed models of LeuT bound to leucine, valine, alanine, and glycine by starting from the PDB crystal structure of LeuT bound to leucine and using the Mutator plug-in within VMD to mutate the leucine substrate to the corresponding amino acid of choice. The simulations were constructed as described in Section 3.1.1.1.1. Simulations, but with reduced box size such that the total number of atoms was approximately ~120,000. This was required to run on the Anton supercomputer at the Pittsburgh Supercomputing Centers, which limits system size. After an initial minimization and equilibration as described in Section 3.1.1.1.1. Simulations, the systems were subjected to microsecond-scale MD simulations on Anton, a special-purpose supercomputer machine[188]. These production runs implemented the same set of CHARMM36 force-field parameters and were carried out in the NPT ensemble under semi-isotropic pressure coupling conditions (using the Multigrator scheme that employs the Martyna-Tuckerman-Klein (MTK) barostat[189] and the Nosé-Hoover thermostat[190]), at 310 K temperature, with 2 fs time-step, and using PME for electrostatic interactions. All the other run parameters were derived from the Anton guesser scripts based on the system chemistry [62].

A mutant construct, F259W bound to glycine, was also constructed from a representative from of the glycine-bound construct in which F259 was in the perpendicular state and simulated using the same protocol for 3 microseconds.

3.1.2.1.2. Experimental Methods

All experiments were performed in the laboratory of Scott Blanchard by Daniel Terry.

3.1.2.1.2.1. TIRF single-molecule fluorescence imaging of LeuT

Microfluidic chambers passivated with polyethylene glycol (PEG) and a small percentage of biotin-PEG[191] were incubated for 5 min with 0.8 μM streptavidin (Invitrogen), followed by 4 nM biotin-tris-NTA-Ni^{2+} [192] in T50 buffer (50 mM KCl, 10 mM Tris-acetate, pH 7.5). LD550/LD650-labeled, His-tagged LeuT molecules were immobilized via the His-tag:Ni^{2+} interaction by incubating for 2 min at ~4 nM concentration. Subsequent imaging experiments were conducted in imaging buffer containing 50 mM Tris/Mes (pH 7.5), 10% glycerol, 0.05% (w/v) DDM, 1 mM β-mercaptoethanol and 200 mM total salt (KCl and NaCl, as specified). An oxygen scavenging system consisting of 0.1% w/v glucose, 0.2 units/mL glucose oxidase (Sigma), and 1.8 units/μL catalase (Sigma) was added to minimize photobleaching. Both enzymes were purified by gel filtration prior to use. Microfluidic chambers were reused up to five times by eluting the protein from the surface with 0.3 M imidazole in imaging buffer.

Single-molecule FRET imaging of LeuT dynamics was performed at 25 °C using a custom-built prism-based total internal reflection (TIR) microscope, as previously described[84,85,193]. Surface-bound LD550 fluorophores were excited by the evanescent wave generated by TIR of an Opus 532 nm solid state laser (Laser Quantum).

Scattered excitation light was removed by a ET555lp filter (Chroma) between the objective and the MultiCam. Synchronization was ensured with an external pulse generator and verified with an oscilloscope. Data were acquired with 2x2 hardware binning using custom software implemented in LabView (National Instruments). Unless otherwise noted, data were recorded at a rate of 10 s^{-1} (100 ms time resolution).

3.1.2.1.2.2. Analysis of smFRET data

Analysis of single-molecule fluorescence data was performed in MATLAB {Juette 2016, Nature Methods}. Single-molecule fluorescence traces were extracted from wide-field movies and corrected for background, spectral crosstalk, and unequal apparent brightness[194]. Each FRET trajectory was calculated as $E_{FRET} = I_A/(I_A+I_D)$, where I_A and I_D are the acceptor and donor fluorescence intensities at each frame, respectively. Traces were selected for further analysis according to the following criteria: (1) single-step photobleaching, (2) signal to background noise ratio > 8, (3) less than four donor blinking events, and (4) FRET efficiency above baseline levels (0.15) for at least 100 frames. Figures were made with Origin software (OriginLab).

To quantify dwell-times in each state, we idealized the smFRET trajectories using the segmental K-means algorithm in QuB[195] with models containing three non-zero FRET states. The model FRET values (0.0, 0.52, 0.65, and 0.82) were obtained by fitting FRET histograms to a sum of three Gaussian functions in Origin (OriginLab).

3.1.2.2. Results

We investigated the nature of the F259-substrate interaction in each of the wild-type trajectories. The F259 side chain can undergo a rotation, which we monitored by calculating the dihedral angle formed by the Cα, Cβ, Cγ, and CD1 (benzyl carbon

ortho in respect to the Cβ). We found that in the WT:Leu trajectory, F259 rotation is constricted and only two rotation events can be observed (see **Figure 28**, top left), between symmetric states of the phenyl ring at approximately 80-110° and -80-110°, which we will call the "perpendicular state", to reflect the orientation relative to the side chain of the substrate. In the WT:Val trajectory, the rotation between thee two symmetric perpendicular states becomes more frequent (see **Figure 28**, bottom left). However, in the WT:Ala and WT:Gly simulations, two new states (symmetric to each other at approximately 90° from the) appear with high frequency (see **Figure 28**, right). We call these symmetric states the "parallel states".

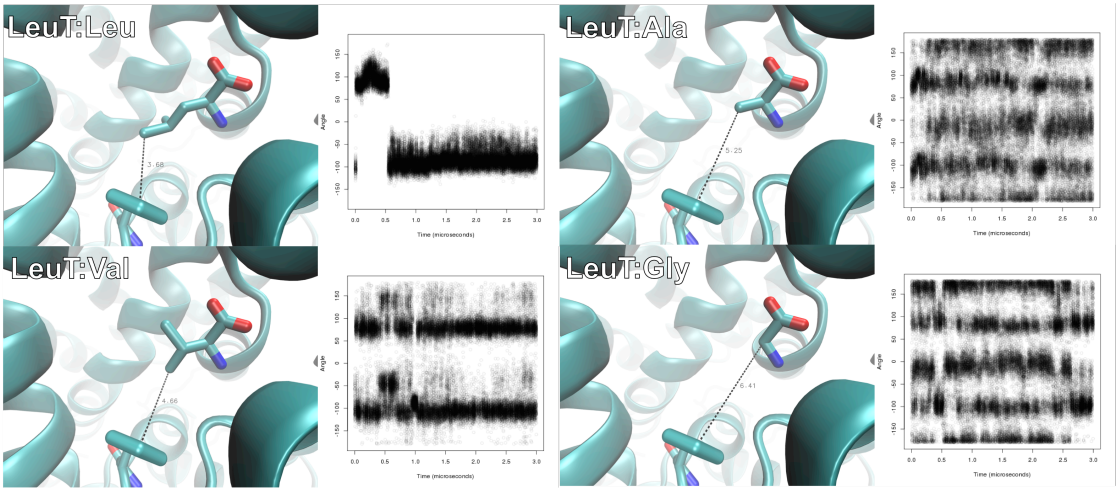


Figure 28. The F259-substrate interaction in various substrate-bound complexes. In each system, the starting state of the F259-substrate interaction is shown with the protein backbone shown in cyan cartoon and the substrate and side chain in licorice representation (carbon in cyan, oxygen in red, and nitrogen in blue). The dynamics of the dihedral angle formed by the CA-CB-CG-CD1 atoms over the course of each 3 microsecond simulation is shown to the right. **Top left**: LeuT:Leu. **Bottom left**: LeuT:Val. **Top right**: LeuT:Ala. **Bottom right**: LeuT:Gly.

We hypothesized that the ligand-specific smFRET behavior previously observed for Leu and Ala was related to their ligand-specific modulation of F259, and thus predicted that Val's effect on intracellular smFRET would resemble that of Leu, and that Gly's effect would resemble that of Ala. The corresponding smFRET experiments, performed by Daniel Terry in the Blanchard lab, are presented in **Figure** 29 and **Figure** 30. Due to improved experimental methodology, three states could be resolved in the smFRET data. In accordance with our predictions, Leu and Val both stabilize the high FRET state (see **Figure** 29), indicating they induce intracellular closing. Additionally, Ala and Gly both stabilize the mid FRET state (see **Figure** 29) and increase the frequency of transitions (see **Figure** 30**)**. This data suggest that ligand-specific modulation of the F259 is correlated with ligand-specific modulation of smFRET distributions and dynamics. However, the data does not necessarily suggest a causal relationship.

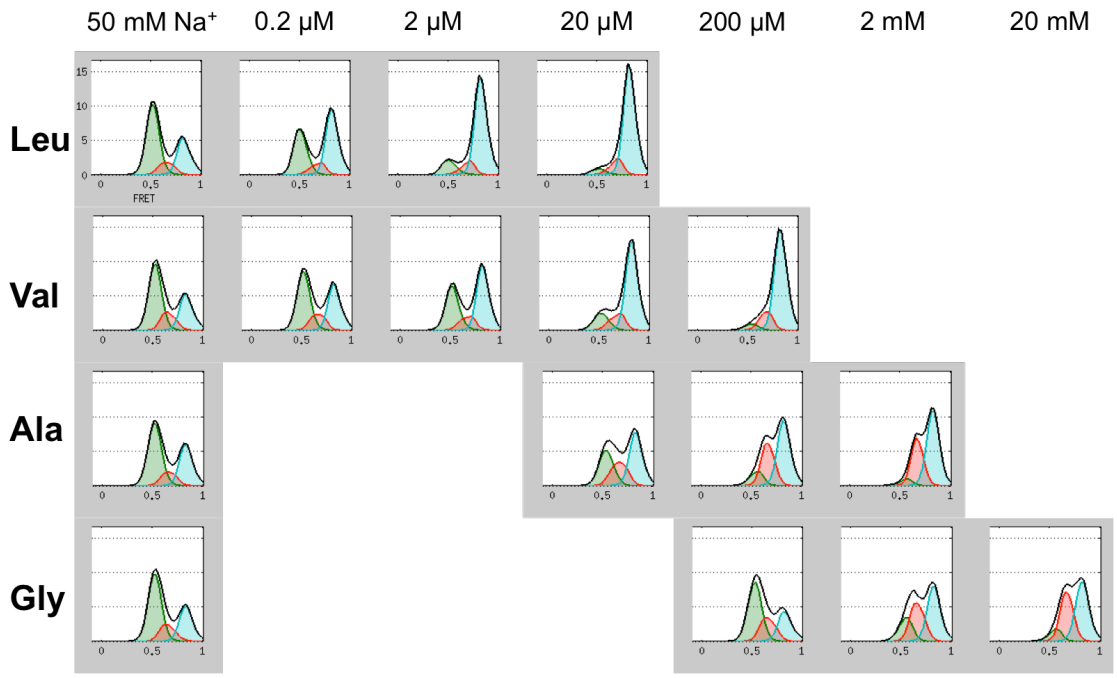


Figure 29. 3-State smFRET distributions as a function of substrate concentration. The idealized population of each state (low FRET in green, mid FRET in red, and high FRET in blue) is shown as a function of substrate concentration for Leu, Val, Ala, and Gly.

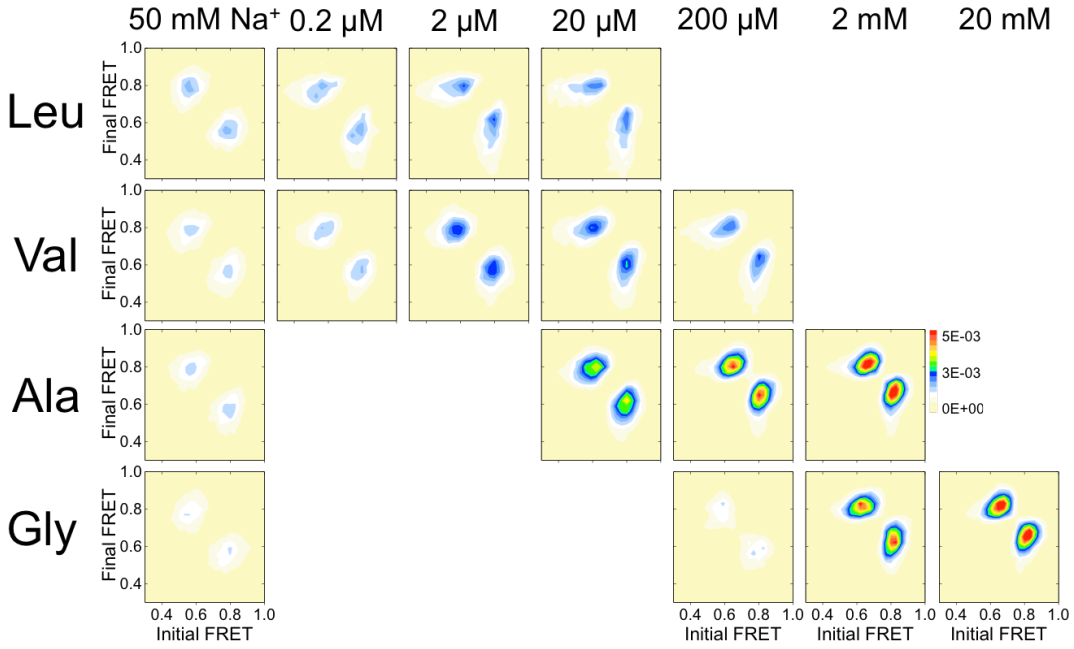


Figure 30. Transition density as a function of substrate concentration.
The transition density, represented as the number of events in which the FRET transitions from a given initial value (x-axis) to a given final value (y-axis), as a function of substrate concentration for Leu, Val, Ala, and Gly.

These results are supported by a recent report of results from fluorescence quenching experiments that indicate that substrate binding in S1 induces conformational changes at the intracellular end of the transporter[196], and that lead to reduced transport. In these experiments, TM5, which undergoes a conformational change during inward opening[72,197], was labeled with a fluorescence tag and the quenching of the fluorescence by the water-soluble reagent potassium iodide was measured in response to ions, substrates, and inhibitors. Substrate binding was found to induce quenching of fluorescence (indicating increased accessibility of the fluorophore), but when measured across a panel of substrates with varying transport efficacy, the magnitude of maximal quenching was *inversely* correlated with transport efficacy (the correlation

between the quenching rate constant and transport efficacy was 0.985). This was taken to indicate that while substrate binding may induce transport dynamics, poorly transported substrates like leucine might actually prevent this induction of dynamics by stabilizing either an inward closed state, or a rate-limiting intermediate in the inward opening process.

Based on these observations, we hypothesized that it would be possible to decrease the occupancy of the inward closed state and increase the rate of transitions if F259 could be locked into a parallel state. To investigate this possiblity, we constructed an F259W mutant bound to Gly, and simulated the new system to determine if this mutant would exclusively sample states in which the tryptophan ring was in a perpendicular state (as it would not fit in the site with the rings in the parallel orientation). Our simulations revealed that over the course of the 3-microsecond trajectory, the F259W side chain did not rotate, suggesting that the mutant was constrained to the perpendicular state (see **Figure 31**). As F259 is a W in the glycine transporter GlyT, we hypothesized that the mutant would be folded and able to transport glycine. However, we did not expect binding or transport of Leu, Val, or Ala as the presence of the bulky F259W side chain would likely produce a steric clash with the substrate side chain.

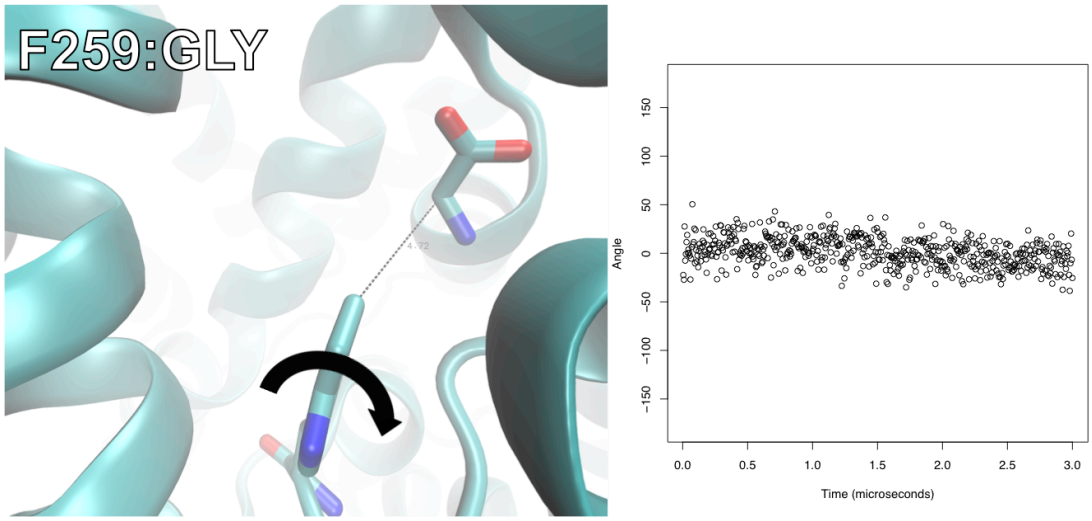


Figure 31. **F259W:Gly is locked in a parallel state.**

The F259W-glycine interaction is shown with the protein backbone shown in cyan cartoon and the substrate and side chain in licorice representation (carbon in cyan, oxygen in red, and nitrogen in blue). The dynamics of the dihedral angle formed by the CA-CB-CG-CD1 atoms over the course of the 3 microsecond simulation is shown to the right.

In related smFRET experiments with the F259W mutant, the apo state was found to exhibit an increased occupancy of the inward open state (see **Figure 32**) and the dynamics were also increased to levels greater than were previously seen when induced by alanine (see **Figure 33**), supporting our hypothesis that the parallel state of the F259 allosterically induces inward opening, whereas the perpendicular state of F259 allosterically induces inward closing.

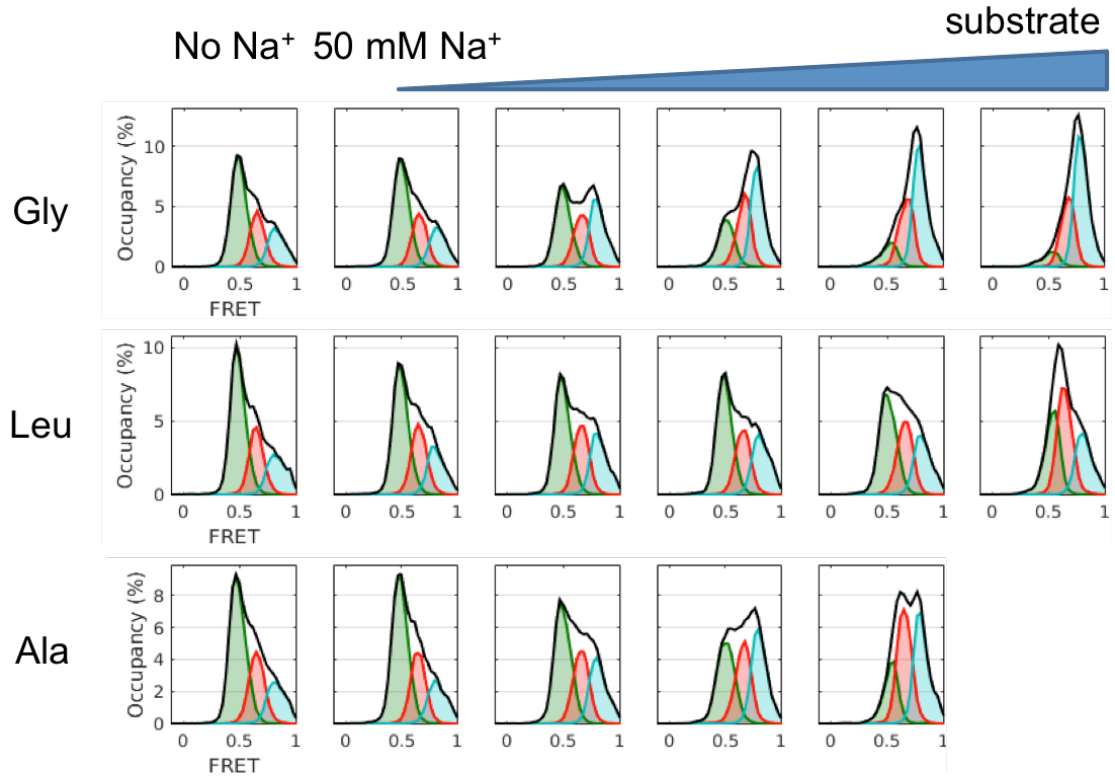


Figure 32. 3-State smFRET distributions for F259W as a function of substrate concentration.

The idealized population of each state (low FRET in green, mid FRET in red, and high FRET in blue) is shown as a function of substrate concentration for Leu, Ala, and Gly.

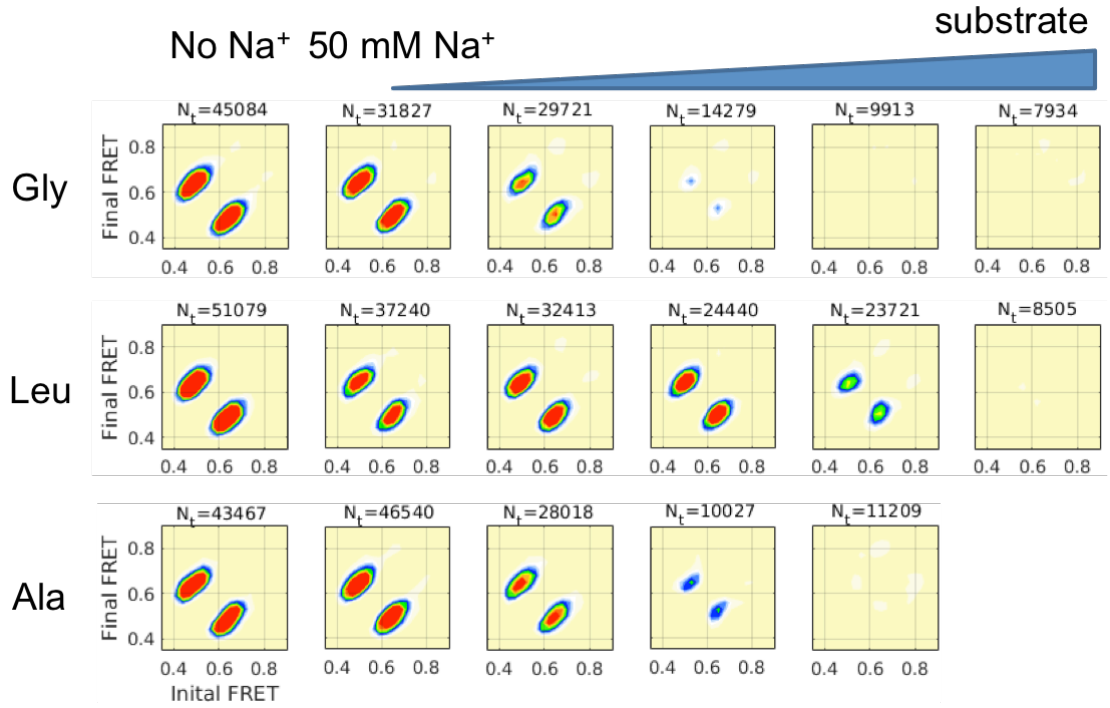


Figure 33. Transition density of F259W as a function of substrate concentration. The transition density, represented as the number of events in which the FRET transitions from a given initial value (x-axis) to a given final value (y-axis), as a function of substrate concentration for Leu, Ala, and Gly.

However, the effect of substrates on F259W smFRET also revealed some unexpected behavior. Whereas in the wild-type transporter the allosteric effect of Gly is similar to that of Ala, in the F259W mutant the allosteric effect of Gly is similar to that of Leu on the wild-type transporter (see **Figure 32** and **Figure 33**). As the F259W side chain and glycine Cα are in close proximity, this may indicate that any constriction of the dynamics of TM6b through interaction with the substrate can lead to inward closing and reduced dynamics, even if the ring can be in the parallel state. Secondly, Leu and Ala still have an allosteric effect on intracellular dynamics, even though they are not expected to be able to bind in S1 in the F259W mutant, which makes the S1 pocket very crowded. In F259W, the Leu allosteric effect is similar to the allosteric effect of

Ala on the wild-type transporter. As F259W should not be able to bind Leu in S1, these results may indicate that the Ala/Gly allosteric effect on the wild-type transporter is actually mediated by binding in the S2 site. Previously, it was shown that L400S and F253A both blocked Ala's induction of dynamics[85]. While the F253A mutant does inhibit the Ala effect, the inhibition may be due to allosteric coupling between S2 and S1 rather than S1-mediated allostery.

3.1.1.1.5.3. Discussion

These results lead us to hypothesize that the smFRET data present two separate ligand-mediated allosteric effects. We propose that allosterically induced intracellular closure, as seen for Leu and Val, is mediated by substrate binding in S1 through the stabilization of the F259 perpendicular state. However, the allosterically induced intracellular dynamics, as seen for Ala and Gly, is mediated by S2 binding. This hypothesis can explain both the wild-type LeuT and F259W mutant results. In wild-type LeuT, Leu interacts with F259 strongly, leading to an S1-dominated phenotype, whereas Ala and Gly do not interact with F259 strongly, leading to S2-dominated phenotypes. However, in F259W, Leu and Ala both do not bind to S1, leading to S2-dominated phenotypes, whereas Gly binds to S1 and interacts with F259W, leading to an S1-dominated phenotype.

To test this hypothesis, it is necessary to confirm the binding stoichiometry in the F259W mutant, as the above described proposal necessitates 1:1 F259W:Leu binding, 1:1 F259W:Ala binding, and 1:2 F259W:Gly binding. Estimating the likely stoichiometry of binding using MD poses some difficulties, as the binding free energies that can be calculated using techniques like free energy perturbation[198] are sensitive to the conformations of both the protein and the ligand. However, as future

work, these calculations will be performed in parallel to stoichiometry experiments. Furthermore, to confirm that the Leu and Ala allosteric effect on F259W is S2-mediated, it is necessary to perform the same smFRET experiments on an F259W/S2 double mutant, such as F259W/L400S.

In the future, it will be necessary to decompose the S2-mediate allosteric modulation of intracellular gating in a similar way as to S1-mediated allosteric modulation was decomposed in the work presented here. While our original analysis suggested residues that were important for communication, the channel was expected to be composed of S1 and TM6b. This suggests a complex mechanism in which S2 somehow modulates the conformation of S1, and potentially F259, to induce dynamics. By further investigating the S2-mediated allosteric induction of transport-relevant dynamics, the role of this secondary site in physiological transport may become clearer.

3.2. Allostery in the Transport Mechanisms of DAT

Much of the content in this section has been adapted with permission from [199].

As mentioned in Section 1.1.2.2. Membrane Transporters, much of the evidence for allosteric modulation in LeuT is mirrored by similar experiments in DAT and other sMATs. However, the N-terminal domains of sMATs are much longer than that of LeuT and are likely to be composed of significant structured segments[200]. These segments have been implicated in key mechanistic elements of NSS function including regulatory phosphorylation[201–206], and the actions of psychostimulants[201,202,207–210]. Indeed, the involvement of the N-terminal domain (N-term) in amphetamine (AMPH)-induced reverse transport (efflux) of the substrate has been well documented for different neurotransmitter transporters[201,207,208,211–214]. In hDAT, specifically, the

AMPH-induced efflux has been shown to be modulated by the first (distal) 22 residues in the hDAT N-term[210], their electrostatic interactions with highly charged phosphatidylinositol 4,5-biphosphate (PIP_2) lipids[215], and by phosophorylation of this region at one or multiple Ser residues[201].

In order to investigate the potential role of the N-term as an allosteric modulator one of more components of the transport process, we performed extensive (> 14 microseconds in total time) unbiased MD simulations of an hDAT homology model in a physiologically relevant lipid membrane environment, and used analysis tools based on NbIT to understand the allosteric coupling between the N-term and other functional domains.

3.2.1. The role of allostery in spontaneous inward opening of hDAT

As the work described in this section has been previously published[199], only an abbreviated version of the Methods and Results, focusing on the application of NbIT, will be presented below.

3.2.1.1. Methods

Several molecular models of the full-length hDAT (residues 1-620) were prepared for all-atom MD simulations in explicit lipid membrane and water environment. Briefly, we used Modeler 9v10[216] and a previously published sequence alignment of the NSS-family proteins[217] to first construct homology models for the transmembrane (TM) part of the hDAT (contained in residues 57-590) based on either recently released structure of the dDAT (PDB code:4M48)[218], or on the high resolution outward-open X-ray structure of the bacterial member of the NSS-family, LeuT (PDB code:3TT1)[72]. The models included the substrate, dopamine (DA), positioned in the central binding

S1 site, two Na^+ ions, positioned equivalently to those in the LeuT crystal structure, and a Cl^- ion coordinated by residues Asn82, Tyr102, Ser321, and Asn353 of hDAT, based on the chloride binding site described previously[219,220].

As described earlier[200], the 3D folds of the structurally unknown N- and C-terminal domains of the hDAT (fragments 1-57 and 591-620, respectively, that lack sequence homology to proteins of known fold) were generated using Rosetta-based *ab initio* structure prediction algorithms. Briefly, different fragments of the termini were subjected to the Rosetta *ab initio* fold prediction routine and for each construct, the predicted structures were clustered under various residue exclusion conditions. Clusters containing the majority of structures were identified. The conformations in the top clusters were evaluated with the RMSDTT iterative fitting algorithm to find regions with the highest structural conservation within each cluster, and the folds with the lowest scores (from the Rosetta energy function) in each cluster were selected.

The predicted structures for the N- and C-termini were docked onto the two models of the hDAT TM bundle described above to complete the full-length hDAT models based on dDAT and LeuT (referred to throughout as $hDAT^{dDAT}$ and $hDAT^{3TT1}$, respectively). For the $hDAT^{3TT1}$ model, two alternative docking poses were considered, resulting in two starting conformations in which the relative positioning of the two termini were different. For the $hDAT^{dDAT}$ model only one docking pose was considered in which the positioning of the C-terminus closely followed that in the dDAT X-ray structure, and the N-terminus was docked so as not to contact any residue in the TM bundle.

$hDAT^{dDAT}$ and $hDAT^{3TT1}$ models were immersed into a pre-equilibrated membrane containing an asymmetric lipid distribution of 451 lipids between the two leaflets so as

to resemble a lipid composition of neuronal cell plasma membranes[221]: 100:40:32:27:29 mixture of POPE/POPC/PIP_2/POPS/Cholesterol on the intracellular leaflet, and 176:29:18 mixture of POPC/DPPC/Cholesterol on the extracellular leaflet.

For each transporter-embedded membrane patch, lipids overlapping with the protein were removed. After solvating with TIP3P water, the transporter-membrane complexes were neutralized with either K^+Cl^- or Na^+Cl^- salt, resulting in a final atom count of ~150,000.

Simulations of the $hDAT^{dDAT}$ and $hDAT^{3TT1}$ constructs in the corresponding membrane environments were carried out with NAMD software version 2.9[178]. During this stage, the backbone of the protein was first fixed and then harmonically constrained. The solvent was initially prevented from entering the lipid-water interface. The constraints on the protein backbone were released gradually in three steps of 300 ps each, changing the force constants from 1, to 0.5, and 0.1 kcal/ (mol $Å^2$), respectively. This step was followed by relatively short (50-100ns) unbiased MD simulations performed with 2fs integration time-step and under the *NPT* ensemble (at T=310K), using the Particle-Mesh-Ewald (PME) method for electrostatics and the Nose-Hoover Langevin piston to control the target 1atm pressure, with Langevin piston period and decay parameters set to 100 fs and 50fs, respectively.

After this equilibration phase, long, microsecond-scale unbiased MD simulations were initiated on the Acellera GPU cluster that runs the specialized MD simulation software ACEMD[222]. ACEMD allows computations with standard CHARMM force fields and for all the runs (including the equilibration phases with the NAMD described above) we used the all-atom CHARMM27 force field for proteins with CMAP corrections[176], the CHARMM36 force field for lipids[177], the TIP3P water model, and the CHARMM-

compatible force-field parameter set for PIP_2 lipids[223]. The simulations with ACEMD implemented the PME method for electrostatic calculations, and were carried out with 4 fs integration time-step. The computations were conducted under the *NVT* ensemble (at T=310K), using the Langevin Thermostat with Langevin Damping Factor set to 0.1.

The temporal correlation between time-dependent variables extracted from the MD simulations was quantified by calculating the Pearson correlation coefficients between the pairs of variables. To cluster the dynamic quantities based on the strength of temporal correlations between them, we then performed *agglomerative mutual-hierarchical clustering*[224] on the matrix of correlation coefficients, using the mutual information as the distance criterion. The linear approximation of the mutual information was calculated, and corrected for dimensionality using the generalized correlation coefficient[156]. Briefly, the clustering algorithm first assigns each variable to its own branch, and then calculates the pairwise generalized correlation coefficient between all branches. Two branches with the highest correlation are then merged, and the generalized correlation coefficient between the new branch and all other branches are recalculated. The algorithm continues until all variables are members of the same branch. In our application here, the dendrogram (tree) is built from the small number of medium-sized clusters consisting of highly correlated variables. The moderately correlated clusters are then merged into one large cluster. Finally, the tree is completed with the remaining small clusters that are weakly correlated to both each other and the large cluster.

In order to quantitatively identify when medium-sized clusters of interest were merged, we used the Fowlkes-Mallows Index (FMI)[225] as a measure of similarity between the different clustering before and after each merger. The FMI is defined as:

$$B = \frac{\sum_i \sum_j m_{ij}^2 - n}{\sqrt{\left(\sum_i \left(\sum_j m_{ij}\right)^2 - n\right)\left(\sum_j \left(\sum_i m_{ij}\right)^2 - n\right)}} \quad (1.178)$$

In the above, n is the number of objects being clustered and M is a matrix with rows equal to the number of clusters in the pre-merger clustering, columns equal to the number of clusters in the post-merger clustering, and elements m_{ij} equaling the number of common members in cluster i of the pre-merger clustering and cluster j of the post-merger clustering. The FMI ranges from 0 to 1, with low values indicating merger of similarly sized clusters that significantly change the clustering, and high values describing mergers that do not change the overall clustering, such as the joining of small clusters (compared to other existing clusters) to each other or to a larger cluster. Therefore, the mergers of the medium-sized clusters will have relatively low FMI, whereas the mergers of small clusters to the large cluster (as is observed at the end of the clustering) will have a high FMI.

3.2.1.2. Results

In this section, results that do not pertain to the NbIT method will be described briefly and can be found in greater detail in the published manuscript[199].

By measuring several variables that are believed to be associated with conformational differencing between gating states of the transporter, we found that spontaneous inward-opening occurred in several simulations. **Figure 34** shows the count of water molecules inside the EC and IC vestibules during the $hDAT^{dDAT}$ simulation, which indicates a rapid transitioning of the transporter from the initially outward-open state

to the occluded state by the loss of hydration in the EC vestibule and increase in water count in the IC vestibule (see **Figure 34**).

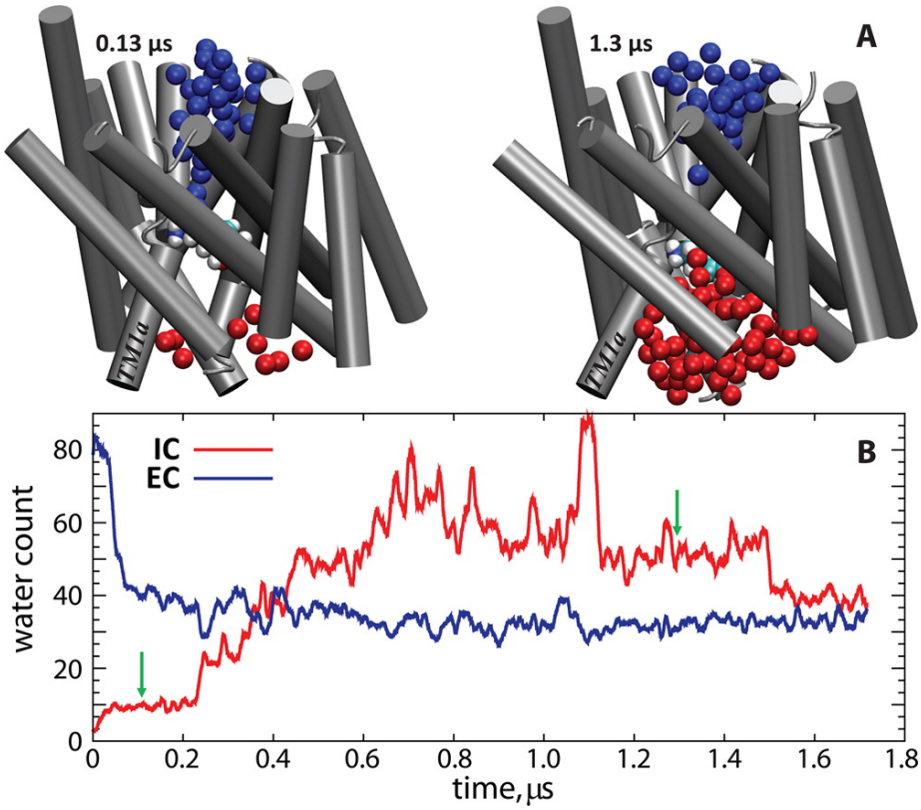


Figure 34. Spontaneous inward opening of hDAT.

(**A**) Snapshots of the hDAT TM bundle (gray cartoon) in the hDATdDAT trajectory at 0.13 and 1.3 µs time-points. Red and blue spheres represent oxygen atoms of the water molecules in the IC and EC vestibules, respectively (see Methods for description of the water count algorithm). The substrate, DA, is shown in van der Waals rendering. The TM1a segment is labeled. (**B**) Time evolution of the number of water molecules in the IC (red) and EC (blue) vestibules in the hDAT/dDAT simulation. The green arrows denote time-points at which the snapshots in panel A were taken.

Distance measurements between various IC regions of the transporter (see **Figure 35A-C**), reveal large-scale concerted motions of the intracellular TM1a, TM6b, and

TM9 segments during the subsequent trajectory interval (~0.2-0.6 μs), whereby TM1a-TM6b, TM1a-TM9, and TM6b-TM9 distances increase by ~4.5Å, 8Å, and 5Å, respectively. As a result, the TM1a and TM9 segments swing away from the TM bundle (see **Figure 35F-G**) and the IC vestibule opens, allowing the large influx of water molecules (see red trace in **Figure 34B**).

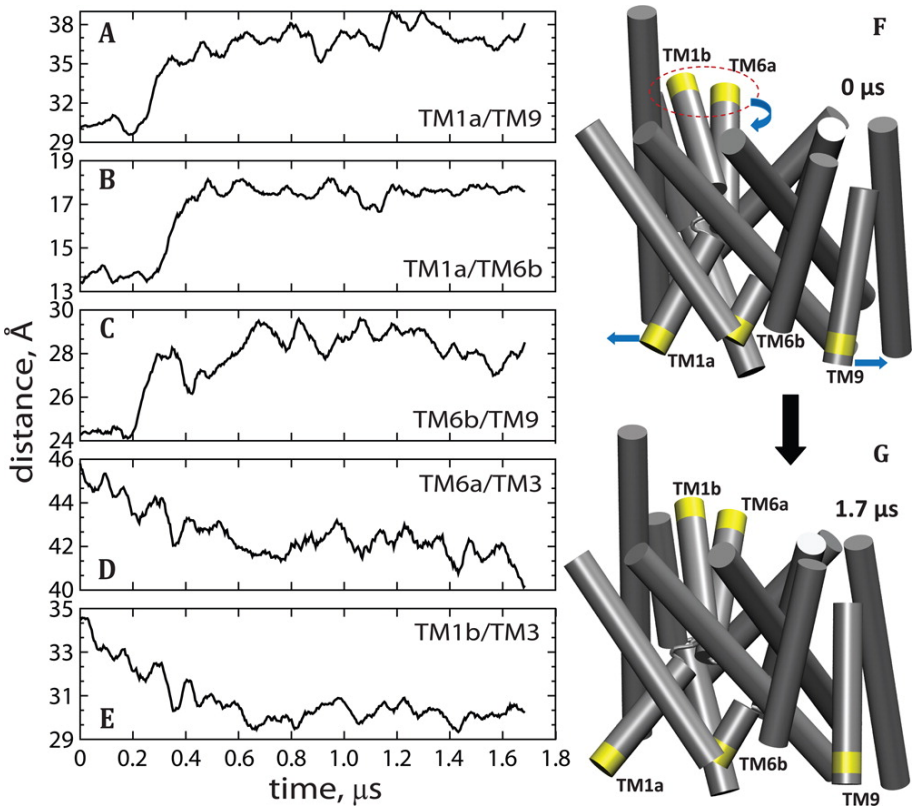


Figure 35. Time evolution in the hDATdDAT simulation of Cβ-Cβ distances between residues in various TM segments.

(A) I67 (in TM1a) and L447 (in TM9); (B) I67 (in TM1a) and S333 (in TM6b); (C) S333 (in TM6b) and L447 (in TM9); (D) E307 (in TM6a) and F171 (in TM3); and (E) F171 (in TM3) and K92 (in TM1b). Panels F and G depict conformations of the hDAT TM bundle (silver) in the initial and final frames of the trajectory. The IC and EC segments from panels A–E are labeled and colored in yellow. Blue arrows in panel F indicate the direction of movement of the different regions in the transition from F to G. Collective motions of TM1b and TM6a segments are highlighted by a red dotted oval.

Collectively, the data indicate that the $hDAT^{dDAT}$ simulation captures the event of the inward-opening in hDAT, which follows dynamic trends which are consonant with the DEER distance measurements[226] in LeuT where TM1a, NT (the fragment of the N-terminus adjacent to TM1a), TM6b, and TM7a segments undergo the most substantial ion- or ligand-dependent movements at the IC side. Computational explorations of dynamics of LeuT[68,82], DAT[227], and SERT[228] showed that the isomerization to the inward-open state induces a destabilization of the ion in Na2. Consistent with these findings, we observe that the transition to the inward-open state in our simulations is accompanied by the spontaneous release of the Na^+ ion from the Na2 site.

We note that a similar isomerization event was detected in the two 4 μs long control simulations initiated from the $hDAT^{3TT1}$ model. Importantly, we found that the inward opening in these simulations followed dynamic trends largely similar to those observed in $hDAT^{dDAT}$.

We additionally found that PIP_2 lipids mediated an interaction between the N-term and ICL4, which appears to be related to the intracellular opening. To establish the relation between the PIP_2-mediated association of the N-term with ICL4, and the sequence of rearrangements leading to the inward-opening transition, we clustered the time-dependent variable describing PIP_2-mediated N-term/ICL4 contacts with the several other dynamics measurements. To determine the temporal relationship between the structural motifs that underlie the isomerization event in the hDAT and the PIP_2-controled N-term/ICL4 dynamics we identified the dominant clusters using the FMI measure described in Methods above. The resulting dendrogram is shown in **Figure 36**.

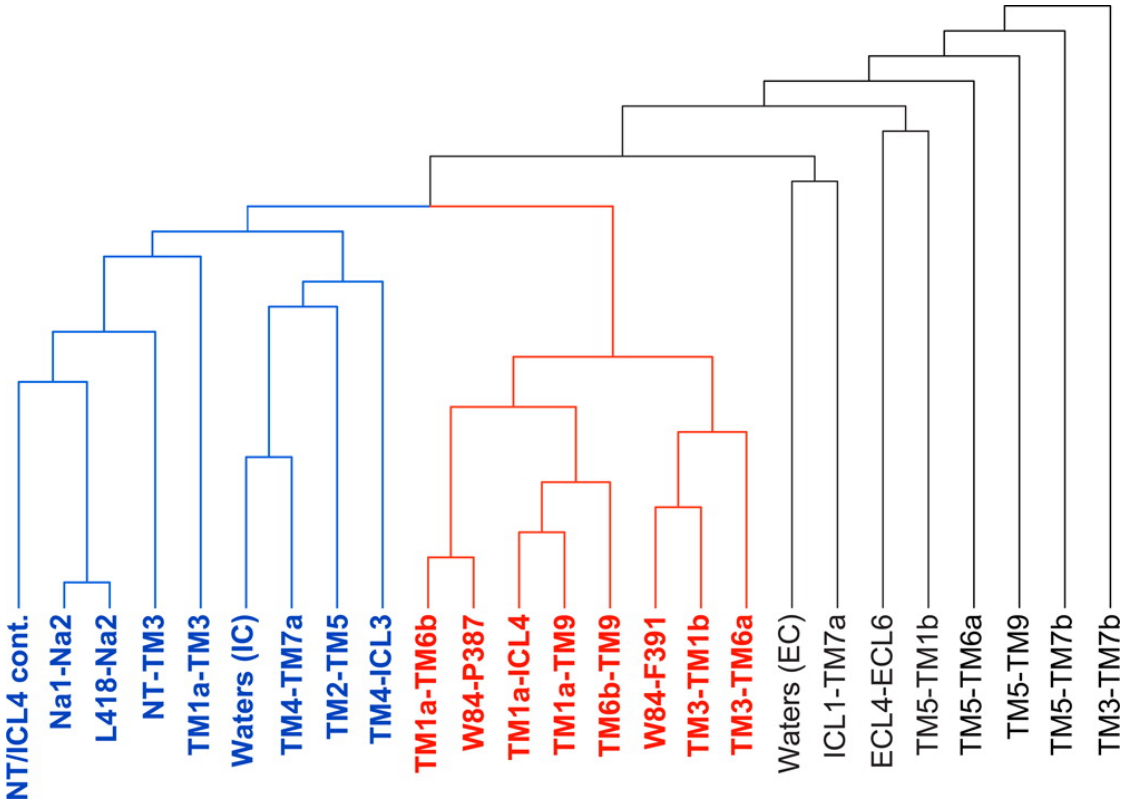


Figure 36. Mutual information dendrogram of several measures of hDAT structure.

The dendrogram shows a merger of two clusters (variables and limbs shown in red and blue) and a connection of the resulting large cluster to smaller branches of the tree (in black). The variables in black are weakly correlated with those belonging to the colored clusters; as assessed by the Fowlkes–Mallows Index, the variables in black do not affect the clustering significantly.

The dendrogram shows a large cluster of highly correlated variables (rendered in red) consisting of distance changes on the IC and EC sides of the transporter. The cluster identifies strong temporal correlations between the dynamics of the TM1a/TM6b/ICL4/TM9 segments on the IC end, and the distance changes related to the movements of TM1b/TM6a regions on the EC vestibule combined with the

movement of the ECL4b region with respect to TM1b (involving residues W84, F387, and F391). Connected to this limb of the dendrogram is another large branch (shown in blue in **Figure 36**), which consists mostly of additional dynamic variables that describe the opening of the IC vestibule during the isomerization. This branch also contains the PIP_2-mediated N-term/ICL4 contacts ("NT/ICL4 cont." in **Figure 36**), as well as the dynamics of the Na2 ion.

Clustering analysis on the same quantities extracted from $hDAT^{3TT1}$Run1 and $hDAT^{3TT1}$Run2 trajectories again revealed merger of two clusters (in red and blue colors) containing various IC and EC structural motifs that describe inward-opening of the transporter (most prominently, dynamics in TM1a/TM6b/ICL4/TM9 segments), movement of the Na2 ion, the extent of PIP_2-mediated N-term/ICL4 contacts, and the dynamics of the ECL4b. Together, the clustering analysis quantitatively establishes coupling between the PIP_2-mediated N-term/ICL4 association and the structural hallmarks related to the inward-opening transition in the hDAT.

The prominent role played by the ICL4 region in the inward opening transition prompted a deeper analysis of the manner in which dynamics in the ICL4 propagates to the functional sites of the transporter. To this end we quantified in the simulated trajectories the total intercorrelation between the ICL4 segment and (i)-the residues that line ion binding Na1 and Na2 sites, and (ii)- the residues in the primary S1 and the presumed secondary S2 substrate binding sites in hDAT.

The time evolution of the total *intercorrelation coefficient* (r_{INTER}), used to quantify the extent of coupling between the collective motions of the ICL4 region and various functional sites, was calculated from the $hDAT^{dDAT}$, $hDAT^{3TT1}$Run1, and $hDAT^{3TT1}$Run2 trajectories and is shown in **Figure 37**.

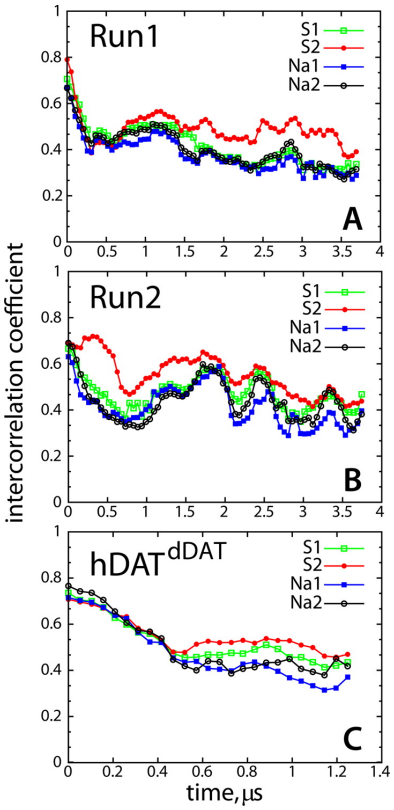


Figure 37. Total intercorrelation coefficient between the residues in ICL4 and in different functional sites of the hDAT.

Substrate binding S1 and S2 sites (green and red, respectively); Na+ ion binding sites Na1 and Na2 (blue and black, respectively). The intercorrelation coefficients are shown separately for simulations Run1 (A), Run2 (B), and hDATdDAT (C) and were obtained as averages over 500 ns time intervals by sliding the analysis windows by 50 ns.

Figure 37 reveals strong coupling (large r_{INTER} values) between the dynamics within the ICL4 segment and that in the functional sites of the transporter in the initial stages of the hDAT^{dDAT}, hDAT^{3TT1}Run1, and hDAT^{3TT1}Run2 simulations. The reduction in coupling that occurs within the first ~0.5μs could be due to collective relaxation dynamics from the starting outward-open model towards the occluded state in these systems. However, since the initial 0.5 μs interval coincides in time with the inward opening in these simulations (see above), the results may also indicate that the isomerization event is preceded by highly coupled motions in the ICL4 and the functional regions. In fact, the same analysis performed on trajectories in which PIP2 was not a component of the membrane (dDATΔPIP and Run2ΔPIPa trajectories), for which the initial protein models where the same as in hDAT^{dDAT} and hDAT^{3TT1}Run2 simulations, respectively, but in which the hDAT did not transition to the inward-open state, revealed r_{INTER} values equivalent to the late time values of hDAT^{dDAT}, hDAT^{3TT1}Run1, and hDAT^{3TT1}Run2. This leads to the inference that the high correlations measured in the trajectories collected in PIP_2-enriched membranes are indeed related to the transition to the inward-open state observed in these systems.

Figure 37 also shows that after the initial decrease in the correlations, the coupling between the ICL4 and some of the functional sites rises again in the hDAT^{dDAT}, hDAT^{3TT1}Run1, and hDAT^{3TT1}Run2 systems. Especially notable is the higher value of r_{INTER} for the S2 site (red traces in **Figure 37**). The S2 site in hDAT includes residues W84, P387, and F391, which are involved in the conformational rearrangements accompanying the inward opening (i.e., the specific motion in ECL4b results in pulling F391 towards W84 while P387 moves away from W84). Thus, we find that the observed strong correlations between the ICL4 and the S2 site is primarily due to

highly coupled motions in the ICL4 segment and the S2 residues from the ECL4b loop, adjacent to P387, that participate in the described pulling motion.

The timing of the increase in correlations shown in **Figure 37**coincides with the event of Na^+ release from the Na2 site (at ~0.75, 1.25, and 1.0 μs time-points in the 3 simulations $hDAT^{dDAT}$, $hDAT^{3TT1}$Run1, and $hDAT^{3TT1}$Run2 systems, respectively). After the release is complete, we observe a slow relaxation of the correlations (see $hDAT^{3TT1}$Run2 in **Figure 37**), with the coupling between ICL4 and the S2 site remaining the highest among all the correlations considered. Collectively, the NbIT analysis demonstrates and quantifies the allosteric coupling between the ICL4 and the functional sites during the PIP_2-mediated inward-opening of the hDAT, and identifies the S2 site as the region with the strongest coupling to the ICL4, suggesting that this distant communication is mechanistically important for the inward-opening transition.

3.2.1.3. Discussion

The analysis of the >14 μs unbiased atomistic MD trajectories of a full-length model of the hDAT in lipid membranes presented here addresses, to our knowledge for the first time, the mechanistic involvement of the N-terminal region of the hDAT in the functionally relevant conformational transitions of the transporter involved in the inward-opening of the hDAT. The results show that the conformational isomerization triggered by the strong tendency of the N-term to associate with the ICL4 segment through PIP_2-mediated electrostatic interactions. The mechanistic consequences of the PIP_2-mediated N-term/ICL4 association that emerge from this analysis are the disruption of a conserved IC network of ionic interactions, which triggers the inward-opening by destabilizing the IC network of ionic interactions, and the associated release of the Na^+ ion from the Na2 site causes destabilization of the substrate DA in

the primary S1 site. The consequences of these interactions for the functional mechanism of the transporter are underscored by our findings showing that *inward opening is accompanied by concomitant movements in the EC vestibule, and that isomerization to the inward-facing state in hDAT results in the release of the Na+ ion from the Na2 site, and the destabilization of the substrate (DA) in the S1 site.*

Our analysis using NbIT found that the collective motions triggered by the N-term/ICL4 association on the intracellular side are strongly coupled to collective motions in the extracellular vestibule and in the substrate and ion binding sites. Further substantiating the mechanistic importance of the PIP_2-mediated N-term/ICL4 interactions, is the clear identification from the mutual information clustering (MIC) results of their effect on the intracellular side. This substantiates the allosteric coupling of these N-term/ICL4 interactions to the functional sites in hDAT involved in inward-opening dynamics. Thus, the MIC revealed a strongly coupled helical bundle (composed of TM1a, TM6b, and TM9) in the intracellular side that was highly correlated to the N-term/ICL4 association, suggesting how the N-term can modulate the stability of this bundle and thus modulate intracellular gating. In combination, these results identify the N-terminus as an important allosteric modulator of the functional inward-opening and ion/substrate release in hDAT.

3.3. The D_2 Dopamine Receptor

3.3.1 The asymmetric D_2 receptor homodimeric signaling complex as an illustration of AIM-based analysis of allosteric coupling mechanisms

The D2 dopamine receptor is know to signal as both a monomer and a homodimer, but a novel experimental construct developed in the Javitch lab [229] was required to make possible the characterization of the dimer as a signaling unit. The results demonstrated

experimentally that rather than signaling through each monomer independently, the D2R homodimer signals through a single protomer at a time (the signaling protomer will be referred to as "protomer A"). Furthermore, the results indicate that the function of the protomers is characterized by negative cooperativity: the stabilization of the *on* state of the non-signaling monomer ("protomer B") by agonist biding decreases signaling by protomer A, whereas the stabilization of the *off* state of protomer B by the binding of an inverse agonist increases signaling by protomer A. Lastly, it is shown in [229] that perturbations known to completely disrupt activation in the monomer, including (i)-ablation of ligand binding, (ii)-removal of intracellular loop 3 (IL3), and (iii)-mutations to (a)-intracellular loop 2 (ICL2), (b)-the conserved DRY motif, and (c)-the conserved NPxxY motif – all disrupt activation in the homodimer when applied to protomer A. Unexpectedly, however, the perturbations in (iii) also disrupt activation when applied to protomer B.

To explain the experimental results in a structural context, a molecular model of the homodimer complex with the G protein that senses the activation of the receptor was constructed in [229] using the active state crystal structure of another GPCR, rhodopsin, bound to its G protein, transducin. In this molecular model the interface of the homodimer involves the 4th transmembrane segment (TM4), and the G protein interacts with the signaling protomer A through IL3, IL2, and helix 8 (H8), while protomer B interacts through its IL2 and H8 (see **Figure 38**). We used AIMs as described below to explore the feasibility of the allosteric properties proposed for this structural model.

Based on the experimental measurements of activation, an AIM representing the homodimer was constructed starting with a model for a signaling monomer (monomer A) and a G protein that can bind this monomer and become activated. Since the

experiments had shown[229] that mutations in IL2, the DRY motif, and the NPxxY motif has identical phenotypes with regard to G protein binding, we represented all three as a single structural component termed *conserved binding motifs* (CBMs), due to their role in G protein activation by the GPCR[230–234]. In the AIM constructed accordingly (see **Figure 38A**), the signaling monomer is composed of the following structural components: a ligand that can bind and unbind, a transmembrane domain, and two intracellular regions (IL3 and the CBMs); the G protein is composed of a structural component that can bind and unbind the signaling monomer, and one that can be activated. The conformational energies of the components of each protomer were chosen to prefer the *off* state ($u^{conf} = 1$), and the interaction energies between all components were negative such that they preferred to be in the same state ($u^{int} = -1$). We find that this coarse grained model responds as expected to agonists, antagonists, and inverse agonists (see **Figure 38B**). To create a homodimer with negative cooperativity, we then added to the AIM a negative interaction between the one monomer that can bind G protein (which is now protomer A) and one that cannot (protomer B), represented as a positive interaction energy between their transmembrane domains (see **Figure 38C**). We then calculated the allosteric efficacy for the homodimer when promoter A was bound to agonist and protomer B was simultaneously bound to either an agonist, an antagonist, or an inverse agonist. This model reproduces the observed negative cooperativity (see **Figure 38D**).

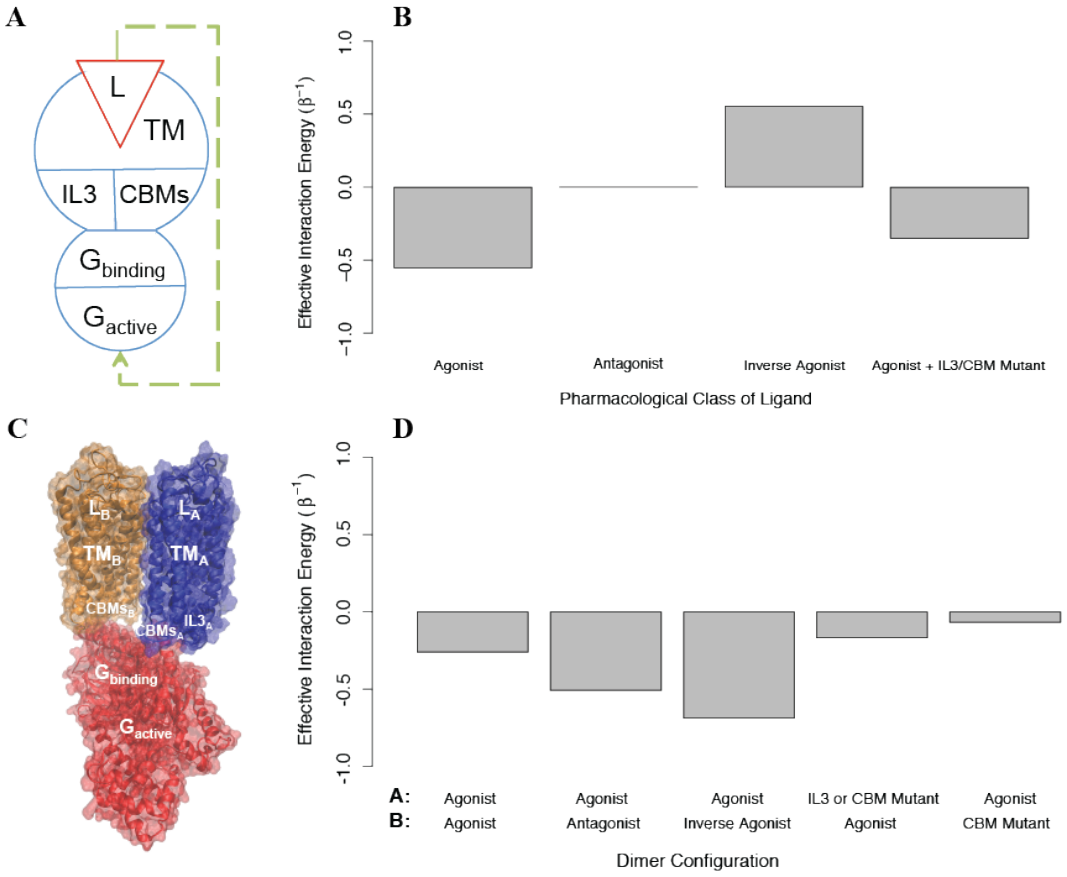


Figure 38. Analysis of the AIM for a well-characterized asymmetric D2 homodimer of the dopamine D2 receptor (D2R).

(A): The D2R monomer AIM. **(B):** The effective interaction energy calculated for the D2R monomer AIM is presented for ligands that are agonists, antagonists, and inverse agonists, and also for the mutation of either IL3 or the *conserved binding motifs* (CBMs). **(C):** A molecular model of the homodimer obtained as described in the text, is shown with each AIM domain labeled in white on the structural representation. Protomer A is in blue, protomer B is in orange, and the G protein is in red. **(D):** The effective interaction energy for the D2R homodimer AIM is presented for different combinations of the states of protomer A (indicated by **A** in the top row) and those of protomer B in the dimer (**B**, bottom row).

To explore the effects of removing IL3 and introducing the CBM mutations, we constructed AIMs with the perturbations modeled as either i) stabilizing the *off* state of the mutated structural component, ii) stabilizing its *on* state, or iii) reducing the interaction energy between the structural component and the G protein to 0. Modeling the two perturbations in protomer A by imposing (i) or (iii), reduced activation as expected. However, stabilizing the *off* state of IL3 in protomer B increases activation in our model when it should have no effect, indicating that treating the IL3 mutation such that it eliminates interaction between IL3 and the G protein is a better model. On the other hand, treating the CBM perturbation in protomer B as stabilizing the *off* state leads to more activation, so that the effect of the mutation cannot be explained without an interaction between the CBM in protomer B and the G protein. To reconcile these effects in the model, we assumed that protomer B and the G protein bind in a state-independent way (the G protein's state independent binding is represented by $u^{conf}_{G_{binding}}$ in the AIM), and modeled the CBM mutation effect as further decreasing state-independent binding. We find that if $u^{conf}_{G_{binding}}$ is increased from 1 to 2, allosteric efficacy is reduced (see **Figure 38D**). The finding that state-independent interactions between the G protein and CBMs on both protomer A and protomer B are required for activation is in full agreement with the structural model of the dimer as presented[229], in which not only protomer A, but also ICL2 and H8 from protomer B interact with the G protein directly. *As this structural information was not used in the construction of the AIMs, the prediction from the allosteric model underscores the ability of the AIMs-based approach in this illustration to connect the representation of allostery with the structural context of the modeled biomolecular systems.*

3.4. 5-HT_{2A}R

Of the fifteen different receptors activated by the neurotransmitter serotonin, the 5-HT_{2A} subtype is of great interest not only because it plays a crucial role in cognitive processing but also because it is the target of a large number of medications including antidepressants and antipsychotics[235–237]. Remarkably, several 5-HT_{2A} agonists, such as LSD[28], are known to display hallucinogenic properties. Indeed, a large body of evidence indicates that the common target of all hallucinogens is the 5-HT_{2A} receptor (5-HT_{2A}R)[28,235,237].

Given that 5-HT_{2A}R agonists and partial agonists can exhibit hallucinogenic properties or not by activating the same receptor, indicates a strong functional selectivity. Functional selectivity by hallucinogenic ligands (HLs) and non-hallucinogenic ligands (NHLs) at the level of PLC and PLA signaling pathway activation has been observed pharmacologically at 5-HT_{2A}R and 5-HT_{2C}R[238,239,240,241]. Furthermore, it has been observed that in a transgenic mouse model with humanized HT_{2A}R, hallucinogenic 5-HT_{2A}R agonists induce a gene expression profile distinct from that elicited by non-hallucinogenic 5-HT_{2A}R agonists[242,235]. In addition, recent computational work by our lab investigated the activation of the 5-HT_{2A} serotonin receptor by endogenous, hallucinogenic, and non-hallucinogenic ligands using Molecular Dynamics (MD) simulations of homology models of 5-HT_{2A}R[233]. This work indicated that these ligands can induce very different structures and dynamics in known functional micro domains similar findings were reported by others for the β_2AR adrenergic receptor using computational[243] and experimental techniques[244,245]. Lastly, very recently, structures of 5-HT_{1B}R and 5-HT_{2B}R bound to the agonist ergoline were solved using x-ray crystallography[246,247]. Ergoline displays no β-arrestin bias in 5-HT_{1B}R while showing substantial bias in 5-HT_{2B}R. While nearly all known activation motifs were in

their "active-like" conformation in both structures, the conserved PIF motif displayed an "inactive-like" conformation in the 5-HT_{2B}R:ergoline structure, providing the first structural evidence that allosteric modulation of known functional microdomains may contribute to functional selectivity. These results support the role of specific structural components in activation and inactivation and also indicate that ligands induce distinct conformational ensembles that may be responsible for functional selectivity.

3.4.1 Identification of Hallucinogen-Specific Allosteric Modulation of 5-HT_{2A}R

The Methods and Results described in this section are an abbreviated adaptation of previously published results[131,248], adapted with permission. This will be followed below by a more complete description of new analysis of these trajectories, which is currently in preparation.

The remarkable functional selectivity of HL compounds on 5-HT_{2A}R[235,249–251], included in the ample literature on the experimentally determined properties of the receptor and of structure-activity relations for its ligands[28], prompted us to investigate structural and dynamical elements associated with the functional selectivity of the HL and cognate NHL 5-HT_{2A}R agonists[235]. To cover a chemically distinct ligand space, we selected 5-HT_{2A}R complexes with the four agonists (*i*-*iv*) described below for extensive unbiased all-atom MD simulations. We chose two HL compounds: (*i*)- the hallucinogenic substituted amphetamine, 2,5-dimethoxy-4-iodoamphetamine (DOI), and (*ii*)- the prototypical hallucinogen L-lysergic acid diethylamide (LSD). DOI has a relatively small and flexible chemical structure, whereas LSD is representative of the larger and more rigid chemical family of ergots. We also selected two cognate NHL compounds: (*iii*)- the endogenous 5-HT_{2A}R ligand serotonin (5-HT), and (*iv*)- the partial agonist R-lisuride (LIS) that belongs to the same chemical family as LSD but

has a very different pharmacological fingerprint with regards to perceptional and cognitional phenotypes[235]. These ligands are shown in **Figure 39**.

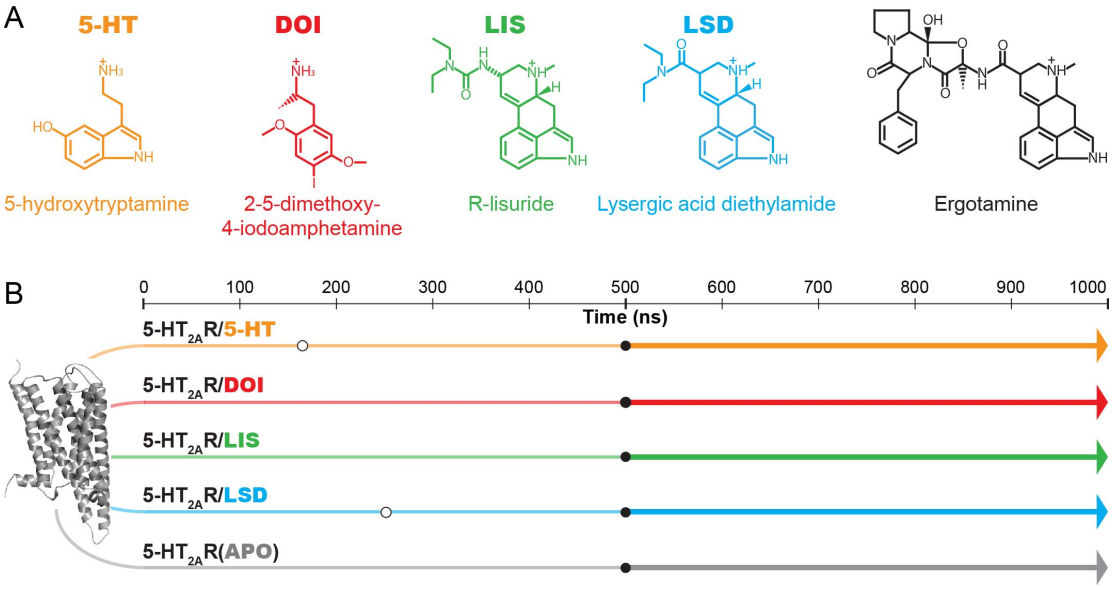


Figure 39. Schematic representation of the MD simulations and different ligands. **(A)** The structures of the four ligand agonists involved in this study, two hallucinogens (LSD and DOI) and two non-hallucinogens (5-HT and LIS), are depicted. For comparison the structure of ergotamine is also depicted. Ergotamine is one of the ligands co-crystallized with the closely related serotonin receptors, 5-HT_{1B}R, and 5-HT_{2B}R. **(B)** Starting from the same 5-HT_{2A}R structure, five different simulations were carried out. Although the production phase consists of 1000 ns, during the analysis just the second half, from 500 ns to 1000 ns, was considered. The midpoint of the simulations is indicated with black dots. Previously, a short segment of the two of the simulations (5-HT and LSD) were included as part of a study from our group[233]. These segments range from the starting point until the point indicated by an open circle in each of the simulations, ~175 ns for 5-HT and ~250 ns for LSD.

All four ligands have been extensively characterized with diverse biophysical and physiological techniques *in vitro* and *in vivo* with respect to serotonergic signaling efficacy across several downstream pathways and hallucinogenic phenotypes[235].

3.4.1.1. Methods

Microsecond unbiased all-atom MD simulations were carried out in the membrane-embedded 5-HT_{2A}R in the unbound form (APO) and in complex with four different agonists: 5-HT, LSD, DOI and LIS. For two of the systems described here (5-HT_{2A}R/5-HT) and (5-HT_{2A}R/LSD), shorter segments of the simulations (relative to the extent of the trajectories presented in this work) were part of a previous study from our group[233]. As a control, MD simulations of two closely related 1B (5-HT_{1B}R) and 2B (5-HT_{2B}R) human serotonin receptors in complex with the 5-HT ligand were carried out (100 ns each). All analyses were performed on the second half of the trajectories.

3.4.1.1.2. 5-HT_{2A}R structure complexes.

The different systems were constructed as described previously[233]. Briefly, the 5-HT_{2A}R model was created with homology modeling using as templates, the high-resolution X-ray crystal structures of the β_2 adrenergic receptor (PDB accession code, 2RH1) and bovine rhodopsin (PDB accession code, 1U19)[252]. The crystal structures of two closely related human serotonin receptors, the 1B (5-HT_{1B}R) and 2B (5-HT_{2B}R) receptors, were solved after the MD simulations presented here were collected[246,253] and thus, they were not considered as template for the 5-HT_{2A}R structure, but were used for validation and controls. The resulting 5-HT_{2A}R structure is comprised of the segment S67 to K400 (a 28-residue segment in the long ICL3, the first 66 N-terminal residues and the last 70 C-terminal residues were not included, see Fig. S3A in *SI*) and was capped at its N- and C-termini by the acetyl and N-methylamide groups,

respectively. A palmitoyl moiety was attached at position C397 based on the structural information of the β_2 adrenergic receptor (PDB accession code, 2RH1). All MD simulations started from the same 5-HT_{2A}R structure and the initial positioning of the agonists in the ligand binding pocket of 5-HT_{2A}R was carried out by using several docking protocols (*i.e.*, Autodock 4[254], Simulated Annealing Docking[255], and Glide and IFD (Schrödinger Inc.)) and were consistent with experimental information[233]. The 5-HT_{2A}R systems were embedded in a physiologically relevant lipid membrane composed of a symmetric 7:7:6 mixture of SDPC (1-stearoyl-2-docosa-hexaenoyl-sn-Glycero-3-phosphocholine):POPC (phosphatidylcholine):Cholesterol, respectively. The GPCR-membrane systems were then hydrated by using the TIP3P water model followed by neutralization of the entire system by introducing ions to generate a NaCl salt concentration of 0.15 M[233].

The parameters for the different ligands were obtained as described previously[233].

3.4.1.1.3. All-atom molecular dynamics simulations.

Details of the 5-HT_{2A}R simulations are as described previously[233]. Briefly, unbiased all-atom MD simulations were performed using NAMD[178] with the all-atom CHARMM27 force field with CMAP corrections for proteins and lipids[176] for trajectories of at least 1 microsecond. Langevin dynamics and the hybrid Nosé-Hoover Langevin piston were used to maintain constant temperature (310 K) and constant pressure (1 atm), respectively. Full electrostatics were evaluated using PME techniques with grid spacing less that 1.0 Å in each dimension and a fourth-order interpolation. Bond lengths involving hydrogen atoms were constrained to their equilibrium values by the SHAKE algorithm[256]. All MD simulations were performed with a 2 fs time step.

3.4.1.1.4. Structural alignment.

For the structural analyses, all the structures were aligned to the structure of the β_2 adrenergic receptor (PDB accession code, 2RH1) oriented with respect to the lipid bilayer according to the Orientations of Proteins in Membranes (OPM) database[257] by using the Cα atoms of the TM helices. Such alignment ensured that the Z-coordinate axis coincided with the helical axis of the TM bundle.

3.4.1.1.5. Principal component analysis.

We used PCA to quantify the major motions in ICL2. Using the Cα and heavy atom covariance matrices, we first found the first principal component (PC1) of the ICL2 movement in each system, which represented a large portion of the variance in all systems except for APO. To investigate differences in ICL2 dynamics in all five systems, we calculated PC1 for each simulation and then calculated the variance across that principal component for each other simulation. Atomic fluctuation correlations were calculated using *carma*[179], and PCA was performed with in-house programs.

3.4.1.2. Results

3.4.1.2.1. ICL2 rigid-body dynamics are modulated upon ligand binding.

The application of the *total intercorrelation* and entropy decomposition measures to characterize rigid-body behavior was illustrated with the analysis of the results from for 1 microsecond MD simulations of 5-HT_{2A}R, in the apo and 5-HT-bound states. The analysis focused on the secondary structure of ICL2 of the $5HT_{2A}$R. Residues I181-F186 were helical within the initial structures of both states, and traditional

secondary structure calculation using *stride* [258] indicates that the interior helical turn, composed of residues H182-R185, is stable throughout our simulations, with the turn being entirely helical for 84.3% of the apo trajectory and 89.7% of the 5-HT-bound trajectory. For the remainder of the analysis, we consider residues I181-F186 as ICL2.

We then calculated generalized correlation coefficients (Eq. (1.140)) using *N-body mutual information* (Eq.(1.141)) and *N-body total intercorrelation* (Eq. (1.168)) for ICL2 in both simulations to quantify the rigid-body behavior of the helical segment and to assess if there were differences in rigid-body behavior between the two states that could not be observed by calculating the secondary structure alone. We found that the apo state displayed weak rigid-body dynamics ($r_{mutual} = 0.30$ and $r_{TCinter} = 0.60$), while the 5-HT bound state displayed stronger rigid-body dynamics ($r_{mutual} = 0.52$ and $r_{TCinter} = 0.89$). These results indicate that there are increased rigid-body motions in the 5-HT bound simulation, although both states have a helical segment in the ICL2.

Using the entropy decomposition framework to analyze the dynamics of IL2, one would expect a high RBF and CO if the helical segment truly behaves as a rigid body helix, and a moderate RBF and low CO if the backbone is behaving like a rigid body but the side chains are not (likely a more accurate expectation based on the previously calculated generalized correlation coefficients). Conversely, if ICL2 were behaving as a completely disordered segment, which is not expected from its helical secondary structure, RBF and CO would be low. We find that while ICL2 is helical when the receptor is unbound, the RBF and CO parameters calculated from both the *mutual information* and *total intercorrelation* are low (see Table 7), indicating that ICL2 contains a very flexible helix. In addition, we find that the RBF increases in the 5-HT bound state of the 5-HT_{2A}R. Interestingly, the comparison of CO_{mutual} to CO_{inter} reveals

different trends upon 5-HAT binding, indicating that the choice of information measure can influence the interpretation of the system's dynamics. Thus, most of the high-order correlation are identified as rigid body using *total intercorrelation*, but not when using *mutual information.* These results indicate that there is a significant increase in the rigidity of the IL2 upon ligand binding although the helical secondary structure is retained and comparable in both states.

Table 7. Rigid-body parameters of the apo and 5-HT-bound 5-HT_{2A}R. The standard error on the mean of 50 bootstraps is displayed in parenthesis.

	Apo		**5-HAT**	
	Mutual	Inter	mutual	inter
r	0.30 (0.002)	0.60 (0.003)	0.52 (0.002)	0.89 (0.002)
RBF	0.16 (0.002)	0.50 (0.002)	0.39 (0.002)	0.90 (0.002)
CO	0.77 (0.001)	1.91 (0.001)	1.11 (0.004)	0.67 (0.005)

Moreover, a greater overall rigidity is indicated for both systems when using *total intercorrelation* as opposed to *mutual information*, as seen in the N-body generalized correlation coefficient and *rigid-body fraction*. We expect this result to be general and apply to other systems as well. However, we find that the RBF and CO parameters are greater when using *mutual information*. Thus, we find that ICL2 of 5-HT_{2A}R transitions from a flexible helix to a more rigid-body helix upon binding the endogenous agonist 5-HT. As previous crystallography data [33,44,259,260] and computational analysis [261,262] have pointed to the helix properties of IL2 in relation to

GPCR activation for different pathways, it is possible that ligands can determine their agonist bias by allosterically modulating the rigid-body properties of IL2 upon binding.

3.4.1.2.2. ICL2 adopts distinct conformations in 5-HT_{2A}R complexes with different ligands.

From the analyses of the microsecond MD simulation trajectories of 5-HT_{2A}R with different ligands, we found that ICL2 conformations favored in the HL-bound systems are different from those favored in the NHL-bound and in the unbound constructs. The distinct conformations were monitored by defining the center of mass of the helical segment on ICL2 as a collective variable, and calculating the root-mean-square deviation (rmsd) of the center of mass of the ICL2 along the trajectories, relative to the center of mass of the ICL2 in the initial structure. The distributions of the rmsd values show two distinct conformations for the ICL2 (see **Figure 40**).

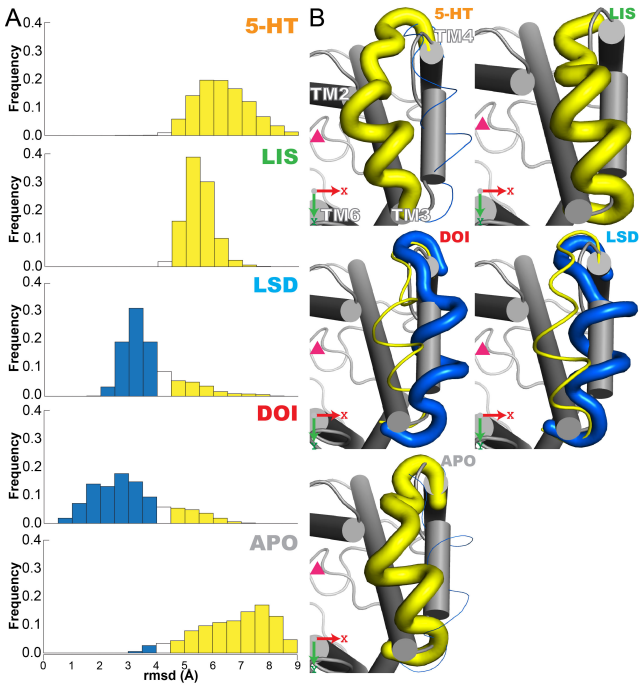


Figure 40. RMSD distribution and representative structures of ICL2.

(A) The distributions of rmsd values, relative to the starting structure, are shown for the five simulated systems, 5-HT_{2A}R/5-HT, 5-HT_{2A}R/LIS, 5-HT_{2A}R/LSD, 5-HT_{2A}R/DOI and 5-HT_{2A}R (APO), respectively. The more outward-upward conformations (blue) are highly favored in just the hallucinogenic systems. **(B)** Representative ICL2 structures for the five simulated systems, as seen from the intracellular side, are shown. As a reference, the initial structure (gray) is also depicted in each case. The more outward-upward ICL2 conformations are colored blue whereas the more inward-downward conformations are colored yellow. In these views, the more outward ICL2 conformations correspond to larger values in the X-axis coordinate. Interestingly, the outward-upward conformations (blue) are preferentially stabilized in the hallucinogenic systems, LSD and DOI. The thickness of the ICL2 representation corresponds to the percentages of the distributions from **(A)**. In the case of LIS, any of the conformations was sorted as part of the "blue" conformations. The helical axis of the TM bundle is represented by a magenta triangle in each case.

The more “outward” and more “upward” oriented ICL2 conformations (colored blue in **Figure 40B**) are seen to be highly favored by HL (DOI and LSD, see middle panel in **Figure 40B**) in contrast to the more “inward” and more “downward” ICL2 conformations (colored yellow in **Figure 40B**) adopted when the NHL (5-HT and LIS) are bound or when the unbound (APO) receptor is simulated. The representative structures of the ICL2 segment conformations in each of the studied systems (Fig. 1B) show that the more outward conformations (favored by HL) situate the ICL2 segment farther away from the axis of the TM helical bundle, whereas more upward conformations place the ICL2 segment closer to the center of the membrane bilayer Representative structures in the 5-HT_{2A}R/DOI complex are also depicted in **Figure 41**. In this particular complex, ICL2 selectively prefers more outward-upward conformations (colored blue), but explores as well the inward-downward ICL2 conformations preferred by the NHL (colored yellow), see **Figure 41**. All ligand-bound receptors exhibited dynamic transitions between states, but with notable preferences related to their pharmacological class (see **Figure 40B**).

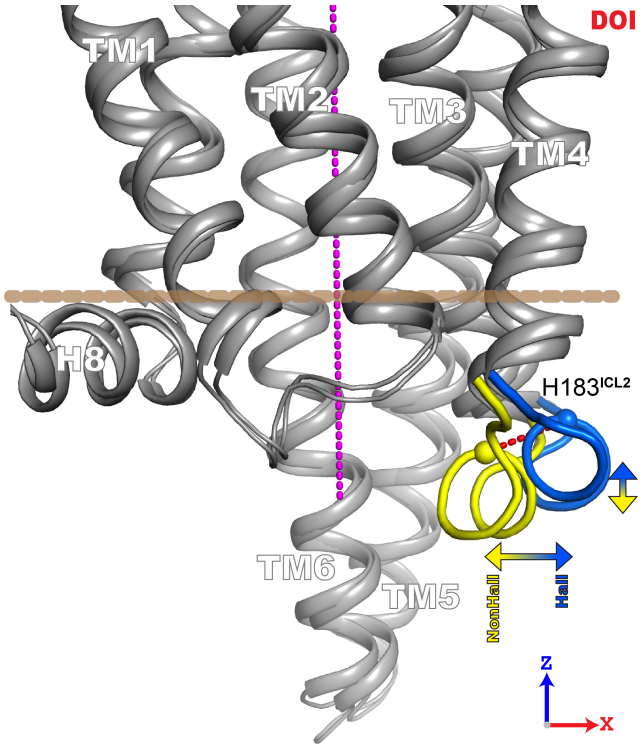


Figure 41. Conformations explored by the ICL2 in the 5-HT$_{2A}$R/DOI complex. Lateral view of two representative structures of the 5-HT$_{2A}$R/DOI system. The more outward (relative to the helical axis of the TM bundle, shown here as a magenta line) and more upward (that is, closer to the center of the lipid bilayer) are preferred in the hallucinogenic systems (DOI and LSD), colored here in blue. In these views, the more outward ICL2 conformations correspond to larger values in the X-axis coordinate while the more upward conformations correspond to larger values in the Z-axis coordinate. The more inward-downward conformations are preferentially sampled in the non-hallucinogenic systems (5-HT and LIS) and in the APO form, colored here in yellow. As a magnitude reference, the Cα atoms of residue H183 are depicted in both structures and the distance for these particular structures is 5.3 Å (indicated as a red line). The predicted intracellular boundary of the bilayer is depicted as a brown line.

In addition, we calculated the generalized correlation coefficient between the center of mass and the first principal component (PC1) of the ICL2 motion, which indicated that the center of mass motion of ICL2 was strongly correlated with the PC1 in each system (see Table 8A). This finding further supports the use of the center of mass as a collective variable, as described above. From the PCA we further found that the PC1 motion in the 5-HT_{2A}R/DOI system accounts for a large fraction of the variation present in 5-HT_{2A}R/DOI and 5-HT_{2A}R/LSD systems, but not in the 5-HT_{2A}R/5-HT, 5-HT_{2A}R/LIS, or 5-HT_{2A}R (APO) systems, indicating that this motion is HL-specific (see Table 8B).

Table 8. The major motion in ICL2 is highly correlated to the COM and discriminates hallucinogens from non-hallucinogens.

A**. Cα motions in ICL2**					
	5-HT	**LSD**	**LIS**	**DOI**	**APO**
r(PC1, COM)	0.894	0.971	0.935	0.964	0.470
$PC1_{5\text{-}HT}$	0.637	0.502	0.452	0.248	0.098
$PC1_{LSD}$	0.498	0.629	0.477	0.469	0.106
$PC1_{LIS}$	0.545	0.573	0.523	0.384	0.104
$PC1_{DOI}$	0.215	*0.473*	0.316	*0.605*	0.092
$PC1_{APO}$	0.057	0.034	0.043	0.024	0.296
B**. Heavy atom motions in ICL2**					
	5-HT	**LSD**	**LIS**	**DOI**	**APO**
r(PC1, COM)	0.871	0.888	0.830	0.880	0.608
$PC1_{5\text{-}HT}$	0.460	0.376	0.225	0.180	0.071

$PC1_{LSD}$	0.351	0.484	0.262	0.355	0.095
$PC1_{LIS}$	0.366	0.437	0.287	0.313	0.082
$PC1_{DOI}$	0.151	*0.345*	0.183	*0.475*	0.071
$PC1_{APO}$	0.068	0.083	0.057	0.062	0.261

The major Cα (A) and heavy atom (B) motions of ICL2 are presented. Row 1 corresponds to the generalized correlation coefficient[9] between the first principal component of the ICL2 motions (PC1) and the center of mass of ICL2 (COM). Rows 2 – 6 correspond to the fraction of total variance in ICL2 that is contributed by a given principal component. $PC1_X$ corresponds to the first principal component of ICL2 motion found in system X.

To identify specific molecular interactions involved in the observed differential conformations of the ICL2 segment, we analyzed comparatively the contacts involving residues in ICL2. The direct interaction between residue D172$^{3.49}$ (from the conserved DRY motif) and H183 (located in the middle of the ICL2) was found to be more extensively maintained in the trajectories of HL systems compared to the NHL counterparts or the APO (see **Figure 42A**). **Figure 42B** shows that the minimal distance between any of the carboxylate oxygen atoms from the side chain of D172$^{3.49}$ with any of the imidazole nitrogen atoms from the side chain of H183^{ICL2} in the HL systems, fluctuates mainly to values ~ 4 Å or shorter. In contrast in the NHL systems the values are mostly larger than 4 Å (the 4 Å is selected as reference distance to match the cutoff distance value used herein to define a molecular contact). This interaction is proposed to play a key role in determining the different conformational and dynamic properties of the ICL2 in the HL versus NHL systems.

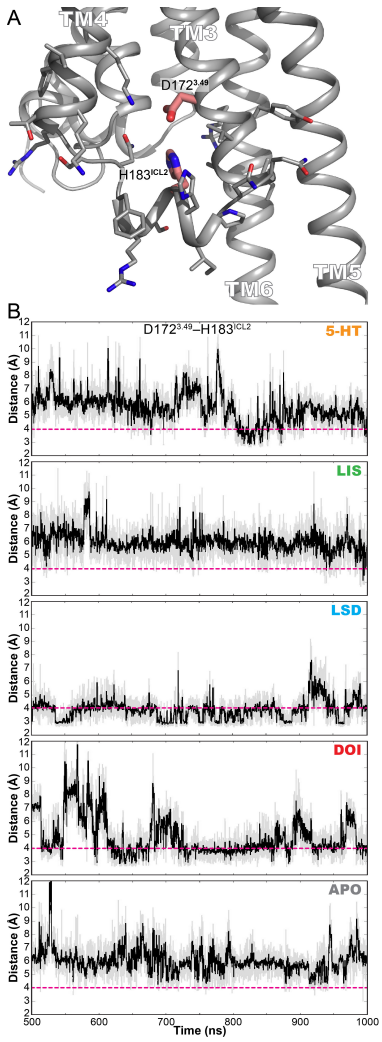


Figure 42. Distances of residues D172 and H183.

(A) Representative structure from the 5-HT_{2A}R/DOI complex where the interaction of $D172^{3.49}$ and ICL2 residue $H183^{ICL2}$, is depicted. In the context of the DRY motif, this position in the ICL2 is a residue that can establish polar interactions by using its side chain and is located in the sequence position ***Z*** in the "$DRY(X)_5P(X)_2$***Z***" motif. **(B)** The minimal distances between any of the carboxylate oxygen atoms from the side chain of $D172^{3.49}$ with any of the imidazole nitrogen atoms from the side chain of H183 are depicted. The distance (gray) and its moving average (black) are displayed. As a reference, a dashed line at 4 Å is also displayed (the same cutoff value used to define a receptor-ligand interaction contact).

3.4.1.2.3. Binding site interactions of 5-HT_{2A}R agonists.

The receptor-ligand contacts were evaluated by considering all positions at which any heavy atom from the ligand comes within 4 Å of any heavy atom from the protein in the course of the trajectory. From this set we identified six contact loci (I210$^{\mathbf{4.60}}$, G238$^{\mathbf{5.42}}$, S239$^{\mathbf{5.43}}$, F243$^{\mathbf{5.47}}$, W336$^{\mathbf{6.48}}$, and N343$^{\mathbf{6.55}}$) that were found in all the simulated complexes, but that exhibited differences in the frequency of contacts for HL versus NHL ligands (Fig. 5A and 5B). An additional position (W151$^{\mathbf{3.28}}$) was also found to have differential contact frequencies between HL and NHL but only in the case of the larger ergoline ligands, LSD and LIS (see **Figure 43A** and **Figure 43B**). **Figure 43A** and **Figure 43B** depict the seven residues in the context of their positions inside the binding site whereas **Figure 43C** displays their respective contact frequencies as the percentage of trajectory time in which each of the positions is in contact with the ligand. The location of this set of residues suggests that HL agonists preferentially interact with residues located in TM6, whereas their NHL counterparts preferentially establish contacts with residues in TM4 and TM3 (see **Figure 43A**). Both classes of compounds interact with residues in TM5 but the HL preferentially contact residues that are located at the helical interface formed with TM6, whereas the NHL contact residues located at the helical interface formed with TM3 and TM4 (see **Figure 43A**). Residues G238$^{\mathbf{5.42}}$ and S239$^{\mathbf{5.43}}$ present an interesting example of this selectivity because they occupy neighboring positions in the vicinity of the indole nitrogen of the 5-HT ligand (or equivalent atoms in the other ligands), see **Figure 43A**. Yet, position 5.42 is preferentially contacted by NHL (94%, 98% for 5-HT, and LIS versus 37% and 79% for DOI and LSD, respectively), whereas position 5.43 is contacted more extensively by HL compounds (24%, 31% for 5-HT, and LIS, versus 82% and 54%, by DOI and LSD, respectively). Interestingly, even though all the

ligands contact the same residues in the orthosteric binding site (albeit with different frequencies), two of the residues preferentially contacted by the HL are large aromatic amino acids that are located deep in the orthosteric binding pocket, *i.e.*, the highly conserved W336$^{\mathbf{6.48}}$, known to be implicated in signal transduction in different GPCRs[263], and F243$^{\mathbf{5.47}}$, known to modulate DOI-dependent downstream signaling in 5-HT_{2A}R[264](see **Figure 43B**).

It is noteworthy that in spite of the minimal chemical and structural similarity of the HL ligands, they both have a positively charged nitrogen atom and an indole-like nitrogen atom (or equivalent) which have long been considered to be particularly important in interacting in the 5-HT_{2A}R orthosteric binding site[264,265]. This is also the case for the NHL ligands. The lack of chemical and structural similarity within the groups, and the much greater similarity of compounds belonging to the different groups (cf. LSD and LIS), accentuates the significance of the identified common set of residues that establish different protein-ligand contacts in the HL versus the NHL systems.

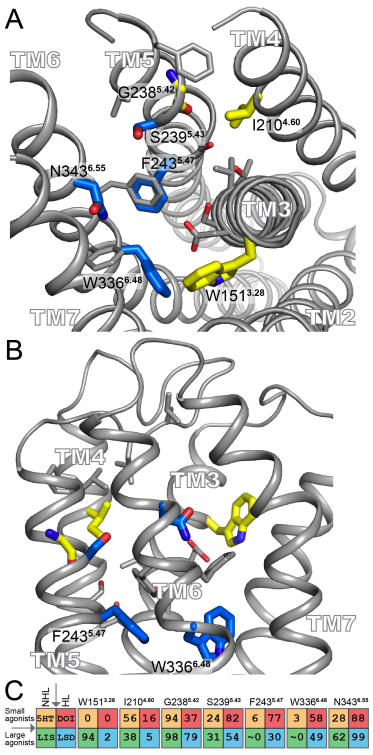


Figure 43. Ligand binding contacts in the 5-HT$_{2A}$R.

(A) Extracellular and **(B*)*** lateral views that show the seven residues (W151$^{3.28}$, I210$^{4.60}$, G238$^{5.42}$, S239$^{5.43}$, F243$^{5.47}$, W336$^{6.48}$, and N343$^{6.55}$) that display preferential frequency contacts between HL (blue) and NHL (yellow) ligands **(C)** The percentage of time that each of the seven positions are in contact with the ligands along the trajectories are shown. Similar color code is used, 5-HT (orange), DOI (red), LIS (green) and LSD (cyan). The different agonist types are arranged: *small* agonists (first row), *large* agonists (second row), NHL (first column) and HL (second column). To discern contact frequency differences between NHL and HL compounds compare data in the different columns in each case. Similarly, by discern contact frequency differences between *small* and *large* agonists compare data in the different rows. The first three residues show a tendency to directly interact with NHL (W151$^{3.28}$ only interacts with the ergoline ligands, LSD and LIS) whereas the other four show a preference for the HL.

3.4.1.3. Discussion

Our findings described here, identifying a role for the second intracellular loop of the 5-HT_{2A}R in discriminant pathway activations, are consistent with previous observations about the signaling of class A GPCRs through various intracellular signaling partners[230,266–268]. Thus, the ICL2 of the 5-HT_{2A}R has been shown to be involved in the interaction with G protein (including desensitization)[269] and with β-arrestin[268], whereas for the related serotonin 1A receptor, ICL2 has been directly implicated in G protein coupling[270]. The more recent structural information, for the β_2 adrenergic receptor complexed with the G_s protein, shows the ICL2 establishing extensive interactions with the β2/β3 loop in the N-terminus of the Gα subunit and with the C-terminus of helix α5[44,271]. In this context, the extensive unbiased MD simulations presented here provide evidence that different ligand classes bound to the 5-HT_{2A}R can produce distinct conformations of the ICL2. Thus, ICL2 favors more *outward-upward* conformations in the HL-bound systems (*i.e.*, the 5-HT_{2A}R/DOI and 5-HT_{2A}R/LSD complexes), while these conformations are not highly explored in the NHL systems, or in the unbound receptor. The spatial distributions of the ICL2 conformations relative to the helical bundle are similar among the HL systems (DOI and LSD), as quantitatively depicted by the calculation of the overlap coefficient of the ICL2 center of mass and the projections of the principal components, and are different from those adopted by the NHL counterparts (5-HT and LIS) or the unbound receptor. This is consonant with previous results from Lefkowitz and coworkers who used quantitative mass spectrometry to identify ligand-specific conformations of the β_2 adrenergic receptor and found that ICL2 adopts distinct conformations that differ between agonists[122].

Our computational analysis shows that the ICL2 conformations are likely to be largely dependent on the extent of the interaction between D172$^{3.49}$, from the conserved DRY motif in TM3, and H183 in the ICL2. Interestingly, interactions of D172$^{3.49}$ with H183-equivalent residues in ICL2 has been observed in the crystal structures of several other GPCRs: in all the opioids and the aminergic muscarinic receptors (with an Arg in the corresponding position)[266], and in the serotonin 1B receptor (with a Tyr in that position)[246]. Moreover, in another related GPCR, the aminergic β_1 adrenergic receptor, a hydrogen bond is formed between D172$^{3.49}$ and a tyrosine residue (Y149) in the equivalent ICL2 position, and introduction of the Y149A mutation, decreases receptor stability[39]. The relevance of this interaction is further emphasized by the fact the in the β_2 adrenergic receptor, the phosphorylation of the equivalent tyrosine (Y141) shifts the conformational equilibrium so as to facilitate active state conformations[272]. In the context of the DRY motif in TM3, this particular ICL2 position is located in the sequence position ***Z*** in the "DRY(X)$_5$P(X)$_2$***Z***" motif, where in all the aforementioned examples position ***Z*** is a residue able to establish side chain polar contacts with D172$^{3.49}$.

3.4.2. Identification of Hallucinogen-Specific Allosteric Modulation of Pairwise Interactions using a Random Forest-based Method

The original comparative analysis of HLs and NHLs identified hallucinogen-specific modulation of ICL2 by manually testing potential differences through trial and error. Additionally, we identified differences in the binding site and local to ICL2, but not a specific mechanism of allosteric transmission. To seek more hallucinogen-specific allosteric modulation that might be more general to all HL, we increased our comparison to include two new HLs, psilocin (PSI) and mescaline (MES), and another NHL, ergotamine (ERG) that has been crystalized in complex with 5HT2BR. We then

analyzed all of the HL and NHL simulations using the 2-step random forest-based method we developed to identify allosteric modulation of pairwise interactions. As we had previously identified a hallucinogen-specific change in the of D172$^{3.49}$ /H183^{ICL2} interaction, the approach was expected to reproduce the finding of our original analysis as well as identify new pairwise interaction that were subject to hallucinogen-specific allosteric modulation.

3.4.2.1. Methods

Systems were prepared and simulated as described in the previous section. All new simulations were initiated from a representative frame of the lisuride-bound complex (see **Figure 44**).

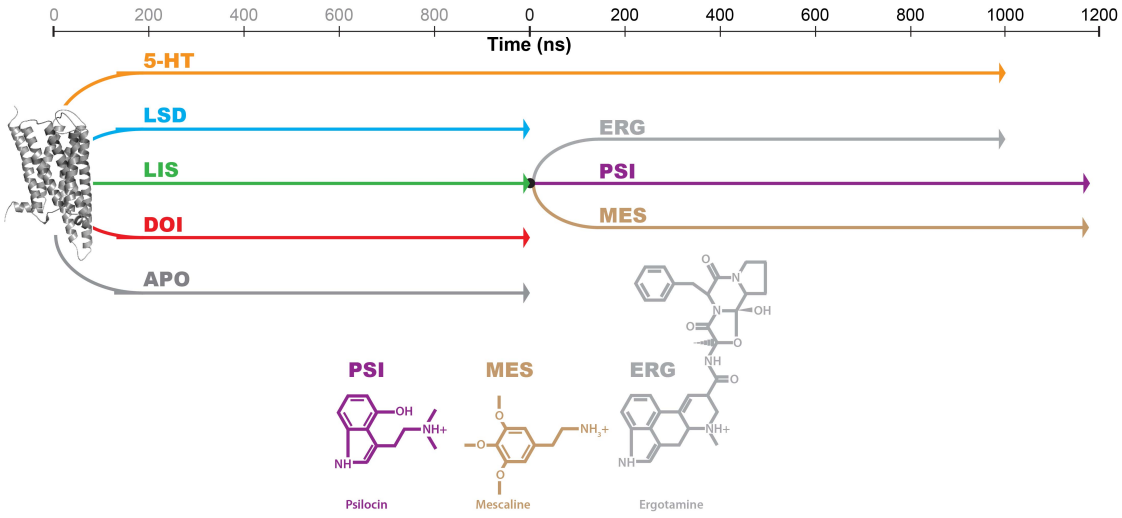


Figure 44. Additional simulations of psilocin, mescaline, and ergotamine.

While previously the last 500 ns of each simulation were analyzed, here we applied a recently described method[273] to identify how much of the initial portion of the trajectories should be discarded. In brief, the goal of the method is to find the time point, t_0, at which discarding all prior time points leads to maximization of the effective number of statistically independent data points, n_{eff}. Because it is impractical to apply this criterion to all pairwise interactions, we used the RMSD of TMs 1-4, after alignment using iterative fitting, as a global measure of convergence. We found that when plotting n_{eff} versus t_0, there was often an increase in n_{eff} after removing most of the trajectory (see **Figure 45**), and this sometimes resulted in a maxima towards the end of the trajectory. This behavior is likely due to a departure from the assumptions of the method (e.g. monotonic convergence to a specific average value), and thus we chose to ignore these end maxima in our choice of t_0.

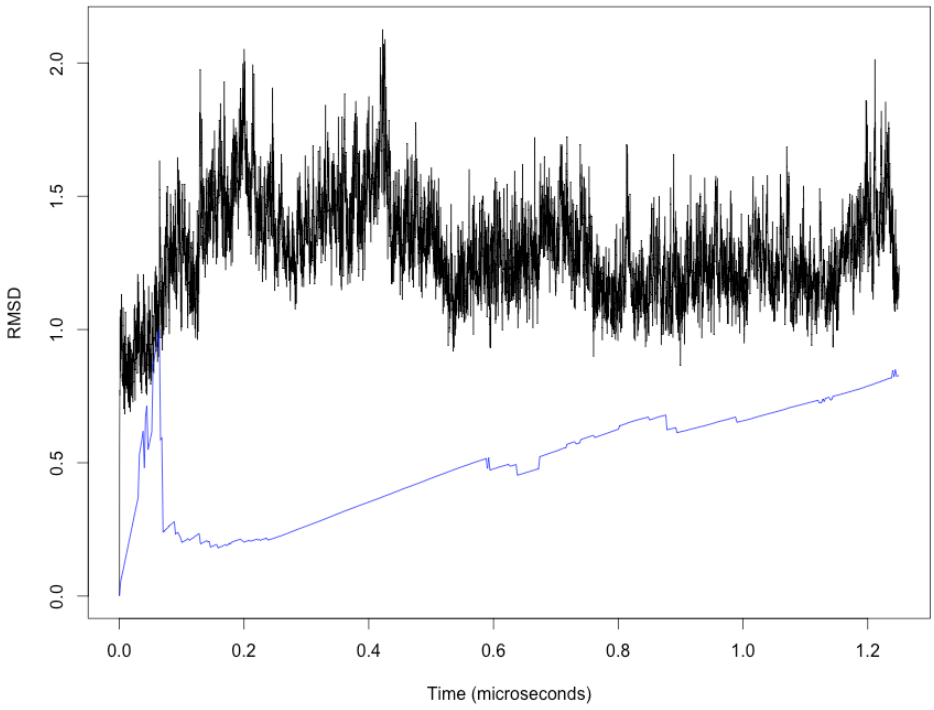


Figure 45. Convergence analysis.

An example RMSD time series is shown over 1.2 microseconds of simulation. N_{eff} is also shown (blue line) as a function of t_0, normalized to its maximum.

In order to identify pairwise interactions (PIs) throughout the trajectories, we used the PI Analyzer software. For this analysis, we used the default distance and geometry-based interaction parameters to identify pairwise interaction between all possible residues, without distinguishing between side chain or backbone interactions. All interactions that were not made at least once in all systems at any point in the simulations (including the discarded region) were removed in order to prevent rarely sampled interactions from dominating the analysis. In addition, since we aim for a classification of the class of ligand that is not limited to the trivial classification from the binding site alone, even if allosteric effects are present, but rather one that reveals

the class of the ligand from specifically identified distant allosteric modulation, the ligand-protein contacts were removed from the analysis.

3.4.2.2. Results

An important parameter in the two-step random forest analysis is k, which sets the cut-off for variable importance in both steps. Our analysis revealed two values of l that result in an α value of greater than 1 (see **Figure 46**). Here, we will present the analysis using the highest k cut-off of that still has $\alpha > 1$. By investigating the raw κ values for both within-class and between-class classification, we find that the lowest cut-off results in nearly no remaining accuracy to predict class ($\kappa = 0.1412$), whereas the highest cut-off that still has $\alpha > 1$ has reasonable accuracy ($\kappa = 0.472$). This is a weakness of using only the ratio to select the cut-off, and an improved automated selection method is required for the future.

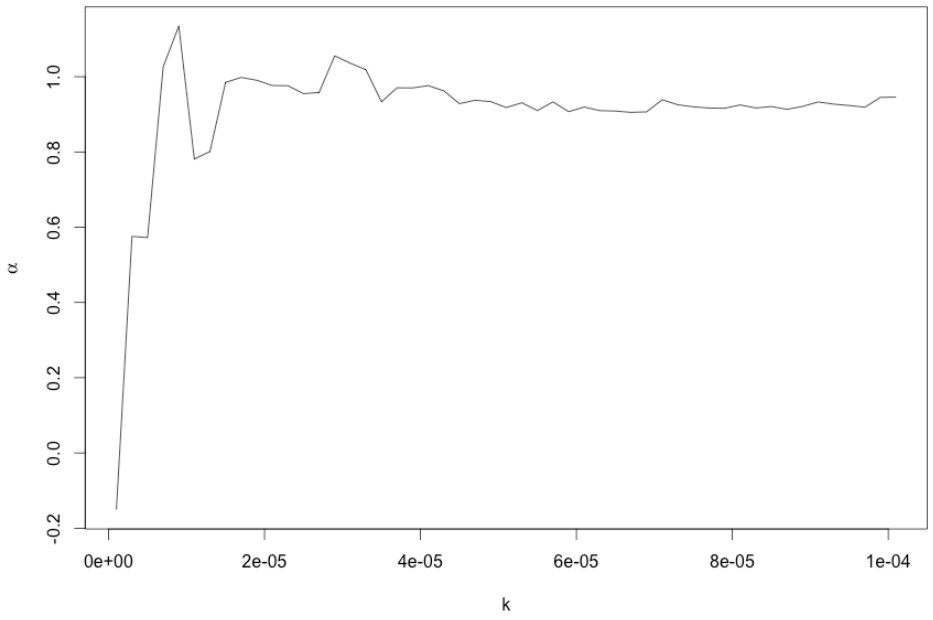


Figure 46. Class-specific classification of ligand as a function of the cut-off k. A representative plot of the class-specific classification score, α, is shown as a function of cut-off used, k.

Our analysis finds that PIs that contribute to discrimination between HL and NHLs occur throughout the TM region of 5-HT_{2A}R. While it may certainly be the case that the PI signature of HLs may span the entirety of the TM region, and recent analysis of NMR data[274] suggests that activation-associated conformational changes are present throughout the TM region, we investigated the top 7 PIs in further detail. As expected, an interaction was identified that involves H183^{ICL2}. However, while we had previously identified an interaction between H183^{ICL2} and D172$^{3.49}$ as being discriminant, the new analysis found the interaction between H183^{ICL2} and E318$^{6.30}$ to be discriminant (see **Figure 47**). Overall, the results reinforce the finding that allosteric modulation of the interaction of ICL2 with the ionic lock / DRY motif region is a characteristic of HLs.

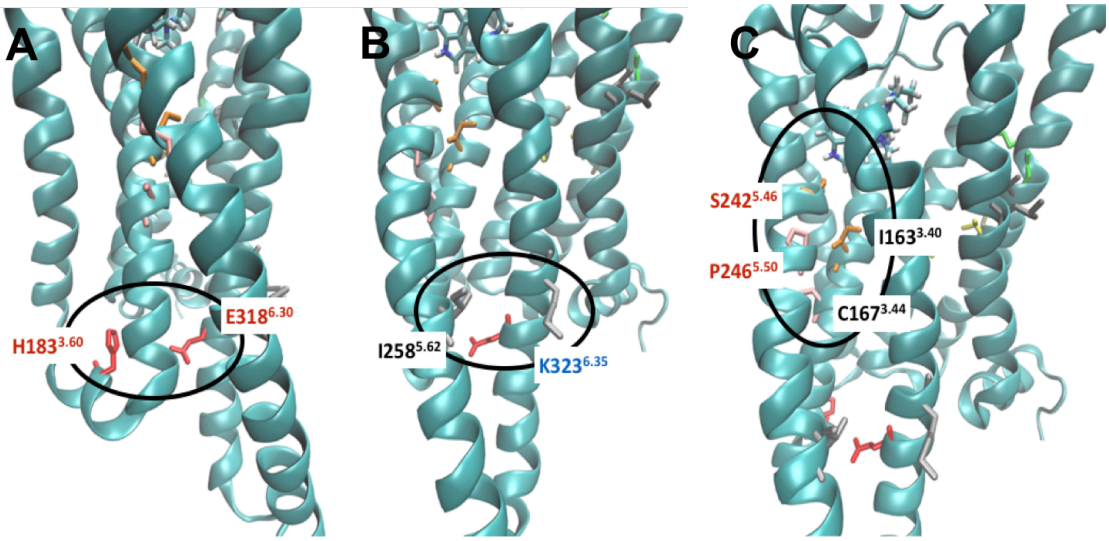


Figure 47. Hallucinogen-specific PIs.

4 of the top 7 interactions that are modulated in a hallucinogen-specific manner are presented. Residues denoted in red have been implicated in function, whereas residues in blue are conserved.

In addition to the H183^{ICL2} / E318$^{6.30}$ PI, an additional TM5 / TM6 interaction, I258$^{5.62}$ / K323$^{6.35}$, was also identified. Notably, K323$^{6.35}$ is strongly conserved among class A GPCRs and is local to E318$^{6.30}$, and conformational changes in TM5 / TM6 are a hallmark of activation.

Finally, two pairs of inter-TM interactions were identified at the interface of TM5 and TM, S242$^{\mathbf{5.46}}$ / I163$^{\mathbf{3.40}}$ and P246$^{\mathbf{5.50}}$ / C163$^{\mathbf{3.44}}$. These inter-TM interactions are both separated by a turn and indicate a HL-specific modulation of this interface. Interesting, P246$^{\mathbf{5.50}}$ and I163$^{\mathbf{3.40}}$ composed the so-called PIF motif that has recently been proposed to be involved in the difference in ergoline-induced activation of 5-HT$_{1B}$R and 5-HT$_{2B}$R[247]. Additionally, interaction with S242$^{\mathbf{5.46}}$ has been proposed to be involved in ligand efficacy[275,276], and the two interactions are local to two residues that we previously identified to make HL-favored contacts, S239$^{\mathbf{5.43}}$ and F243$^{\mathbf{5.47}}$, which are reproduced in the new analysis.

3.4.2.3 Discussion

Based on the new analysis, we hypothesize a process in which HL-specific engagement of TM5 triggers the propagation through TM5 to TM6 and TM3, which leads to allosteric modulation of the PIs between ICL2 and TM6 and the release of ICL2 into its previously described hallucinogen-specific conformation. These results provide a clear mechanism of hallucinogen-specific activation of 5-HT$_{2A}$R, in which HLs engage the existing activation mechanism (involving conserved and functionally relevant residues in TM5) in a ligand-specific manner leading to the stabilization of a functionally selective conformation of ICL2. Indeed, mutations in ICL2 have recently been shown to induce receptor bias[277], and comparison of x-ray structures of β_2AR bound to a heterotrimeric G protein complex and rhodopsin bound to arrestin-1

suggest that these downstream effectors differentially engage ICL2 and thus their binding would be differentially affected by hallucinogen-specific modulation of ICL2.

4. Expanding Allostery Past the Two-State Model

Little theory is available to describe allostery rigorously outside of the two-state models described in Section 1.1.1.2. Theoretical Background. However, there is no reason to assume systems will behave in a strictly two-state manner, and some systems are know to be more complex, as it was shown for the intracellular gate of LeuT in Section 3.1.2.2. Results. Thus, it is important to develop of a more general theory of allostery, even at the phenomenological level. Here, we derive a statistical mechanical form of the allosteric efficacy between collective variables that are either continuous or discrete.

4.1. Derivation

We would like to derive an analogous allosteric efficacy for the transformations of continuous or discrete variables. Let $\mathbf{r} \in \mathbb{R}^N$ represent the coordinates of a system that define the microstate. The microstates are distributed according to the Boltzmann distribution:

$$p(\vec{r}) = \frac{e^{-\beta U(\vec{r})}}{\sum_{\vec{r} \in R} e^{-\beta U(\vec{r})}} \tag{1.179}$$

where U(r) is the potential energy function. The free energy of this distribution is:

$$A\left[p(\vec{r})\right] = -\frac{1}{\beta} \log\left(\sum_{\vec{r} \in R} e^{-\beta U(\vec{r})} \right) \tag{1.180}$$

We consider a collective variable (CV), X(r), which is a function of the coordinates. The probability density f(x) expresses the probability that X(r) takes value x.

$$p(x) = \frac{\sum_{\vec{r}\in R} \delta_{X(\vec{r})-x} e^{-\beta U(\vec{r})}}{\sum_{\vec{r}\in R} e^{-\beta U(\vec{r})}} \tag{1.181}$$

where

$$\delta_{X(\vec{r})-x} \begin{cases} 1 & X(\vec{r}) - x = 0 \\ 0 & |X(\vec{r}) - x| > 0 \end{cases} \tag{1.182}$$

An analogous probability density function can be written for another CV, Y(r). For each CV, we can calculate the free energy of the distribution conditional on a value of the CV as:

$$A\left[p\left(\vec{r}\middle|X(\vec{r}) = x\right)\right] = -\frac{1}{\beta}\log\left(\sum_{\vec{r}\in R} \delta_{X(\vec{r})-x} e^{-\beta U(\vec{r})}\right) \tag{1.183}$$

Equation (1.183) can be rewritten as:

$$A\left[p\left(\vec{r}\middle|X(\vec{r}) = x\right)\right] = -\frac{1}{\beta}\log\left(p(x)\right) + A\left[p(\vec{r})\right] \tag{1.184}$$

We can also write a joint probability mass function for the two CVs:

$$p(x,y) = \frac{\sum_{\vec{r}\in R} \delta_{X(\vec{r})-x}\delta_{Y(\vec{r})-y} e^{-\beta U(\vec{r})}}{\sum_{\vec{r}\in R} e^{-\beta U(\vec{r})}} \tag{1.185}$$

And then an analogous free energy conditional on values of both CVs:

$$A\left[p\left(\vec{r}\middle|X(\vec{r}) = x, Y(\vec{r}) = y\right)\right] = -\frac{1}{\beta}\log\left(\sum_{\vec{r}\in R} \delta_{X(\vec{r})-x}\delta_{Y(\vec{r})-y} e^{-\beta U(\vec{r})}\right) \tag{1.186}$$

or

$$A\left[p\left(\vec{r}\middle|X(\vec{r}) = x, Y(\vec{r}) = y\right)\right] = -\frac{1}{\beta}\log\left(p(x,y)\right) + A\left[p(\vec{r})\right] \tag{1.187}$$

We would now like to see if it is possible to calculate an allosteric efficacy between transformations of these collective variables, given the equilibrium joint probability distribution is known. Therefore, we will calculate the allosteric efficacy for transformations in which the CVs are constrained to specific values:

$$\begin{array}{ccc} p(\vec{r}) & \xrightleftarrows{\Delta A_1} & p\left(\vec{r}\middle|X(\vec{r})=x\right) \\ \xrightleftarrows{\Delta A_4} & \begin{array}{c}\Delta\Delta A=\Delta A_2-\Delta A_1\\ \Delta\Delta A=\Delta A_3-\Delta A_4\end{array} & \xrightleftarrows{\Delta A_3} \\ p\left(\vec{r}\middle|Y(\vec{r})=y\right) & \xrightleftarrows{\Delta A_2} & p\left(\vec{r}\middle|X(\vec{r})=x,Y(\vec{r})=y\right) \end{array} \tag{1.188}$$

We will refer to this class of thermodynamic cycles as "allosteric cycles". The thermodynamic coupling in this cycle can be calculated as:

$$\Delta\Delta A(x,y)=A\left[p\left(\vec{r}\middle|X(\vec{r})=x,Y(\vec{r})=y\right)\right]-A\left[p\left(\vec{r}\middle|X(\vec{r})=x\right)\right]-A\left[p\left(\vec{r}\middle|Y(\vec{r})=y\right)\right]+A\left[p(\vec{r})\right] \tag{1.189}$$

Equation (1.189) simplifies to:

$$\Delta\Delta A(x,y)=-\frac{1}{\beta}\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1.190}$$

Interestingly, $\Delta\Delta A$ in (1.190) is proportional to the *pointwise mutual information* (PMI).

$$PMI(x,y)=\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1.191}$$

From the perspective of information theory, the PMI is quantifies the loss in information gained from measuring one variable given that one has already measured another variable, when the measurements are those specified. The measure is symmetric, i.e. the order of variables does not matter. In fact, the mutual information is the average PMI, weighted by the equilibrium probability mass function.

$$I[p(x,y)] = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (1.192)$$

One can immediately see the weakness of using the mutual information between two CVs as a description of their allosteric coupling. Functionally significant perturbations to allosterically coupled collective variables are not required to drive the system towards a region of the collective variable phase space that is already high probability at equilibrium. In fact, perturbations such as ligands generally drive the system away from its unbound equilibrium (e.g. from the inactive state to the active state). Thus, the mutual information does not necessarily capture the allosteric couplings that determine the response to physiological or synthetic modulators of function. It is instead preferable to analyze the entire 2-dimensional coupling surface, which we call the "allosteric landscape", as it contains information regarding the allosteric efficacy for all possible perturbations to the distribution of those CVs.

4.2. Illustration on Alanine Dipeptide

To illustrate the utility of this representation of allostery, we analyzed the allosteric landscape of alanine dipeptide. The alanine dipeptide free energy landscape is generally described by two CVs – the ϕ and ψ angles around the peptide bond – and is a popular model system for free energy methods. However, the irregular free energy surface indicates that these CVs are thermodynamically coupled in a non-trivial way. Thus, alanine dipeptide is an ideal model allosteric system.

4.2.1. Methods

The alanine dipeptide (N-Acetyl-Alanine-N'-Methyl amide, see **Figure 48A**) was modeled with the all-atom charmm36[278] force field and solvated in explicit TIP3P water molecules. Molecular dynamics simulation were performed using the Charmm

port[279] in the Gromacs 4.5 program[280] with particle-mesh Ewald[281] treatment of electrostatics and Lennard-Jones interactions switched off between 10Å and 12Å.

The systems were maintained at temperature T=300K with Nosé-Hoover chain thermostats[190]. Enhanced sampling was achieved with the driven adiabatic free energy dynamics[282,283] (dAFED), also known as temperature accelerated molecular dynamics[284] (TAMD), implemented in the PLUMED plugin[285]. Two collective variables (CVs), defined as the backbone dihedral angles ϕ and ψ were coupled (harmonic constant 1000 kJ/mol/rad^2) to heavy fictitious particles (pseudo-mass 50 amu•nm^2/rad^2) held at temperature T_s=600K or T_s=1000K by generalized Gaussian Moment thermostats (order 2)[286]. After a standard equilibration phase, simulations were conducted in five independent replicates of 50ns each. Free energy surfaces (FESs) in the (ϕ,ψ) plane were reconstructed[287] using the reweighted histogram smoothed with multivariate Gaussian kernel regression in Matlab (release 2014b, The MathWorks, Inc., Natick, Massachusetts, United States). A cutoff of 50 kJ/mol was used for the FESs, above which sampling was too poor for reliable surface estimation.

In principle, estimating an observable from a dAFED/TAMD simulation requires binning the observable values in the CV space, and reweighting each bin by a function of the FES at this point[288]. However, the allosteric coupling depends only on the density at 300K in the CV space, $p(\phi,\psi)$. This can be obtained directly from the density obtained from the dAFED/TAMD simulation, $p_{obs}(\phi,\psi)$, by rescaling and re-normalizing:

$$p\left(\Phi,\Psi\right) \propto \left[p_{adb}\left(\Phi,\Psi\right) \right]^{\frac{T_s}{T}} \tag{1.193}$$

Due to the surface smoothing steps, propagation of uncertainties is not practical to estimate confidence intervals on the allosteric landscape. Instead, we use the

bootstrapping approach. Specifically, because observations from MD time series are notoriously not independent, we use block bootstrapping[289], i.e. we generate artificial samples by drawing at random (with replacement) segments of trajectory of 1 ns in length. We then compute local standard deviations on the allosteric landscape calculated from each of these samples.

4.2.1. Results

The free energy landscape was recovered using the histogram method is shown in **Figure 48B.**

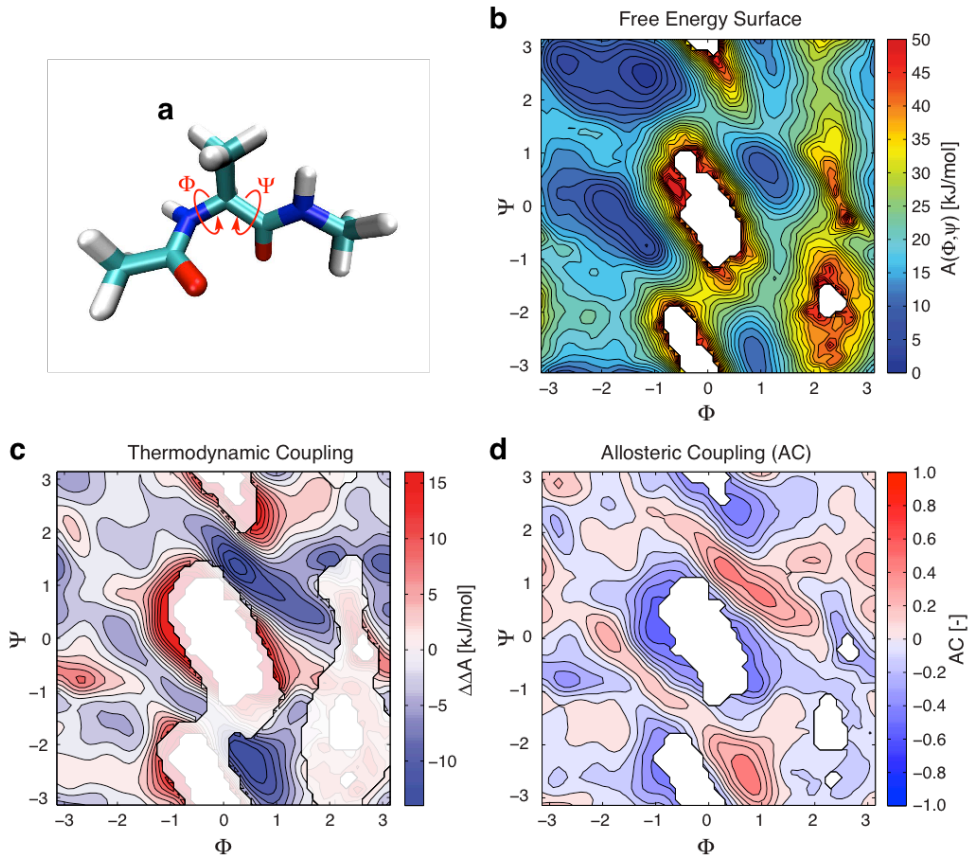


Figure 48. The allosteric couplings in alanine dipeptide.

(A) Alanine dipeptide, colored by atom type (carbon in cyan, hydrogen in white, nitrogen in blue, oxygen in red). (B) The ϕ,ψ free energy surface. (C) The allosteric landscape calculated from the thermodynamic coupling between ϕ,ψ for different perturbations of the equilibrium distribution. (D) The normalized allosteric landscape using AC.

We find significant allosteric couplings in the regions of the left-handed α-helix and the C_{7ax} (see **Figure 48**C), indicating that if ϕ is driven to the 0° to 120° region, the transition of ψ to the 0° to 120° and -60° to -180° region becomes thermodynamically more favorable. While these regions are low probability, the allosteric coupling may account for the small populations of the left-handed α-helix and C_{7ax} conformations that appear at equilibrium. We also see significant allosteric coupling in the disallowed regions, indicating that an energetically unfavorable interaction is involved in the high free energy.

It should be noted that the allosteric landscape has a natural normalization. If the two CVs are maximally coupled, constraining one CV will fully constrain the other. Thus, at maximum coupling,

$$A\left[p\left(\vec{r}\middle|X(\vec{r})=x\right)\right]=A\left[p\left(\vec{r}\middle|Y(\vec{r})=y\right)\right]=A\left[p\left(\vec{r}\middle|X(\vec{r})=x,Y(\vec{r})=y\right)\right] \tag{1.194}$$

and thus

$$\Delta\Delta A_{max}(x,y)=A\left[p(\vec{r})\right]-A\left[p\left(\vec{r}\middle|X(\vec{r})=x,Y(\vec{r})=y\right)\right] \tag{1.195}$$

We can then normalize (1.190) to this upper bound to find the normalized allosteric coupling, AC:

$$\frac{-\Delta\Delta A(x,y)}{A\left[p\left(\vec{r}\middle|X(\vec{r})=x,Y(\vec{r})=y\right)\right]-A\left[p(\vec{r})\right]}=\frac{\log\left(p(x)p(x)\right)}{\log\left(p(x,y)\right)}-1=AC(x,y) \tag{1.196}$$

The AC ranges from 1 to -1 and matches the convention of commonly used positive and negative allostery; positive values indicate that constraining one variable reduces the free energy required to constrain the other, whereas negative values indicates that constraining one variables increases the free energy required to constrain the other. In

essence, the AC describes what fraction of potential positive or negative allostery is contributing to the free energy of the joint state. In the AC landscape of alanine dipeptide (see **Figure 48D**), both the left-handed α-helix and C_{7ax} regions have ACs of around 0.5, indicating a substantial amount of their stability is due to allostery. Similarly, the regions sampled around the highly unfavorable mid-ϕ region indicate ACs around -0.5, indicating that a substantial amount of the instability is due to allostery.

4.3. Conclusions

We have derived the generalized form of the allosteric coupling between continuous and discrete collective variables. We find that it is related to the pointwise mutual information, and is best represented in the form of an allosteric landscape to demonstrate the allosteric response to all possible perturbations of the CVs. Our calculation of the allosteric landscape of alanine dipeptide reveals positive allosteric coupling between the ϕ and ψ angles, which appear to stabilize the left-handed α-helix and C_{7ax} conformations, and negative allosteric coupling due to steric clashes, which defines the unpopulated regions of CV space. This method is applicable to larger systems, and should be a strong tool in understanding allosteric molecular mechanisms and identifying novel allosteric sites for the modulation of functionally important CVs and reaction coordinates.

5. Concluding Remarks

Allostery is a ubiquitous biophysical phenomenon that plays a crucial role in many cellular processes. Despite this ubiquity, the study of allostery has primarily been phenomenological in nature and limited to the quantification of specific observations of allosteric behavior rather than the construction of the detailed molecular mechanisms that give rise to those observations. In this dissertation, several theoretical models and computational methods have been presented that were constructed with the goal of creating the framework required for the study of allostery to move from an observational science to a mechanistic science.

An essential component of the dissertation is the rigorous definition of the relationship between allostery and information theory. Information theory has been invoked to describe allostery, and cellular signaling in general, mostly due to the intuition that human communication systems and cellular signaling systems are likely to share essential features. However, it is important to recognize that prior to the use of information theory as a framework for allostery, simple covariance and correlation was the dominant language of allostery. The choice of the framework used to describe allostery has historically been pragmatic and empirical, not theory-driven. By seeking to bridge the relationship between mutual information and allosteric efficacy, it has become clear that while allostery can be described in the language of information theory, the intuitive information theoretical measure, mutual information, is actually misleading. By deriving the allosteric efficacy for coupled perturbations away from equilibrium, we find that the mutual information is actually the average allosteric efficacy over all perturbations. This ensemble average can be misleading, as functional perturbations, such as the binding of a ligand, generally push the system away from the unbound equilibrium, towards regions of conformational space that are low

probability and thus have little contribution to the mutual information prior to perturbation. This finding indicates that the NbIT method, which was built based on the intuition that a fundamental relationship existed between allostery and mutual information, must be recast into the language of coupled perturbations to the 3-dimensional free energy landscape.

In addition, analytical approaches similar to the Allosteric Ising Models must be constructed based on the analytical form of the allosteric efficacy for discrete and continuous distributions described here. While the two-state models described here are useful in providing a qualitative conceptual basis for more complex systems, it should be possible to relax the current assumptions and approximations in order to make the models more directly applicable to real, allosteric protein systems.

The illustrations of the methodology and models to two essential membrane protein systems, transporters and receptors, revealed that allostery plays an essential role in much of the complex, previously unexplained ligand-specific behavior that has been documented over the last decade. By identifying specific residues that play crucial roles in ligand-specific allostery, it becomes clear overly generalized descriptions of the functional architecture of proteins, such as the "binding site", the "allosteric site", or the "channel", must be replaced with a more detailed description of the structural components that compose these proteins and a physical model describing their thermodynamic couplings. A ligand or substrate does not simply bind to a binding site; it engages multiple partners whom may have thermodynamic couplings to other structural components distant within the protein. In the case of LeuT, differential engagement with a single residue can have substantial effects on the transporter's rate of transport for a substrate. Additionally, a differential engagement within a small fraction of the ligand binding site in 5-HT_{2A}R can initiate a downstream signaling

cascade resulting in the remarkable hallucinogenic phenotype. Functional differences of these magnitudes, initiated by structural differences that are often regarded as minor, demonstrate the importance of building a more complete understanding of the implicit allosteric properties of these proteins.

Finally, most of the analysis present in this dissertation was performed using ensembles generated from MD. However, the analysis is not in anyway specific to a method of ensemble estimation. As experimental techniques such as smFRET, EPR, and cryoEM expand our ability to estimate multi-dimensional ensemble, it will be of great importance to be able to identify allosteric couplings directly from the experimental data. By combining experimentally derived ensembles with physics-based models of allostery, it will be possible to truly the mechanism underlying protein function at the molecular level.

6. References

(1) LeVine, M. V., Cuendet, M. A., Khelashvili, G., and Weinstein, H. (2016) Allosteric Mechanisms of Molecular Machines at the Membrane: Transport by Sodium-Coupled Symporters. *Chem. Rev.* acs.chemrev.5b00627.

(2) LeVine, M., and Weinstein, H. (2015) AIM for Allostery: Using the Ising Model to Understand Information Processing and Transmission in Allosteric Biomolecular Systems. *Entropy 17*, 2895–2918.

(3) Clarke, A. C. (1972) Hazards of Prophecy. *Futur.*

(4) Bechtel, W., and Richardson, R. C. (1998) Vitalism. *Routledge Encycl. Philos.*

(5) Wöhler, F. (1828) Ueber künstliche Bildung des Harnstoffs. *Ann. Phys. 88*, 253–256.

(6) Bedau, M. a, Medicine, M., and Cleland, C. E. (2010) The Nature of Life : Classical and Contemporary Perspectives from Philosophy and Science. *Nat. Life Class. Comtemporary Perspect. from Philos. Sci.* 437.

(7) Allen, D. W., Guthe, K. F., and Wyman, J. J. (1950) Further studies on the oxygen equilibrium of hemoglobin. *J. Biol. Chem. 187*, 393–410.

(8) Shi, Y. (2014) A Glimpse of Structural Biology through X-Ray Crystallography. *Cell 4*, 995–1014.

(9) Markwick, P. R. L., Malliavin, T., and Nilges, M. (2008) Structural biology by NMR: Structure, dynamics, and interactions. *PLoS Comput. Biol. 4*.

(10) Bai, X., McMullan, G., and Scheres, S. H. . (2015) How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci. 40*, 49–57.

(11) Orengo, C. A., Todd, A. E., and Thornton, J. M. (1999) From protein structure to function. *Curr.Opin.Struct.Biol. 9*, 374.

(12) Redfern, O. C., Dessailly, B., and Orengo, C. A. (2008) Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol. 18*, 394–402.

(13) Hegyi, H., and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol 288*, 147–164.

(14) Bohr, C., Hasselbalch, K., and Krogh, A. (1904) Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen

Sauerstoffbindung übt. *Skand. Arch. Physiol. 16*, 402–412.

(15) Monod, J., Changeux, J.-P., and Jacob, F. (1963) Allosteric proteins and cellular control systems. *J. Mol. Biol. 6*, 306–329.

(16) Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., Wang, Q., Shi, T., Zhao, Y., Wang, Y., Li, W., Li, Y., Chen, H., Chen, G., and Zhang, J. (2011) ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res. 39*, D663–D669.

(17) Gunasekaran, K., Ma, B., and Nussinov, R. (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins 57*, 433–43.

(18) Monod, J., Wyman, J., and Changeux, J. P. (1965) On the Nature of Allosteric Transitions: a Plausible Model. *J. Mol. Biol. 12*, 88–118.

(19) Tsai, C., and Nussinov, R. (2014) A Unified View of "How Allostery Works." *PLoS Comput. Biol. 10*, e1003394.

(20) Leff, P. (1995) The two-state model of receptor activation. *Trends Pharmacol. Sci. 16*, 89–97.

(21) Fenton, A. W. (2008) Allostery: an illustrated definition for the "second secret of life." *Trends Biochem. Sci. 33*, 420–425.

(22) Michaelis, L., and Menten, M. L. (1913) Die Kinetik der Invertinwirkung. *Biochem. z 49*, 333–369.

(23) Glennon, T. M., Villa, J., and Warshel, a. (2000) How does GAP catalyze the GTPase reaction of ras?: A computer simulation study. *Biochemistry 39*, 9641–9651.

(24) Lienhard, G. E. (1973) Enzymatic catalysis and transition-state theory. *Science 180*, 149–154.

(25) Eyring, H. (1935) The Activated Complex in Chemical Reactions. *J. Chem. Phys. 3*, 107–115.

(26) Schöneberg, T., Schulz, A., Biebermann, H., Hermsdorf, T., Römpler, H., and Sangkuhl, K. (2004) Mutant G-protein-coupled receptors as a cause of human diseases. *Pharmacol. Ther. 104*, 173–206.

(27) Betke, K. M., Wells, C. a, and Hamm, H. E. (2012) GPCR mediated regulation of synaptic transmission. *Prog. Neurobiol. 96*, 304–21.

(28) Nichols, D. E. (2004) Hallucinogens. *Pharmacol. Ther. 101*, 131–81.

(29) Przewłocki, R., and Przewłocka, B. (2001) Opioids in chronic pain. *Eur. J.*

Pharmacol. 429, 79–91.

(30) Seeman, P., Schwarz, J., and Chen, J. (2006) Psychosis pathways converge via D2high dopamine receptors. *Synapse 346*, 319–346.

(31) Boghog. Inverse agonist 3.

(32) Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. a, Motoshima, H., Fox, B. a, Trong, I. Le, Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M., and Miyano, M. (2000) Crystal Structure of Rhodopsin: A G Protein - Coupled Receptor. *Science 289*, 739–745.

(33) Choe, H., Kim, Y., Park, J., and Morizumi, T. (2011) Crystal structure of metarhodopsin II. *Nature 471*, 651–5.

(34) Zhou, X. E., Melcher, K., and Xu, H. E. (2012) Structure and activation of rhodopsin. *Acta Pharmacol. Sin. 33*, 291–299.

(35) Nakamichi, H., and Okada, T. (2006) Crystallographic analysis of primary visual photochemistry. *Angew. Chemie - Int. Ed. 45*, 4270–4273.

(36) Ruprecht, J. J., Mielke, T., Vogel, R., Villa, C., and Schertler, G. F. X. (2004) Electron crystallography reveals the structure of metarhodopsin I. *EMBO J. 23*, 3609–20.

(37) Salom, D., Lodowski, D. T., Stenkamp, R. E., Trong, I. Le, Golczak, M., Jastrzebska, B., Harris, T., Ballesteros, J. A., Palczewski, K., Le Trong, I., Golczak, M., Jastrzebska, B., Harris, T., Ballesteros, J. A., and Palczewski, K. (2006) Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *PNAS 103*, 16123–8.

(38) Hanson, M. A., Cherezov, V., Griffith, M. T., Roth, C. B., Jaakola, V.-P., Chien, E. Y. T., Velasquez, J., Kuhn, P., and Stevens, R. C. (2008) A Specific Cholesterol Binding Site Is Established by the 2.8 Å Structure of the Human β2-Adrenergic Receptor. *Structure 16*, 897–905.

(39) Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G. W., Tate, C. G., and Schertler, G. F. X. (2008) Structure of a β1-adrenergic G-protein-coupled receptor. *Nature 454*, 486–491.

(40) Jaakola, V., Griffith, M., and Hanson, M. (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science 322*, 1211–7.

(41) Rasmussen, S. G. F., Choi, H.-J., Fung, J. J., Pardon, E., Casarosa, P., Chae, P. S.,

Devree, B. T., Rosenbaum, D. M., Thian, F. S., Kobilka, T. S., Schnapp, A., Konetzki, I., Sunahara, R. K., Gellman, S. H., Pautsch, A., Steyaert, J., Weis, W. I., and Kobilka, B. K. (2011) Structure of a nanobody-stabilized active state of the β(2) adrenoceptor. *Nature 469*, 175–80.

(42) Farrens, D. L., Altenbach, C., Yang, K., Hubbell, W. L., and Khorana, H. G. (1996) Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science 274*, 768–770.

(43) Dunham, T. D. (1999) Conformational Changes in Rhodopsin. *J. Biol. Chem. 274*, 1683–1690.

(44) Rasmussen, S. G. F., DeVree, B. T., Zou, Y., Kruse, A. C., Chung, K. Y., Kobilka, T. S., Thian, F. S., Chae, P. S., Pardon, E., Calinski, D., Mathiesen, J. M., Shah, S. T. a, Lyons, J. a, Caffrey, M., Gellman, S. H., Steyaert, J., Skiniotis, G., Weis, W. I., Sunahara, R. K., and Kobilka, B. K. (2011) Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature 477*, 549–55.

(45) Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002) Molecular Biology of the Cell, 4th edition.

(46) Wright, E. M., Hirayama, B. A., and Loo, D. F. (2007) Active sugar transport in health and disease. *J. Intern. Med. 261*, 32–43.

(47) Wood, I. S., and Trayhurn, P. (2003) Glucose transporters (GLUT and SGLT): expanded families of sugar transport proteins. *Br. J. Nutr. 89*, 3.

(48) Christensen, H. N. (1990) Role of amino acid transport and countertransport in nutrition and metabolism. *Physiol. Rev. 70*, 43–77.

(49) Greger, R. (2000) Physiology of renal sodium transport. *Am. J. Med. Sci. 319*, 51–62.

(50) Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth, A. J. (2014) Principles of Neural Science, Fifth Edition. *Neurology*.

(51) Broer, S., and Wagner, C. (2003) Membrane Transporter Diseases.

(52) Ravna, A. W., Sager, G., Dahl, S. G., and Sylte, I. (2008) Membrane Transporters: Structure, Function and Targets for Drug Design, in *Top Med Chem*, pp 15–51.

(53) Pedersen, P. L. (2007) Transport ATPases into the year 2008: a brief overview related to types, structures, functions and roles in health and disease. *J. Bioenerg. Biomembr. 39*, 349–355.

(54) Tanford, C. (1983) Mechanism of free energy coupling in active transport. *Annu. Rev. Biochem. 52*, 379–409.

(55) Rees, D. C., Johnson, E., and Lewinson, O. (2009) ABC transporters: the power to change. *Nat. Rev. Mol. Cell Biol. 10*, 218–227.

(56) Morth, J. P., Pedersen, B. P., Buch-Pedersen, M. J., Andersen, J. P., Vilsen, B., Palmgren, M. G., and Nissen, P. (2011) A structural overview of the plasma membrane Na+,K+-ATPase and H+-ATPase ion pumps. *Nat. Rev. Mol. Cell Biol. 12*, 60–70.

(57) Gadsby, D. C. (2007) Structural biology: ion pumps made crystal clear. *Nature 450*, 957–9.

(58) Jardetzky, O. (1966) Simple allosteric model for membrane pumps. *Nature 211*, 969–970.

(59) Huang, Y., Lemieux, M. J., Song, J., Auer, M., and Wang, D.-N. (2003) Structure and Mechanism of the Glycerol-3-Phosphate Transporter from Escherichia coli. *Science 301*, 616–620.

(60) Wilbrandt, W. (1977) The Asymmetry of Sugar Transport in the Red Cell Membrane, pp 204–211.

(61) Lingrel, J. B., and Kuntzweiler, T. (1994) Na+,K+-ATPase. *J. Biol. Chem 269*, 19659–19662.

(62) Dean, M., Hamon, Y., and Chimini, G. (2001) The Human ATP-Binding Cassette transporter superfamily. *J. Lipid Res. 42*, 1007–1017.

(63) Higgins, C. F., and Linton, K. J. (2004) The ATP switch model for ABC transporters. *Nat. Struct. Mol. Biol. 11*, 918–926.

(64) Hollenstein, K., Dawson, R. J., and Locher, K. P. (2007) Structure and mechanism of ABC transporter proteins. *Curr. Opin. Struct. Biol. 17*, 412–418.

(65) Yamashita, A., Singh, S. K., Kawate, T., Jin, Y., and Gouaux, E. (2005) Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters. *Nature 437*, 215–23.

(66) Khafizov, K., Staritzbichler, R., Stamm, M., and Forrest, L. R. (2010) A study of the evolution of inverted-topology repeats from LeuT-fold transporters using alignMe. *Biochemistry 49*, 10702–10713.

(67) Quick, M., and Javitch, J. A. (2007) Monitoring the function of membrane

transport proteins in detergent-solubilized form. *PNAS 104*, 3603–8.

(68) Shi, L., Quick, M., Zhao, Y., Weinstein, H., and Javitch, J. A. (2008) The mechanism of a neurotransmitter:sodium symporter--inward release of Na+ and substrate is triggered by substrate in a second binding site. *Mol. Cell 30*, 667–77.

(69) Claxton, D. P., Quick, M., Shi, L., de Carvalho, F. D., Weinstein, H., Javitch, J. A., and McHaourab, H. S. (2010) Ion/substrate-dependent conformational dynamics of a bacterial homolog of neurotransmitter:sodium symporters. *Nat. Struct. Mol. Biol. 17*, 822–9.

(70) Kazmier, K., Sharma, S., Quick, M., Islam, S. M., Roux, B., Weinstein, H., Javitch, J. A., and McHaourab, H. S. (2014) Conformational dynamics of ligand-dependent alternating access in LeuT. *Nat. Struct. Mol. Biol. 21*, 472–9.

(71) Beuming, T., Shi, L., Javitch, J. A., and Weinstein, H. (2006) A comprehensive structure-based alignment of prokaryotic and eukaryotic neurotransmitter/Na+ symporters (NSS) aids in the use of the LeuT structure to probe NSS. *Mol. Pharmacol. 70*, 1630–1642.

(72) Krishnamurthy, H., and Gouaux, E. (2012) X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature 481*, 469–74.

(73) Singh, S. K., Yamashita, A., and Gouaux, E. (2007) Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. *Nature 448*, 952–6.

(74) Zhou, Z., Zhen, J., Karpowich, N., and Goetz, R. (2007) LeuT-desipramine structure reveals how antidepressants block neurotransmitter reuptake. *Science 317*, 1390–3.

(75) Quick, M., Winther, A.-M. L., Shi, L., Nissen, P., Weinstein, H., and Javitch, J. A. (2009) Binding of an octylglucoside detergent molecule in the second substrate (S2) site of LeuT establishes an inhibitor-bound conformation. *PNAS 106*, 5563–8.

(76) Quick, M., Shi, L., Zehnpfennig, B., Weinstein, H., and Javitch, J. A. (2012) Experimental conditions can obscure the second high-affinity site in LeuT. *Nat. Struct. Mol. Biol. 19*, 207–11.

(77) Piscitelli, C. L., Krishnamurthy, H., and Gouaux, E. (2010) Neurotransmitter/sodium symporter orthologue LeuT has a single high-affinity substrate site. *Nature 468*, 1129–32.

(78) Wang, H., and Gouaux, E. (2012) Substrate binds in the S1 site of the F253A mutant of LeuT, a neurotransmitter sodium symporter homologue. *EMBO Rep. 13*,

861–6.

(79) Khelashvili, G., LeVine, M. V, Shi, L., Quick, M., Javitch, J. A., and Weinstein, H. (2013) The membrane protein LeuT in micellar systems: aggregation dynamics and detergent binding to the S2 site. *J. Am. Chem. Soc. 135*, 14266–75.

(80) Reyes, N., and Tavoulari, S. (2011) To be, or not to be two sites: that is the question about LeuT substrate binding. *J. Gen. Physiol. 138*, 467–471.

(81) Cheng, M. H., and Bahar, I. (2013) Coupled Global and Local Changes Direct Substrate Translocation by Neurotransmitter-Sodium Symporter Ortholog LeuT. *Biophys. J. 105*, 630–639.

(82) Shaikh, S. A., and Tajkhorshid, E. (2010) Modeling and dynamics of the inward-facing state of a Na+/Cl- dependent neurotransmitter transporter homologue. *PLoS Comput. Biol. 6*.

(83) Cheng, M. H., and Bahar, I. (2014) Complete Mapping of Substrate Translocation Highlights the Role of LeuT N-terminal Segment in Regulating Transport Cycle. *PLoS Comput. Biol. 10*, e1003879.

(84) Zhao, Y., Terry, D., Shi, L., Weinstein, H., Blanchard, S. C., and Javitch, J. A. (2010) Single-molecule dynamics of gating in a neurotransmitter transporter homologue. *Nature 465*, 188–93.

(85) Zhao, Y., Terry, D. S., Shi, L., Quick, M., Weinstein, H., Blanchard, S. C., and Javitch, J. A. (2011) Substrate-modulated gating dynamics in a Na+-coupled neurotransmitter transporter homologue. *Nature 474*, 109–13.

(86) Kniazeff, J., Shi, L., Loland, C. J., Javitch, J. A., Weinstein, H., and Gether, U. (2008) An intracellular interaction network regulates conformational transitions in the dopamine transporter. *J. Biol. Chem. 283*, 17691–17701.

(87) Roux, B., and Islam, S. M. (2013) Restrained-ensemble molecular dynamics simulations based on distance histograms from double electron-electron resonance spectroscopy. *J. Phys. Chem. B 117*, 4733–4739.

(88) Wheeler, D. D., Edwards, A. M., Chapman, B. M., and Ondo, J. G. (1993) A model of the sodium dependence of dopamine uptake in rat striatal synaptosomes. *Neurochem. Res. 18*, 927–936.

(89) Cool, D. R., Leibach, F. H., and Ganapathy, V. (1990) Modulation of serotonin uptake kinetics by ions and ion gradients in human placental brush-border membrane vesicles. *Biochemistry 29*, 1818–1822.

(90) Penmatsa, A., Wang, K. H., and Gouaux, E. (2013) X-ray structure of dopamine transporter elucidates antidepressant mechanism. *Nature 503*, 85–90.

(91) Penmatsa, A., Wang, K. H., and Gouaux, E. (2015) X-ray structures of Drosophila dopamine transporter in complex with nisoxetine and reboxetine. *Nat. Struct. Mol. Biol. 22*, 506–509.

(92) Wang, K. H., Penmatsa, A., and Gouaux, E. (2015) Neurotransmitter and psychostimulant recognition by the dopamine transporter. *Nature 521*, 322–327.

(93) Shan, J., Javitch, J. A., Shi, L., and Weinstein, H. (2011) The substrate-driven transition to an inward-facing conformation in the functional mechanism of the dopamine transporter. *PLoS One 6*, e16350.

(94) Khelashvili, G., Stanley, N., Sahai, M. A., Medina, J., LeVine, M. V, Shi, L., De Fabritiis, G., and Weinstein, H. (2015) Spontaneous Inward Opening of the Dopamine Transporter Is Triggered by PIP 2 -Regulated Dynamics of the N-Terminus. *ACS Chem. Neurosci. 6*, 1825–1837.

(95) Ferrer, J. V, and Javitch, J. A. (1998) Cocaine alters the accessibility of endogenous cysteines in putative extracellular and intracellular loops of the human dopamine transporter. *PNAS 95*, 9238–9243.

(96) Loland, C. J., Grånäs, C., Javitch, J. A., and Gether, U. (2004) Identification of Intracellular Residues in the Dopamine Transporter Critical for Regulation of Transporter Conformation and Cocaine Binding. *J. Biol. Chem. 279*, 3228–3238.

(97) Reith, M. E. A., Berfield, J. L., Wang, L. C., Ferrer, J. V., and Javitch, J. A. (2001) The Uptake Inhibitors Cocaine and Benztropine Differentially Alter the Conformation of the Human Dopamine Transporter. *J. Biol. Chem. 276*, 29012–29018.

(98) Chen, N., Ferrer, J. V., Javitch, J. A., and Justice, J. B. (2000) Transport-dependent accessibility of a cytoplasmic loop cysteine in the human dopamine transporter. *J. Biol. Chem. 275*, 1608–1614.

(99) Androutsellis-Theotokis, a., Ghassemi, F., and Rudnick, G. (2001) A Conformationally Sensitive Residue on the Cytoplasmic Surface of Serotonin Transporter. *J. Biol. Chem. 276*, 45933–45938.

(100) Androutsellis-Theotokis, A., and Rudnick, G. (2002) Accessibility and conformational coupling in serotonin transporter predicted internal domains. *J. Neurosci. 22*, 8370–8378.

(101) Norregaard, L., Frederiksen, D., Nielsen, E. Ø., and Gether, U. (1998) Delineation of an endogenous zinc-binding site in the human dopamine transporter. *EMBO J. 17*, 4266–4273.

(102) Loland, C. J., Norregaard, L., Litman, T., and Gether, U. (2002) Generation of an activating Zn(2+) switch in the dopamine transporter: mutation of an intracellular tyrosine constitutively alters the conformational equilibrium of the transport cycle. *PNAS 99*, 1683–1688.

(103) Gudermann, T., Kalkbrenner, F., and Schultz, G. (1996) Diversity and Selectivity of Receptor-G Protein Interaction. *Annu. Rev. Pharmacol. Toxicol. 36*, 429–459.

(104) DeWire, S. M., Ahn, S., Lefkowitz, R. J., and Shenoy, S. K. (2007) β-Arrestins and Cell Signaling. *Annu. Rev. Physiol. 69*, 483–510.

(105) Shenoy, S. K., Drake, M. T., Nelson, C. D., Houtz, D. a, Xiao, K., Madabushi, S., Reiter, E., Premont, R. T., Lichtarge, O., and Lefkowitz, R. J. (2006) beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. *J. Biol. Chem. 281*, 1261–73.

(106) Urban, J. D., Clarke, W. P., Zastrow, M. Von, Nichols, D. E., Kobilka, B., Weinstein, H., Javitch, J. A., Roth, B. L., Christopoulos, A., Sexton, P. M., Miller, K. J., Spedding, M., and Mailman, R. B. (2007) Functional Selectivity and Classical Concepts of Quantitative Pharmacology *320*, 1–13.

(107) Kenakin, T. (2011) Functional selectivity and biased receptor signaling. *J. Pharmacol. Exp. Ther. 336*, 296–302.

(108) Rajagopal, S., and Ahn, S. (2011) Quantifying ligand bias at seven-transmembrane receptors. *Mol. Pharmacol. 80*, 367–377.

(109) Gregory, K. J., Sexton, P. M., Tobin, A. B., and Christopoulos, A. (2012) Stimulus bias provides evidence for conformational constraints in the structure of a G protein-coupled receptor. *J. Biol. Chem.*

(110) Singh, S. K., Piscitelli, C. L., Yamashita, A., and Gouaux, E. (2008) A competitive inhibitor traps LeuT in an open-to-out conformation. *Science 322*, 1655–61.

(111) LeVine, M. V., and Weinstein, H. (2014) NbIT - A New Information Theory-Based Analysis of Allosteric Mechanisms Reveals Residues that Underlie Function in the Leucine Transporter LeuT. *PLoS Comput. Biol. 10*, e1003603.

(112) Hilser, V. J., Wrabl, J. O., and Motlagh, H. N. (2012) Structural and energetic basis of allostery. *Annu. Rev. Biophys. 41*, 585–609.

(113) Ising, E. (1925) Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Phys. A Hadron. Nucl.*

(114) Hopfield, J. (1982) Neural networks and physical systems with emergent collective computational abilities. *PNAS 79*, 2554–2558.

(115) Machta, B. B., Papanikolaou, S., Sethna, J. P., and Veatch, S. L. (2011) Minimal model of plasma membrane heterogeneity requires coupling cortical actin to criticality. *Biophys. J. 100*, 1668–77.

(116) Muñoz, V., Thompson, P. a, Hofrichter, J., and Eaton, W. a. (1997) Folding dynamics and mechanism of beta-hairpin formation. *Nature 390*, 196–199.

(117) Vorov, O. K., Livesay, D. R., and Jacobs, D. J. (2009) Helix/coil nucleation: A local response to global demands. *Biophys. J. 97*, 3000–3009.

(118) Vorov, O. K., Livesay, D. R., and Jacobs, D. J. (2011) Nonadditivity in conformational entropy Upon molecular rigidification reveals a universal mechanism affecting folding cooperativity. *Biophys. J. 100*, 1129–1138.

(119) Bray, D., and Duke, T. (2004) Conformational spread: the propagation of allosteric states in large multiprotein complexes. *Annu. Rev. Biophys. Biomol. Struct. 33*, 53–73.

(120) Graham, I., and Duke, T. (2005) Dynamic hysteresis in a one-dimensional Ising model: Application to allosteric proteins. *Phys. Rev. E 71*, 061923.

(121) Perez, D. M., and Karnik, S. S. (2005) Multiple signaling states of G-protein-coupled receptors. *Pharmacol. Rev. 57*, 147–161.

(122) Kahsai, A. W., Xiao, K., Rajagopal, S., Ahn, S., Shukla, A. K., Sun, J., Oas, T. G., and Lefkowitz, R. J. (2011) Multiple ligand-specific conformations of the β2-adrenergic receptor. *Nat. Chem. Biol. 7*, 692–700.

(123) Sethi, A., Eargle, J., Black, A. a, and Luthey-Schulten, Z. (2009) Dynamical networks in tRNA:protein complexes. *PNAS 106*, 6620–5.

(124) Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. (2012) Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J. 26*, 868–81.

(125) Bowman, G. R., and Geissler, P. L. (2012) Equilibrium fluctuations of a single

folded protein reveal a multitude of potential cryptic allosteric sites. *PNAS 109*, 11681–11686.

(126) Ming, D., and Wall, M. E. (2005) Allostery in a coarse-grained model of protein dynamics. *Phys. Rev. Lett. 95*, 1–4.

(127) Su, J. G., Qi, L. S., Li, C. H., Zhu, Y. Y., Du, H. J., Hou, Y. X., Hao, R., and Wang, J. H. (2014) Prediction of allosteric sites on protein surfaces with an elastic-network-model-based thermodynamic method. *Phys. Rev. E 90*, 1–10.

(128) Gasper, P., and Fuglestad, B. (2012) Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. *PNAS 109*, 21216–22.

(129) Ian H. Witten, Eibe Frank, and Hall, M. A. (2005) Data Mining Practical Machine Learning Tools and Techniques.

(130) del Sol, A., Tsai, C.-J., Ma, B., and Nussinov, R. (2009) The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure 17*, 1042–50.

(131) LeVine, M. V, Perez-Aguilar, J., and Weinstein, H. (2014) N-body Information Theory (NbIT) Analysis of Rigid-Body Dynamics in Intracellular Loop 2 of the 5-HT2A Receptor, in *Proceedings IWWBIO 2014*, pp 1190–1201.

(132) Gleitsman, K. R., Shanata, J. a P., Frazier, S. J., Lester, H. a, and Dougherty, D. a. (2009) Long-range coupling in an allosteric receptor revealed by mutant cycle analysis. *Biophys. J. 96*, 3168–78.

(133) Wriggers, W., Stafford, K. A., Shan, Y., Piana, S., Maragakis, P., Lindorff-Larsen, K., Miller, P. J., Gullingsrud, J., Rendleman, C. A., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2009) Automated event detection and activity monitoring in long molecular dynamics simulations. *J. Chem. Theory Comput. 5*, 2595–2605.

(134) Stolzenberg, S., Quick, M., Zhao, C., Gotfryd, K., Khelashvili, G., Gether, U., Loland, C. J., Javitch, J. A., Noskov, S., Weinstein, H., and Shi, L. (2015) Mechanism of the Association between Na + Binding and Conformations at the Intracellular Gate in Neurotransmitter:Sodium Symporters. *J. Biol. Chem.* jbc.M114.625343.

(135) del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol. 2*, 2006.0019.

(136) Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol. 344*, 1135–46.

(137) De Ruvo, M., Giuliani, A., Paci, P., Santoni, D., and Di Paola, L. (2012) Shedding light on protein-ligand binding by graph theory: The topological nature of allostery. *Biophys. Chem. 165-166*, 21–9.

(138) Böde, C., Kovács, I. A., Szalay, M. S., Palotai, R., Korcsmáros, T., and Csermely, P. (2007) Network analysis of protein dynamics. *FEBS Lett. 581*, 2776–82.

(139) Doncheva, N. T., Klein, K., Domingues, F. S., and Albrecht, M. (2011) Analyzing and visualizing residue networks of protein structures. *Trends Biochem. Sci. 36*, 179–182.

(140) Bahar, I., Atilgan, A. R., and Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des. 2*, 173–181.

(141) Atilgan, a R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J. 80*, 505–15.

(142) Kolan, D., Fonar, G., and Samson, A. O. (2013) Elastic network normal mode dynamics reveal the GPCR activation mechanism. *Proteins* 1–8.

(143) Bahar, I., Lezon, T. R., Bakan, A., and Shrivastava, I. H. (2010) Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. *Chem. Rev. 110*, 1463–1497.

(144) Stolzenberg, S., Khelashvili, G., and Weinstein, H. (2012) Structural Intermediates in a Model of the Substrate Translocation Path in the Bacterial Glutamate Transporter Homologue GltPh. *Biophys. J. 102*, 519a.

(145) Lezon, T. R., and Bahar, I. (2012) Constraints imposed by the membrane selectively guide the alternating access dynamics of the glutamate transporter GltPh. *Biophys J 102*, 1331–1340.

(146) Swope, W. C., Andersen, H. C., Berens, P. H., and Wilson, K. R. (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys. 76*, 637–649.

(147) VanWart, A. T., Eargle, J., Luthey-Schulten, Z., and Amaro, R. E. (2012) Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput. 8*, 2949–2961.

(148) Eargle, J., and Luthey-Schulten, Z. (2012) NetworkView: 3D display and

analysis of protein·RNA interaction networks. *Bioinformatics 28*, 3000–1.

(149) Van Wart, A. T., Durrant, J., Votapka, L., and Amaro, R. E. (2014) Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput. 10*, 511–517.

(150) Newman, M. E. J., and Girvan, M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. 69*, 1–15.

(151) Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell Syst. Tech. J. 27*, 623–656.

(152) Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.*

(153) McClendon, C. L., Friedland, G., Mobley, D. L., Amirkhani, H., and Jacobson, M. P. (2009) Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput. 5*, 2486–2502.

(154) Dubay, K. H., Bothma, J. P., and Geissler, P. L. (2011) Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone. *PLoS Comput. Biol. 7*, e1002168.

(155) Cover, T. M., and Thomas, J. A. (1991) Differential Entropy, in *Elements of Information Theory*, pp 224–238. John Wiley & Sons.

(156) Lange, O. F., and Grubmüller, H. (2006) Generalized correlation for biomolecular dynamics. *Proteins 62*, 1053–61.

(157) Rivalta, I., Sultan, M. M., Lee, N., Manley, G. A., Loria, J. P., and Batista, V. S. (2012) Allosteric pathways in imidazole glycerol phosphate synthase *109*.

(158) Matsuda, H. (2000) Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E 62*, 3096–3102.

(159) Bell, A. J. (2003) The co-information lattice. *4th Int. Sympoosium Indep. Compon. Anal. Blind Signal Sep.* 921g926.

(160) McGill, W. J. (1954) Multivariate Information Transmission. *Inf. Theory, Trans. IRE Prof. Gr. 4*, 93–111.

(161) Sakaguchi, M. (1967) Interaction information in multivariate probability distributions. *Kodai Math. Semin. Reports 19*, 147–155.

(162) Jaynes, E. (1963) Information Theory and Statistical Mechanics (Notes by the lecturer). *Stat. Phys. 3*.

(163) Reva, B., Antipin, Y., and Sander, C. (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol. 8*, R232.

(164) McClendon, C. L., Hua, L., Barreiro, A., and Jacobson, M. P. (2012) Comparing Conformational Ensembles Using the Kullback-Leibler Divergence Expansion. *J. Chem. Theory Comput. 8*, 2115–2126.

(165) Wolfe, K. C., and Chirikjian, G. S. (2012) Quantitative comparison of conformational ensembles. *Entropy 14*, 213–232.

(166) Yang, S., Salmon, L., and Al-Hashimi, H. (2014) Measuring similarity between dynamic ensembles of biomolecules. *Nat. Methods 11*, 552–4.

(167) Lindorff-Larsen, K., and Ferkinghoff-Borg, J. (2009) Similarity measures for protein ensembles. *PLoS One 4*.

(168) Leighty, R., and Varma, S. (2013) Quantifying Changes in Intrinsic Molecular Motion Using Support Vector Machines. *J. Chem. Theory … 9*, 868–875.

(169) Kniazeff, J., Shi, L., Loland, C. J., Javitch, J. A., Weinstein, H., and Gether, U. (2008) An intracellular interaction network regulates conformational transitions in the dopamine transporter. *J. Biol. Chem. 283*, 17691–701.

(170) Zhao, Y., Terry, D. S., Shi, L., Quick, M., Weinstein, H., Blanchard, S. C., and Javitch, J. a. (2011) Substrate-modulated gating dynamics in a Na+-coupled neurotransmitter transporter homologue. *Nature 474*, 109–13.

(171) Andersen, J., Taboureau, O., Hansen, K. B., Olsen, L., Egebjerg, J., Strømgaard, K., and Kristensen, A. S. (2009) Location of the antidepressant binding site in the serotonin transporter: importance of Ser-438 in recognition of citalopram and tricyclic antidepressants. *J. Biol. Chem. 284*, 10276–84.

(172) Bakan, A., and Bahar, I. (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *PNAS 106*, 14349–14354.

(173) Mondal, S., Khelashvili, G., Shi, L., and Weinstein, H. (2013) The cost of living in the membrane: a case study of hydrophobic mismatch for the multi-segment protein LeuT. *Chem. Phys. Lipids 169*, 27–38.

(174) Chae, P. S., Rasmussen, S. G. F., Rana, R. R., Gotfryd, K., Chandra, R., Goren, M. A., Kruse, A. C., Nurva, S., Loland, C. J., Pierre, Y., Drew, D., Popot, J., Picot, D., Fox, B. G., Guan, L., Gether, U., Byrne, B., Kobilka, B., and Gellman, S. H. (2010) Maltose–neopentyl glycol (MNG) amphiphiles for solubilization, stabilization and

crystallization of membrane proteins. *Nat. Methods 7*, 1003–1008.

(175) Chung, K. Y., Kim, T. H., Manglik, A., Alvares, R., Kobilka, B. K., and Prosser, R. S. (2012) The role of detergents on conformational exchange of a G protein-coupled receptor. *J. Biol. Chem.* 1–19.

(176) Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem. 30*, 1545–614.

(177) Klauda, J. B., Venable, R. M., Freites, J. A., O'Connor, J. W., Tobias, D. J., Mondragon-Ramirez, C., Vorobyov, I., MacKerell, A. D., and Pastor, R. W. (2010) Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B 114*, 7830–43.

(178) Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem. 26*, 1781–802.

(179) Glykos, N. M. (2006) Software news and updates. Carma: a molecular dynamics analysis program. *J. Comput. Chem. 27*, 1765–8.

(180) Piscitelli, C. L., and Gouaux, E. (2012) Insights into transport mechanism from LeuT engineered to transport tryptophan. *EMBO J. 31*, 228–35.

(181) Moradi, M., and Tajkhorshid, E. (2013) Mechanistic picture for conformational transition of a membrane transporter at atomic resolution. *PNAS 110*, 18916–21.

(182) Celik, L., Schiøtt, B., and Tajkhorshid, E. (2008) Substrate binding and formation of an occluded state in the leucine transporter. *Biophys. J. 94*, 1600–12.

(183) Koldsø, H., Autzen, H. E., Grouleff, J., and Schiøtt, B. (2013) Ligand induced conformational changes of the human serotonin transporter revealed by molecular dynamics simulations. *PLoS One 8*, e63635.

(184) Khelashvili, G., Levine, M. V, Shi, L., Quick, M., Javitch, J. a, and Weinstein, H. (2013) The membrane protein LeuT in micellar systems: Aggregation dynamics and detergent binding to the S2 site. *J. Am. Chem. Soc. 135*, 14266–14275.

(185) Krishnamurthy, H., and Gouaux, E. (2012) X-ray structures of LeuT in

substrate-free outward-open and apo inward-open states. *Nature 481*, 469–74.

(186) Wang, H., Goehring, A., Wang, K. H., Penmatsa, A., Ressler, R., and Gouaux, E. (2013) Structural basis for action by diverse antidepressants on biogenic amine transporters. *Nature 503*, 141–145.

(187) Stolzenberg, S., Quick, M., Zhao, C., Gotfryd, K., Khelashvili, G., Gether, U., Loland, C. J., Javitch, J. A., Noskov, S., Weinstein, H., and Shi, L. (2015) Mechanism of the Association between Na+ Binding and Conformations at the Intracellular Gate in Neurotransmitter:Sodium Symporters. *J. Biol. Chem.* 1–28.

(188) Shaw, D. E., Chao, J. C., Eastwood, M. P., Gagliardo, J., Grossman, J. P., Ho, C. R., Lerardi, D. J., Kolossváry, I., Klepeis, J. L., Layman, T., McLeavey, C., Deneroff, M. M., Moraes, M. A., Mueller, R., Priest, E. C., Shan, Y., Spengler, J., Theobald, M., Towles, B., Wang, S. C., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., and Bowers, K. J. (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM 51*, 91.

(189) Martyna, G. J., Klein, M. L., and Tuckerman, M. (1992) Nose–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys. 97*, 2635–2643.

(190) Hoover, W. G. (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A 31*, 1695–1697.

(191) Blanchard, S. C., Kim, H. D., Gonzalez, R. L., Puglisi, J. D., and Chu, S. (2004) tRNA dynamics on the ribosome during translation. *PNAS 101*, 12893–12898.

(192) Lata, S., and Piehler, J. (2005) Stable and functional immobilization of histidine-tagged proteins via multivalent chelator headgroups on a molecular poly(ethylene glycol) brush. *Anal. Chem. 77*, 1096–1105.

(193) Akyuz, N., Altman, R. B., Blanchard, S. C., and Boudker, O. (2013) Transport dynamics in a glutamate transporter homologue. *Nature 502*, 114–118.

(194) Roy, R., Hohng, S., and Ha, T. (2008) A practical guide to single-molecule FRET. *Nat. Methods 5*, 507–16.

(195) Qin, F. (2004) Restoration of single-channel currents using the segmental k-means method based on hidden Markov modeling. *Biophys. J. 86*, 1488–1501.

(196) Billesbølle, C. B., Krüger, M. B., Shi, L., Quick, M., Li, Z., Stolzenberg, S., Kniazeff, J., Gotfryd, K., Mortensen, J. S., Javitch, J. A., Weinstein, H., Loland, C. J., and Gether, U. (2015) Substrate-induced unlocking of the inner gate determines the catalytic efficiency of a neurotransmitter:sodium symporter. *J. Biol. Chem.*

jbc.M115.677658.

(197) Malinauskaite, L., Quick, M., Reinhard, L., Lyons, J. A., Yano, H., Javitch, J. A., and Nissen, P. (2014) A mechanism for intracellular release of Na(+) by neurotransmitter/sodium symporters. *Nat. Struct. Mol. Biol. 21*, 1006–1012.

(198) Beveridge, D. L., and DiCapua, F. M. (1989) Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem. 18*, 431–92.

(199) Khelashvili, G., Stanley, N., Sahai, M. A., Medina, J., LeVine, M. V, Shi, L., De Fabritiis, G., and Weinstein, H. (2015) Spontaneous Inward Opening of the Dopamine Transporter Is Triggered by PIP2-Regulated Dynamics of the N-Terminus. *ACS Chem. Neurosci. 6*, 1825–37.

(200) Khelashvili, G., Doktorova, M., Sahai, M. A., Johner, N., Shi, L., and Weinstein, H. (2015) Computational modeling of the N-terminus of the human dopamine transporter and its interaction with PIP2 -containing membranes. *Proteins 83*, 952–969.

(201) Khoshbouei, H., Sen, N., Guptaroy, B., Johnson, L., Lund, D., Gnegy, M. E., Galli, A., and Javitch, J. A. (2004) N-terminal phosphorylation of the dopamine transporter is required for amphetamine-induced efflux. *PLoS Biol 2*, E78.

(202) Fog, J. U., Khoshbouei, H., Holy, M., Owens, W. A., Vaegter, C. B., Sen, N., Nikandrova, Y., Bowton, E., McMahon, D. G., Colbran, R. J., Daws, L. C., Sitte, H. H., Javitch, J. a., Galli, A., and Gether, U. (2006) Calmodulin Kinase II Interacts with the Dopamine Transporter C Terminus to Regulate Amphetamine-Induced Reverse Transport. *Neuron 51*, 417–429.

(203) Giambalvo, C. (1992) Protein kinase C and dopamine transport. 2. Effects of amphetamine in vitro. *Neuropharmacology 31*, 1211–22.

(204) Giambalvo, C. T. (2003) Differential effects of amphetamine transport vs. Dopamine reverse transport on particulate PKC activity in striatal synaptoneurosomes. *Synapse 49*, 125–133.

(205) Foster, J. D., Pananusorn, B., Cervinski, M. A., Holden, H. E., and Vaughan, R. A. (2003) Dopamine transporters are dephosphorylated in striatal homogenates and in vitro by protein phosphatase 1. *Mol. Brain Res. 110*, 100–108.

(206) Foster, J. D., Pananusorn, B., and Vaughan, R. A. (2002) Dopamine transporters are phosphorylated on N-terminal serines in rat striatum. *J. Biol. Chem. 277*, 25178–25186.

(207) Kahlig, K. M., Javitch, J. a., and Galli, A. (2004) Amphetamine regulation of dopamine transport: Combined measurements of transporter currents and transporter imaging support the endocytosis of an active carrier. *J. Biol. Chem. 279*, 8966–8975.

(208) Kahlig, K. M., Binda, F., Khoshbouei, H., Blakely, R. D., McMahon, D. G., Javitch, J. a, and Galli, A. (2005) Amphetamine induces dopamine efflux through a dopamine transporter channel. *PNAS 102*, 3495–3500.

(209) Dipace, C., Sung, U., Binda, F., Blakely, R. D., and Galli, A. (2007) Amphetamine induces a calcium/calmodulin-dependent protein kinase II-dependent reduction in norepinephrine transporter surface expression linked to changes in syntaxin 1A/transporter complexes. *Mol Pharmacol 71*, 230–239.

(210) Binda, F., Dipace, C., Bowton, E., Robertson, S. D., Lute, B. J., Fog, J. U., Zhang, M., Sen, N., Colbran, R. J., Gnegy, M. E., Gether, U., Javitch, J. A., Erreger, K., and Galli, A. (2008) Syntaxin 1A interaction with the dopamine transporter promotes amphetamine-induced dopamine efflux. *Mol Pharmacol 74*, 1101–1108.

(211) Foster, J. D., Cervinski, M. A., Gorentla, B. K., and Vaughan, R. A. (2006) Regulation of the Dopamine Transporter by Phosphorylation. *Springer-Verlag Berlin Heidelb. 175*, 197–214.

(212) Thwar, P. K., Guptaroy, B., Zhang, M., Gnegy, M. E., Burns, M. A., and Linderman, J. J. (2007) Simple transporter trafficking model for amphetamine-induced dopamine efflux. *Synapse 61*, 500–514.

(213) Torres, B., and Ruoho, a. E. (2014) N-terminus regulation of VMAT2 mediates methamphetamine-stimulated efflux. *Neuroscience 259*, 194–202.

(214) Cremona, M. L., Matthies, H. J., Pau, K., Bowton, E., Speed, N., Lute, B. J., Anderson, M., Sen, N., Robertson, S. D., Vaughan, R. A., Rothman, J. E., Galli, A., Javitch, J. A., and Yamamoto, A. (2011) Flotillin-1 is essential for PKC-triggered endocytosis and membrane microdomain localization of DAT. *Nat Neurosci 14*, 469–477.

(215) Hamilton, P. J., Belovich, A. N., Khelashvili, G., Saunders, C., Erreger, K., Javitch, J. A., Sitte, H. H., Weinstein, H., Matthies, H. J., and Galli, A. (2014) PIP2 regulates psychostimulant behaviors through its interaction with a membrane protein. *Nat Chem Biol 10*, 582–589.

(216) Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2007) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci. Chapter 2*, Unit 2 9.

(217) Beuming, T., Beuming, T., Shi, L., Shi, L., Javitch, J. a, Javitch, J. a, Weinstein, H., and Weinstein, H. (2006) A Comprehensive Structure-Based Alignment of Prokaryotic and Eukaryotic Neurotransmitter/Na. *Mol. Pharmacol. 70*, 1630–1642.

(218) Penmatsa, A., Wang, K. H., and Gouaux, E. (2013) X-ray structure of dopamine transporter elucidates antidepressant mechanism. *Nature 503*, 85–90.

(219) Zomot, E., Bendahan, A., Quick, M., Zhao, Y., Javitch, J. A., and Kanner, B. I. (2007) Mechanism of chloride interaction with neurotransmitter:sodium symporters. *Nature 449*, 726–30.

(220) Kantcheva, A. K., Quick, M., Shi, L., Winther, A.-M. L., Stolzenberg, S., Weinstein, H., Javitch, J. a, and Nissen, P. (2013) Chloride binding site of neurotransmitter sodium symporters. *PNAS 110*, 8489–94.

(221) Ariga, T., Macala, L. J., Saito, M., Margolis, R. K., Greene, L. a, Margolis, R. U., and Yu, R. K. (1988) Lipid composition of PC12 pheochromocytoma cells: characterization of globoside as a major neutral glycolipid. *Biochemistry 27*, 52–8.

(222) Harvey, M. J., Giupponi, G., and Fabritiis, G. De. (2009) ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput. 5*, 1632–1639.

(223) Lupyan, D., Mezei, M., Logothetis, D. E., and Osman, R. (2010) A molecular dynamics investigation of lipid bilayer perturbation by PIP2. *Biophys. J. 98*, 240–247.

(224) Kraskov, A., Stögbauer, H., Andrzejak, R. G., and Grassberger, P. (2005) Hierarchical clustering using mutual information. *Europhys. Lett. 70*, 278–284.

(225) Fowlkes, E., and Mallows, C. (1983) A Method for Comparing Two A Method for Hierarchical Clusterings. *J. Am. Stat. Assoc. 78*, 553–569.

(226) Kazmier, K., Sharma, S., Quick, M., Islam, S. M., Roux, B., Weinstein, H., Javitch, J. A., and McHaourab, H. S. (2014) Conformational dynamics of ligand-dependent alternating access in LeuT. *Nat Struct Mol Biol 21*, 472–479.

(227) Shan, J., Javitch, J. A., Shi, L., and Weinstein, H. (2011) The substrate-driven transition to an inward-facing conformation in the functional mechanism of the dopamine transporter. *PLoS One 6*.

(228) Koldsø, H., Autzen, H. E., Grouleff, J., and Schiøtt, B. (2013) Ligand induced conformational changes of the human serotonin transporter revealed by molecular dynamics simulations. *PLoS One 8*, e63635.

(229) Han, Y., Moreira, I. S., Urizar, E., Weinstein, H., and Javitch, J. A. (2009)

Allosteric communication between protomers of dopamine class A GPCR dimers modulates activation. *Nat. Chem. Biol. 5*, 688–95.

(230) Moro, O., Lameh, J., Högger, P., and Sadée, W. (1993) Hydrophobic amino acid in the i2 loop plays a key role in receptor-G protein coupling. *J. Biol. Chem. 268*, 22273–22276.

(231) Ballesteros, J. a., Jensen, A. D., Liapakis, G., Rasmussen, S. G. F., Shi, L., Gether, U., and Javitch, J. a. (2001) Activation of the β2-Adrenergic Receptor Involves Disruption of an Ionic Lock between the Cytoplasmic Ends of Transmembrane Segments 3 and 6. *J. Biol. Chem. 276*, 29171–29177.

(232) Fritze, O., Filipek, S., Kuksa, V., Palczewski, K., Hofmann, K. P., and Ernst, O. P. (2003) Role of the conserved NPxxY(x)5,6F motif in the rhodopsin ground state and during activation. *PNAS 100*, 2290–2295.

(233) Shan, J., Khelashvili, G., Mondal, S., Mehler, E. L., and Weinstein, H. (2012) Ligand-Dependent Conformations and Dynamics of the Serotonin 5-HT 2A Receptor Determine Its Activation and Membrane-Driven Oligomerization Properties. *PLoS Comput. Biol. 8*.

(234) Han, D. S., Wang, S. X., and Weinstein, H. (2008) Active state-like conformational elements in the beta2-AR and a photoactivated intermediate of rhodopsin identified by dynamic properties of GPCRs. *Biochemistry 47*, 7317–21.

(235) González-Maeso, J., Weisstaub, N. V, Zhou, M., Chan, P., Ivic, L., Ang, R., Lira, A., Bradley-Moore, M., Ge, Y., Zhou, Q., Sealfon, S. C., and Gingrich, J. a. (2007) Hallucinogens recruit specific cortical 5-HT(2A) receptor-mediated signaling pathways to affect behavior. *Neuron 53*, 439–52.

(236) Roth, B. L. (2011) Irving Page Lecture: 5-HT2A serotonin receptor biology: Interacting proteins, kinases and paradoxical regulation. *Neuropharmacology 61*, 348–354.

(237) Meltzer, H. Y., and Huang, M. (2008) In vivo actions of atypical antipsychotic drug on serotonergic and dopaminergic systems. *Prog. Brain Res. 172*, 177–197.

(238) Berg, K. A., Clarke, W. P., Maayani, S., and Goldfarb, J. (1998) Pleiotropic Behavior of 5-HT2A and 5-HT2C Receptor Agonists. *Ann. N. Y. Acad. Sci. 861*, 104–110.

(239) Moya, P., and Berg, K. (2007) Functional selectivity of hallucinogenic phenethylamine and phenylisopropylamine derivatives at human 5-hydroxytryptamine (5-HT) 2A and 5-HT2C receptors. *JPET 321*, 1054–1061.

(240) Cussac, D., Boutet-Robinet, E., Ailhaud, M.-C., Newman-Tancredi, A., Martel, J.-C., Danty, N., and Rauly-Lestienne, I. (2008) Agonist-directed trafficking of signalling at serotonin 5-HT2A, 5-HT2B and 5-HT2C-VSV receptors mediated Gq/11 activation and calcium mobilisation in CHO cells. *Eur. J. Pharmacol. 594*, 32–8.

(241) Raote, I., Bhattacharyya, S., and Panicker, M. M. (2013) Functional selectivity in serotonin receptor 2A (5-HT2A) endocytosis, recycling, and phosphorylation. *Mol. Pharmacol. 83*, 42–50.

(242) González-Maeso, J., and Yuen, T. (2003) Transcriptome fingerprints distinguish hallucinogenic and nonhallucinogenic 5-hydroxytryptamine 2A receptor agonist effects in mouse somatosensory cortex. *J. Neuro 23*, 8836–8843.

(243) Provasi, D., Artacho, M. C., Negri, A., Mobarec, J. C., and Filizola, M. (2011) Ligand-induced modulation of the free-energy landscape of G protein-coupled receptors explored by adaptive biasing techniques. *PLoS Comput. Biol. 7*, e1002193.

(244) Zocher, M., Fung, J. J., Kobilka, B. K., and Müller, D. J. (2012) Ligand-Specific Interactions Modulate Kinetic, Energetic, and Mechanical Properties of the Human β2 Adrenergic Receptor. *Structure* 1–12.

(245) Liu, J. J., Horst, R., Katritch, V., Stevens, R. C., and Wüthrich, K. (2012) Biased signaling pathways in β2-adrenergic receptor characterized by 19F-NMR. *Science 335*, 1106–10.

(246) Wang, C., Jiang, Y., Ma, J., Wu, H., Wacker, D., Katritch, V., Han, G. W., Liu, W., Huang, X.-P., Vardy, E., McCorvy, J. D., Gao, X., Zhou, X. E., Melcher, K., Zhang, C., Bai, F., Yang, H., Yang, L., Jiang, H., Roth, B. L., Cherezov, V., Stevens, R. C., and Xu, H. E. (2013) Structural basis for molecular recognition at serotonin receptors. *Science 340*, 610–4.

(247) Wacker, D., Wang, C., Katritch, V., Han, G. W., Huang, X.-P., Vardy, E., McCorvy, J. D., Jiang, Y., Chu, M., Siu, F. Y., Liu, W., Xu, H. E., Cherezov, V., Roth, B. L., and Stevens, R. C. (2013) Structural Features for Functional Selectivity at Serotonin Receptors. *Science 340*, 615–619.

(248) Perez-Aguilar, J. M., Shan, J., LeVine, M. V, Khelashvili, G., and Weinstein, H. (2014) A Functional Selectivity Mechanism at the Serotonin-2A GPCR Involves Ligand-Dependent Conformations of Intracellular Loop 2. *J. Am. Chem. Soc. 136*, 16044–16054.

(249) Berg, K. a, Maayani, S., Goldfarb, J., Scaramellini, C., Leff, P., and Clarke, W. P. (1998) Effector pathway-dependent relative efficacy at serotonin type 2A and 2C

receptors: evidence for agonist-directed trafficking of receptor stimulus. *Mol. Pharmacol. 54*, 94–104.

(250) Kurrasch-orbaugh, D. M., Watts, V. a L. J., Barker, E. L., and Nichols, D. E. (2003) Phospholipase C and Phospholipase A 2 Signaling Pathways Have Different Receptor Reserves. *J Pharmacol Exp Ther. 304*, 229–237.

(251) Schmid, C. L., Raehal, K. M., and Bohn, L. M. (2008) Agonist-directed signaling of the serotonin 2A receptor depends on beta-arrestin-2 interactions in vivo. *PNAS 105*, 1079–84.

(252) Shan, J., Weinstein, H., and Mehler, E. (2010) Probing the structural determinants for the function of intracellular loop 2 in structurally cognate G-protein-coupled receptors. *Biochemistry 49*, 10691–701.

(253) Wacker, D., Fenalti, G., Brown, M. A., Katritch, V., Abagyan, R., Cherezov, V., and Stevens, R. C. (2010) Conserved Binding Mode of Human 2 Adrenergic Receptor Inverse Agonists and Antagonist Revealed by X-ray Crystallography. *Communications 42*, 11443–11445.

(254) Goodsell, D. S., Morris, G. M., and Olson, A. J. (1996) Automated Docking of Flexible Ligands : Applications of AutoDock *9*, 1–5.

(255) Niv, M. Y., and Weinstein, H. (2005) A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. *J. Am. Chem. Soc. 127*, 14072–14079.

(256) Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys. 23*, 327–341.

(257) Lomize, M. A., Lomize, A. L., Pogozheva, I. D., and Mosberg, H. I. (2006) OPM: Orientations of proteins in membranes database. *Bioinformatics 22*, 623–625.

(258) Frishman, D., and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Genet. 23*, 566–579.

(259) Palczewski, K., Kumasaka, T., and Hori, T. (2000) Crystal structure of rhodopsin: AG protein-coupled receptor. *Science 289*, 739–745.

(260) Kuhn, P., Choi, H.-J., Kobilka, B. K., Rasmussen, S. G. F., Cherezov, V., Stevens, R. C., Weis, W. I., Kobilka, T. S., Thian, F. S., Hanson, M. a, and Rosenbaum, D. M. (2007) High-Resolution Crystal Structure of an Engineered Human 2-Adrenergic G Protein Coupled Receptor. *Science 318*, 1258–1265.

(261) Shan, J., Weinstein, H., and Mehler, E. L. (2010) Probing the structural determinants for the function of intracellular loop 2 in structurally cognate G-protein-coupled receptors. *Biochemistry 49*, 10691–701.

(262) Visiers, I., Hassan, S., and Weinstein, H. (2001) Differences in conformational properties of the second intracellular loop (IL2) in 5HT2C receptors modified by RNA editing can account for G protein coupling. *Protein Eng. 14*, 409–14.

(263) Deupi, X., and Standfuss, J. (2011) Structural insights into agonist-induced activation of G-protein-coupled receptors. *Curr. Opin. Struct. Biol. 21*, 541–551.

(264) Shapiro, D. a, Kristiansen, K., Kroeze, W. K., and Roth, B. L. (2000) Differential modes of agonist binding to 5-hydroxytryptamine(2A) serotonin receptors revealed by mutation and molecular modeling of conserved residues in transmembrane region 5. *Mol. Pharmacol. 58*, 877–86.

(265) Almaula, N., Ebersole, B. J., Zhang, D., Weinstein, H., and Sealfon, S. (1996) Mapping the binding site pocket of the serotonin 5-HT2A receptor. *J Biol Chem 271*, 14672–14675.

(266) Venkatakrishnan, a J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F., and Babu, M. M. (2013) Molecular signatures of G-protein-coupled receptors. *Nature 494*, 185–94.

(267) Burstein, E., Spalding, T., and Brann, M. (1998) The second intracellular loop of the m5 muscarinic receptor is the switch which enables G-protein coupling. *J. Biol. Chem. 273*, 24322–24327.

(268) Marion, S., Oakley, R. H., Kim, K.-M., Caron, M. G., and Barak, L. S. (2006) A beta-arrestin binding determinant common to the second intracellular loops of rhodopsin family G protein-coupled receptors. *J. Biol. Chem. 281*, 2932–8.

(269) Gray, J. A., Compton-Toth, B. A., and Roth, B. L. (2003) Identification of two serine residues essential for agonist-induced 5-HT2A receptor desensitization. *Biochemistry 42*, 10853–10862.

(270) Kushwaha, N. (2006) Molecular Determinants in the Second Intracellular Loop of the 5-Hydroxytryptamine-1A Receptor for G-Protein Coupling. *Mol. Pharmacol. 69*, 1518–1526.

(271) Preininger, A. M., Meiler, J., and Hamm, H. E. (2013) Conformational flexibility and structural dynamics in GPCR-mediated G protein activation: a perspective. *J. Mol. Biol. 425*, 2288–98.

(272) Valiquette, M., Parent, S., Loisel, T., and Bouvier, M. (1995) Mutation of tyrosine-141 inhibits insulin-promoted tyrosine phosphorylation and increased responsiveness of the human beta 2-adrenergic receptor. *Embo J 14*, 5542–5549.

(273) Chodera, J. D., Elms, P. J., Swope, W. C., Prinz, J., Bustamante, C., and No, F. A robust approach to estimating rates from time-correlation functions 1–6.

(274) Isogai, S., Deupi, X., Opitz, C., Heydenreich, F. M., Tsai, C.-J., Brueckner, F., Schertler, G. F. X., Veprintsev, D. B., and Grzesiek, S. (2016) Backbone NMR reveals allosteric signal transduction networks in the β1-adrenergic receptor. *Nature 530*, 237–241.

(275) Ebersole, B. J., Visiers, I., Weinstein, H., and Sealfon, S. C. (2003) Molecular basis of partial agonism: orientation of indoleamine ligands in the binding pocket of the human serotonin 5-HT2A receptor determines relative efficacy. *Mol. Pharmacol. 63*, 36–43.

(276) Swaminath, G., Deupi, X., Lee, T. W., Zhu, W., Thian, F. S., Kobilka, T. S., and Kobilka, B. (2005) Probing the ??2 adrenoceptor binding site with catechol reveals differences in binding and activation by agonists and partial agonists. *J. Biol. Chem. 280*, 22165–22171.

(277) Evron, T., Peterson, S. M., Urs, N. M., Bai, Y., Rochelle, L. K., Caron, M. G., and Barak, L. S. (2014) G protein and ??-arrestin signaling bias at the ghrelin receptor. *J. Biol. Chem. 289*, 33442–33455.

(278) Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and Mackerell, A. D. (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles. *J. Chem. Theory Comput. 8*, 3257–3273.

(279) Lindahl, E., Bjelkmar, P., Larsson, P., Cuendet, M. A., and Hess, B. (2010) Implementation of the charmm force field in GROMACS: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *J. Chem. Theory Comput. 6*, 459–466.

(280) Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., Van Der Spoel, D., Hess, B., and Lindahl, E. (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics 29*, 845–854.

(281) Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method. *J Chem Phys 103*, 8577–8593.

(282) Abrams, J. B., and Tuckerman, M. E. (2008) Efficient and Direct Generation of Multidimensional Free Energy Surfaces via Adiabatic Dynamics without Coordinate Transformations. *J. Phys. Chem. B 112*, 15742–15757.

(283) Rosso, L., Mináry, P., Zhu, Z., and Tuckerman, M. E. (2002) On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *J. Chem. Phys. 116*, 4389.

(284) Maragliano, L., and Vanden-Eijnden, E. (2006) A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett. 426*, 168–175.

(285) Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., Donadio, D., Marinelli, F., Pietrucci, F., Broglia, R. A., and Parrinello, M. (2009) PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun. 180*, 1961–1972.

(286) Liu, Y., and Tuckerman, M. E. (2000) Generalized Gaussian moment thermostatting: A new continuous dynamical approach to the canonical ensemble. *J. Chem. Phys. 112*, 1685.

(287) Cuendet, M. A., and Tuckerman, M. E. (2014) Free energy reconstruction from metadynamics or adiabatic free energy dynamics simulations. *J. Chem. Theory Comput. 10*, 2975–2986.

(288) Cuendet, M. A., and Tuckerman, M. E. (2012) Alchemical Free Energy Differences in Flexible Molecules from Thermodynamic Integration or Free Energy Perturbation Combined with Driven Adiabatic Dynamics. *J. Chem. Theory Comput. 8*, 3504–3512.

(289) Kunsch, H. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Stat. 17*, 1217–1241.