INTEGRATIVE MODELING OF SECONDARY ACTIVE TRANSPORTERS

By

Diego del Alamo

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

December 18, 2021

Nashville, Tennessee

Approved:

Tina M. Iverson, Ph.D.

Hassane S. Mchaourab, Ph.D.

Jens Meiler, Ph.D.

Lauren Jackson Parker, Ph.D.

Erkan Karakas, Ph.D.

Carlos F. Lopez, Ph.D.

ACKNOWLEDGMENTS

The last six years at Vanderbilt University have been an unforgettable experience, and I have many people to thank for guiding me along the way.

My two mentors, Dr. Hassane S. Mchaourab and Dr. Jens Meiler, gave me a chance in their laboratories and kept encouraging me even when things weren't working.

My thesis committee, Dr. Tina M. Iverson, Dr. Lauren Jackson Parker, Dr. Erkan Karakas, and Dr. Carlos F. Lopez, offered me guidance and never seemed to shy away from asking me the tough questions that helped me grow as a researcher.

Members of both laboratories were always incredibly generous and patient, even when I would disappear to the other side of campus for weeks at a time. During my first two years in the Meiler lab, Dr. Axel W. Fischer taught me almost everything I know about Linux and the C++ programming language and answered every one of the near-bottomless supply of questions I asked. Throughout the second half of graduate school, Dr. Kevin L. Jagessar guided me through the ins and outs of membrane protein purification and was always willing to help me find the dip while tuning the "old pulse" EPR spectrometer.

I am further indebted to Dr. Eric J. Hustedt, Dr. Andrew Leaver-Fay, Dr. Rocco Moretti, Dr. Marion F. Sauer, and Dr. Davide Sala for working with me on each of the various algorithms and programs I developed during graduate school, and to Dr. Derek P. Claxton, Dr. Reza Dastvan, Lillian DeSousa, Dr. Smriti Mishra, and Dr. Suhaila Rahman for providing invaluable assistance on the experimental side. Their willingness to help was exemplary and inspires me to pay it forward whenever possible.

My parents and siblings provided unwavering support from the moment I had the idea to start this crazy journey.

Finally, my partner Selena, with whom I have shared many adventures. These last few years, and especially these last 18 months, would have been impossible without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Alternating access in LeuT-fold transporters

1.1 Introduction

Secondary active transporters tap the potential energy stored in electrochemical gradients to allow cells to import and export nutrients and cytotoxic drugs as needed [44]. Despite a divergence in topologies and ligands [98, 373], these transporters are believed to operate via alternating access, a generic term referring to the exposure of the substrate-binding site to no more than one side of the membrane at a time [182, 289]. This mechanism allows transporters to couple conformational changes critical to substrate binding and/or release while minimizing the uncoupled flux or leaking of ions down their electrochemical gradients. Although this basic working model of transport has been devised over sixty years ago [182], characterizing the precise molecular mechanisms relating substrate binding to translocation in individual proteins of interest presents a formidable scientific challenge. High-resolution structures cannot identify the movements underpinning alternating access unless a protein is captured in multiple distinct conformations [168, 316]. As such, solution-state measurements, carried out in the absence or presence of various substrates under conditions permitting conformational interconversion, directly report on the transporter's energy landscapes and allows high-resolution snapshots to be assigned to specific intermediate states observed in the functional cycle [199, 241, 274, 301]. Unfortunately, but not unexpectedly, few proteins have been so exhaustively characterized.

The bacterial Neurotransmitter-Sodium Symporter (NSS) homolog LeuT found in the thermophilic archaea *Aquifex aeolicus* presents a rare example of a transporter that has been studied to this extent (Figure 1.1) [455]. LeuT couples the import of small aliphatic amino acids such as leucine and alanine to an inward sodium gradient and an outward proton gradient [475]. Nearly two decades of persistent investigation has produced a library of atomic-resolution crystal structures capturing conformational intermediate states [120, 147, 224, 259, 381, 455] linked by solution-state

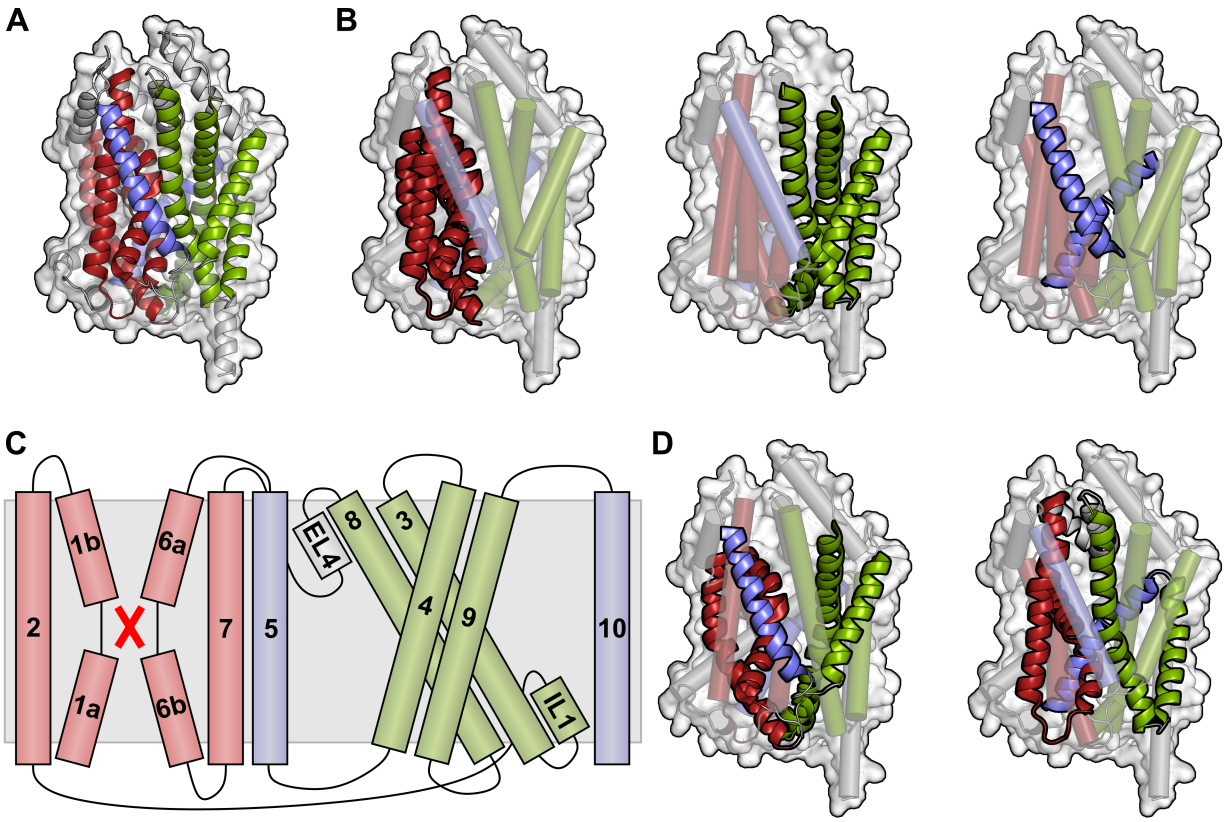

Figure 1.1: Architecture of the bacterial amino acid transporter LeuT. (A) The sodium-coupled amino acid transporter LeuT. (B) The ten-transmembrane helix core consists of a bundle domain (highlighted in red), a hash domain (green), and a pair of gating helices (blue). (C) Cartoon depiction of the conserved transport domain showing the pseudo-twofold symmetry of the ten-helix core. Substrate-binding site indicated by the red X. (D) Inverted repeats in helices 1-5 and 6-10.

experimental data collected using various techniques [6, 32, 33, 76, 213, 281, 474]. Collectively, this body of knowledge describes in exquisite detail the molecular and energetic basis of ligand-dependent conformational changes driving sodium-coupled amino acid transport.

In this chapter, we explore the extent to which structural homologs of LeuT, many of which are found in humans and whose dysfunction contribute to a range of diseases, undergo similar ligand-dependent structural rearrangements [247]. These homologs include mammalian NSSs, such as the serotonin transporter (SERT) and the dopamine transporter (DAT) [27, 77, 311], as well as a multitude of more distant homologs that have been identified across nearly a dozen families (Tables 1.1 and 1.2). As many proteins with the LeuT-fold have been the subject of extensive study over

the course of many decades and have been the subject of numerous recent reviews [51, 67, 69, 112, 159, 223, 309, 345, 380, 480], this chapter focuses on comparing their conformational changes, structural dynamics, and energy landscapes. Necessarily, emphasis is placed on NSSs which are the most extensively studied family of transporters with this topology.

1.2 Functional and structural diversity within the fold

1.2.1 The ten transmembrane helix topology

Proteins in the LeuT-fold belong to the Amino Acid-Polyamine-Organocation (APC) superfamily of transporters [423] and share a core transport structure consisting of ten transmembrane helices (TMH)s related by pseudo-twofold symmetry [136, 211]. These ten helices can be subdivided into three domains: the bundle domain (TMH 1, 2, 6, and 7), the hash domain (TMH 3, 4, 8, and 9), and gating helices (TMH 5 and 10; this Chapter uses this canonical numbering scheme and ignores N-terminal transmembrane helices). Additional N- and/or C-terminal helices uninvolved in transport flank this ten-helix core and vary across individual transporter families comprising the LeuT fold. Characteristically, TMH 1 and 6 contain conserved unwound regions, observed near the geometric center of the ten-helix core, and their involvement in ligand coordination is a hallmark of the fold [364]. Importantly, structural similarity is not accompanied by sequence similarity: few structural motifs are identifiable at the sequence level [32, 258, 394], which prevented these protein families from being co-categorized prior to the determination of their structures [27].

The LeuT fold is perhaps best defined by its evolutionary persistence across families of transporters with sequences that appear unrelated (Figure 1.2) [4]. In addition to NSS [67], the LeuT fold has been found to be adopted by members of the Sodium-Solute Symporter (SSS) [159], Cation-Chloride Cotransporter (CCC) [69], APC transporter [112], Natural resistance-associated macrophage protein (Nramp) [51], Nucleobase-Cation Symporter (NCS1) [309], and Amino Acid-Auxin Permease (AAAP) families. The same topology also describes representatives of proteins families unique to prokaryotes, such as those in the Betaine/Carnitine/Choline Transporter (BCCT) [480], Alanine/Glycine-Cation Symporter (AGCS), and Potassium Uptake Permease (KUP) fami-

Table 1.1: Curated library of LeuT-fold protein structures. Conformations are assigned to either outward-facing open (OFOp), outward-facing occluded (OFOc), fully occluded (OO), inward-facing occluded (IFOc), or inward-facing open (IFOp). Structures are marked if they were mutated (†) or bound to inhibitors (‡) or antibodies (§). Redundant structures have been omitted.

Protein *(Organism)*	**Conf.**	**PDB ID**	**Resolution**	**Substrates**	**Reference**
Neurotransmitter-sodium symporter family					
b^0AT1 *(Homo sapiens)*	OFOc	6M17	2.90 Å	Apo	[458]
DAT *(Drosophila melanogaster)*	OFOp	4XP1§	2.89 Å	$2Na^+/Cl^-$/Dopamine	[312]
GlyT1 *(Homo sapiens)*	IFOp	6ZPL†‡§	3.94 Å	$2Na^+/Cl^-$/Benzoylisoindoline	[369]
LeuT *(Aquifex aeolicus)*	OFOp	3TT1†§	3.10 Å	$2Na^+$	[224]
		3F3A‡	2.00 Å	$2Na^+$/Tryptophan	[381]
		7DIXb	3.49 Å	Na^+/Leucine	[120]
	OFOc	2A65	1.65 Å	$2Na^+$/Leucine	[455]
		5JAE	2.50 Å	Apo	[259]
	IFOc	6XWM†	2.60 Å	$2Na^+$/Phenylalanine	[147]
	IFOp	3TT3†§	3.22 Å	Apo	[224]
MhsT *(Bacillus halodurans)*	IFOc	4US3	2.10 Å	$2Na^+$/Tryptophan	[258]
SERT *(Homo sapiens)*	OFOp	5I6Z†§	4.53 Å	Apo	[77]
		5I6X†‡§	3.14 Å	Paroxetine	[77]
	IFOc	6DZZ‡§	3.60 Å	Ibogaine	[78]
Amino acid-polyamine-organocation transporter family					
AdiC *(Escherichia coli)*	OFOp	3OB6	3.00 Å	Arginine	[221]
		5J4I	2.21 Å	Apo	[174]
		5J4N	2.59 Å	Agmatine	[174]
	OFOc	3L1L†	3.00 Å	Arginine	[122]
AdiC *(Salmonella typhimurium)*	OFOp	3NCY	3.20 Å	Apo	[122]
ApcT *(Methanococcus janaschii)*	IFOc	3GIA	2.32 Å	Apo	[367]
$b^{(0,+)}$AT1 *(Homo sapiens)*	IFOp	6LI9	2.30 Å	Arginine	[457]
		6LID	2.70 Å	Apo	[457]
BasC *(Carnobacterium* sp. AT7*)*	IFOp	6F2W§	3.40 Å	2-Aminoisobutyrate	[111]
		6F2G§	2.92 Å	Apo	[111]
GadC *(Escherichia coli)*	IFOp	4DJI	3.19 Å	Apo	[254]
GkApcT *(Geobacillus kaustophilus)*	IFOc	5OQT	2.86 Å	Alanine	[201]
		6F34	3.13 Å	Arginine	[201]
Lat1 *(Homo sapiens)*	OFOp	7DSQ‡	3.40 Å	Diiodotyrosine	[456]
	IFOp	6IRS‡	3.30 Å	JPH203	[459]
		6IRT‡	3.50 Å	BCH	[459]
		6JMQ§	3.31 Å	Apo	[230]
Lat2 *(Homo sapiens)*	IFOp	7CMH	3.40 Å	Tryptophan	[460]
xCT *(Homo sapiens)*	IFOp	7CCS†	6.20 Å	Apo	[300]
Cation-chloride cotransporter family					
NKCC1 *(Homo sapiens)*	IFOp	6PZT	3.46 Å	Apo	[464]
NKCC1 *(Danio rerio)*	IFOp	6NPL	2.90 Å	$K^+/2Cl^-$	[68]
KCC1 *(Homo sapiens)*	IFOp	6KKT	2.90 Å	$K^+/2Cl^-$	[246]
		6KKR	2.90 Å	Apo	[246]
KCC2 *(Homo sapiens)*	IFOp	7D8Z	3.40 Å	Apo	[452]
KCC3 *(Homo sapiens)*	IFOp	6M22‡	2.70 Å	DIOA	[70]
		7D90	3.60 Å	Apo	[452]
KCC4 *(Homo sapiens)*	IFOp	7D99	2.90 Å	Apo	[452]
KCC4 *(Mus musculus)*	IFOp	6UKN	3.65 Å	K^+/Cl^-	[337]

Table 1.2: Curated library of LeuT-fold protein structures (continued).

Protein *(Organism)*	Conf.	PDB ID	Resolution	Substrates	Reference
Betaine/carnitine/choline transporter family					
BetP *(Corynebacterium glutamicum)*	OFOp	4DOJb†	3.25 Å	Apo	[316]
	OFOc	4DOJa†	3.25 Å	Apo	[316]
	OO	4AINa	3.10 Å	Apo	[316]
	IFOc	2WIT	3.35 Å	$2Na^+$/Betaine	[339]
	IFOp	3P03	3.35 Å	$2Na^+$/Choline	[315]
CaiT *(Escherichia coli)*	IFOp	2WSX	3.50 Å	γ-Butyrobetaine	[362]
		3HFX	3.15 Å	Carnitine	[402]
CaiT *(Proteus mirabilis)*	IFOp	2WSW	2.29 Å	Apo	[362]
Natural resistance-associated macrophage protein family					
ScaDMT *(Staphylococcus capitis)*	IFOp	5M94†§	3.10 Å	Apo	[106]
		5M95†§	3.40 Å	Mn^{2+}	[106]
EcoDMT *(Eremococcus coleocola)*	OFOp	5M8K†	3.60 Å	Apo	[107]
		5M87†	3.30 Å	Mn^{2+}	[107]
DraNramp *(Deinococcus radiodurans)*	OFOp	6D91†	2.36 Å	Apo	[52]
		6BU5†	3.30 Å	Mn^{2+}	[52]
	IFOc	6C3I†	2.40 Å	Apo	[52]
	IFOp	5KTE†§	3.94 Å	Apo	[50]
		6D9W†§	3.94 Å	Apo	[52]
Sodium-solute symporter family					
vSGLT *(Vibrio parahaemolyticus)*	IFOc	3DH4	2.70 Å	$2Na^+$/Galactose	[118]
	IFOp	2XQ2†	2.70 Å	Apo	[438]
SiaT *(Proteus mirabilis)*	OFOp	5NV9	1.95 Å	$2Na^+$/Neuraminic acid	[428]
		5NVA	2.26 Å	Apo	[428]
Potassium uptake permease family					
KimA *(Bacillus subtilis)*	IFOp	6S3K	3.70 Å	$3K^+$	[404]
Amino acid-auxin permease family					
DrSLC38A9 *(Danio rerio)*	IFOp	6C08†§	3.17 Å	Arginine	[231]
		7KGV†§	3.40 Å	Apo	[232]
Nucleobase-cation symporter-1 family					
Mhp1 *(Microbacterium tumefaciens)*	OFOp	2JLN	2.85 Å	Apo	[444]
	OFOc	4D1B	3.80 Å	Na^+/Benzylhydantoin	[377]
	IFOp	2X79	3.80 Å	Apo	[374]
Alanine/glycine-cation symporter family					
AgcS *(Methanococcus maripaludis)*	OO	6CSE§	3.24 Å	Na^+/Alanine	[255]
		6CSF§	3.30 Å	Na^+/D-Alanine	[255]

lies. Additionally, several transporter families lacking representatives in the PDB are predicted to have this fold (Figure 1.3) [349, 418, 423].

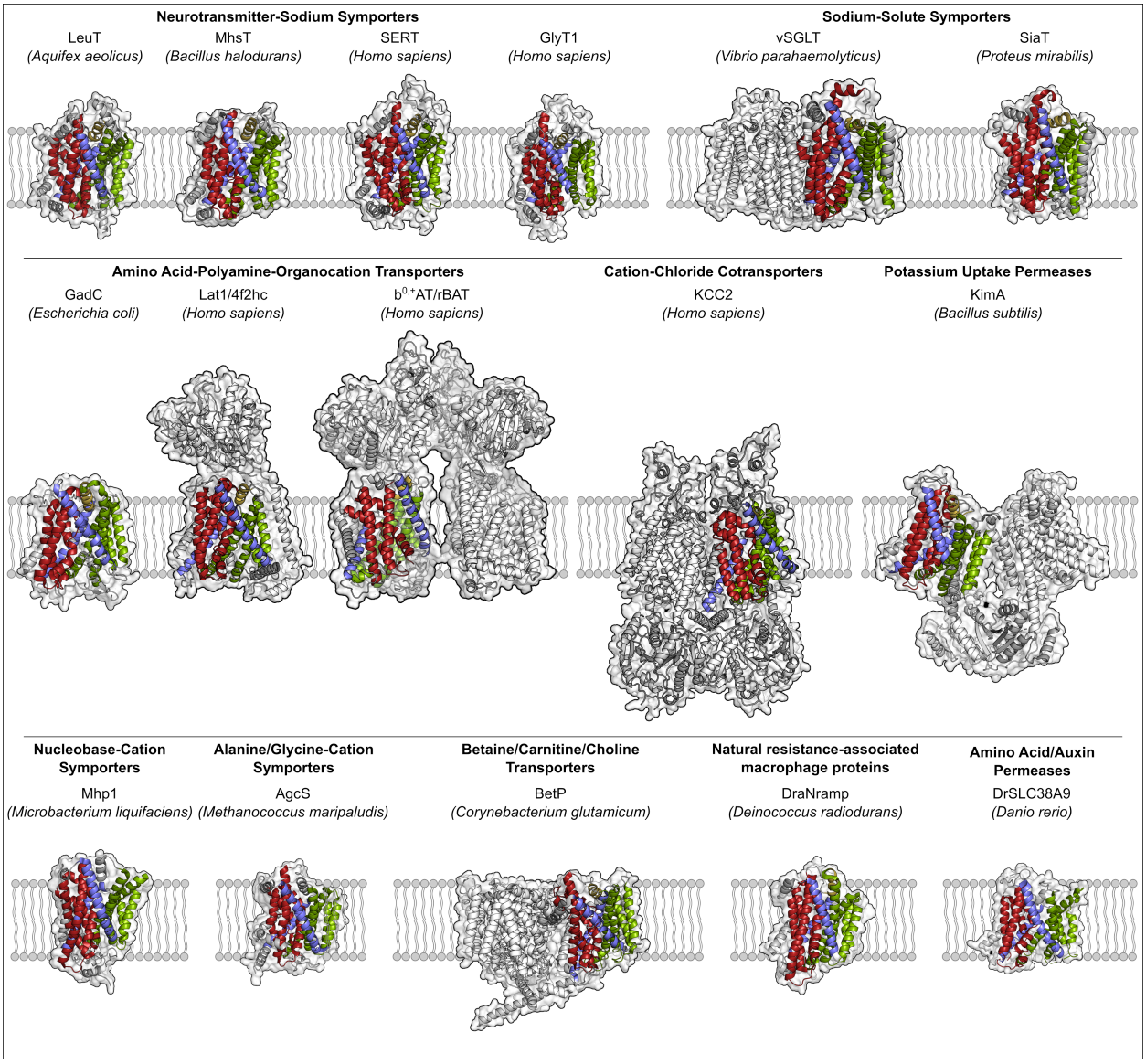

Figure 1.2: Structural diversity within the LeuT fold. The LeuT fold is adopted by proteins found in a range of transporter families. PDB IDs, starting from top left: 2A65, 4US4, 5I6X, 6PZL, 3DH4, 5NVA; 4DJI, 6IRS, 6LI9, 6M23, 6S3K; 2JLN, 6CSE, 2WIT, 6D91, 6C08.

1.2.2 Functional variation within the LeuT fold

Retention of this ten-helix core is all the more remarkable considering the extent to which the functions, ligands, and sequences of these proteins differ. Although the majority of LeuT-fold proteins studied thus far cotransport their substrates and driving ions, others exchange them in opposite directions (symport and antiport, respectively [134]). The centrally located substrate-binding site, shared by symporters and antiporters, accommodates ligands ranging in size and charge from halo-

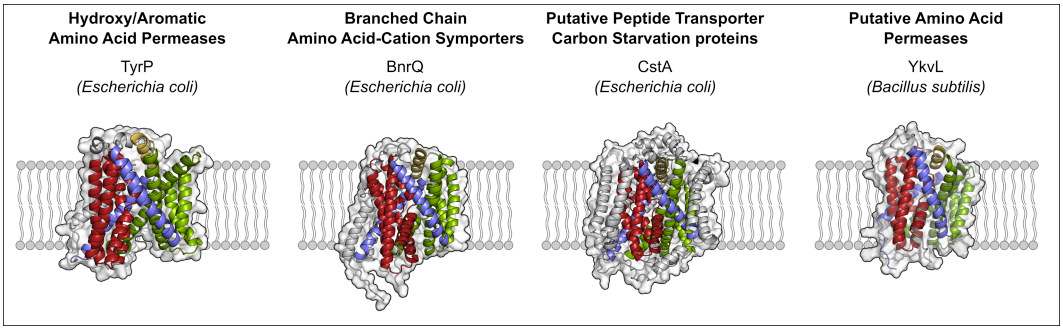

Figure 1.3: Predicted structural models of proteins belonging to LeuT-fold families with no representatives in the PDB. Transporter families are defined by the Transporter Classification Database, and each model was generated using AlphaFold2.

gen ions and divalent metals to sugars and aromatic amino acids. Its manipulation by mutagenesis has been shown to alter substrate specificity profiles in LeuT (NSS) [320, 482], GAT-1 (NSS) [482], and BetP (BCCT) [314], highlighting the involvement of this region in identifying and trafficking ligands.

Beyond plasticity at this specific site, structural features decorating the core transporter further contribute to functional specialization. A remarkable example is SLC38A9, which both exports amino acids from the lysosome and activates the regulatory complex mTORC1 under nutrient-rich conditions [335, 432]. An N-terminal domain elegantly couples these two functions by binding to the cytoplasmic cavity [232] and, following displacement by the transported substrate arginine, releases and binds to GTPases involved in downstream signaling [139]. In an interesting case of convergent evolution, the C-terminal domain of the pH-dependent glutamic acid (Glu)/γ-aminobutyric acid (GABA) exchanger GadC arrests transport by binding to the intracellular cavity at neutral pH in a nearly identical conformation [254]. Autoinhibition by disordered terminal domains has also been directly visualized in several potassium-chloride symporters [70, 452]. In other proteins, such as eukaryotic NSSs, disordered termini instead regulate transport by interacting with a range of cytoplasmic proteins [82, 206]. In BetP, a cytoplasmic C-terminal helical domain regulates transport in response to osmotic stress [149, 222]. These domains often go unobserved in structural studies [77, 246, 311] due to truncation or intrinsic disorder, prompting speculation regarding their role in transport.

1.2.3 Quaternary structures adopted by LeuT-fold transporters

Recent structures obtained by cryogenic electron microscopy (cryo-EM) have bolstered the diversity of quaternary structures known to be adopted by these proteins. Whereas X-ray crystallography demonstrated that LeuT-fold transporters can assemble as homodimers (AdiC, vSGLT) [118, 368], homotrimers (BetP, CaiT) [339, 362], or heterodimers (GkApcT) [201], due to technical limitations only compact binding interfaces were observed. By contrast, oligomers visualized by cryo-EM reveal a multitude of weaker, more flexible, and in some cases asymmetric binding interfaces [68, 230, 300, 404, 457, 458, 459, 460]. For example, the eukaryotic APC transporters Lat1, Lat2, and xCT each associate with 4f2hc (also called CD98hc) [183, 300, 459, 460], a protein with a large extracellular domain and a single transmembrane helix that is uninvolved in transport (Figure 1.2). The homologous APC transporter $b^{(0,+)}$AT1 further assembles into a dimer of dimers with rBAT, each of which resemble the Lat/4f2hc complexes [449, 457]. This dimer-of-dimers arrangement was even observed in the NSS b^0AT1, which associates with Angiotensin-converting enzyme 2 (the experimental structure of this tetramer was determined as part of a larger complex involving the SARS-CoV-2 spike protein) [458]. Notably, the N-terminal transmembrane helices of rBAT and 4f2hc bind to the hash domains of $b^{(0,+)}$AT1 and the Lats, respectively, at a similar position as the C-terminal transmembrane helix of ACE2 to b^0AT1. Separately, the eukaryotic CCCs [68, 71, 246, 337, 464] and the bacterial potassium transporter KimA [404] both fold as homodimers with large domain-swapped cytoplasmic regions. Divergence in both the sequences of these protein families and the structures of their cytoplasmic domains may highlight a recurrent quaternary assembly mechanism, the extent of which has not yet come to light.

Finally, in many cases, the oligomeric interfaces observed by cryo-EM appear to be weaker and more flexible than suggested by the crystal structures. In both Lat1 and KCC1, for example, substantial interdomain movements have been reported when comparing their respective inward-facing apo and outward-facing inhibitor-bound states [246, 456, 459, 460, 476]. Molecular dynamics simulations of the homodimer KimA suggested that contacts between the two transport domains are transient and fleeting, as these domains are tethered to one another only by their intertwined cy-

toplasmic domains [404]; similar observations were experimentally made in CCCs [70]. Such arrangements sharply contrast with the interfaces observed in vSGLT [118, 438], BetP [339], and other oligomers determined by crystallography, and hint at future discoveries regarding how these transporters interact as part of larger complexes.

1.2.4 Recurring elements of substrate binding

These unique structural features and arrangements surround a highly conserved ten-helix architecture shown in Figure 1.1.C that, in several cases, have been shown to retain ligand-binding modes across distantly related proteins (Figure 1.4). A widely discussed example is the conserved sodium site, termed Na2, found in the majority of sodium-coupled symporters [118, 339, 428, 444, 455]. In fact, the Na2 site's recurrence in this fold prompted Chew *et al.* to assign its position to the sodium-binding site in the sodium/potassium/chloride symporter NKCC1, which was subsequently corroborated by molecular dynamics simulations and mutagenesis experiments that severely abrogated transport [68, 181]. Among symporters that bind two sodium ions (NSSs, SiaT [428], BetP [214]), no such conservation is observed in the position of the other sodium ion. In proton-coupled symporters and amino acid exchangers, positively-charged residues occupy this position (lysine in ApcT, GkApcT, and BasC, and arginine in CaiT [111, 368, 201, 362, 402]), highlighting the malleability of substrate coupling throughout the fold. A noteworthy exception is the sodium-coupled amino acid symporter AgcS, which coordinates its only sodium at a position equivalent to LeuT's Na1 site, leaving the Na2 site unoccupied [255]. This is despite its alanine binding site overlapping nearly perfectly with the substrate binding sites of unrelated amino acid transporters from the NSS and APC families (Figure 1.4.C). To our knowledge, no cations besides sodium ions and protons have been observed in this site, and no comparable degree of structural conservation has observed at other ligand-binding sites, such as those involved in binding potassium (transported by SERT, as well as the KUP and CCC families) or chloride (transported by NSSs and CCCs).

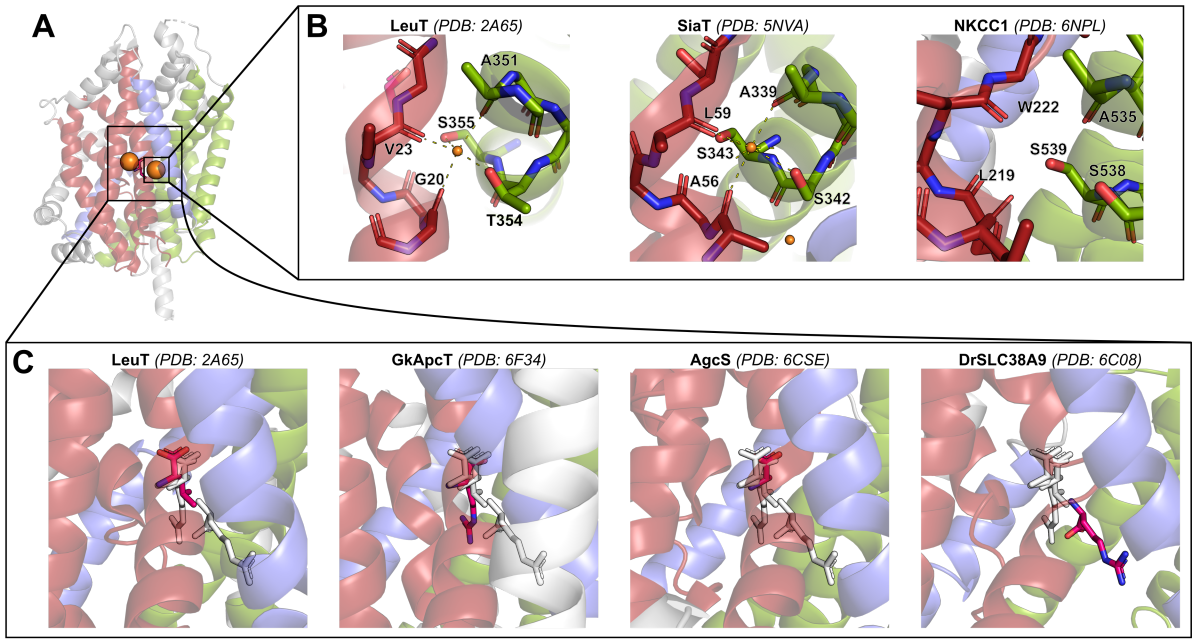

Figure 1.4: Examples of conserved ligand coordination. (A) Sodium ions (orange) and leucine (pink) in LeuT. (B) Conservation of the Na2 site. (C) Partial recurrence of amino acid binding modes. Substrates colored white are shared in all four panels.

1.3 Alternating access inferred from crystal and cryo-EM structures

At the molecular level, alternating access involves the opening and closing of gates providing passage to the substrate-binding site from either the intracellular or extracellular spaces [182] For transporters with the LeuT-fold, this principally manifests as isomerization between outward-facing (OF) and inward-facing (IF) states using a ”rocking bundle” mechanism that forbids substrate entry and exit from the cytoplasmic or periplasmic side of the membrane, respectively [134]. Despite their co-classification, however, closer examination at structural changes in these proteins reveals a striking lack of consensus over the molecular details of alternating access. Conformational divergence as the rule, not the exception, became apparent nearly a decade ago with the publication of high-resolution structures of Mhp1 [374, 444], BetP [339], and LeuT [211, 224, 455], and has since been reinforced by similar studies in SERT [77, 78], DraNramp [50, 52], and Lat1/4f2hc (Figure 1.5) [456, 230, 456]. Additional structures of AdiC [122, 368] and vSGLT [118, 438] in both open and occluded conformations, though limited to OF and IF states, respectively, further expand the

ways in which these transporters grant access to the substrate-binding site. Overall, comparison of pairs of structures reveals fundamental differences in which helices move and which stay fixed (Figure 1.6).

1.3.1 Movement of gating helix TMH5

Along with the intracellular loop preceding it, TMH5 ranks among the most consistently mobile and dynamic regions in the transporters studied so far [394]. In OF conformations, TMH5 nestles against the bundle domain helices TMH1a and TMH6b, forming the highly ordered intracellular "thick gate". In the IF state, by contrast, opening of the intracellular vestibule is driven by rearrangements that vary across families and even individual proteins within families. The contribution of TMH5 to alternating access has been most extensively studied in NSSs [394]. A G_XNP sequence, strictly conserved within the family and partially conserved throughout the fold, putatively mediates both bending and unfolding motions instrumental to the initiation of substrate release (Figure 1.7). Mutagenesis of glycine or proline severely abrogates transport [258], highlighting the importance of the dynamic processes facilitated by this motif. First observed in a substrate-bound IF-occluded conformation of MhsT [258], partial unwinding of TMH5 has been corroborated in several NSSs by hydrogen-deuterium exchange/mass spectrometry (HDX/MS) studies under conditions promoting the IF conformation of each protein [6, 281, 292, 298]. However, although IF-open structures of LeuT and SERT show this helix protruding out from the rest of the transporter [77, 147, 224], orthogonal measurements in LeuT both suggest that under IF-promoting conditions it adopts conformations where the intracellular cavity is occluded, rather than open [212, 372, 474, 475]. As is elaborated below in Section 1.4.1 below, however, these results are qualified by the frequent use of leucine, which has a low transport rate and nanomolar binding affinity [381]. Subsequent solution-state experiments bound to different amino acids found that quenching of fluorescent probes attached to the intracellular half of TMH5 inversely correlated with transport rate [32], suggesting that this IF-occluded conformation may be less stable, relative to IF-open, when transporting substrates with higher turnover rates such as alanine. Nevertheless, in conjunction with other findings

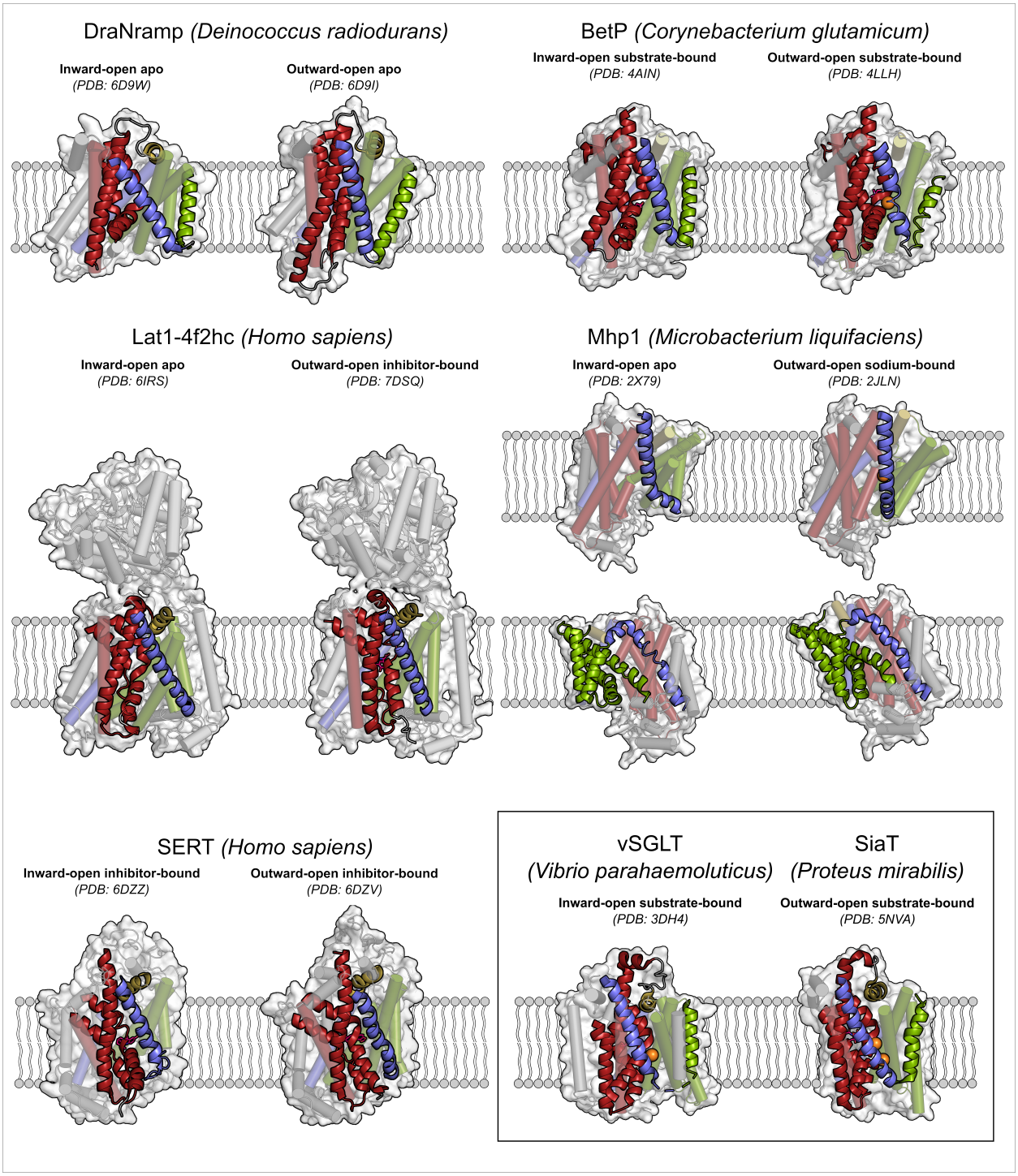

Figure 1.5: Variations in structural dynamics within the LeuT-fold. Conformational dynamics of LeuT-fold transporters show striking differences in how alternating access is carried out. Dynamic and static helices are depicted as ribbons and cylinders, respectively. Bottom left: No individual SSS has been characterized in both OF and IF conformations.

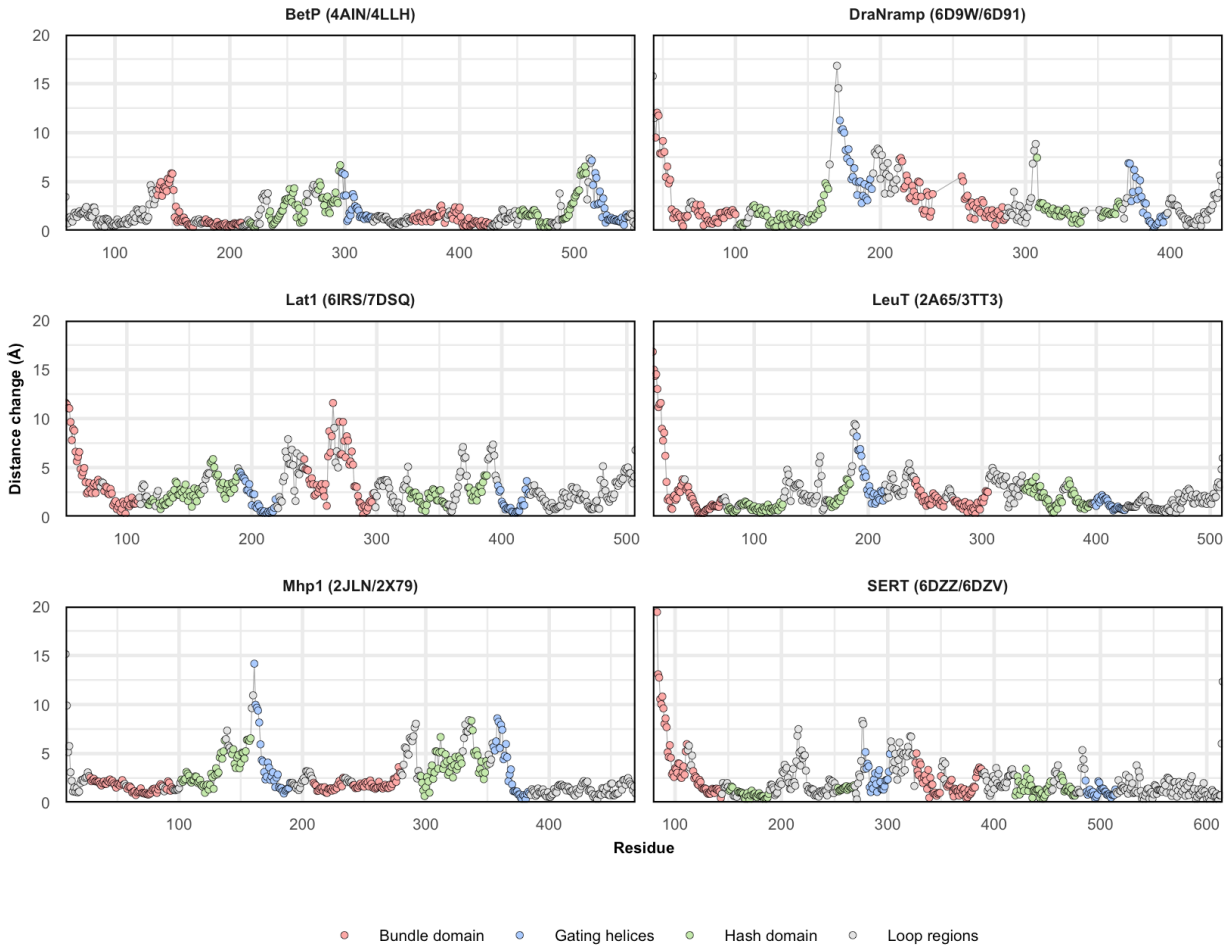

Figure 1.6: Residue-level movements during IF-to-OF isomerization in various LeuT-fold proteins. Unresolved residues omitted from the plot. All structures were aligned using TM-Align [453, 471].

discussed below in Section 1.4.1, this points to a mechanism in which TMH5 preferentially adopts the partially unwound occluded conformation when bound to its substrate but transiently bends to release substrates.

It is notable that TMH5 adopts a similar, but not identical, conformation in IF-open Mhp1, which shares this G_XNP sequence [374]. Despite this agreement, electron paramagnetic resonance (EPR) measurements revealed a degree of disorder in TMH5 altogether absent from similar measurements carried out on LeuT in the presence of leucine [213]. Interestingly, ApcT also shares a LeuT-like bend despite lacking a proline in TMH5 at the equivalent position [367]. Since its

structure has only been determined in a single conformation, and since its homologs such as GadC and BasC maintain a straight conformation of this helix [201, 254], the extent to which the aforementioned NSS movements occur in ApcT and its homologs is unclear [371]. Finally, although TMH5 is also involved in opening the intracellular cavity in DraNramp [52], which also lacks the conserved mid-helical proline, it undergoes a rigid-body up-and-out translation rather than bending and unfolding. Many other IF structures, such as those observed in vSGLT and GkApcT, lack a fully resolved stretch of residues corresponding to intracellular loop (IL) 2, located between TMHs 4 and 5, indicating a high degree of heterogeneity in the crystal lattice or cryo-EM grid [201, 438]. Ultimately, the conformational variation observed across the fold in this loop and helix, combined with solution-state data indicative of local disorder, suggests that the protruded conformation observed in some proteins, though perhaps physiologically relevant and likely fundamental to the transport cycle, may not represent a well-defined low-energy state.

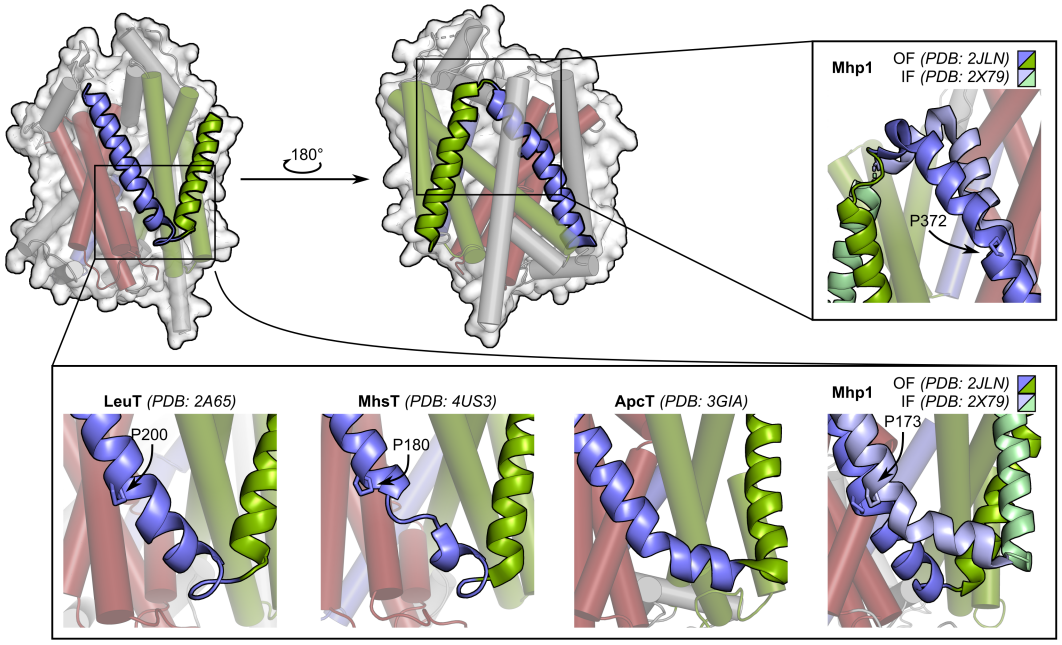

Figure 1.7: Pivoting of TMH5 and TMH10 is observed in a subset of LeuT-fold transporters. Top left: LeuT with TMH4/5 and TM9/10 highlighted. Bottom: Movements of TMH5 observed in NSSs, ApcT, and DraNramp in the IF state. Conserved proline residues are highlighted in LeuT, MhsT and Mhp1. Top right: Movement of TMH0 in Mhp1.

1.3.2 Movement in gating helix TMH10

Movement in TMH10, despite its pseudosymmetry to TMH5, is less frequently observed (Figure 1.7.B). In NSSs, for example, no evidence has been collected that show involvement in either ligand-dependent conformational dynamics, or partial unwinding [213, 281, 292, 298]. Mhp1 shows some partial symmetry of TMH10 to TMH5 in both sequence and structure, with comparable increases in conformational heterogeneity detected by EPR under OF-stabilizing experimental conditions (Figure 1.7) [212, 374, 444]. However, the movement inferred from crystal structures is less dramatic than that of TMH5. In both BetP and DraNramp, differences between their OF-closed and open conformations in this region, though less drastic than in Mhp1, are nonetheless unmistakable; indeed, the corresponding proline in TMH10 facilitating this bend is strictly conserved in the BCCT family and partially conserved among Nramps [51, 362]. Although the APC transporter AdiC both shares this specific residue and shows evidence of this structural movement, its structural similarity to the eukaryotic homolog Lat1, which instead has a cysteine at the equivalent position, indicates that placement of the proline halfway across TMH10 may be coincidental [122, 456].

1.3.3 Helical pivoting in the bundle domain

LeuT's twofold pseudosymmetry initially appeared to imply that a rigid-body rotation of the bundle domain relative to the rest of the structure mediates alternating access [136]. This proposal, although elegant, failed to predict subsequent structural evidence in two key respects. First, the contribution of this domain to alternating access, although prominent in some proteins, is far from universal. Movement in ancillary helices and loop regions has been observed in every protein studied thus far. Second, the bundle domain virtually never moves as a rigid body. The exception, Mhp1, locks the bundle domain into place and instead pivots the hash domain and gating helices around this scaffold [444] (see Section 1.3.4 below).

Movements in TMH1a embody the variable intradomain dynamics observed in these helices. Its apparent dissociation from the rest of the intracellular vestibule, observed in X-ray and cryo-EM structures of NSSs and Nramps [52, 78, 106, 224], has been verified in solution (both families, it

should be noted, lack N-terminal helices and oligomeric interfaces capable of restricting the dynamics of TMH1a; see Figure 1.6). Particular controversy surrounds the relevance of the signature 45° pivot observed in LeuT, which has been attributed to both the use of short-chain detergents commonly used in membrane protein crystallography [291, 383], as well as alanine mutagenesis of a conserved tyrosine residue essential for function [224, 248]. Molecular dynamics simulations of LeuT's IF-open crystal structure in a lipid bilayer later revealed the steep energetic cost of this movement into a more physiological membrane environment [383]. Although this brought attention to the contribution of the membrane mimetic (along with a high-affinity antibody) in stabilizing such an extreme conformer, these findings, alongside experimental measurements obtained using both luminescence resonance energy transfer (LRET) [383] and HDX/MS [6] in lipid environments, nonetheless corroborated the more general hypothesis that TMH1a becomes conformationally disordered in the IF-open state. Nevertheless, as these experiments were executed on similar tyrosine-to-alanine mutants, they do not address the extent to which this IF-open conformation is sampled by the wildtype protein in solution. For example, EPR measurements on equivalent tyrosine-to-alanine mutants recorded a comparable degree of disorder in TMH1a; by contrast, no such dynamics were observed in variants without this mutation [213, 281]. Follow-up experiments in SERT would paint a similar picture, with ibogaine, a ligand used to stabilize the IF-open structure during cryo-EM studies, taking the role of this tyrosine-to-alanine mutation in LeuT [78, 292]. Overall, the data suggest that the IF-open conformations of NSSs, and perhaps other proteins with similar conformational dynamics, transiently sample disordered TMH1a states as part of their function, dovetailing with the conclusions on TMH5 discussed above.

A similar pattern of increased disorder under IF-open-promoting conditions has also been observed in the intracellular side of TMH7. In addition to the movements observed crystallographically, TMH7 in LeuT appears to partially unfold under IF-favoring experimental conditions [281]. For example, the eight N-terminal residues of this helix were not assigned to electron density in IF-open apo DraNramp [52]. The most pronounced motion of TMH7 is likely found in Lat1, which swings over 10 Å to close its intracellular cavity [456]. As is discussed below, such a motion was

suggested by, but not directly observed in, its bacterial homologs in the APC family [111, 254]. Interestingly, no comparable movements are observed on the extracellular side of TMH7. Unfortunately, the absence of data reporting the dynamics of transporters in the APC family prevents any conclusions regarding increases in disorder in this helix from being established.

TMH1b and TMH6a, located on the extracellular sides of the protein, consistently undergo smaller scale but nonetheless significant dynamics essential to opening of the extracellular vestibule [381]. These helices appear to open the extracellular vestibule by moving in concert with the conserved helix extracellular loop (EL) 4, located between TMH7 and TMH8, in NSSs, APC transporters, and Nramps. While the observed movements of EL4 appear minor when compared to TMH1a on the intracellular side, EPR spectroscopy data on LeuT indicate that crystal structures may understate the true extent to which these helices move [212]. Interestingly, on the intracellular side, no equivalent coupling between TMH1a, TMH6b, and IL1 has been detected to our knowledge. Indeed, unlike EL4, IL1 appears to be firmly stapled to the hash domain.

1.3.4 The hash domain generally acts as a rigid body

Relative to movements outlined above, independent helical movement within the hash domain are relatively rare. Mhp1 stands out in rocking this domain, alongside bending in TMH5 and TMH10, to fully mediate alternating access [213, 374, 444]. Similar movements were recorded in vSGLT using EPR [310], although these coincided with additional movement distributed throughout the rest of the structure. As a point of contrast, other proteins limit their movements to bending of TMH4 on the intracellular side and TMH9 on the extracellular side to complement aforementioned movements of TMH5 and TMH10, respectively. The contribution of ancillary helices, which are frequently found adjacent to the hash domain, in explaining this phenomenon is unclear. In an interesting twist, a preprint publication describing the IF-to-OF transition in KCC1 proposes that alternating access is purely mediated by movement of TMH3 and TMH8, while TMH4 and TMH9 remain fixed [246, 476].

1.3.5 Conformational stabilization complicates the interpretation of these structures

Importantly, many of these transporters, with Mhp1 and BetP being noteworthy exceptions, could only be coaxed into specific conformations using mutations, antibodies, or high-affinity transport inhibitors detrimental to function (see Tables 1.1 and 1.2). In addition to the controversial use of a transport-abolishing tyrosine-to-alanine mutation in LeuT discussed above (Section 1.3.3) [224], stabilization of its IF-open state was also achieved by mutating a conserved tryptophan similarly found to be essential for function [147]. Crystallographic capture of DraNramp in OF and IF conformations required glycine-to-arginine and glycine-to-tryptophan mutations near the unwound regions of TMH1a and TMH6a, respectively, that prevented isomerization by obstructing closure of the appropriate vestibule [50, 52]. In human Nramps, the equivalent missense mutation in TMH1a is correlated with severely reduced iron uptake *in vivo* [21], highlighting the extent to which transport function is impaired. Similarly, crystallization of IF-open vSGLT resulted from a lysine-to-alanine mutation that prevented ligand binding and showed no transport activity [438]. Capture of the OF-occluded conformation of AdiC, achieved using an aspartate-to-alanine mutation, may have played a role in stabilizing a ligand pose distinct from those observed in subsequent ligand-bound crystal structures of the wildtype protein [368]. Equivalent studies of the eukaryotic transporters SERT [78], Lat1 [456], and KCC1 [476] have employed a broad panel of potent inhibitors that preferentially bind to specific conformations. Whereas apo SERT readily crystallized in an OF conformation, capture of its IF conformation required the small molecule ibogaine [78]. Similarly, both Lat1 and KCC1 were structurally characterized in IF conformations in the absence of ligands [230, 246, 452, 457] but could only be described in OF conformations using inhibitors [456, 476]. In each of these cases, the introduction of small molecules and/or inactivating mutations arrested transport by stabilizing conformers off-path with respect to the protein's functional cycle.

Nevertheless, assuming the physiological relevance of these structures, their comparison naturally prompts speculation regarding the evolutionary and/or functional basis for the variation in alternating access mechanisms. One proposal, made by the authors that determined the OF structure of KCC1, is the size of the ligand; only small-scale movements, limited to TMH3 and TMH8,

are necessary for symport of the relatively small ions potassium and chloride [476]. This hypothesis implies that larger substrates require larger movements and is supported by the observation that outward-locked DraNramp can transport protons, but not metals [52]. However, the small size of metals nonetheless raises questions about why DraNramp's OF-to-IF transition consists of such large-amplitude movements. Similarly, the relatively minor conformational changes observed in the transport cycle of BetP [316], the substrates of which are comparable in size to amino acids, have been justified by function-specific adaptations. Ultimately, the insufficient data prevent any conclusions from being established in this respect.

1.4 Connecting the dots: from structures to landscapes

Proteins rarely navigate the conformational space accessible to them in the stepwise fashion suggested by depictions such as the one shown in Figure 1.8. Mechanistic models of transport must necessarily, therefore, also map the conditions under which specific conformations occur. Secondary active transporters lack a molecular motor such as an ATPase and must therefore rely on the energy input provided by ions and ligands to undergo forward transport [45]. The central question concerns how transporters harness this energy to undergo reconfigurations that ultimately result in productive transport. Unfortunately, the outstanding structural record of the LeuT-fold overrepresents static states amenable to structural characterization [294, 414]. As a result, although the resulting structures enrich mechanistic models of transport, such as the glide-symmetry symport mechanisms proposed over two decades ago for NSSs [344] and CCCs [346], they are ill-equipped to directly test them. An example that will not be discussed further is the possible existence of an allosteric binding site in LeuT and other prokaryotic NSSs, which remains controversial despite over a decade of structural and experimental research [131, 237, 281, 321, 330, 329, 371, 405].

1.4.1 Characterizing dynamics in the NSSs

The energy landscapes of NSSs are far better characterized than thoose of LeuT-fold proteins in other families and point to a striking degree of conservation. Measurements carried out in solution consistently demonstrate that sodium stabilizes OF conformations, ligands stabilize occluded con-

formations, and absence of either promotes flexible interconversion between IF-open and OF-open [212, 281, 407, 472, 473]. These data provide additional context for these structures by reporting on how ligand-binding events bias the conformational ensemble, which can hint at the drivers of the transport cycle. At the same time, the data can reveal steps unanticipated by canonical symport and/or antiport mechanisms. For example, recent data suggesting that potassium stabilizes IF-open LeuT hint at a step in which intracellular potassium ions indirectly participate in the transport cycle by competing with sodium and accelerating their release from the central binding site [33, 281] (facilitation of substrate release by allosterically bound ions has since been reported in KCC1 [246] and KCC2 [469] and suggested for KCC4 [337]).

These studies further highlight how homologs might diverge due to functional specialization. Despite their structural similarity, LeuT, SERT and DAT were observed to have slight variations in their conformational dynamics. Whereas LeuT adopts an IF-occluded conformation when bound to leucine [213, 281], SERT fluctuates between IF-occluded and IF-open [292]. More intriguingly, the helical unwinding in IL2 and TMH5 initially proposed by the IF-occluded structure of MhsT [258], although plausible in LeuT and SERT, was altogether inconsistent with data collected in DAT suggesting a lack of cooperative movement in this region [298]. Comparable dynamics were instead observed in IL4, a nearby region that was previously found to be critical to IF-opening in other eukaryotic NSSs but static in LeuT [154, 213] (lack of coverage in HDX/MS studies of SERT prevented this region from being studied [292]). However, the presence of lipids and cholesterol in DAT samples presents a confounding factor when attempting to directly compare these results to those collected in SERT, which was studied in detergent micelles. This suspicion is supported by previous studies that reported modulation of conformational dynamics in both eukaryotic NSSs and other secondary active transporters by detergent and lipids [78, 87, 180, 267, 292, 311, 468].

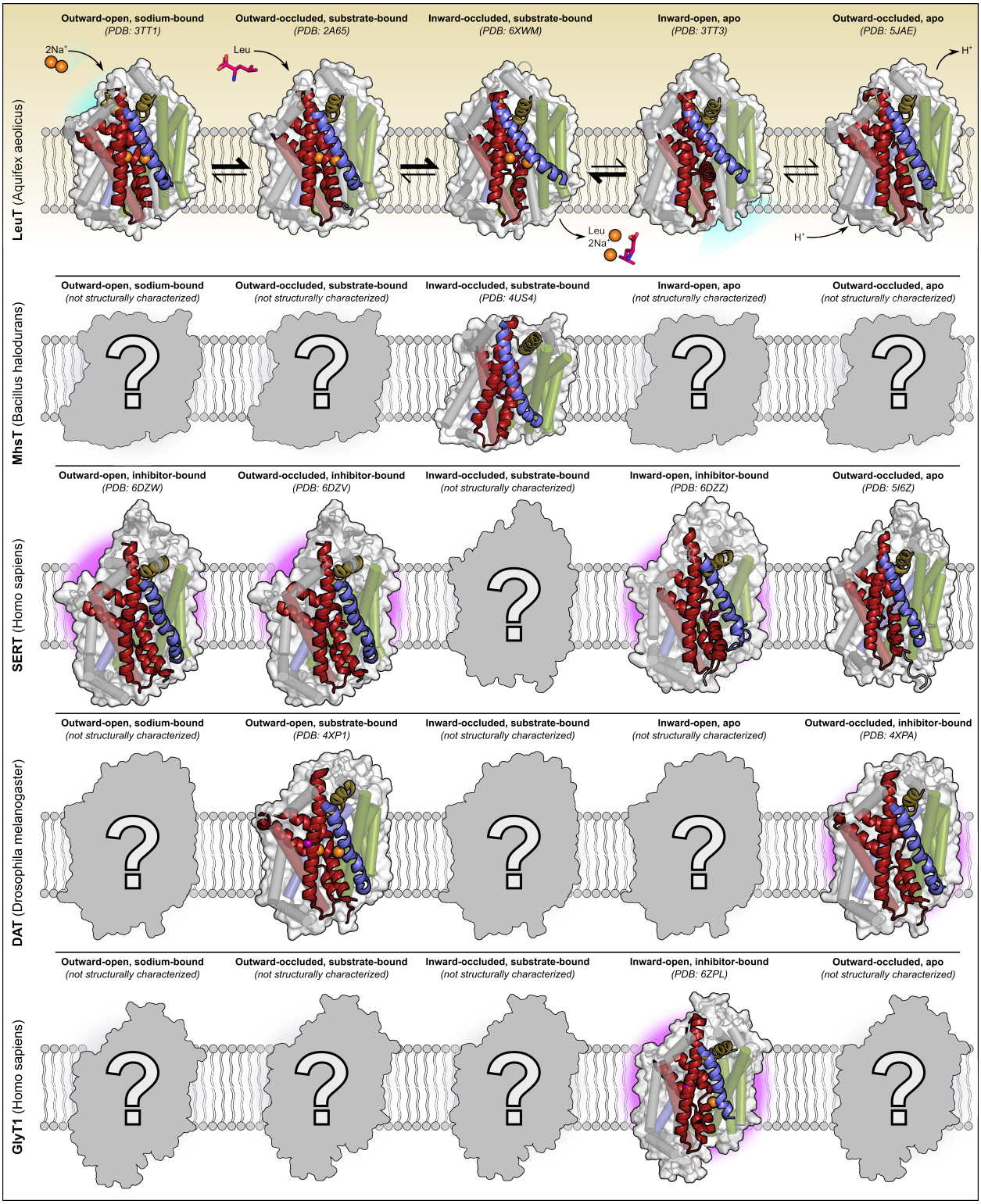

Figure 1.8: Conformational dynamics of neurotransmitter-sodium symporters inferred from crystal structures. Top: LeuT couples the import of small aliphatic amino acids, such as leucine, to the inward and outward electrochemical gradients of sodium and protons, respectively. Bottom: Incomplete transport cycles of other NSSs. Inhibitor-bound states are highlighted in purple.

Single-molecule visualization of conformational changes using Förster resonance energy transfer (FRET) has added a layer of detail entirely missed by these ensemble-level measurements [199, 361]. A recent study employing this technique showed how LeuT appears to undergo uncoupled movement on the intracellular and extracellular sides of the membrane in the absence of substrate, including sampling a channel-like conformation simultaneously open to both sides [407]. Although canonical symport mechanisms of alternating access forbid this arrangement [134], LeuT appears to avoid uncoupled sodium flux by sealing the intracellular cavity in response to sodium binding. Additionally, this study corroborated previous experimental and computational studies on LeuT suggesting that substrate dissociation, and specifically ligand-dependent sodium dissociation from the Na2 site (see Section 1.2.4 above) [32, 333], is the rate-limiting step in transport. However, as mentioned above, this phenomenon may be unique to LeuT. A subsequent study on wildtype MhsT, in which soluble amino acid-binding proteins labeled with pairs of complementary fluorescent probes were cleverly introduced to the interior of MhsT-containing proteoliposomes, determined that the substrate-free IF-to-OF transition was instead rate-limiting [131]. Electrophysiology studies in human NSSs led to similar conclusions [28], suggesting that these discrepancies may be attributable to lower rates of transport and/or higher ligand-binding affinity observed in the thermophilic protein LeuT relative to transporters adapted to function at lower temperatures.

1.4.2 Differential dynamics in the SSS family

The differences in conformational dynamics among NSSs, although not trivial, are dwarfed by those distinguishing them from SSSs such as the eukaryotic sodium/glucose symporter SGLT1, the prokaryotic sodium/galactose symporter vSGLT, and the prokaryotic sodium/proline symporter PutP. While not as well studied, these proteins traverse an energy landscape that is distinct from those of NSSs and indicative of the challenges inherent to the interpretation of solution-state dynamics data. EPR measurements of vSGLT [310] and PutP [331] as well as fluorescent labeling and cysteine accessibility measurements in SGLT1 [249, 250, 351] suggest that ligand-dependent conformational dynamics were effectively inverted relative to NSSs, with apo and/or sodium-rich

conditions favoring IF conformations and substrate binding stabilizing the OF conformation. As with NSSs discussed above, key differences between vSGLT and PutP were observed: no conformational response to sodium binding was detected in the former [310], whereas sodium binding to the latter led to closing of EL4 and increased labeling of residues lining the intracellular vestibule [194, 331, 440]. Importantly, similar sodium-invariant conformational dynamics were independently reported in the unrelated bacterial transporter Mhp1 using both EPR and cysteine accessibility measurements [212, 62, 443], and sodium-driven stabilization of an IF state was also suggested by EPR data in BetP [235].

Data from EPR studies on vSGLT prompted the conclusion that the sodium gradient is a critical driver of transport [310]. Consistent with this hypothesis, accessibility measurements and fluorescent labeling data collected in human SGLT1 in cells that actively maintain a sodium gradient showed a conformational landscape nearly identical to NSSs: unrestricted isomerization between OF and IF in the apo state [351] and stabilization of the OF conformation in the presence of sodium and absence of glucose [249, 278]. Critically, the OF-promoting effect of sodium diminished when the electrochemical gradient was decreased [249]. However, while these data support the hypothesis that the gradient may play a similar role in prokaryotic SSSs, a critical difference with unknown significance is SGLT1's 2:1 sodium-to-glucose stoichiometry, which contrasts with the 1:1 stoichiometry of the prokaryotic model systems discussed above.

1.4.3 Energy landscapes in antiporters

Missing from our knowledge of transporters with this fold is a detailed accounting of conformational dynamics data in antiporters. Despite their ubiquity in all domains of life and their extensive structural study by crystallography and cryo-EM (Figure 1.9), their energy landscapes remain virtually uncharacterized. Whereas canonical symport mechanisms, such as those broadly defining NSSs and SSSs, contain a substrate-free isomerization step, canonical antiport mechanisms facilitate substrate exchange by forbidding this conformational change [134]. Instead, the second half of the antiport cycle involves the translocation of a second substrate in the opposite direction. In

non-LeuT-fold antiporters, import of one molecule is proposed to power the energetically unfavorable export of another. For example, mechanisms of alternating access in unrelated transporters that expel toxic drugs frequently involve the cation-dependent stabilization of an IF conformation [108, 180, 270]. The relevance of these findings to understanding LeuT-fold antiporters, however, is unclear, since many of them are not coupled to electrochemical ion gradients and instead function as substrate exchangers facilitating downhill translocation in opposite directions [122, 254, 362, 402].

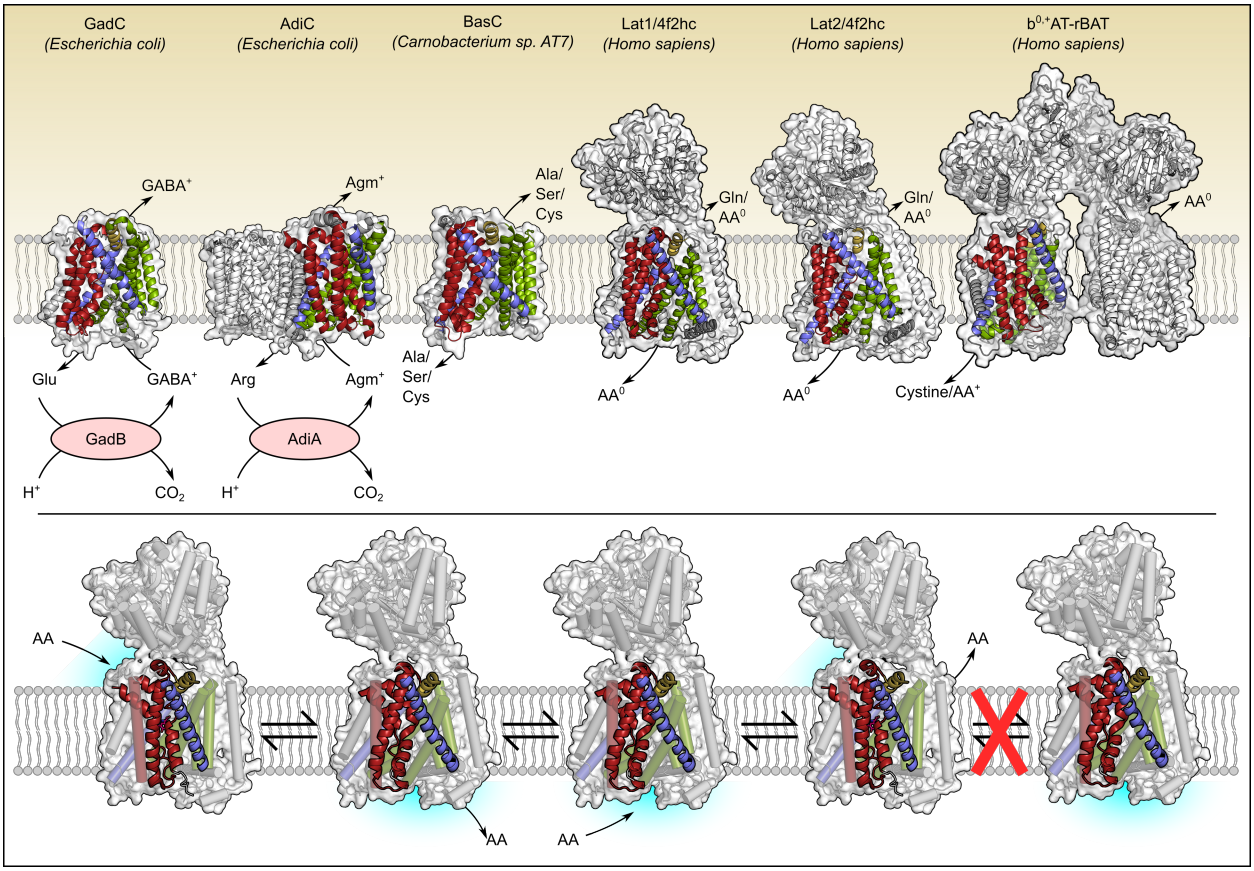

Figure 1.9: Structures and transport cycles of amino acid exchangers in the APC family. Top: The pH-activated precursor-product exchangers GadC and AdiC are co-transcribed with decarboxylases GadB and AdiA, respectively. Bottom: Canonical mechanisms of antiport forbid substrate-free conformational isomerization.

A range of pH-dependent amino acid/polyamine antiporters in the APC family, sometimes called ”virtual proton pumps”, import and export the precursors and products, respectively, of proton-consuming amino acid decarboxylases with which they are cotranscribed (Figure 1.9) [122, 137, 204, 223, 254]. Activation of these decarboxylases under extreme acidic conditions drains the

intracellular concentration of the appropriate amino acids and raises that of the cognate polyamines, thus ensuring that both halves of the antiport cycle are energetically favorable. In fact, the two most well-characterized transporters, AdiC and GadC, are capable of forward and reverse transport of both substrates under equilibrium conditions [254, 253, 416, 417]. In contrast, they are inactive when their intracellular sides, but not extracellular sides, are exposed to protons, or in the presence of a negative-inside electric potential [416, 417]. One hypothesis proposed from Molecular Dynamics (MD) simulations of AdiC posits that protonation of a conserved glutamate leads to the rate-limiting substrate dissociation from the central binding site [481], similar to the proposed role of potassium-induced substrate release on the intracellular side of LeuT (see section 1.4.1).

Homologs with broader substrate specificities such as BasC and Lat1 exchange amino acids in accordance with cellular needs while maintaining high intracellular amino acid concentrations [111, 230, 459]. A critical adaptation in these proteins is their asymmetric binding affinity, with apparent Km values in the micromolar and millimolar range during out-to-in and in-to-out transport, respectively [23]. This is hypothesized to address the disparity between amino acid concentrations on the intracellular and extracellular side, which differ by several orders of magnitude. A secondary form of asymmetry in amino acid exchangers of the APC family is the selective import and export of charged substrates; $b^{(0,+)}$AT1 selectively imports and exports cationic and neutrally charged amino acids, respectively, although neither the structures nor subsequent studies have shed light on the structural basis of this observation [449, 457].

Unfortunately, the recent explosion of structures in the APC transporter family has not been accompanied by studies into these proteins' energy landscapes. To our knowledge, the only study of dynamics in antiporters in the APC family has been in the bacterial serine/threonine antiporter SteT [338] using single-molecule dynamic force spectroscopy, which reports on a protein's kinetic barriers [36]. This study found that conformational flexibility in SteT increased under substrate-bound conditions relative to apo, consistent with the substrate-dependent conformational movement predicted by a canonical model of antiport. However, the data do not report on the protein's thermodynamics, which leaves critical questions about the protein's energy landscape unanswered.

1.5 Comparison to homologous proteins

Overall, the outstanding evidence suggests that proteins in the LeuT-fold share few, if any, signature motifs of alternating access. To answer the question posed in the Introduction, differences in the structures and conformational dynamics of transporters both within and across families suggest that functional specialization has contributed to a degree of evolutionary divergence that prevents our rich knowledge of NSSs in general and LeuT in particular from being directly applied to less well-known families such as KUPs or AAAPs. However, the data also suggest that transporters in the same family are more similar to each other, both in structure and dynamics, than to proteins in other families. Thus, the question is one of evolutionary conservation at the family-level.

1.5.1 Structural similarity within families of transporters

From the perspective of structural similarity, the outstanding data suggest that homologs within a protein family show a degree of structural conservation not shared by other proteins with this fold. During their transport cycles, individual proteins sample conformations that differ by 3 Å C_α root mean squared deviation (RMSD) [323]. Structures of homologous proteins within and across families, meanwhile, differ by around 3 Å and 5 Å C_α RMSD, respectively. This divergence has complicated the development of unified mechanistic models of transport for any protein or family because structural variation between different proteins could either reflect differences in sequence and function, or represent distinct steps in the transport cycle. Indeed, minor structural differences are even observed when comparing the structures of the same protein across different species, such as CaiT from *Proteus mirabilis* and *Escherichia coli* [362], AdiC from *Salmonella typhimurium* and *E. coli* [122], NKCC1 from *Homo sapiens* [464, 469] and *Danio rerio* [68], and KCC4 from *H. sapiens* [452] and *Mus musculus* [337].

An instructive example that was discussed in Section 1.3.1 above is the varying position of TMH5 in NSSs. The ligand-bound IF-occluded structure of MhsT [258, 394] was interpreted as evidence that unwinding of TMH5 is a feature preceding substrate release in all NSSs. In contrast, the conformation of TMH5 in IF-occluded LeuT, which was corroborated by FRET, was bent and

not unwound, which provided comparatively greater access to the intracellular vestibule [147, 407], while the conformation of TMH5 in IF-occluded SERT was described as ”halfway” between that of LeuT and MhsT [78]. It remains unclear if these structures represent distinct steps in a shared transport cycle, or if they reflect more fundamental differences between proteins resulting from functional specialization, evolutionary divergence, and/or environment-specific adaptations.

1.5.2 Structural similarity between prokaryotic and eukaryotic transporters

Nevertheless, the striking degree of structural correlation observed between distantly related proteins, particularly between bacterial and eukaryotic proteins, has been a recurring theme throughout studies of proteins with this fold. The first eukaryotic LeuT-fold transporter to be structural determined, the dopamine transporter DAT from *Drosophila melanogaster* [311], bore a remarkable resemblance to LeuT, obtained from a thermophilic archaeum found in hot springs. Similarly, the structure of the APC transporter GadC in an IF conformation [254] aligns well with those determined for the eukaryotic transporters Lat1/4f2hc [230, 456, 459], Lat2/4f2hc [183, 460], and $b^{(0,+)}$AT1 [449, 457]. This observation is all the more intriguing given that GadC's structure putatively represents an auto-inhibited state with no mechanistic equivalent in the transport cycles of these eukaryotic proteins [254, 253]. Equally fascinating is the correspondence between the OF-open inactive conformations of Lat1 [456] and AdiC [174]. Unfortunately, comparison of eukaryotic and bacterial homologs is only possible in the NSS and APC families, as the structurally determined proteins comprising every other family of transporters are either exclusively prokaryotic or exclusively eukaryotic.

Regarding the energy landscapes of these proteins, a dearth of dynamics data prevents straightforward comparisons from being made. It is nonetheless remarkable that the structural dynamics of vSGLT are more similar to those of Mhp1, which has identical 1:1 sodium-to-substrate stoichiometry despite being unrelated at the sequence level, than those of its homolog SGLT1, which instead mirror those of NSSs sharing its 2:1 sodium-to-substrate stoichiometry [212, 249, 310]. At the same time, given the variation in dynamics data observed among NSSs [6, 213, 281, 292], small

divergences in how SSSs respond to their substrates at the structural level appear to be expected.

1.5.3 Implications of structural similarity and divergence on modeling

In 2008, vSGLT became the second protein, after LeuT itself, to be observed with the LeuT-fold [118]. Publication of its crystal structure in an IF conformation was complemented by an OF model generated from the structure of LeuT that attempted to predict which helices are involved in alternating access. Though this comparison is, with the benefit of hindsight, somewhat inappropriate, modeling of alternate conformers has since been a staple of structural studies that has guided experimental design [142, 147, 310, 295, 467], generated starting points in simulations [37, 215, 333], and contextualized experimental findings [213, 310, 407]. The previous discussion emphasized the risk of assuming that conformations observed in one protein are relevant to others, modeling has been particularly effective at extracting mechanistic insights into eukaryotic proteins from bacterial model systems [37, 136, 135, 142, 431, 482].

Shortly after structure determination of LeuT in 2005, modeling studies led to the identification of the chloride-binding site of eukaryotic NSSs [135, 482]. Chloride ions are not native ligands of LeuT, which instead exchanges two extracellular sodium ions and an amino acid with an intracellular proton [475]. Nevertheless, by identifying a negatively-charge side chain near the sodium binding site exclusive to chloride-independent NSSs [27], this chloride-transporting phenotype could be introduced by mutating native glutamate and aspartate near the sodium-binding site of LeuT and the bacterial NSS Tyt1, respectively [482]. Likewise, chloride-independent transport could be introduced in the GABA transporter GAT1 by replacing the serine in the same position with a glutamate. In parallel, the chloride-binding site in SERT was identified by aligning its sequence to LeuT, and the resulting homology model predicted the chloride-binding position in SERT with astonishing detail [136].

Several drug discovery studies of Lat1 employed homology models, generated from IF-occluded ApcT [367] and OF-open AdiC [368], to guide rational design of inhibitors and ligands [142, 295, 380, 467]. Although small details in the substrate-binding sites of these models were subsequently

found to be inconsistent with the cryo-EM structures [230, 456, 459], they nonetheless facilitated the identification of multiple novel inhibitors. One of these inhibitors was later used to trap Lat1 in an OF-open conformation for cryo-EM studies [456], and the resulting structures verified the ”carboxylate-up” orientation initially predicted by the homology model that had been observed in other transporters (Figure 1.4).

Finally, structural models have been invaluable in guiding restraint selection using sparse experimental data. A model of OF vSGLT, generated from the homologous SiaT [428], was used to determine which measurements to pursue using EPR [310]. Importantly, the OF model provided a context for the relatively sparse distance data and highlighted the flexibility of TMH10 that was unanticipated. Additionally, prior to the structural determination of NKCC1 by cryo-EM, cross-linking experiments were guided by computational models generated from AdiC; the conformation and register of TMHs 10-12 predicted by this model was subsequently validated by cryo-EM [293, 464].

1.6 Scope of this dissertation

The objectives of this dissertation are twofold.

The first objective focuses on studying the structural dynamics of the glutamate/GABA antiporter GadC. At low pH, GadC serves as a model system for other LeuT-fold antiporters, particularly those in the APC sub-family, while at high pH it putatively adopts an inactive conformation. As was discussed in section 1.4.3, these transporters are far less well studied than homologous sodium-coupled symporters, and the extent to which their mechanisms of alternating access are conserved is unclear. Symporters such as LeuT, Mhp1, PutP, vSGLT, and SERT have been shown to undergo ligand-dependent changes in their conformational equilibria; whether GadC or homologous exchangers found in eukaryotes do the same is unknown. These questions are explored in Chapter 5 using EPR spectroscopy and computational modeling.

As will be discussed in Chapter 2, the experimental data collected in GadC can report on changes in the distribution of distances between two spin labels but are local in nature and must be com-

plemented with computational modeling to obtain global, fold-level structural insights. Thus, the second objective of this dissertation focuses on developing novel computational methods to model the structures of these proteins using these data. The Markov Chain Monte Carlo approach used throughout this text separates the modeling process into two steps, sampling and scoring. As is discussed in subsequent chapters, existing sampling and scoring methods do not effectively leverage the experimental data, leading to unacceptable losses in modeling precision and unnecessary increases in computation time. Therefore, this dissertation describes and discusses advancements in both halves of the modeling process. However, more attention is paid to the development of scoring approaches; the novel sampling approach is discussed further in Appendix C. These sampling and scoring methods are combined to attempt to model conformational changes in three homologs of GadC using experimental EPR data.

One thread discussed in chapters 3 and 4 of this dissertation, unrelated to these two objectives, explores the extent to which the analysis of DEER data and its use for computational modeling can be integrated. In general, the time-domain data are first interpreted as distance distributions, which are in turn used as modeling restraints (see Chapter 2 for details). While several recent reports have begun to couple interpretation of the time domain data with structural modeling, the benefits of this approach have not been studied. Chapters 3 and 4 explore these approaches in greater detail for two specific tasks, predicting the folds of protein structures using sparse DEER and determining the positions of nitroxide rotamers, respectively. The goal of these two chapters is to determine the extent to which these two steps can be coupled. As is discussed in Chapter 3, this has the potential to advance the integration of computation and spectroscopy in a manner analogous to the prediction of protein structures using unassigned nuclear magnetic resonance (NMR) chemical shifts or raw SAXS scattering profiles. An overview of methods to analyze and simulate DEER data is provided in the following chapter.

CHAPTER 2

Analysis and modeling applications of DEER data

This Chapter presents an overview of the methods available to analyze DEER data and apply the resulting distance distributions to model protein structures. Particular attention is paid to interpreting these data using Tikhonov regularization and Gaussian mixture models. Distributions converted from the time domain using these methods can then be used to guide structural modeling. Commonly used strategies for integrating these experimental data as restraints are discussed. This Chapter concludes by speculating on how the analysis of DEER data and its application for modeling protein structures can be coupled.

2.1 Introduction

Among the tools available to the field of structural biology, DEER spectroscopy is uniquely suited to monitor the dynamic properties of proteins [171, 186, 274]. DEER, also called PELDOR, measures nanometer-scale distance distributions between two paramagnetic probes attached to the protein's surface. By resolving and reporting full distributions, rather than just average distance values, DEER can reveal conformational heterogeneity and intermediate states that may be inaccessible to crystallography and cryo-EM. The contribution of this technique to the derivation of mechanistic inferences has been bolstered in recent years by its integration with computational modeling [190]. Distributing experimental measurements throughout the structure of a protein allows qualitative conclusions to be synthesized into quantitative structural models. Computational modeling allows one-dimensional distance data to be interpreted in the context of a three-dimensional structural models, thus facilitating further hypothesis testing.

Nonetheless, computational modeling and DEER spectroscopy are each areas of research under active development. Their integration in the literature is highly nonstandard, with customized protocols often being used on a case-by-case basis. As will become clear in this chapter, best practices

are far from established. Individually, each method provides opportunities to make unwarranted inferences; in tandem, they present a risk of overfitting the data and arriving at spurious conclusions. Perhaps as a result, most studies employ the data conservatively and deliberately underleverage some of the DEER technique's advantages, such as its ability to reveal minor populations. This hinders the development of one's understanding of a protein's structural dynamics. Nonetheless, recent methodological advancements in both DEER data analysis and macromolecular modeling promise to mitigate this possibility.

This chapter discusses the analysis of four-pulse DEER data [304] and its interpretation by structural modeling. We focus our attention on the common experimental scenario where proteins are labeled with two flexible spin-$\frac{1}{2}$ nitroxide probes per macromolecule and flash-frozen prior to measurement. Our discussion is limited with respect to more exotic applications, include alternative pulse sequences [41, 55, 280, 386], labeling with lanthanide ions [145, 271, 324, 454] or noncanonical amino acids [54, 357, 358], deliberate introduction of orientation effects [48, 110, 264], specialized sample preparation conditions for long-distance measurements [109, 360], and interpretation of measurements performed at room temperature [148, 227, 283] or in cells [17, 198, 379, 462]. The scope of this chapter nonetheless encompasses the vast majority of integrative modeling studies. That being said, the DEER technique is certain to continue evolving; it is not difficult to envision room-temperature experiments capable of in-cell measurements in proteins with many conserved cysteines.

2.2 Analysis of DEER data

2.2.1 Composition of the DEER signal

Pulse EPR methods measure the amplitude of spin echoes caused by the successive application of microwave-frequency pulses to samples containing unpaired electrons in the presence of an external magnetic field [286, 287]. The signal obtained from a four-pulse DEER experiment reflects time-dependent spin-spin coupling within a macromolecule ($S(t)$) and across macromolecules ($B(t)$):

$$\frac{V(t)}{V_0} = B(t) * (1 - \lambda(1 - S(t))) + \epsilon \tag{2.1}$$

Here $\frac{V(t)}{V_0}$ denotes the normalized signal amplitude and ϵ is normally distributed experimental noise in the signal [104]. The modulation depth λ reports the spin inversion efficiency and is also affected by the parameters of the DEER experiment. The background coupling signal is most commonly modeled using the following stretched exponential function $B(t) = \exp\left(-k|t|^{\frac{d}{3}}\right)$ and relates the background spin concentration k and the intermolecular coupling dimensionality d (with d=3 except under circumstances where, for example, membrane proteins are reconstituted into lipid environments). Alternative background functions can be used to account for an excluded volume effect observed when the size of the molecule under study forbids short-distance intermolecular coupling [209].

Ultimately the spectroscopist seeks to isolate and extract the distance information encoded by $S(t)$. Although the background coupling parameters k, d, and λ report biologically meaningful information [185], they are frequently treated as nuisance parameters during the analysis of DEER data. Nevertheless, their identification is critical to the accurate recovery of experimental distance data, and failure to disentangle $S(t)$ from background contributions to the signal can corrupt the resulting distance distribution by, for example, introducing spurious long-distance peaks [192].

2.2.2 Intramolecular contributions to the experimental signal

The isolation of $S(t)$ and its conversion into a distance distribution $P(r)$ is at the heart of substantial research and methods development. The two are related by the following kernel function:

$$S(t) = \int_0^\infty K(t,r)\, P(r)\, \mathrm{d}r \tag{2.2}$$

$$K(t,r) = \int_0^{\frac{\pi}{2}} \sin\theta \cos\left(\frac{\left(1 - 3\cos^2\theta\right)\mu_0 \mu_B^2 g_X^2 t}{4\pi\hbar r^3}\right)\mathrm{d}\theta \tag{2.3}$$

Here μ_B is the Bohr magneton, μ_0 is the vacuum permeability constant, g is the electron g-

factor, r is the interspin distance in nanometers, t is the timing of the third pulse in microseconds, and θ is the angle between the interspin vector and the bulk magnetization vector (we emphasize the distinction of θ from ϑ, which denotes the parameters of a model and is used in Equations 2.6, 3.3, 4.1, 4.2, 4.5, and 4.6 below). Detailed derivations of 2.3 are available [448]. In practice, the DEER signal S and distance distribution P are discretized into time points and distance bins, respectively. The relationship between the intramolecular component of the signal S and the probability distribution P to K can be denoted by the matrix-vector multiplication $S = KP$. Unfortunately, P cannot be obtained by matrix inversion, as K is close to singular [104, 115]; consequently, the problem is ill-posed. The resultant distributions are spiky, unstable, overly sensitive to the noise in the data, and almost certainly not representative of actual distance values between unpaired electrons in the sample. Their shapes contrast with our expectation of smoothness from distributions of distances between flexible nitroxide probes attached to flexible macromolecules. In short, obtaining distributions using ordinary least squares is not an option.

Thus, the analysis of DEER data must overcome two problems. First, the background component of the signal must be correctly identified, and second, the intramolecular component must be interpreted into distance data without overfitting noise in the signal. Both problems, particularly the latter, are subjects of active research and have been addressed using a wide range of mathematical strategies. Several approaches have been developed over the past two decades and are listed in Table 2.1 and shown in Figure 2.1. For brevity, the following discussion is limited to Tikhonov regularization and model-based fitting, which are perhaps the two most widely used approaches in the literature. We note that for high-quality data containing an unambiguous background coupling component

Table 2.1: Commonly used methods for analyzing DEER data.

Method	**Reference**
Pake transformation	[193]
Tikhonov regularization	[73, 185]
Osher's Bregman iterative regularization	[115]
Integral Mellin Transform	[272]
Neural networks	[448]
Monte Carlo	[102]
Wavelet denoising	[388]
Maximum entropy/Tikhonov	[72]
Sum-of-gaussians model-based fitting	[53, 391]

and a high signal-to-noise ratio (SNR), the distributions reported by these methods are nearly identical. Methodological idiosyncrasies become more pronounced as the fidelity of the experiment signal decreases and the intramolecular coupling component $S(t)$ becomes difficult to isolate.

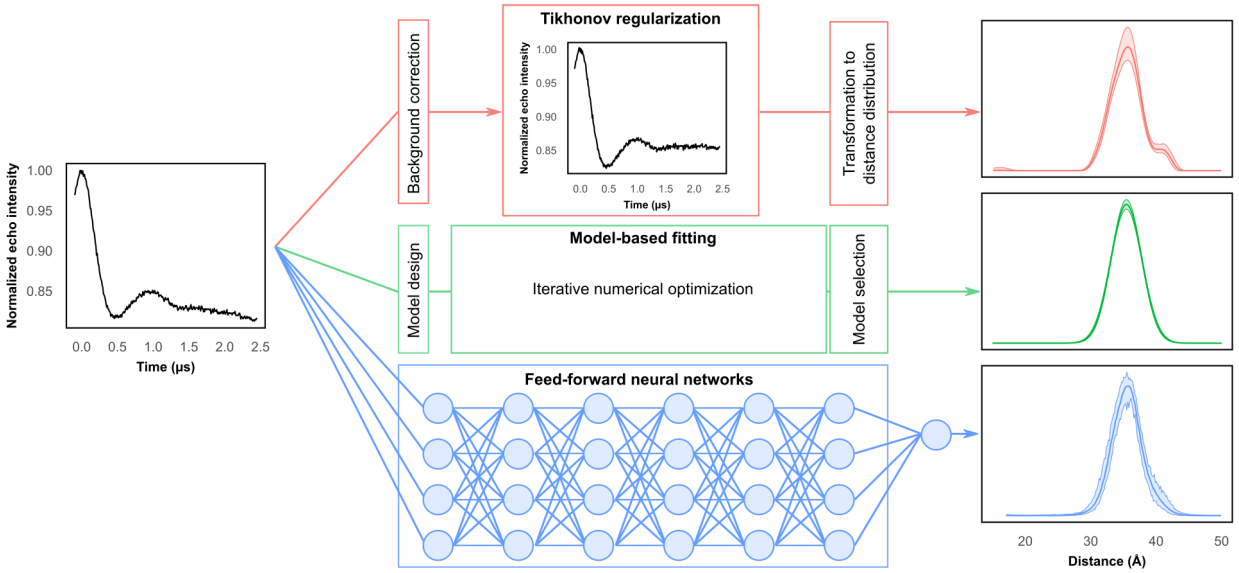

Figure 2.1: Commonly used methods for analyzing DEER data.

2.2.3 Tikhonov regularization

A widely used approach to stabilizing ill-posed linear problems is to include a penalty term, which is denoted by L and is specific for the problem at hand, that quantifies our expectations of what a reasonable solution should look like. Tikhonov regularization provides a general framework for integrating this penalty term into ordinary least-squares problems [413]. The general strategy calculates the optimal probability distribution $\hat{P}$ by minimization of both the sum of squared residuals $\|KP - S\|^2$ and the penalty term $\|LP\|^2$:

$$\hat{P} = \underset{P \geq 0}{\arg\min} \left\{ \|KP - S\|^2 + \alpha^2 \|LP\|^2 \right\} \tag{2.4}$$

Ultimately, the goal of this approach is to generate a smooth distribution with gradual changes in amplitudes between distances fractions of an angstrom apart. For example, it seems unreasonable to expect the peak of a distribution to be at 34.1 Å if the probability at 34.0 Å is zero. An effective

penalty term would therefore restrain the derivative of the distribution, rather than the distribution itself. For this reason, the most widely used matrix penalizes changes in the second derivative [105, 115, 185]:

$$L_2 = \begin{bmatrix} 1 & -2 & 1 & & & 0 \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ 0 & & & 1 & -2 & 1 \end{bmatrix} \tag{2.5}$$

This applies the smoothness constraint uniformly throughout the distribution, thus avoiding the spikiness observed by ordinary least-squares fitting. The weight of this term, relative to that of the least-squares term, is determined by the regularization parameter α, which does not reflect an experimental variable and is only used to stabilize the probability distribution in solution. In general, lower-quality data require more regularization, and thus larger α values. For any given value of α, one of several iterative non-negative least-squares algorithms can be used to obtain an optimal solution [56, 244].

Not surprisingly, determining the most appropriate value for α is critical to the accurate recovery of distance data. It should be high enough to avoid overfitting noise in the time domain, but no higher than necessary to minimize information loss in the distance domain. Prior to 2018, the most common approach for selecting a value for α was to use an L-curve, in which the distribution is calculated using many values, and the resulting logarithm of the least-squares term of each fit is plotted as a function of the logarithm of the penalty term [73, 185]. However, a recent benchmark did not find the widespread practice of selecting the value of α from the "elbow" of this plot to be particularly effective [105]. Instead, the authors proposed using either the Akaike Information Criterion (AIC) [7] or generalized cross-validation [252], and since then the use of the former has become standard practice [117]. A follow-up study by Fábregas-Ibáñez and Jeschke found that iterative regularization methods can further improve the quality of distributions generated this way [115]. Although they proposed several alternative approaches to integrate the penalty matrix into

the cost function, in practice the Tikhonov functional continues to be widely employed.

As evidenced by its widespread adoption among EPR spectroscopists, this strategy provides informative distributions without the problems encountered by purely least-squares fitting. Additionally, and in contrast with the model-based paradigm discussed in the following section, it is guaranteed to return the best possible fit given the choice of the penalty matrix L and the regularization parameter α. However, its ability to report physiologically meaningful distributions is contingent on the universal application of the smoothness constraint throughout the distribution. As a result, Tikhonov is ill-equipped to handle substantial variation within the distribution, such as mixtures of broad and narrow components, particularly in the presence of noise [391]. Sharp components may be smoothed over and broad components may be split. Moreover, artificial peaks may be introduced throughout the distribution to improve the fit, even if there is no physiological basis for their existence [66, 186].

One aspect of this optimization approach for the analysis of DEER data warrants further discussion. The least-squares portion is defined by the matrix-vector multiplication KP, effectively preventing the nonlinear background contribution of the DEER signal from being integrated into this step of the analysis. As a result, the intermolecular coupling signal, modeled by $B(t)$ as described above, must be determined and removed in advance [116, 185]. This is not a problem when the time collection window is sufficiently large that the intramolecular signal eventually decays to zero, such as when measuring distance distributions that are either short or broadly distributed. Less ideal circumstances may prevent the contribution of background coupling from being readily identified, which can lead to the introduction of fitting artifacts in the distance domain. A secondary result of this *a priori* correction is its effect on the apparent SNR near the end of the data collection window, termed the "noise explosion" [116], which can challenge the assumption fundamental to least-squares fitting that noise values are independently and identically distributed. This can be detrimental to the accurate recovery of short-distance components, and in practice this issue is often sidestepped by signal truncation after background correction. In part to address these concerns, an iterative algorithm was developed that alternates between determining the distribution from the

background-corrected data using Tikhonov regularization and refining the background parameter values using nonlinear least-squares fitting [117]. The method promises to overcome many of these challenges, but as of writing has not been widely used because of its novelty.

2.2.4 Model-based fitting

The fragmented analytical pipeline required by Tikhonov regularization contrasts with nonlinear least-squares minimization, a one-step approach that can be achieved using a parametric model [59]. The multi-Gauss model, for example, represents the experimental distribution using one or more Gaussian distributions. These parameters, when combined with the background coupling parameters in Equation 2.1, constitute a parametric model (denoted by ϑ) that can potentially recreate the DEER signal. As such, their values can be optimized by directly comparing the simulated DEER trace $V_{\mathrm{sim}}(\vartheta)$ to the experimental data in the time domain without any background correction:

$$\hat{\vartheta} = \arg\min_{\vartheta} \left\| V_{\mathrm{exp}} - V_{\mathrm{sim}}(\vartheta) \right\|^2 \tag{2.6}$$

Standard nonlinear optimization algorithms can minimize the deviation between the simulated and experimental DEER traces. These include the Levenberg-Marquardt [53, 236, 266, 391], Interior Point [173, 325], and Trust Region Reflective [61, 117, 288] algorithms, as well as approaches such as Hamiltonian [166, 400] and random-walk Monte Carlo [102, 296], Gibbs sampling [104], and particle swarm optimization [173]. Many of the shortcomings discussed above for Tikhonov do not arise with model-based fitting; the background component need not be identified *a priori*, and the use of Gaussian distributions ensures smoothness in the distance domain. Moreover, the only constraint placed upon the model parameters is that the components' amplitudes each exceed zero and total one. This allows the Gaussian mixture model to accurately fit a mix of broad and narrow components that may be challenging for Tikhonov regularization. Finally, the resulting distributions typically lack many of the spurious side peaks and long-distance components that define distributions obtained using Tikhonov and may not be borne out of the signal.

However, these advantages come at the cost that the best solution is no longer guaranteed to be

reached. Whereas linear least-squares problems can be solved analytically, the aforementioned numerical methods required to solve numerical methods may not necessarily converge upon the global minimum. Instead, the solution obtained is the result of iterative improvements of an initial guess provided by the user. Depending on the number of parameters in the model, starting from several unique guesses may be sufficient to identify a reasonable solution [117]. A further consideration is the choice of how many Gaussian components constitute the model, which can be guided by several statistical criteria [7, 395, 424] but must ultimately be chosen by the end user. Collectively, these constraints place a greater burden on the practitioner to avoid the overinterpretation of suboptimal fits.

2.3 Structural interpretations of distance distributions

The intrinsic dynamics and equilibrium states of protein structures come into focus when carrying out multiple orthogonal distance measurements under identical conditions. Proteins may alternate between a discrete number of conformations as part of their function, and the relative proportions of these conformations may be affected by experimental conditions. Whereas an individual DEER distribution might resemble a featureless smear, a series of measurements in the same pair might reveal distinct distance components that ebb and flow under different conditions. This approach allows discrete conformational states to be extracted from distributions that may otherwise be difficult to interpret. With small datasets and/or simple conformational landscapes, unique conformations can be identified by eye alone [22, 99, 213, 261].

However, when underlying components are not obvious, specialized approaches may be warranted. In a study of the sodium-coupled aspartate transporter GltPH, which isomerizes between three conformations, eight restraints were measured under six experimental conditions [143]. These measurements were subsequently fit using three Gaussian components, one for each predicted conformation, under the assumption that the underlying components would retain their means and widths, but change in amplitude. This effort benefited from the determination of several crystal structures in distinct conformations, which provided initial guesses for the distance of each compo-

nent. This allowed each component across different distance distributions to be linked to specific structures, thus revealing the energy landscape of the transporter. Variations of this approach have been used elsewhere. In a study of HIV-1 protease, the collection of multiple experimental distance distributions allowed the effect of various drugs on specific distance components to be quantified [38, 66]. Similarly, weighted ensembles of 5-NT in either open, intermediate, or closed conformations were determined under either apo or ligand-bound conditions using a Monte Carlo reweighting approach [226]. Again, both cases were aided by a large library of determined conformations.

Similar procedures may nonetheless be carried out even without the information provided by high-resolution protein structures. In their study of the Angiotensin receptor, Wingler *et al* measured ten distance restraints across ten experimental conditions and used non-negative matrix factorization (NNMF) to assign specific distance components to four conformations from the entire pool of experimental distance data [446]. Unlike the previous examples, no structures were known *a priori*; in fact, the conformation-specific distances obtained by NNMF were ultimately used for molecular dynamics simulations (see Section 2.6 below).

Each of these cases relied on distance distributions obtained using Tikhonov regularization to make inferences about a protein's structure and dynamics. As was previously mentioned, Tikhonov can at times introduce artifacts and smooth over sparsely populated components in the distribution. Information lost during analysis cannot be recovered by NNMF or Gaussian fitting. One solution is to simultaneously fit the time domain data collected between the same spin-labeled residues under the assumption that certain distance components are conserved across conditions [53]. As with the examples mentioned above, the amplitudes of distinct components will increase and decrease across various conditions, while their means and widths remain fixed. Individual conformations consist of one or more distance components, and their relative energetics can be tracked and monitored under different conditions. The population data obtained this way can reveal details such as the energetics of protein-ligand interactions [79, 90] or conformational interconversion [89, 180, 267, 425], and facilitate the development of kinetic models of protein function [79, 310].

2.4 Integrative modeling of protein structures using DEER data

The previous sections discussed the challenges of transforming one-dimensional data in the time domain to one-dimensional distributions in the distance domain. The remainder of this chapter discusses the more daunting task of converting multiple distance distributions into detailed models of protein structures and ensembles. If properly executed, the integration of these data with modeling can reveal the binding and interaction interfaces of multiple proteins or subunits, the conformational heterogeneity of protein substructures under different conditions, and even the topology or fold of protein structures that have not been previously determined to high resolution.

2.4.1 General principles of integrative modeling using experimental data

Quantitative models aim to explain or justify past observations, and/or predict future observations [167]. Several recent reviews focused on integrative modeling of protein structures [342, 354] have outlined five possible uses for experimental data:

1. **Choosing a representation of the protein structure.** In the context of DEER data, which reports distance distributions between flexible spin labels, atomic-detail information is unavailable. Thus, absent other sources of experimental information, only low-resolution fold-level models are generally feasible.

2. **Scoring candidate structural models.** A model's agreement with experimental data must be quantified for direct comparison to other models.

3. **Constraining the search space.** Given both the complexity of protein structures and the fundamental sparseness of the DEER data, exhaustive searches of the fold space are impossible. This is closely related to the choice of which sampling method to use [273].

4. **Filtering of models after sampling.** Agreement with experimental data may be use to filter models *ex post facto*, which is often necessary if consistency with experimental data cannot be rapidly quantified.

5. **Model validation.** Finally, the soundness of structural models can be verified using experimental data that may be difficult to quantitatively incorporate.

Most recent advancements surrounding the integration of DEER data into modeling pipelines can be categorized as improvements in either scoring or filtering approaches. Nevertheless, all modeling studies must first address how to adequately constrain the search space.

2.4.2 Working with sparse data

Even the smallest proteins contain thousands of atoms, and even the coarsest depictions must account for at least two rotatable bonds per residue. If a model can sample these degrees of freedom without restriction, then the DEER technique's low throughput all but ensures that model parameters outnumber distance restraints (Figure 2.2). The problem can be simplified if the structures of certain regions are known and forced to remain static: rigid body docking of N discrete domains has $6(N-1)$ degrees of freedom [100], or $4(N-1)$ for symmetric homooligomers [164]. Indeed, countless examples exist in the literature of DEER data being used to identify either a docking interface or the relative spatial arrangements of multiple domains. As an added benefit, the search space of the problems shrinks to the point that it may be searched nearly exhaustively [14, 219, 398].

If intradomain movement cannot be ruled out, then soft restraints can be used to limit the extent to which a model reconfigures away from a known conformation. For example, Evans *et al.* modeled a conformational change in SthK by restraining C_α atoms to distances observed in the starting conformation using a sigmoid-like function that increased the penalty of deviations up to but not beyond 1 Å [114]. This limited the changes introduced to the model to regions that were inconsistent with the experimental data. Alternatively, the source of the conformational change can be determined by eye. Kazmier *et al.* found that conformational changes in the transporter LeuT predominately map to four transmembrane helices. This allowed the remainder of the protein structure to be restrained, preventing unnecessary movement [213]. Finally, such restraints can be introduced implicitly, such as by starting from a template [35, 114].

Without prior structural information, DEER distance restraints play a supplementary role to

either energy functions and/or alternative sources of experimental data such as NMR restraints. Their integration is discussed in Section 2.5.5.

2.5 Evaluating a model's agreement with experimental DEER data

2.5.1 Simulation of nitroxide side chains

Accurately modeling a protein structure is only possible when the target conformation recapitulates the experimental data more effectively than alternative conformers. The reliance on flexible spin labels complicates the problem of identifying the conformer of interest using DEER data. Distances are measured between stably bound unpaired electrons far from the protein backbone; for example, in the widely used methanethiosulfonate spin label (MTSSL), they are separated from the C_β atom by 6 Å and five χ angles. Ultimately the linkers anchoring these unpaired electrons to the protein backbone determine how distances between unpaired electrons relate to distances between backbone atoms. Their conformations are generally not known in advance (although they can be determined using specialized algorithms, discussed in Section 2.6 below). Therefore, accurate models of protein structures require accurate distance distributions, which in turn require accurate predictions of spin labels rotamers.

However, the introduction of spin labels into the protein model is not straightforward. For example, several problems become apparent even in simple cases where protein structures being simulated in an MD environment are modified to have nitroxides at residues that have been experimentally labeled [43, 49, 57, 83, 157, 175, 176, 262, 263, 343]. First, whereas the two spin labels attached to any individual experimentally expressed double-cysteine mutant are unlikely to clash, an *in silico* model attempting to integrate data from many DEER experiments may contain many explicitly modeled nitroxide side chains, some of which may be in close proximity to one another [176]. Second, the identity of the residue being labeled can provide valuable modeling information about which environments it preferentially occupies [8]. For example, mutation of a tryptophan in a membrane protein to a nitroxide prevents it from informing the simulation software about structural characteristics such as membrane depth. Finally, and perhaps most critically, the compu-

tational cost of atomic-detail molecular dynamics simulations precludes its use as a screening tool that quantifies any given structural model's agreement with DEER data.

Monte Carlo (MC) modeling paradigms provide one alternative approach. In this modeling paradigm, the spin label configurations are randomly picked from rotamer libraries, the probabilities of which are computed in advance. However, whereas rotamer statistics for canonical amino acids can be obtained directly from the PDB, those for nitroxide spin labels must instead be computed from detailed quantum mechanical calculations, a procedure described in detail elsewhere [10, 103, 322, 357, 464]. After obtaining these rotamers, MC facilitates the rapid sampling a residue's local environment, and pairwise measurements can then be compared to experimental distances. The accuracy of distances obtained this way is comparable to those obtained by traditional MD [220] and in many cases are able to recover the rotamers observed experimentally [10]. However, MC methods have the advantage of being several orders of magnitude faster.

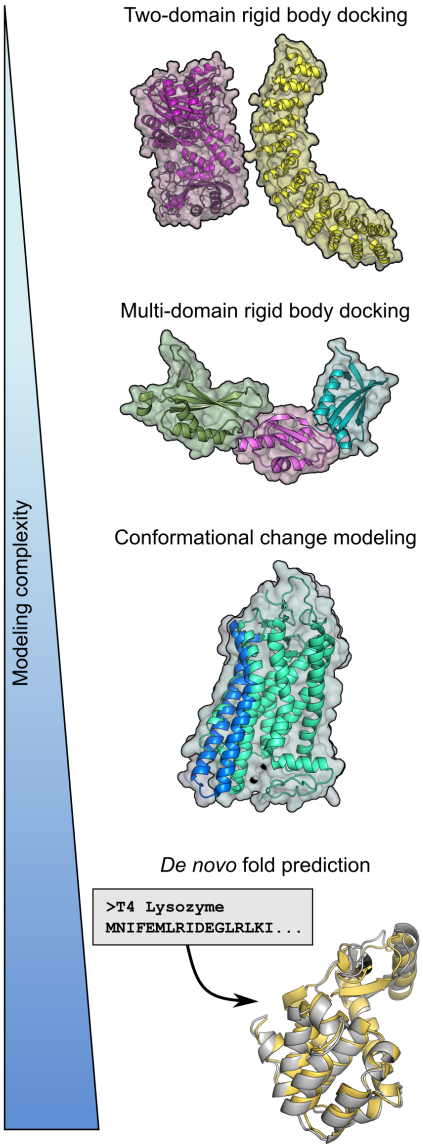

Figure 2.2: Protein modeling applications ranked by complexity. More complex problems, such as fold prediction or homology modeling, generally require more restraints to obtain accurate.

MDDS, an alternative developed by Islam and Roux, allows MD methods to rapidly calculate distance distributions from protein models in minutes to hours while avoid modeling spin labels to atomic detail [176]. By using a set of over fifty experimental distance restraints collected in T4 Lysozyme, they generalized the position of the unpaired electron with respect to the protein backbone. They then converted these positions into force fields between backbone atoms and a dummy

atom representing the unpaired electron, allowing them to discard the rest of the nitroxide side chain.

2.5.2 Simulation of DEER distance distributions

Whereas experimental distance measurements are carried out between ensembles of molecules, computational methods generally model only a single structure at a time. To model the distribution of distance measured in ensembles of molecules, MDDS uses large numbers of noninteracting dummy atoms to build distance distributions [213, 332]. Marinelli and Faraldo-Gomez instead used a time window to build their distributions from MD simulations between explicit nitroxide rotamers [173, 262, 263]. Alexander *et al.* generated thousands of models using full-atom MC modeling and assembled distributions from those with the lowest energy [10].

By far the most common strategy in the literature models ensembles of possible nitroxide conformations. Methods such as MMM [189, 322], Pronox [156], Nasnox [24, 327, 403] and TagDock [103] introduce every possible configuration stored in of a rotamer library one at a time and calculate Boltzmann energy values for those that do not clash with the rest of the protein. Some methods, such as TagDock, allow neighboring side chains to be repacked in response [103]; others, such as MMM, can be applied to individual frames within an MD simulation to visualize the contribution of backbone dynamics [392, 408]. By contrast, MtsslWizard [152, 153] uses MC to sample rotamers until the number of either accepted rotamers or clashes reaches a predetermined threshold. A benchmark by Klose *et al.* established that distributions simulated using these rotamer libraries are comparable in accuracy to those simulated using MD, with the distribution's average values deviating by about 3.0 Å from those observed experimental in each case [220].

However, neither this approach, nor to our knowledge MD, can simulate distributions with widths that correlate with experimental values [187], with the former consistently overstating spin label dynamics while neglecting backbone dynamics [88]. The only method that has achieved any correlation with experimental widths is a sampling-intensive MC refinement approach that permits changes in the protein backbone [10], which suggests that coupling of backbone and side chain

dynamics in solution may contribute to the width and shape of experimental distance distributions. Such a case has been crystallographically observed in the chaperone Spa15 when a loop containing a spin-labeled residue was found to slightly reconfigure compared to the wildtype structure [242].

One explanation for this discrepancy between experimental and simulated distribution widths is the fact that DEER measurements are almost always carried out in protein samples following the addition of cryoprotectants (such as glycerol) and flash-freezing. The common practice of submerging microliter-volume samples in liquid nitrogen gives spin labels up to hundreds of milliseconds to reconfigure and converge upon low-energy conformers [359]. Studies with T4 Lysozyme [144] and hemoglobin [20] have observed that the speed at which a sample freezes affects the width of the resulting experimental distribution. It is unclear if this effect could be corroborated *in silico*, since the timescales of such a simulation are not currently feasible. Most likely, the experimental distribution likely involves a complex interplay between the spin label's local environment, the sample conditions and choices involved in its preparation, and the speed at which the sample was frozen.

2.5.3 Scoring functions

How can the target conformation be identified if it cannot be expected to recreate the experimental data? Unfortunately, no consensus exists on this topic in the literature. Average values of distributions are widely used for the simple reason that their values correlate reasonably well with computational predictions made from structural models. As was just discussed, distribution widths cannot be recapitulated and are frequently ignored altogether, while distribution shapes are far more difficult to glean from the raw data and as a consequence rarely factor into modeling.

One strategy that has seen relatively widespread use is to model the nitroxide rotamers using a rotamer library such as MMM and to measure distances from the unpaired electron closest to the ensemble average [88, 99, 100, 101, 124, 331, 436]. Comparing these distance values to experimental averages allows a model to be scored using the sum of squared residuals. This retains the benefits of using the entire ensemble of spin label conformations while removing potentially unreliably information from the distance distribution. Nevertheless, alternative scoring functions sometimes are

used that factor the width of the distribution. For example, in several studies, the deviation between the simulated and experimental average values has been divided by the experimental distribution's width. Presumably, this accounts for backbone heterogeneity by less aggressively penalizing deviations between residues with wider distributions [188, 317], although to our knowledge no benchmarks have been carried out to ascertain the merit of this fact. Alternatively, the entire shape can be retained, and the score of a model can be either its percent overlap [157, 191, 213, 332] or the area between the integrals of the two distributions [226].

It is important to note that regardless of the scoring function, the spin label's flexibility means that the direct application distance restraints to the unpaired electron or nitroxide bond to be largely ineffective at obtaining high-precision protein models, as the rotamers can easily reconfigure without any backbone changes. For example, when dummy atoms are used to model conformational changes modeled using MDDS, they have been found to absorb changes in the distance distribution, leaving the backbone untouched [213, 332]. Applying distance restraints to the nitroxide bond of the rotamers in an MC simulation is fraught with similar challenges: for example, Herrick *et al.* constrained the first two chi angles of MTSSL after finding that rotamers would otherwise adopt "unrealistic orientations" [161, 228]. Similar observations were made when modeling 5-NT [226] and Omp85 [88]. An exception is when data has also been collected using paramagnetic relaxation enhancement (PRE) paramagnetic relaxation enhancement NMR, which measures distances between an unpaired electron and backbone nuclei. These distance data can be used as additional restraints on nitroxide rotamers to prevent them from reconfiguring [285, 350, 450]. However, the rotamers that contribute to the experimental signals may differ between the two datasets, since DEER samples are measured after flash-freezing whereas PRE samples are measured at room temperature.

2.5.4 Limits of modeling protein structures using explicitly modeled rotamers and rotamer libraries

Distance distributions in non-MD environments can typically be simulated in seconds or tenths of a second, with the majority of computation time typically devoted to calculating clashes with the rest of the protein. Unfortunately, even these speeds still preclude the use many modeling applications. Most studies therefore use these methods to screen models *ex post facto*, rather than restrain them directly during sampling. Some studies circumvent these cost restrictions by introducing rotamers prior to modeling and retaining them throughout sampling without recalculating clashes, thus skipping the most computationally expensive step. Such a strategy was appropriate when docking CDB3 and Ankyrin [103], as the local environment of the spin labels were not expected to change throughout the docking simulation. Alternatively, a number of studies only retain the rotamer closest to the center of the ensemble: by fixing it in space relative to the backbone, it could be restrained using distance data without risking its reconfiguration [31, 99, 101, 331, 436].

2.5.5 Direct integration of DEER data during sampling using restraints between backbone atoms

Because of this computational cost issue, these data are more frequently integrated as restraints between backbone atoms during sampling. Restraints between C_β atoms are more common, although C_α atoms have also been used [308, 352]. While trivial to implement in most modeling packages, this approach requires that distance values between spin labels be transformed into distance restraints between backbone atoms. Unfortunately, these measurements do not always line up perfectly [165]. Yang *et al.* demonstrated the danger of taking these experimental distances at face value when predicting the structure of the homodimer Dsy0195 [463]. The RMSD of their model improved when using a restraint with an experimental DEER distance that matched the C_β-C_β distance, but worsened when the two values differed by as little as 5 Å. As has been previously noted [8, 30, 42, 128, 143, 187, 353], deviations of this magnitude are not uncommon. In globular proteins, spin labels tend to face away from each other, leading to experimental distances values

that exceed the corresponding C_β-C_β distance by a median of 6 Å to 7 Å. By contrast, the deviations between these values across protein-protein interfaces are lower but nonetheless significant [30, 128, 219]. Notably, these values exceed the 3 Å deviation observed between the average values of simulated and experimental DEER distance distributions, highlighting the loss of precision that comes with this scoring approach [353].

Backbone potentials frequently score a model's agreement with experimental data using wide and flat-bottomed harmonic potentials. These have be introduced as NOE-like restraints into programs such as CYANA [150], Xplor-NIH [363], CNS [58], Rosetta [229, 234, 378], and BCL [205, 447]. Both Alexander *et al.* and MacCallum *et al.* successfully predicted the structures of T4 Lysozyme and aA-crystallin *de novo* using this type of restraint [8, 256]. In both cases, a score of zero was assigned to C_β-C_β distances (in angstroms) ranging from $\mu_{SL} - \sigma_{SL} - 12.5$ to $\mu_{SL} + \sigma_{SL} + 2.5$, where μ_{SL} and σ_{SL} are the average values and standard deviations of the experimental distribution. Similarly, Kim *et al.* determined the docking interface of CDB3 and Ankyrin using a criterion in which any model was kept if its C_β-C_β distances deviated by less than 14 Å from the experimental average distance [219]. Alternatively, the edges of the distribution have been used as restraints bounds [31].

These scoring functions, although broad, can nonetheless respond poorly to outliers. Bhatnagar *et al.* noted when docking CheA and CheW that a pair of C_β-C_β restraints whose distance deviated substantially from experimental interspin distance values prevented low-RMSD models from being obtained; their removal improved the RMSD of the docking interface from over 10 Å to 2.6 Å [30]. However, they used a more restrictive potential with lower and upper bounds of $\mu_{SL}-5.0$ Å and $\mu_{SL}+1.0$ Å, respectively, which may increase the scoring function's sensitivity to outliers [29].

A similar strategy was employed by Evans *et al.* when modeling conformational changes in the ion channel SthK [114]. They took advantage of a starting conformation, as well as a set of DEER data for that conformation, by applying the magnitude of the distance change as a harmonic restraint between C_β atoms with a lower and upper bound of $\mu_{SL}-1.0$ Å and $\mu_{SL}-1.0$ Å, respectively. This strategy has the benefit of reducing variation among generated models but assumes that domains

and spin labels do not rotate or substantially reconfigure with respect to one another.

Hirst *et al.* generated a custom cubic spline function called the motion-on-a-cone (CONE) model that, unlike a flat-bottomed harmonic function, accounts for the fact that the experimental distance is more likely, but not necessarily guaranteed, to be longer than the C_β-C_β distance [165]. Its use improved the number of high-resolution models folded *de novo* for T4 Lysozyme, and it has since been applied to problems involving *de novo* folding [126, 127, 128], docking [9, 88, 128, 410], and conformational change modeling [88, 217, 226, 422].

None of these approaches can determine unequivocally the structures of proteins or complexes. For example, Kim *et al.* found that the incorporation of twenty experimental restraints when docking CDB3 and Ankyrin led to a wide range of possible configurations [219]. This highlights the breadth of the solution space, even in cases where the experimental restraints outnumber the degrees of freedom by more than threefold. Ultimately this pool of models was further trimmed using an orthogonal set of EPR data that measured the solvent accessibility of individual spin-labeled residues.

This hierarchical approach contrasts with the direct integration of flat-bottomed harmonic restraints into sampling, which has the intended effect of directing the modeling protocol to a conformational subspace encompassing the target structure. Optimization within this subspace proceeds using additional criteria, such as energy functions or experimental NMR restraints, which can evaluate structural features with greater precision. The MELD structure prediction program, for example, navigates this subspace using AMBER forcefields and replica exchange molecular dynamics [256, 313], whereas Rosetta relies on its coarse-grained energy function and MC sampling [8, 210]. By contrast, Ling *et al.* used Xplor-NIH to model the structure of YagP by combining DEER data with EPR membrane depth restraints [245]. In each case, the flat-bottomed potentials do not interfere with optimization within this region; instead, they guide the conformational search away from false minima that fall outside their bounds. As a result, the weights of these restraints, relative to other experimental restraints or score terms, is not generally optimized or fine-tuned.

By contrast, the CONE model developed by Hirst *et al.* does introduce a bias into the search

procedure by assuming that the experimental distance will usually, but not always, exceed the C_β distance. This assumption modifies the function landscape and must therefore be carefully integrated into the modeling protocol. During the development of this energy potential, Hirst *et al.* determined by grid search a weight for this energy potential that maximized the proportion of high-accuracy models of T4 Lysozyme. This weight has since been used in virtually all cases outlined above.

We note that although these restraints are imprecise, they are often followed by the use of higher-precision scoring functions in conjunction with atomic-detail depictions of spin label rotamers more amenable to the identification of accurate models [355]. For example, structural modeling of EmrE was achieved using the CONE model in BCL::Fold followed by refinement with Modeller using the rotamer libraries available in MMM [88]. The structure of the C-terminus of ExoU was similarly modeled *de novo* using the CONE model, followed by *in silico* spin labeling with explicit MTSSL side chains using Rosetta [127]. The CheA/CheW binding interface was determined by generating a set of initial models using C_β-C_β restraints followed by finer-grain selection using rotamer libraries [30]. Thus, the use of coarse backbone restraints can set the stage for finer-grain modeling.

2.5.6 Error analysis

Our discussion has extensively mentioned the limit of DEER restraints in modeling structures: even docking interfaces cannot be determined with absolute certainty using these data. Fortunately, quantification of the uncertainty of these models is increasingly widely used. Cross-validation requires that multiple subsets of these restraints be used for modeling, and the results compared to check the stability of the solutions. This approach has been used to model the NhaA homodimer [164], 5-NT [226], and RmsE/RmsZ [30]. The docking interface of the NhaA homodimer, for example, was determined using each of the 36 available combinations of seven restraints from nine restraints available, leading to an RMSD of 0.45 Å among each of the thirty-six best models from each set [164]. Bowen and colleagues took this idea a step further and quantified the uncertainty resulting from the choice of rotamer library [48]. Alternatively, if a starting structure is available,

the error calculated for that structure can be used to inform the precision with which certain features and/or substructures of a protein can be modeled with high confidence. Bliecken *et al.* calculated the deviations among restraints in the soluble Bax monomer, which had been initially determined using NMR, and used those values to inform the accuracy of their model for the membrane-bound Bax dimer [39]. More commonly, uncertainty is depicted informally, for example by presenting a handful of the best-scoring models [114, 146, 219]. Far more often it is ignored altogether.

2.5.7 Integrating DEER restraints with other types of experimental data

Flat-bottomed harmonic functions, although imprecise, nonetheless allow DEER data to be elegantly integrated into a modeling protocol involving other types of data. When modeling distributions using programs such as MMM, it is less clear how best to jointly consider agreement with both DEER data and, for example, small-angle X-ray scattering (SAXS) data. One option mentioned above is to use hierarchical approaches, where models are effectively filtered out if they fail to satisfy each of several experimental criteria. Bowman, Boura, and Sundaramoorthy combined DEER and SAXS data using this approach for rigid-body modeling of Vps75/Nap1, ECSRT-I, and Chd1, respectively [46, 47, 48, 398]. In each case, models needed to be docked in such a way that both the DEER data and SAXS density were satisfied. Several examples in the literature exist in which this is done informally, such that a solution obtained using DEER data is simply validated using experimental data from another source. Hilger *et al.* compared their model for the dimeric NhaA antiporter to previously published low-resolution 2-D cryo-electron microscopy data [164], whereas Sung *et al.* docked their model of Bax into SAXS density [399].

A preferable approach is the direct integration of the two during sampling. This, however, requires that scores be balanced in such a way that avoid overemphasizing data from either method. To our knowledge, no single widely used approach exists. In the literature, ESCRT-II was modeled using both SAXS and DEER data using a χ^2 potential for each experimental technique [46]. Peter *et al.* also used χ^2 potentials to model YopO using both SAXS and DEER, but simply compared the average values of the simulated and experimental distributions [317].

2.5.8 When simulation guides experiment: choosing restraints using starting structures

Several application-specific algorithms have been developed that recommend experimental restraints. These recommendations are typically application-specific, as the demands of each application change the information being sought by the experimental data. As an example, *de novo* folding problems benefit from restraints between residues that are distant in sequence but close in space. The MC algorithm developed by Kazmier *et al.* that chooses restraints for *de novo* folding uses two criteria: it tries to ensure that the pairs of residues are far apart in sequence space while diversifying their placements to avoid collecting redundant information, for example between the same pair of helices [210]. By contrast, when a structure is known but its loops are not, networks of four restraints per spin-labeled loop residue have been proposed in a procedure analogous to triangulation [188]. For conformational change problems, both Hays *et al.* [158] and Mittal and Shukla [290] developed restraint-picking methods for conformational change modeling problems that prioritize the reduction of redundant information by selecting residue pairs whose movements are minimally correlated; both rely on short nanosecond MD simulations to obtain an initial guess for these structural dynamics. If for some reason such a simulation is impossible, an alternative method [184] predicts which regions of a protein structure will move using normal mode analysis and a modification of the Zheng-Brooks algorithm [477]. The same author proposed a separate criterion for rigid body docking problems, which require far fewer restraints to define analytically [191]. Since only three residues need to be spin labeled per rigid body, an intuitive choice is to choose the three residues in each rigid body that generate the largest nearly-equilateral triangle. This minimizes the propagation of errors throughout the docked structure.

2.6 Towards the analysis of DEER data by structural modeling

The first half of this chapter discussed the difficulty of extracting distance data from DEER traces in the time domain, while the second half discussed how best to use those distance data for modeling protein structures. In principle, the two steps can be directly integrated by attempting the simulate the raw data using the same approach summarized earlier. Briefly, the conformation determines

the distance distributions being fitted and reconfigures to improve the goodness-of-fit in the time domain. Several studies have attempted to recapitulate the raw data from protein models as early as 2007 [30, 164]. The previously mentioned ESCRT-II, which was modeled using the raw DEER traces alongside SAXS data using a χ^2 potential to balance the two [46, 47]. Marinelli and Fiorin modeled a conformational change in VcSiaP by restraining the spin labels directly using the raw data [263]. Notably, both the study of VcSiaP and NhaA also attempted to fit the background contribution to the signal, which was added to the simulated data using a "fit-within-a-fit" procedure. By contrast, the study of ESCRT-II, as with others [336], background-corrected the data before using it as a restraint.

Additionally, we mentioned in this chapter that native or correctly folded models are not expected to perfectly recapitulate the experimental DEER distance distribution. Failure to account for this expectation can cause the data to be overfitted. In fact, whereas this expectation can easily be encoded into the scoring function when scoring models using data in the distance domain - for example, using broad, flat-bottomed functions (see Section 2.5.3 above) - it is far more difficult to anticipate the magnitude of these deviations in the time domain. This imprecision prevents minor contributions to the experimental signal from being resolved, removing one of the fundamental advantages of the DEER technique.

How can this issue be addressed? One might intuit that less flexible spin labels, the positions of which can more precisely be simulated from the backbone, may be positioned to improve the precision of simulated distributions enough to permit modeling using data in the time domain. Several options are explored in the literature. The label IDSL (also called V1 or RSSR), for example, has far less rotameric freedom than MTSSL and leads to sharper distance distributions [19, 437]. Other alternatives that have been used for modeling include bifunctional spin labels [119, 175, 347] and paramagnetic copper ions [86, 282] that are conjugated to or coordinated by pairs of alpha carbons. By reducing or altogether eliminating the contribution of rotamer dynamics to the width of the distribution, these methods can potentially isolate the contribution of backbone dynamics. Moreover, it avoids the risk of uncritical accepting the width of the rotamer distribution obtained

from methods such as MMM, which may cause backbone dynamics to be understated (see Section 2.5.1).

A second option to reduce uncertainty is to determine the positions of flexible spin labels in advance. This triangulation procedure, however, requires a conformation of the protein to be known *a priori* that is consistent with a set of experimental data. The positions of the spin labels can be determined by singular value decomposition [152], non-linear least squares minimization [39], or by eye [446]. An instructive example is provided in the study of the Angiotensin receptor, where distance distributions attributed to a known conformation were used to determine which rotamers were sampled in solution[446]. The distance data from other conformations were then used for conformational change modeling using MD. Similarly, the model of dimeric membrane-embedded Bax was obtained by first determining the MTSSL rotamers from the monomeric model [39]. However, the constraints placed upon rotamer triangulation - that a conformation is both known to high resolution and consistent with a set of experimental data - make it uncommon, and this technique has not yet been used to model proteins using data in the time domain. More often, it is used to localize a paramagnetic ligand or ion within the same structure [140, 465].

What would the key benefit be of working with the raw data? To answer this question, it may be prudent to take a global perspective on the state of the art, as outlined in Section 2.2 above. DEER data are commonly background corrected, then analyzed; DEER distance distributions are partitioned among several conformations, with the average peak of each component frequently isolated for the purposes of modeling. As discussed above, information lost during each of these four steps cannot later be recovered. It may, however, be accessible to one-step approaches that simulating the raw data from an ensemble of models. It is in precisely this direction that we hope the field moves.

CHAPTER 3

Rapid simulation of unprocessed DEER decay data for protein fold prediction

The contents of this chapter have been previously published [97].

Despite advances in sampling and scoring strategies, Monte Carlo modeling methods still struggle to accurately predict *de novo* the structures of large proteins, membrane proteins, or proteins of complex topologies. Previous approaches have addressed these shortcomings by leveraging sparse distance data gathered using site-directed spin labeling (SDSL) and EPR spectroscopy to improve protein structure prediction and refinement outcomes. However, existing computational implementations entail compromises between coarse-grained models of the spin label that lower the resolution and explicit models that lead to resource-intense simulations. These methods are further limited by their reliance on distance distributions, which are calculated from a primary refocused echo decay signal and contain uncertainties that may require manual refinement. Here, we addressed these challenges by developing RosettaDEER, a scoring method within the Rosetta software suite capable of simulating DEER distance distributions and decay traces between spin labels fast enough to fold proteins *de novo*. We demonstrate that the accuracy of resulting distance distributions match or exceed those generated by more computationally intensive methods. Moreover, decay traces generated from these distributions recapitulate intermolecular background coupling parameters, even when the time window of EPR data collection is truncated. As a result, RosettaDEER can discriminate between poorly folded and native-like models using decay traces that cannot be accurately converted into distance distributions using regularized fitting approaches. Finally, using two challenging test cases, we demonstrate that RosettaDEER leverages these experimental data for protein fold prediction more effectively than previous methods. These benchmarking results confirm that RosettaDEER can effectively leverage sparse experimental data for a wide array of modeling applications built into the Rosetta software suite.

3.1 Introduction

Structural biology increasingly relies on integrated methods to model the structure and dynamics of proteins and protein assemblies [393, 451]. Multiple complementary experimental methodologies can describe the structure and dynamics of proteins that elude determination from a single technique, such as integral membrane proteins, conformationally flexible proteins, and those that fall outside the size limitations of solution-state nuclear magnetic resonance and cryo-EM. By integrating experimental data from multiple approaches, computational modeling can build accurate models in regions with sparse experimental data. One promising source of high-resolution experimental data for integrated structural biology combines SDSL and EPR [190, 348]. Previous studies have employed SDSL-EPR and computation in tandem to predict protein structures *de novo* [8, 126, 127, 128, 165, 210, 245, 463], model conformational changes [212, 213, 263, 332], and dock rigid-bodies [30, 103, 164].

Existing modeling methods largely focus on data gathered using four-pulse DEER [304], which can report on distances of up to 60 Å to 80 Å between stable unpaired electrons conjugated to the protein backbone by SDSL [186, 274]. However, incorporation of these distances as interatomic restraints for modeling purposes is confounded by the conformational freedom of these paramagnetic probes. The central challenge is to convert inter-spin distance information into structural restraints that report on the protein backbone [2, 9, 177]. Additionally, the need to incorporate two spin labels into the protein sequence per restraint results in sparse coverage of the experimental data that can introduce ambiguities into computational modeling [210]. As a result, only a few experimental restraints are generally available to describe the protein fold.

These sparse datasets have nonetheless been leveraged for protein structure prediction and refinement by a range of computational modeling approaches that represent the spin labels either implicitly or explicitly. Implicit models such as the CONE model [8] use knowledge-based potentials to translate inter-spin distance values into backbone restraints, typically between C_β atoms [353]. Introducing these restraints led to measurable improvements in *de novo* structure prediction benchmarks by programs employing Monte Carlo sampling strategies [8, 126, 127, 128, 165, 210],

gradient minimization [245, 463], and molecular dynamics [256]. However, because these potentials fail to account for the environment or the relative orientations of the spin labels, they tend to be ambiguous [353]. Explicit methods, by contrast, model spin labels as either individual side chains [10, 88, 226, 263, 262], ensembles of side chains [153, 156, 164, 322], or ensembles of dummy atoms [176, 212, 213]. The added detail improves accuracy of modeling but makes implementations too computationally intensive for *de novo* protein structure prediction and limits the utility of these methods to modeling small-scale conformational changes [212, 213, 263].

Despite their diversity, these methods largely share a common limitation in their reliance on distance distributions, rather than the primary spectroscopic readout. Other computational methodologies directly incorporate primary experimental data, such as two-dimensional NMR spectr [277] and cryo-EM electron density maps [433] to fold and refine proteins. The feasibility of using DEER dipolar coupling decay traces as modeling restraints has only recently been explored [263]. Whereas processing spectroscopic decay traces into distance distributions risks introducing ambiguities and artifacts [53, 173, 192, 365, 448], simulating a decay trace from a distance distribution is well-described and mathematically straightforward [169, 173, 186, 263].

Here we introduce RosettaDEER, a method in the macromolecular modeling suite Rosetta capable of rapidly simulating distance distributions and DEER decay traces between spin labels as well as evaluating a model's agreement with experimental data. RosettaDEER's computational efficiency enables prediction of protein structures *de novo* with greater accuracy than the default energy function or the CONE model. Owing to Rosetta's Monte Carlo sampling strategy [229], the experimental data can be used directly without analysis or background-correction. Thus, as with other forward modeling approaches[391], the quality of the primary spectroscopic data can be significantly poorer than what would ordinarily be required for rigorous transformation into distance distributions using common fitting strategies. This method reinforces the utility of DEER in conjunction with computational modeling to accurately model protein structures.

3.2 Materials and Methods

3.2.1 Assembly of diverse experimental datasets

RosettaDEER was implemented in the Rosetta software suite [229, 234], trained on distance data gathered in T4 Lysozyme obtained from the laboratory of Hassane S. Mchaourab, and tested and cross-validated using both raw spectroscopic and analyzed distance data gathered in five laboratories (Table 3.1). Data for the ExoU C-terminus [127], Bax [39], and Mhp1 [213] were obtained from and analyzed by the laboratories of Dr. Jimmy Feix, Dr. Enrica Bordignon, and Dr. Hassane S. Mchaourab, respectively. New ExoU double-cysteine mutants were purified, spin labeled, measured and analyzed as previously described (Figure E.1) [127]. Raw data for CDB3 [479] and bovine rhodopsin [15] were obtained from the laboratories of Dr. Albert Beth and Dr. Wayne Hubbell, respectively, and were analyzed using DEERAnalysis2016 [192]; the last 200 ns and 500 ns were removed from experimental decay traces shorter and longer than 1.5 µs, respectively. The distribution of restraints is shown in Figure E.2.

3.2.2 Generation of DEER distance distributions

The accuracy of various methods that simulate distance distributions between spin labels were compared using Bax (PDB: 1F16, NMR state 8), ExoU (PDB: 3TU3), CDB3 (PDB: 1HYN chains R/S), Rhodopsin (PDB: 1GZM chain A), and Mhp1 (PDB: 2JLN). The methods compared were MMM [322], MDDS [176], MtsslWizard [153], Pronox [156], and TagDock [103] (See Figure 3.1 and Table 3.1). MMM2017 was run locally on both cryogenic 175 K and ambient mode 298 K with default settings. MDDS was run using the CHARMM-GUI web server [195]. MtsslWizard was run locally from PyMol 1.7.2.1 using tight fitting unless no rotamers could be placed, in which case loose fitting was used (Mhp1 residue 324 could not be labeled using loose fitting). Pronox was run from the USC web server using a bias of 0.9 and a van der Waals radius scaling factor of 0.75, the latter of which was reduced to 0.4 if rotamers could not be placed. TagDock was run locally with SCWRL4 [225] and a bump radius of 0.85. Measurements using the CONE model [8, 165] were determined by adding 1.79 Å to the C_β-C_β distance.

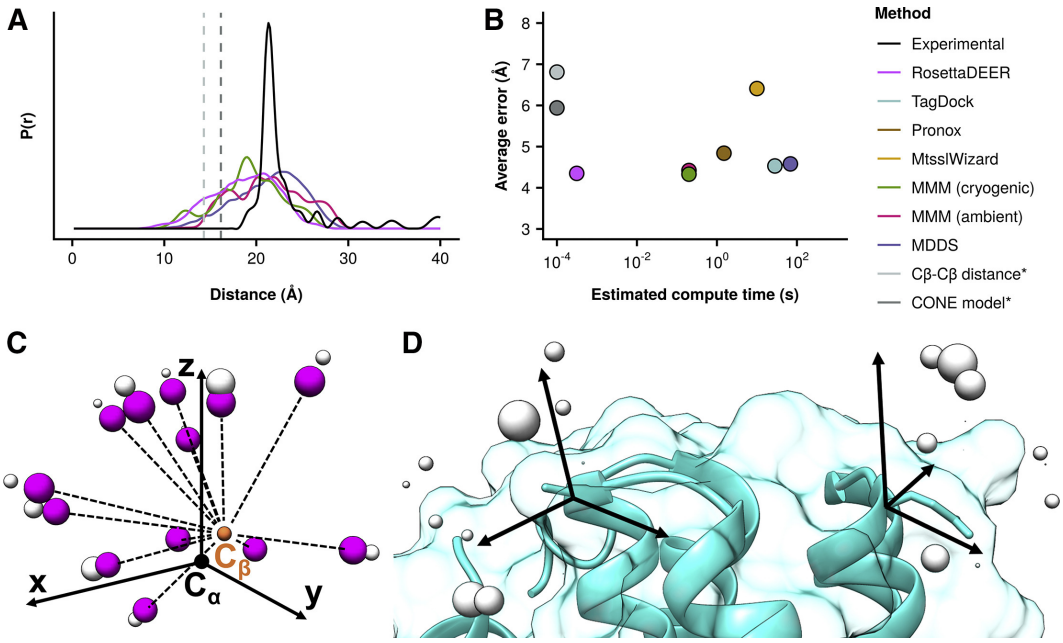

Figure 3.1: Simulations of distance distributions between nitroxide probes using RosettaDEER. (A) An example of an experimentally observed distance distribution in apo Mhp1 51/278, shown in black. Distance distributions were simulated using RosettaDEER, MMM, and MDDS from the occluded Mhp1 structure (PDB: 2JLN). The average distance between C_β atoms and the average distance calculated using the CONE model shown in light gray and dark gray, respectively. (B) The estimated average time required to simulate distance distributions (*the lower limit of quantification exceeded the C_β-C_β distance compute time). (C) Coarse-grained rotameric ensemble representation of the MTSSL. Centers of mass, shown in purple, are used for clash evaluation, whereas electron coordinates, shown in gray, serve as measurement coordinates. (D) Distance distributions between residues are simulated by superimposing coordinates, evaluating clashes and measuring all resulting pairwise distances.

3.2.3 RosettaDEER method description

The Rosetta rotamer library for the paramagnetic probe MTSSL [10] served as the basis for the coarse-grained rotameric ensemble used in this study. For each of fifty-four possible rotameric configurations, the unpaired electron was assumed to occupy the nitroxide bond midpoint; it was from these coordinates that distances would be measured. These coordinates were consolidated into a common frame defined by the C_α atom at the origin, the backbone carbonyl carbon along the Z-axis, and the backbone nitrogen in the Y-Z plane (Figure 3.1.C). The remainder of each rotamer was represented by a single pseudo-atom with a radius of 2.4 Å that was placed at 87.5% of the distance between each nitroxide bond midpoint coordinate and an idealized C_β coordinate; if this

pseudo-atom clashed with the protein model, its corresponding electron coordinate was not used for distance measurements. The placement of this pseudo-atom coincides with the center of mass of the nitroxide ring of MTSSL (all spin-labeled proteins in the PDB are listed in Table E.1 and Figure E.3). In cases where the steric environment prevented the placement of any coordinates, the van der Waals radius of this pseudo-atom was gradually lowered until at least one rotamer could be accommodated. Distance distributions between two residues reflect all pairwise distance measurements between their respective coordinates after evaluating clashes; we smoothed each of these distance values into gaussian distributions with a 0.5 Å standard deviation. The resulting distance distributions were then binned to 0.5 Å.

Table 3.1: Benchmark set for the evaluation of RosettaDEER.

Protein	**Organism**	**Restraints**	**PDB ID**	**Reference**
Bax	*Homo sapiens*	21	PDB: 1F16 model 8	[39]
ExoU	*Pseudomonas aeruginosa*	11	PDB: 3TU3	[127]
CDB3	*Homo sapiens*	15	PDB: 1HYN R/S	[479]
Rhodopsin	*Bos taurus*	14	PDB: 1GZM A	[15]
Mhp1	*Microbacterium tumefaciens*	18	PDB: 2JLN	[444]

The resulting coordinate frame, which consisted of 54 unweighted coordinates and their positions with respect to protein backbone, did not account for the dynamics of the spin label (e.g. the configurations and positions it preferentially occupies) and was highly redundant, with coordinates often being placed less than 1.0 Å apart (Figure E.4). We addressed both issues using a scheme outlined in Figure E.4. Neighboring coordinates were merged using *k*-means clustering to generate a series of coordinate sets ranging from 3 positions to 53 total positions. The weights of these resulting positions were then optimized using 49 previously published experimental distance distributions between 37 residues gathered in T4 Lysozyme [176]. During each of half a million iterations, a Monte Carlo Metropolis algorithm randomly modified the weight of a coordinate and either accepted or rejected the change based on the improved agreement with the

experimental T4 Lysozyme distance data. This algorithm was carried out on each set of clustered coordinates one thousand times. The resulting set of weights with the best agreement consisted of seventeen coordinates, four of which were fit to be zero. This set was introduced as the default set of coordinates for RosettaDEER and was used for all subsequent experiments described here.

3.2.4 Simulation of DEER dipolar coupling decay traces and comparison to experimental values

Because the simulation of DEER decay traces has been extensively described [173, 186, 192, 263], here we limit our discussion to their generation from distance distributions for the purpose of evaluating protein structural models. The traces simulated by RosettaDEER (V_{sim}) are shown in Figure E.5 and reflect coupling between spin labels attached to the same macromolecule (V_{intra}), as well as an intermolecular "background" component reflecting coupling between spin labels across different macromolecules:

$$V_{\text{sim}}\left(t_{\text{i}},\lambda,k,\vec{r},\vec{w}\right) = \exp\left(-k|t_{\text{i}}|\right) * \left(1-\lambda\left(1-V_{\text{intra}}\left(t_{\text{i}},\vec{r},\vec{w}\right)\right)\right) \tag{3.1}$$

This background is assumed to be homogeneous across three dimensions and is modeled using a slope k and a modulation depth λ. The simulated distribution consists of a vector of distances r (in nanometers) and their corresponding amplitudes w. Simulated traces obtained this way are converted into scores (S_{DEER}) by comparing them to the corresponding experimental spectra (V_{exp}) using the following cost function:

$$S_{\text{DEER}} = \frac{1}{n}\sum_{i=1}^{n}\left(V_{\text{exp}}\left(t_{\text{i}}\right) - V_{\text{sim}}\left(t_{\text{i}},\lambda,k,\vec{r},\vec{w}\right)\right)^2 \tag{3.2}$$

where n is the number of time points in the data.

To convert a distance distribution into a spectroscopic signal that can be compared to experimental data, RosettaDEER first simulates V_{intra} for each 0.5 Å bin j between 15 Å to 100 Å:

$$V_{\text{intra}}\left(t_{\text{i}},\vec{r},\vec{w}\right)=\sum_{j=1}^{m}w_{\text{j}}*\int_{0}^{\frac{\pi}{2}}\sin\theta\cos\left(\frac{\left(1-3\cos^{2}\theta\right)*\mu_{0}\mu_{\text{B}}^{2}g_{\text{A}}g_{\text{B}}t_{\text{i}}}{4\pi\hbar r_{\text{j}}^{3}}\right)\text{d}\theta \tag{3.3}$$

where t is the time point of a trace in microseconds, μ_{B} is the Bohr magneton, μ_0 is the vacuum permeability constant, g_{X} is the g-factor of electron X, θ is the angle between the interelectron vector and the bulk magnetic field, and m is the number of distance bins.

Background parameters k and λ are then determined and optimized in two stages. Initial values for both parameters were first determined by incrementing λ with step size 0.01 and log-transforming 3.1 to determine k using linear regression:

$$\hat{k}=\sum_{\text{i}=1}^{n}\left(t_{\text{i}}*\ln\left(\frac{V_{\text{exp}}\left(t_{\text{i}}\right)}{1-\lambda\left(1-V_{\text{intra}}\left(t_{\text{i}},\vec{r},\vec{w}\right)\right)}\right)\right)*\left(\sum_{\text{i}=1}^{\text{n}}t_{\text{i}}^{2}\right)^{-1} \tag{3.4}$$

Subsequent attempts to fit simulated intramolecular decay traces were achieved using gradient minimization to solve for λ and linear regression to solve for k. Convergence was reached when $|\Delta\lambda|<0.0025$. The iterative strategy used to obtain the initial guess was repeated in cases where λ exceeded reasonable values, the lower and upper bounds of which are defined by default as 0.02 and 0.50. This range corresponds to modulation depth values that would ordinarily be obtained from Q-band DEER on well-labeled double-cysteine mutants without using an arbitrary waveform generator. Deviations from experimentally observed values for these two parameters was found to frequently occur during the initial stages of extended chain *de novo* folding, where simulated distance distributions deviated drastically from experimental values and lead to erroneous background parameter results.

3.2.5 Rosetta model generation and evaluation

Rosetta models were generated with two approaches to sample a large conformational space but also ensure native-like models at a high density. The native-like models were generated with RosettaCM [384] using either full-length or truncated native models as inputs. Coverage of a large conformational space was accomplished by *de novo* protein folding. Bax, ExoU, and CDB3 were scored

using the *ref2015* energy function [13], and Rhodopsin and Mhp1 were scored using RosettaMembrane [466]. The transmembrane regions for Rhodopsin and Mhp1 were predicted using OCTOPUS [426]. These models were evaluated using $\text{RMSD}_{100\text{SSE}}$, which measures the size-normalized RMSD over residues in secondary structures [65]. Enrichment of these models was evaluated as $\log_{10} \frac{TP_{\text{P,Score}} * N_{\text{Total}}}{P * N_{\text{Total}}}$ where N_{Total} refers to the total number of models being considered, P refers to the proportion of models considered native-like by C_α $\text{RMSD}_{100\text{SSE}}$, and $TP_{\text{P,Score}}$ refers to the number of true positives identified in the top $P * N_{\text{Total}}$ models by score [126]. We treated the top 10% of models as native-like (P=0.1), thus scaling the metric from -1 (none of the top 10% of models by $\text{RMSD}_{100\text{SSE}}$ were in the top 10% by score) to 1 (all of the top 10% of models by $\text{RMSD}_{100\text{SSE}}$ were also in the top 10% by score), with a value of 0 indicating that the number of native-like models found in the top 10% by score was equal to what is expected by chance.

Oscillation frequencies of decay traces in microseconds for distributions with an average distance r_{avg} (in angstroms) were calculated as $\frac{r_{\text{avg}}}{5.2 * 10^4}$ [173]. Decay traces with fewer than three oscillations were not used to evaluate enrichment as a function of decay trace duration.

3.2.6 *De novo* protein structure prediction benchmark

The protein structure prediction protocol we used largely follows a previously published template [302] and consists of three stages. In the first stage, 10,000 models were generated using extended chain AbInitio with either RosettaDEER restraints, CONE model restraints [8], or no restraints. This protocol relies on the insertion of fragments obtained from a July 2011 copy of the Protein Data Bank and was obtained from the Robetta online server [216]; homologous protein structures were excluded from these fragment libraries. The contribution of the RosettaDEER score term was adjusted so that its dynamic range was similar to that of the Rosetta energy function [302]. Since the proportion of DEER restraints relative to the protein length was comparable for Bax and ExoU, the impact of the number of restraints on the weight of the score term was not considered [441].

Models generated this way were then clustered to a radius of 7.5 Å C_α RMSD_{100} using Durandal [25]. Each cluster was evaluated by scoring its models using both RosettaDEER and the full-atom

Rosetta energy function [13], obtaining the cluster averages for both values, and adding their Z-scores with respect to those of other clusters. After discarding sparsely-populated clusters (smaller than 5% of the size of the largest cluster), the top ten best-scoring models by combined Z-score were selected from the five best-scoring clusters for subsequent modeling.

An additional 1,200 models were generated from these 50 models using RosettaCM [384], which also relies on fragment insertion but ensures that the input model's topology is retained throughout the modeling process. The scripts were obtained from a recently published refinement protocol [307], and no experimental restraints were used. Models generated during this stage were again clustered to 7.5 Å C_α $RMSD_{100}$ and scored, except only the RosettaDEER score was used to evaluate the quality of these models.

During the third and final stage, models in the best-scoring cluster were minimized using FastRelax [81], which introduces and repacks side chains while performing gradient descent on a full-atom depiction of the entire model. Models generated at this stage were scored exclusively using the native Rosetta energy function, with the lowest-scoring model selected as the output model.

3.3 Results

3.3.1 Modeling nitroxide spin labels using RosettaDEER

A strategy to model proteins using DEER data must reliably simulate distance distributions between spin-labeled residues. To quantify the computational cost and efficiency of this task, we considered a panel of five proteins, listed in Table 3.1, where both atomic-detail structures and experimental DEER data were available [15, 39, 127, 212, 479]. Distance distributions between residue pairs that have been previously measured experimentally were simulated using a number of methods, and the resulting error was quantified as the difference between the average values of the simulated and experimental distance distributions (example shown in Figure 3.1.A). In addition, we measured how rapidly each program calculated these distance distributions (Figure 3.1.B). Consistent with previous results [128, 153, 353], the average values of experimental distance distributions gathered in monomeric proteins, but not the homodimer CDB3, agree more closely with those of simulated

distributions than their corresponding C_β-C_β distances, from which restraints such as the CONE model are derived [8]. By contrast, none of the methods examined here reliably reproduced the width of the distance distributions. This is likely attributable to oversampling of available conformational space of the spin label, which results from the exclusive use of van der Waals repulsive energies to limit possible rotameric configurations. Finally, the data revealed how simulation times varied substantially between these methods.

These results further illustrate that increasing computational complexity did not lead to more accurate distance distributions. We hypothesized that, for the same reason, decreasing the computational complexity would not lead to less accurate distance distributions. Therefore, RosettaDEER's design prioritized computational efficiency (see section 3.2). Rather than measure distances from full-atom rotamers or mobile dummy atoms, RosettaDEER uses a probability density function to capture high-occupancy electron positions that would be explored by MTSSL and map them onto the protein structure (Figures 3.1.C and D). For each of these coordinates, an evaluation of a potential van der Waals overlap was performed between a pseudo-atom representing the nitroxide ring's center of mass and the rest of the protein. Placing this pseudo-atom at an idealized location, consistent with spin-labeled protein structures in the Protein Databank (Figures E.3 and Table E.1), reduced the number of atoms for this evaluation to one per rotamer, thus maximizing computational efficiency. Figures 3.1.A and B demonstrate that RosettaDEER's simplified representation of the spin label allows the generation of distance distributions three to five orders of magnitude faster than other approaches but with comparable accuracy.

3.3.2 Comparison of simulated with experimental DEER decay traces

Most existing methods that leverage DEER experimental data for structural modeling require the primary spectroscopic readout first be processed into a distance distribution. A conventional approach, such as the Rosetta CONE model, is outlined in Figure 3.2.A. This involves 1) manually identifying and removing the "background" signal, which corresponds to coupling between spin labels across macromolecules; 2) using Tikhonov regularization to convert the remaining intramolec-

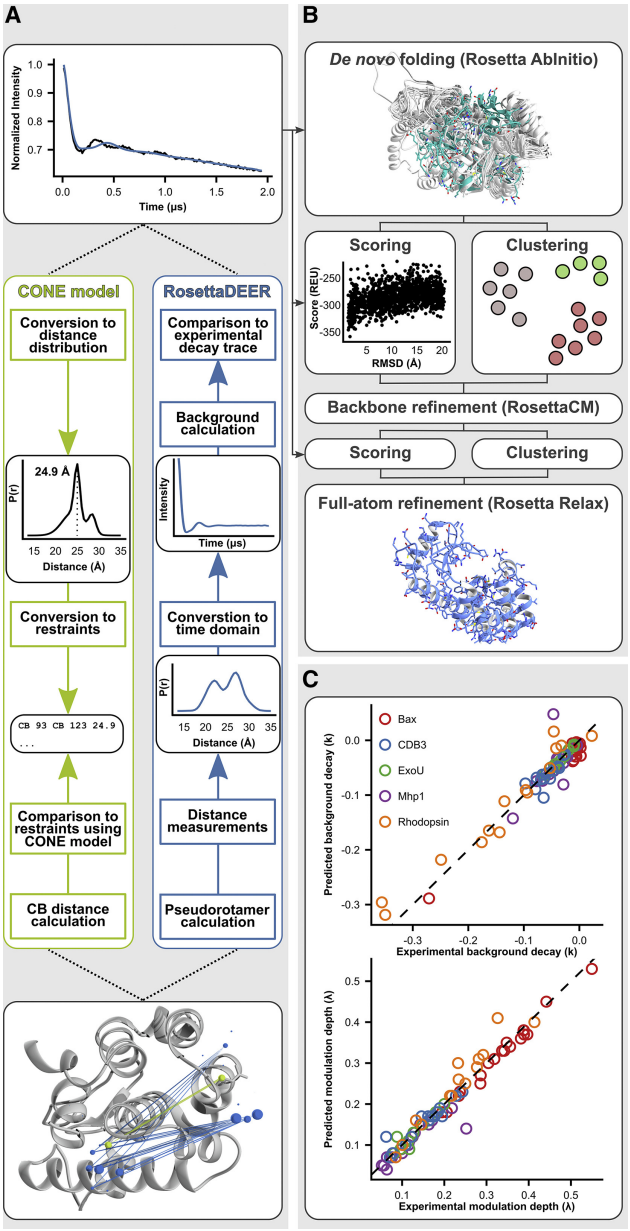

Figure 3.2: RosettaDEER simulations of distance distributions and decay traces. The forward approach taken by RosettaDEER contrasts with the preprocessing required by the CONE model. (A) A flowchart illustrating how both the CONE model and RosettaDEER use experimental DEER data to model proteins (example shown is T4 lysozyme residues 93 and 123). (B) Incorporation of DEER experimental restraints into Rosetta structure prediction pipeline. (C) Recovery of experimental background coupling and modulation depth parameter values.

ular signal into a distance distribution; and 3) selecting a single distance value from this distribution to restrain the modeling process. An additional bias is often required to convert these distance data into backbone restraints [8, 10, 352, 463].

We reasoned that these preprocessing steps could be avoided by simulating a spectroscopic signal from candidate models for direct comparison to the experimental data. As with other forward approaches to fitting DEER data [173, 391], the steps are as follows: 1) the model is used to generate a distance distribution; 2) this distance distribution is converted into a spectroscopic signal consisting solely of the effect of coupling between spin labels attached to the same macromolecule; and 3) the slope of the "background" coupling and depth of modulation needed to optimally fit the simulated and experimental decay traces is determined.

This final step represents the outstanding challenge in the proposed pipeline, as most modeling programs, including Rosetta, focus on isolated protein structural models. We instead used a two-parameter exponential function to simulate the background coupling (k) and modulation depth (λ) (see Section 3.2). The values of these parameters were determined by minimizing the sum of the squared residuals. The optimum values obtained strongly correlated with those obtained using DeerAnalysis [192], with r^2 values exceeding 0.90 for both parameters (Figure 3.2.C), despite the fact that the inaccuracies in the distance distributions affected the fit (Figure E.5). In fact, we found that this correspondence correlated less strongly with the goodness-of-fit in the distance domain than it did with the quality of the experimental data in the time domain (Figure E.6).

3.3.3 Enrichment of native-like models using experimental decay traces

Being able to simulate DEER traces from candidate structural models without any pre-processing offers the possibility to reframe the problem currently faced by translating the DEER traces into distance distributions. Whereas methods such as Tikhonov regularization convert individual DEER traces into distance distributions, RosettaDEER, in conjunction with Monte Carlo modeling, would instead seek to determine the structural model most consistent with both an energy function and the experimental data. To investigate whether unprocessed DEER traces can be used to discriminate native-like models from incorrectly-folded models, we generated a series of 1000-2000 misfolded models for each of the five proteins in our test set and scored their agreement with experimental DEER data. In addition, we generated 1000 docked models of the homodimer CDB3 that retained

the native fold for the protomer, but not the oligomeric interface. Similarity to the native model was measured by C_α $RMSD_{100SSE}$ [65], which is the size-normalized root means squared deviation across secondary structural elements (Figure 3.3). RosettaDEER's effectiveness at this task was measured by the enrichment parameter, which is defined in the Section 3.2.5 section and quantifies a scoring function's ability to discriminate native-like models from incorrectly-folded models.

RosettaDEER consistently scored native-like models of the monomeric proteins more favorably than poorly-folded models (Figure 3.3). This was also observed with correctly-docked models of CDB3. Moreover, it generally outperformed the CONE model in enriching native-like models (Figure E.7). Perhaps unsurprisingly, the simultaneous use of Rosetta's energy function often improved enrichment, since it overwhelmingly considers short-range interactions and is therefore expected to complement the evaluation of longer ranger, fold-level information provided by DEER restraints (Figure E.7) [13]. We note that RosettaDEER could not effectively identify misfolded models of CDB3, which we attribute to the fact that DEER restraints reflect distances across the center of symmetry, rather than within the protomer. Nevertheless, these results suggest that RosettaDEER's inability to perfectly recreate the experimental DEER data did not impede its ability to identify correctly folded models, suggest-

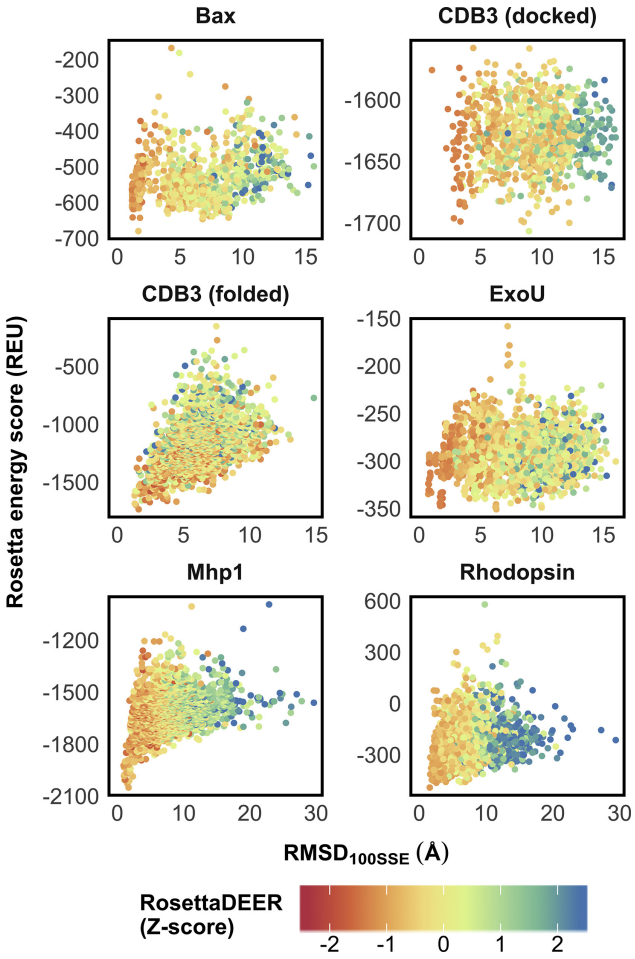

Figure 3.3: Evaluation of models using DEER decay traces. Models with C_α $RMSD_{100SSE}$ ranging from 0.5 Å to 30.0 Å were scored using both the Rosetta energy function and RosettaDEER.

ing that it could be effectively used for structure prediction.

The fact that structural models are scored based on their consistency with the primary spectroscopic data led us to hypothesize that they could be evaluated using lower-quality data than what would be necessary for conversion into precise distance distributions. We were specifically interested in evaluating the importance of the experimental data's time window, which must undergo roughly 0.8 and 1.6 oscillations for Tikhonov regularization to accurately identify a distance distribution's average and standard deviation, respectively [186]. This hypothesis was tested by artificially truncating the experimental data in the time domain and measuring enrichment as a function of how many oscillations were included (see 3.2 and Figure E.7). Strikingly, RosettaDEER could enrich native-like models of Bax, ExoU, Rhodopsin, and Mhp1 with highly truncated data (< 0.8 oscillations), albeit to a reduced degree. We found that the addition of data in the time domain beyond one oscillation failed to lead to any measurable improvements in enrichment, despite its importance in allowing RosettaDEER to identify the correct background coupling parameters (Figure E.6). These results suggest that RosettaDEER is more permissive than Tikhonov regularization with respect to the effect of data quality on protein structural modeling.

3.3.4 *De novo* folding of Bax and ExoU

To further illustrate RosettaDEER's capability to identify native-like models, we folded Bax and ExoU *de novo* using experimental DEER decay data. These two proteins were chosen because native-like models cannot be identified using the default Rosetta energy function alone (Figures 3.3 and E.7). The structure prediction protocol we used is similar to one used to model proteins using other types of sparse data [216, 302] and is illustrated in Figure 3.2.B and described in detail in 3.2. We first generated an initial set of ten thousand models using Rosetta AbInitio folding supplemented by either experimental restraints through RosettaDEER, experimental restraints through the CONE model [8], or no restraints. These models were then clustered, and models from the best-scoring clusters were refined and recombined into one thousand two hundred new models without using experimental data. After a second round of clustering, models from the cluster with the best agree-

ment to the experimental data were refined and minimized, and the model with the best Rosetta energy score was returned as the predicted model.

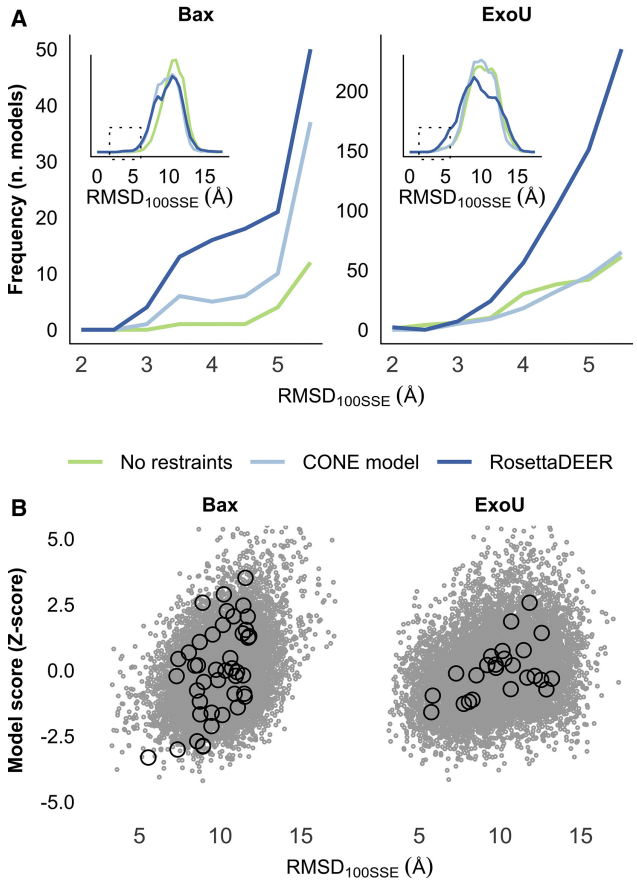

Figure 3.4: Structure prediction of Bax and ExoU using experimental decay data. (A) *De novo* protein folding of native-like models using DEER decay restraints with RosettaDEER, $C_\beta - C_\beta$ distance restraints with the CONE model, or no restraints. Inset: spread of all models generated using these three methods. (B) Accuracy of *de novo* folded models (gray dots) and clusters (black circles) as a function of combined DEER and Rosetta Z-score.

In the absence of experimental restraints, few of the models generated by AbInitio folding resembled the native fold (Figure 3.4.A). Perhaps strikingly, providing DEER restraints with the CONE model had no effect on the proportion of native-like models of ExoU generated this way (a measurable improvement was observed when folding Bax). This contrasts with the proportion of native-like models generated using RosettaDEER, which was substantially higher in the case of both proteins.

Although agreement between models and experimental structures loosely correlated with both RosettaDEER score and Rosetta energy score for both proteins, an abundance of incorrectly-folded models obscured this trend (Figure 3.4.B; RosettaDEER and Rosetta energy scores were jointly considered by adding the Z-scores of each). As a result, we were unable to identify native-like models for either Bax of ExoU from score values alone. The ten best-scoring models by these metrics were generally incorrectly folded (5 Å to 10 Å C_α $RMSD_{100SSE}$) and buried amphipathic features found on the surface of the native model. This shortcoming is typically addressed by clustering, since native-like models are more likely to be found near the centers of large

clusters with favorable average scores [375]. We therefore clustered Bax and ExoU models with a radius of 7.5 Å and evaluated these clusters by taking the Z-scores of both the average Rosetta energy and RosettaDEER score and adding them together (Figure 3.4.B). In the case of both proteins, this step placed native-like models in the best-scoring clusters. Focusing our attention on the five best-scoring clusters allowed us to discard 85.3% of the Bax models and 61.3% of the ExoU models, while retaining a majority of the native-like models in each case.

Each cluster at this stage represented a broad population of models that satisfied the DEER data. To test whether refining models without experimental restraints would reveal the native fold, ten models from each of the top five clusters were refined and recombined using RosettaCM [384]. This step retained the topology of the input models, but permitted minor backbone rearrangements that allowed misfolded models to optimize away from conformations consistent with the experimental data. As a result, the cluster with the most native-like models after this resampling stage scored the most favorably by RosettaDEER. After minimization of models in this cluster [81], the best-scoring model by Rosetta score for both Bax and ExoU had near-native folds (<3.5 Å C_α $RMSD_{100SSE}$; Figure 3.5 and E.8).

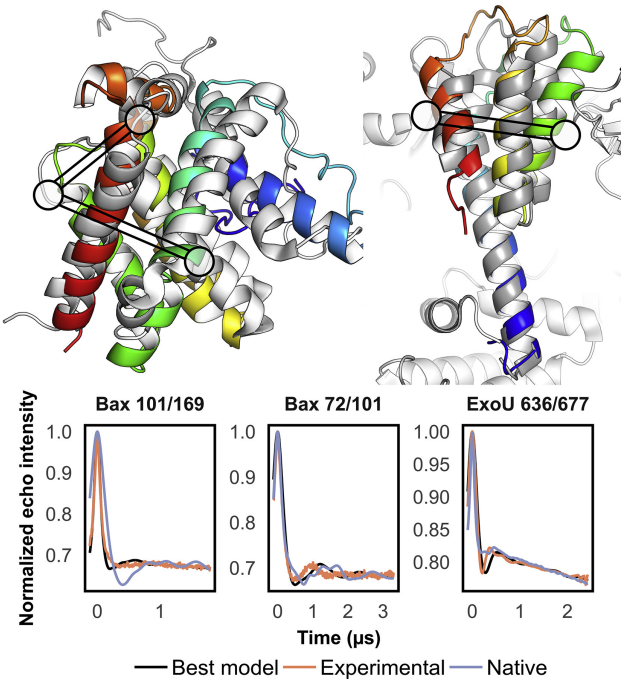

Figure 3.5: Predicted models of Bax and ExoU generated using DEER data. Best-scoring models of Bax and ExoU had an accuracy of 3.2 Å and 2.1 Å C_α $RMSD_{100SSE}$, respectively.(Top) Models were obtained from 10,000 *de novo* folded models, the best-scoring of which were refined into 1200 additional models. Native models shown in white. (Bottom) Example DEER traces in which the best model outperformed the native. Corresponding residues indicated as circles in (A) and (B).

3.4 Discussion

RosettaDEER predicts and refines protein structures by integrating DEER spectroscopy data and Rosetta computational modeling protocols. The novel aspects of this method are a simplified representation of the commonly used spin label MTSSL and a strategy to rapidly simulate DEER decay traces for comparison to uncorrected experimental traces. The robustness of the method was demonstrated by benchmarking every step on five sparse datasets. Despite the simplified spin label representation, the distance distributions simulated by RosettaDEER are comparable to those generated using more computationally complex rotamer library approaches. Moreover, even though simulated spectra fail to perfectly fit experimental DEER traces, this integrated approach efficiently identifies conformations that simultaneously satisfy the data and the Rosetta energy function. Our findings illustrate how RosettaDEER can complement similar methods that are more computationally intensive but able to use DEER decay data to perform high-resolution refinement of protein structures [263].

The *de novo* folding benchmark with the small soluble proteins ExoU and Bax highlights the success of this strategy. Both proteins possess surface-exposed amphipathic regions that insert into the membrane. Bax transitions from a soluble monomer into a membrane-bound oligomer using its C-terminal helix [39], whereas ExoU is hypothesized to move into the membrane using a flexible loop between its two C-terminal helices [409]. Consistent with previous results [127, 128], the Rosetta energy function favored models that packed these substructures in the protein core, leading to incorrectly folded models and lack of correlation between the Rosetta score and model accuracy. As a result, orthogonal experimental data that define the structure are critical to *de novo* folding. Our folding benchmark suggests that RosettaDEER more effectively leverages the experimental data than the C_β-based CONE model. Moreover, even low-quality data can be used to discriminate native-like from incorrectly-folded models. We appreciate that, for larger proteins, structure determination from DEER experiments alone would require extensive experimental data. Integrating RosettaDEER with other types of sparse experimental data could therefore reduce the number of DEER restraints required for accurate modeling.

The strategy of RosettaDEER to predict the structures of these two proteins leverages the experimental data by folding and optimizing protein structures with and without restraints, respectively. The first step leads to a substantial reduction in the search space and a concomitant increase in the number of models that satisfy the restraints, although not all of these models are correctly folded. After clustering the models to remove those that correspond to narrow energy minima, the second step, optimization without restraints, allows clusters with incorrectly folded models to reach energy minima inconsistent with the data. This filtering procedure restores the experimental data's ability to identify native-like models, since the most native-like models of Bax and ExoU at this stage were not identifiable using the Rosetta energy function. Overall, this protocol decreases both the number of incorrectly-folded structures that fit the data and the conformational search space inherent to the protein folding problem.

Despite its success illustrated here, the current implementation of RosettaDEER assumes that a single protein conformation describes the data. For example, the distance distributions of Mhp1, the most conformationally flexible protein examined in this dataset, were generally more poorly simulated using available methods than those collected in other proteins. Experimental applications of the DEER technique often focus on monitoring ensembles of protein conformations. They can therefore be effectively complemented by computational methods that interpret this data with the capability to generate multiple models and examine their consistency with sparse experimental data.

3.5 Acknowledgements

The authors would like to thank Dr. Christian Altenbach, Dr. Enrica Bordignon, and Dr. Eric Hustedt for providing experimental data used in this study and Dr. Rocco Moretti, Dr. Axel Fischer, and Dr. Andrew Leaver-Fay for helpful discussions on designing and implementing RosettaDEER. Research was funded by the National Institutes of Health (R01 GM080403, R01 GM073151, R01 GM114234, R01 GM077659, R01 HL122010, and R01 HL144131).

CHAPTER 4

Methodology for rigorous modeling of protein conformational changes by Rosetta using DEER distance restraints

The contents of this chapter have been previously published [96].

We describe an approach for integrating distance restraints from DEER spectroscopy into Rosetta with the purpose of modeling alternative protein conformations from an initial experimental structure. Fundamental to this approach is a multilateration algorithm that harnesses sets of interconnected spin label pairs to identify optimal rotamer ensembles at each residue that fit the DEER decay in the time domain. Benchmarked relative to data analysis packages, the algorithm yields comparable distance distributions with the advantage that fitting the DEER decay and rotamer ensemble optimization are coupled. We demonstrate this approach by modeling the protonation-dependent transition of the multidrug transporter PfMATE to an inward facing conformation with a deviation to the experimental structure of less than 2 Å C_α RMSD. By decreasing spin label rotamer entropy, this approach engenders more accurate Rosetta models that are also more closely clustered, thus setting the stage for more robust modeling of protein conformational changes.

4.1 Introduction

Distance measurements between pairs of spin labels by DEER spectroscopy have been utilized extensively to investigate the structures and dynamics of proteins [79, 114, 212, 288] and the assembly of protein-protein complexes [30, 218, 219, 410]. At the fundamental level, DEER measures magnetic dipolar coupling to infer the distributions of distances between two or more spin labels [186, 274]. A two-step process typically interprets these distances as spatial restraints describing the protein backbone structure. First, the echo-decay time traces are transformed into distributions consisting of distance components characterized by a mean and width [116, 173, 185, 304, 391]. Second, these distributions are compared to those predicted using one of several strategies, ranging

from generic rotamer libraries[153, 322, 336], explicitly modeled pseudoatoms[212, 332], or explicitly modeled spin label side chains [10, 88, 226, 263, 353, 385]. However, these strategies tend to overestimate the dynamics of flexible probes such as the commonly used MTSSL. Therefore, the predicted distributions are broad relative to the experimental ones [153, 156, 176, 187, 220], which hinders DEER-based evaluation of protein structures or complexes as well as mapping of protein conformational changes. The latter can be obscured entirely if modeled distribution widths exceed distance changes observed between spin labels.1 Another layer of complications in modeling of conformational changes arises if the ensemble of spin label rotamers is allowed to reconfigure, hence providing a low energy pathway to account for changes in distance distributions that originate from backbone movements. Collectively, these caveats limit the accuracy and precision of molecular models generated from DEER restraints.

Several algorithms have recently been developed to refine ensembles of spin label rotamers by employing multilateration [1, 140, 152, 157, 189, 336]. Multilateration refers to the determination of an object's position in three-dimensional space given its distance from a constellation of points; common applications include the positioning of electronic devices using the Global Positioning System and of earthquakes epicenters using time-of-arrival data [121]. To utilize this approach to position spin label rotamers requires both a high-resolution starting structure and a set of DEER distance data consistent with that structure. However, a unique challenge in this endeavor is that spin labels are flexible relative to the protein backbone. As a result, the ensembles characterizing their positions must be refined simultaneously for all spin labels in a given protein model.

Molecular dynamics simulations have been used to determine a set of optimized rotamers from explicitly modeled spin labels restrained by experimental distance distributions [173, 262, 263, 343]. Alternatively, rotamer libraries have been precomputed and reweighed using either Monte Carlo [140, 157], singular value decomposition [152], or nonlinear least-squares minimization [189]. The positions of these labels can, in turn, be used to more precisely locate paramagnetic ligands or metal ions [1, 3, 140, 465] as well as make small-scale refinements to protein structures [336, 446]. To our knowledge, however, none of these methods have demonstrated that these

optimized rotamers can lead to improvements in modeling conformational changes.

Furthermore, these methods generally do not address unique factors confounding multilateration of spin labels. First, the width of a distribution reflects disorder in the solid state as a result of backbone and spin label side chain dynamics at room temperature. Existing multilateration methods generally ignore the former, by assuming the distribution is explained entirely by spin label dynamics [140, 336], or both, by extracting the peak distance from the distribution and discarding the width [446]. Second, relying on distance distributions rather than time domain data propagates assumptions intrinsic to the method used for the transformation of the latter. Depending on the noise level of the experimental measurements, this step can distort true components or introduce ghost components to the distribution. Finally, although DEER distributions are often reported with confidence bands to reflect the uncertainty inherent to this transformation [104, 116, 173], they are generally taken at face value when used for rotamer multilateration. This incorrectly implies that experimental uncertainty is uniformly distributed across the dataset and can lead to rotamers that over- or underfit the DEER distributions. Collectively, these obstacles prevent the straightforward positioning of spin label ensembles in three-dimensional space and complicate the confidence with which such ensembles can be used for subsequent modeling purposes.

To address these issues, we developed and implemented, as part of the RosettaDEER module, an algorithm that combines rotamer multilateration [3, 140, 189] for pairs sharing common spin labeling sites with direct analysis of DEER time traces. The algorithm calculates a weighted distribution of "pseudo-rotamers", or inflexible coarse-grained side chains, capable of recapitulating large experimental datasets collected using DEER. Importantly, this algorithm goes beyond comparable methods by refining these ensembles using raw data in the time domain, rather than distance distributions calculated *a priori*, thus avoiding the loss of information that can occur as result of data transformation. Using experimental collected in the model system T4 Lysozyme and the multidrug transporter PfMATE, we demonstrate that this algorithm is able to fit time domain data as effectively as widely-used DEER data analysis programs. Integrated with Rosetta, these rotamers ensembles yield substantial improvements in both accuracy and precision of modeling the outward-to-inward

isomerization of the multidrug transporter PfMATE, thus reinforcing the notion that coupling analysis of primary data with rotamer optimization is a superior approach for restrained modeling of protein conformational states.

4.2 Results and Discussion

4.2.1 Overview of the multilateration algorithm

The algorithm capitalizes on the concept of pseudo-rotamers, which are simplified representations of the spin label designed to maximize computational efficiency. A pseudo-rotamer models the spin label side chain as a centroid atom representing the nitroxide ring and its unpaired electron, yielding predicted distance distributions that are comparable to full-atom depictions. Unlike explicit depictions of the spin label used in all-atom simulations, ensembles of pseudo-rotamers do not interact with one another; as a result, the dynamics of spin labels close in space are fully independent. However, in principle, any rotamer library can be used for the multilateration strategy described here [10, 119, 153, 156, 322, 385].

The transformation of DEER data to distance distributions is an ill-posed mathematical problem necessitating the use of either regularization [73, 116, 192], parametric modeling [116, 173, 391], neural networks [448], or other methods [104, 304, 388]. Because these methods have intrinsic approximations which could interfere with rotamer ensemble determination, we elected to fit the raw experimental data directly using an iterative simulated annealing strategy that 1) measures all pairwise distances between pseudo-rotamers, 2) converts each distance distribution into a DEER decay, and 3) calculates the intermolecular dipolar coupling contribution by nonlinear least-squares minimization (Figure 4.1). Different levels of noise between DEER traces linked by multilateration were normalized using estimates obtained from each signal's corresponding imaginary component [263]. The algorithm prioritized the generation of parsimonious ensembles by minimizing the total number of pseudo-rotamers with nonzero weights using the Akaike Information Criterion-corrected (AICc) [7, 395]. This metric, which allows for regularization in rotamer space rather than the distance domain, was guided by the heuristic that the flash-freezing process sharpens the distribution

of rotamers that contribute to the DEER signal [20, 144]. Finally, to account for backbone heterogeneity and the expectation of smoothness in the distance domain, simulated distributions were broadened by a magnitude corresponding to the residues' intrinsic flexibility, as reported by their respective crystallographic B-factor values [397, 461].

4.2.2 Data analysis benchmark

We benchmarked this method using experimental DEER data collected in two model proteins, T4 Lysozyme [176, 439] (PDB: 2LZM) and the Multidrug and Toxin Extrusion (MATE) transporter PfMATE [180, 401] in its outward-facing conformation (PDB: 6GWH). The extracellular and intracellular spin label pairs of PfMATE were treated independently since they did not share residues in common. These three DEER datasets consisted of 65 restraints between 47 residues; a subset of the restraints in T4 Lysozyme is shown in 4.1.A. We note that unlike the benchmarks used in other multilateration methods, these restraints were highly interconnected; half of the residues were spin labeled in three or more DEER pairs, and in the most extreme case, two residues in T4 Lysozyme were spin labeled across seven pairs (Figure F.1). For each of the three datasets, the RosettaDEER multilateration algorithm was executed for 1000 replicas, with each replica yielding refined pseudo-rotamer ensembles at every spin labeled site.

We compared the resulting fits to those obtained using GLADDvu [173], DeerAnalysis [192], and DeerNet [448], which are programs that analyze DEER data using Gaussian mixture models, Tikhonov regularization, and feed-forward neural networks, respectively. Although other analysis methods are available, we believe these represent a sufficiently diverse range of analytical approaches for the purposes of comparison. We found that the optimum rotamer ensembles, selected by the AICc, could recapitulate the experimental DEER traces as effectively as each of these programs (Figures 4.1.B and C, F.2, F.3, F.4, and F.5). The mean squared errors obtained by the best fit were not statistically different from those obtained by any of these three methods, or from the noise estimated from the imaginary component (Student's paired one-tailed t-test with Bonferroni correction). However, unlike the latter methods, the interconnectedness of the spin label pairs al-

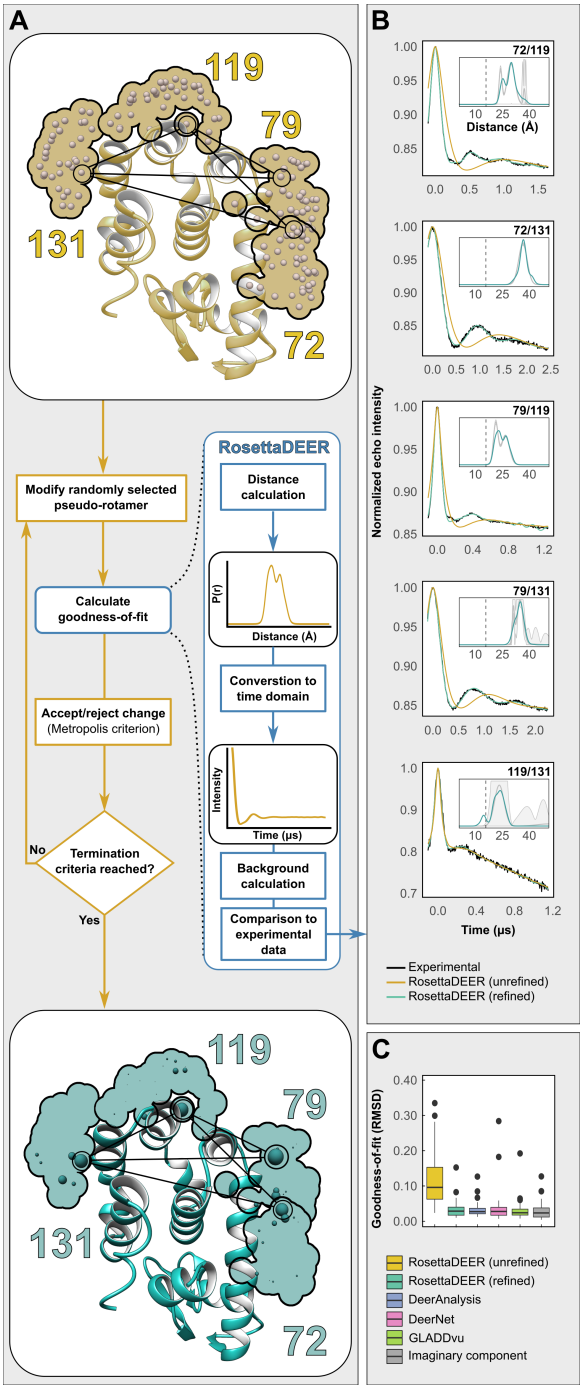

Figure 4.1: Scheme of the RosettaDEER multilateration algorithm. (A) Refinement of pseudo-rotamers using the RosettaDEER multilateration algorithm. (B) Representative DEER traces prior to and following refinement. Insets: Distributions with 95% confidence bands. (C) Comparison of RosettaDEER to other analysis programs.

lowed our algorithm to couple pseudo-rotamer parametrization to the analysis of DEER data in the time domain.

4.2.3 Distance distribution benchmark

We anticipated that the analysis of DEER data by multilateration would yield distance distributions similar to those obtained using traditional methods. Consistent with this expectation, distributions between refined pseudo-rotamers in both T4L and PfMATE showed remarkable agreement with those obtained using the three methods mentioned above (see insets in Figure 4.1.B for examples and F.3, F.4, and F.5 for all distributions). For example, the average values of these distributions were within 0.5 Å of those obtained using GLADDvu for 60 of the 65 restraints (Figure 4.2). Additionally, the widths of 52 of these restraints were within 0.5 Å of those obtained using GLADDvu. Discrepancies occurred for broad distributions or long distances (because the information content in the time domain is not as well-defined) or components less than 15 Å (because these distances minimally contribute to the DEER signal). Additionally, we uncovered differences when comparing the widths of these distributions to those obtained using DeerAnalysis, likely resulting from small "ghost" side peaks fre-

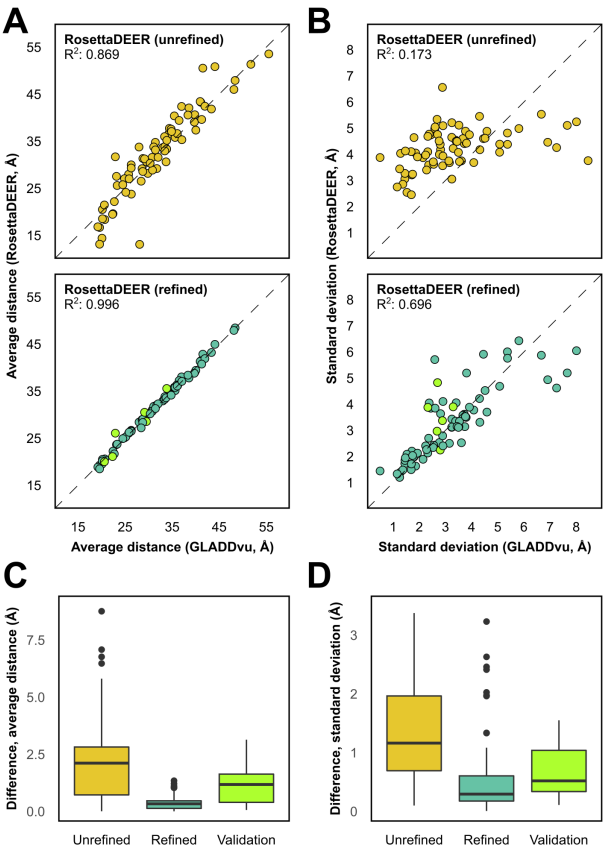

Figure 4.2: Evaluation of distance distributions by multilateration. Average distances (A) and widths (B) between pseudo-rotamers prior to (top) and following (bottom) refinement by multilateration. T4 lysozyme distributions omitted from multilateration are shown in light green. (C and D) Boxplots showing the difference between values obtained using GLADDvu and values simulated between pseudo-rotamer ensembles prior to and following refinement.

quently observed in regularization. Discrepancies were also observed when comparing these distributions to those obtained using DeerNet, which yielded widths clustered between 2.5 Å to 4.5 Å (Figure F.6).

Finally, the uncertainty of these distributions was calculated from the five pseudo-rotamer ensembles with the lowest AICc values. The resulting confidence bands, which capture 95% of the variation in the distance distributions, are qualitatively comparable to those obtained using GLADDvu, DeerAnalysis, and DeerNet (Figure F.7).

To further validate the algorithm, we simulated distance distributions for six T4L spin label pairs which were excluded from the multilateration dataset. We observed that the median error between the average distance values fell by 50% (Figure 4.2) using the refined rotamers. By contrast, the standard deviations did not significantly sharpen, and their values are similar to those observed prior to refinement. Notably, the uncertainty of these distributions is greater than those of the distributions included in the training set.

4.2.4 Conformational change modeling in PfMATE using refined pseudo-rotamers

While the results above demonstrate the robustness of the multilateration algorithm in identifying optimal spin label pseudo-rotamer ensembles, the central question is whether these provide superior restraint quality for modeling conformational changes. To address this question, we modeled the isomerization of PfMATE between OF and IF conformations (shown in Figure 4.3.A and B, respectively), both of which were determined by X-ray crystallography [401, 468]. The two conformations differ primarily in the relative orientations of the N- and C-terminal domains resulting from changes in the backbone dihedral angles of TMH7. Of direct relevance to the question addressed here, distance distributions between pairs of spin labels measured at pH 7.5 and pH 4.0 were shown to be consistent with the OF and IF conformations, respectively [180].

We generated several thousand models, using Rosetta [229, 234] without DEER restraints, by perturbing TMH7 and found that none of the built-in membrane protein scoring functions [11, 12, 442, 466] could identify the IF state by score alone (Figure F.9) even if it was included in the initial

model set. Thus, from an MC modeling perspective, the OF-to-IF conformational transition can be sampled, but not necessarily identified, without experimental data.

Table 4.1: Restraints used to score PfMATE models.

Set	1	2	3	4	5	6	7	8	9	10
1	120/269	215/318	120/413	215/394	134/269	215/442	186/269	240/318	186/348	240/442
2	120/413	215/318	134/269	215/394	186/348	215/442	186/269	240/318	196/348	240/442
3	240/442	186/348	215/394	196/348	215/442	120/269	215/318	186/269	240/318	120/413
4	215/318	186/269	240/442	120/413	240/318	134/269	186/348	196/348	215/442	120/269

To test the notion that DEER restraints interpreted with the refined pseudo-rotamers can drive convergence of Rosetta modeling, we identified spin label pairs where the EPR lineshape showed minimal changes upon a pH shift from 7.5 to 4.0, supporting the approximation that the spin label rotamer ensembles are invariant and thus were not allowed to reconfigure during Rosetta modeling (Table 4.1). From these pairs, 40 sets of restraints were generated, each of which consisted of one to ten spin label pairs. Using scoring functions to assess the agreement with the DEER restraints (see section 4.3), the OF-to-IF conformational transition was modeled by perturbing the dihedral angles of TMH7. DEER distributions were simulated using either the pseudo-rotamers ensembles refined by multilateration or the unrefined

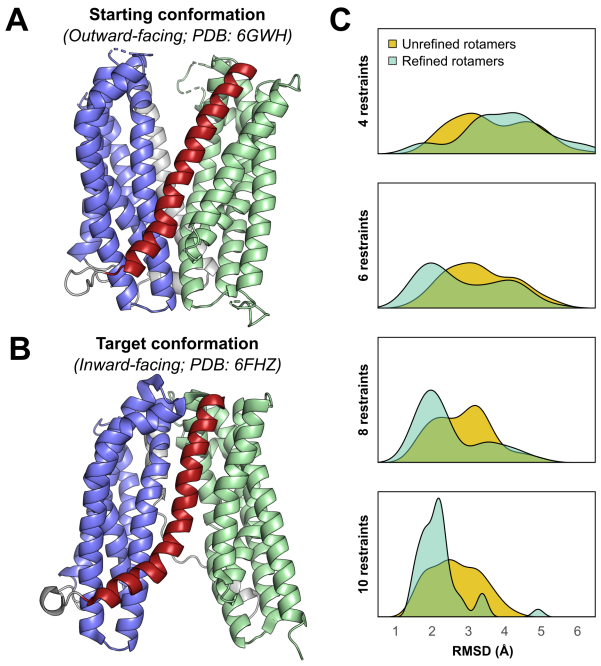

Figure 4.3: Modeling conformational changes in the multidrug transporter PfMATE. (A) OF and (B) IF crystal structures of PfMATE. N- and C-terminal domains shown in purple and green, respectively, with TMH7 in red. (C) RMSD values of ten best-scoring models relative to the IF conformation using pseudo-rotamers either refined by multilateration (teal) or available by default (yellow).

ensembles available to RosettaDEER by default. Agreement with the experimental distributions was evaluated by the overlap between the experimental and simulated distance distributions. Similarity to the inward-facing crystal structure was quantified by the RMSD of the alpha carbons excluding TMHs 1 and 7.

We observed a striking contrast between the effectiveness of the refined and unrefined ensembles (Figure 4.3.C). The default rotamer library did not effectively improve the average RMSD values of the ten lowest-scoring models beyond 2.0 Å to 3.5 Å. By contrast, the use of multilaterated pseudo-rotamers converged upon inward-facing models to 1.5 Å to 2.5 Å C_α RMSD using restraints obtained from the same spin label pairs.

Alongside these improvements in accuracy, the sharper range of RMSD values suggested that multilateration improved model precision. Distributions of representative distances in the intracellular and extracellular sides of the top ten models (Figures 4.4.A and B) revealed that, when using the default pseudo-rotamers, a majority of these models failed to close the extracellular cavity and were less inward-open than the OF structure (Figure 4.4.C), even when ten restraints were used. By contrast, models obtained using refined pseudo-rotamers deviated less drastically from the crystal structure. Nonetheless, these models were virtually all less inward-open than the crystal structure, consistent with shorter-than-expected experimental DEER measurements on the intracellular side at pH 4.0 (Figure 4.4.D) [180] .

4.2.5 Concluding remarks

Our results highlight a strategy to improve the quality of models obtained from EPR restraints. We envision that the main application of this strategy is to model alternate conformational states starting from an experimental structure and a set of interconnected DEER data. By implementing this algorithm in Rosetta, we hope to encourage its use for a wide variety of modeling applications, such as protein-protein docking and *de novo* folding. Moreover, further development of this approach, as well as extensive use of multilateration in the design of spin label pairs, will open the door to modeling proteins where conformational changes are defined by more complex modes of motion.

4.3 Materials and Methods

4.3.1 Overview of the model-based approach

The objective of the RosettaDEER multilateration algorithm is to fit a set of DEER data by weighting the nitroxide pseudo-rotamers available to each spin-labeled residue in a protein structural model. Each replicate of the algorithm independently generates a unique set of pseudo-rotamer ensembles for each spin-labeled residue. For clarity throughout this text, we will refer to these outputs as ”coordinate models”, to differentiate them from the starting structural models. The space accessible to the unpaired electron of each residue's spin label is divided into fifty discrete pseudo-rotamers, which are shown as small spheres in Figure 4.1.A. RosettaDEER then identifies and removes pseudo-rotamers that clash with the protein backbone. Each residue's ensemble of pseudo-rotamers represents a probability density function of the space accessible to the unpaired electron of that residue's spin label. As a result, following refinement using this algorithm, the weights of a coordinate model's pseudo-

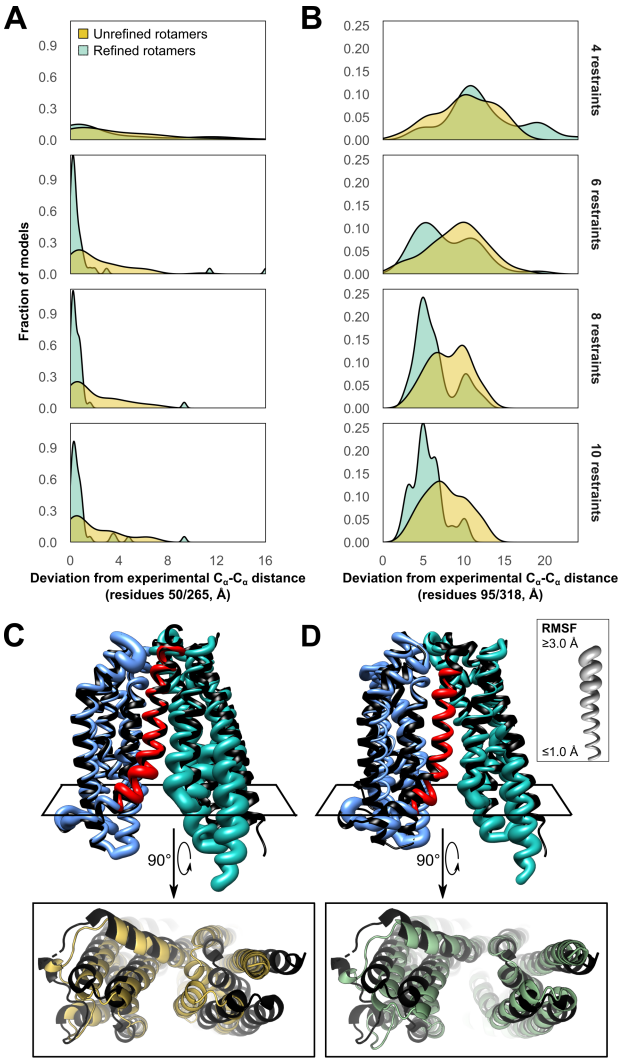

Figure 4.4: Models obtained using multilaterated rotamers more closely resemble the IF structure. Deviation between experimental and predicted C_α-C_α distances observed between pairs on the A) extracellular and B) intracellular sides. Models obtained using ten restraints either with C) default or D) refined rotamers (IF structure in black).

rotamers for any given residue are tightly coupled to those of other residues.

In this study we focus our attention on coordinate models with high parsimony. For example, coordinate models capable of recapitulating DEER traces using only one pseudo-rotamer per residue are prioritized over those with two or more. However, if the DEER trace indicates a broad and multimodal distribution, additional pseudo-rotamers may be necessary to improve the goodness-of-fit. The total number would ideally be no greater than the minimum required to fit the data, and multiple combinations of pseudo-rotamers may be equally consistent with the data. We identified parsimonious coordinate models using the AICc [7, 59, 395]:

$$AICc = -2 * \ln(L(\hat{\vartheta}|D) \tag{4.1}$$

This metric balances two competing objectives of 1) fitting the experimental data as well as possible and 2) simplifying the model as much as possible. The leftmost term, goodness-of-fit, is expressed as the maximum likelihood estimate of the coordinate model with parameters ϑ given the experimental DEER data D and is described below. The middle and rightmost term express the complexity of the model, with the variable K corresponding to the total number of parameters in the coordinate model and n_{total} corresponding to the total number of time points in the experimental DEER data. K includes the number of pseudo-rotamers with nonzero weights, as well as the number of parameters required to fit the intramolecular DEER data in the time domain. The rightmost term, which converges to zero as the data-to-parameter ratio increases, serves as further regularization in modeling cases where less experimental data is available (in this case corresponding to the number of time points in all DEER traces). Overall, the AICc quantifies the expectation that few spin label rotamers contribute to the distance distribution.

4.3.2 Detailed description of the multilateration algorithm

The multilateration algorithm is implemented in Rosetta [229, 234] as part of the RosettaDEER package and can be run using RosettaScripts [132]. It uses an iterative simulated annealing approach and is therefore non-deterministic. As a result, it obtains diverse sets of solutions when executed

multiple times. However, there is no guarantee that the global minimum solution is obtained using this algorithm.

The positions of the pseudo-rotamers are kept fixed in space throughout the duration of the algorithm, e.g., they are reweighted, rather than moved. Initial positions are obtained from the nitroxide bond midpoints of each rotamer in the Rosetta MTSSL rotamer library following clash evaluation [10]. At the start of the algorithm, one of these pseudo-rotamers is randomly chosen for each residue and has its weight set to 1; the rest have weights set to zero.

The algorithm then proceeds as follows:

- The weight of a randomly chosen pseudorotamer is modified by a randomly chosen number. Initially this value ranges uniformly from -0.1 to 0.1.
- The weight change is applied, and the resulting sum-of-squared residuals is calculated as discussed below.
- Any move that decreases the sum-of-squared residuals is accepted, while any move that increases it is accepted with the following probability (with iter being the current iteration):

$$p_{accept} = \exp\left(-\frac{\ln\left(L\left(\vartheta_{\text{iter}+1}|D\right)\right) - \ln\left(L\left(\vartheta_{\text{iter}}|D\right)\right)}{k_{\text{B}}T}\right) \tag{4.2}$$

- The Boltzmann temperature $k_{\text{B}}T$ starts at 1.5 and asymptotically approaches zero with each iteration as the algorithm proceeds. A total of 2500 trials per round are performed per DEER trace in the dataset. However, each round is aborted if 500 consecutive trials fail to sample an improvement.
- At the end of each round, the temperature $k_{\text{B}}T$ is raised to 1.5. If no improvements were sampled, the magnitude of the weight changes made to coordinates is reduced by a factor of $\sqrt{10}$. Once this magnitude reaches 10^{-4}, the algorithm is concluded.

For PfMATE, we used a non-three-dimensional background model to fit the intermolecular contribution of the experimental signal. This required a modification to the algorithm in which the

first round of optimization was performed using a three-dimensional background. The first time $k_B T$ was reset to 1.5, this restriction was removed. Otherwise, the dimensionality of the intermolecular background coupling was found to immediately drop to a value of 2, trapping the solution in a local minimum.

4.3.3 Simulation of DEER distance distributions

To simulate distance distributions between two spin-labeled residues u and v, pairwise distances were measured between all coordinates belonging to each residue. To account for backbone heterogeneity, each of these measurements were then broadened by a value equal to the pairwise root mean squared fluctuation (RMSF) as inferred from the crystallographic isotropic B-factor of the residues' C_α atoms:

$$RMSF_u = \sqrt{\frac{3B_{u,C_\alpha}}{8\pi^2}} \tag{4.3}$$

$$RMSF_{uv} = \sqrt{RMSF_u^2 + RMSF_v^2} \tag{4.4}$$

The result is equivalent to the convolution of the original distribution with a Gaussian distribution with a width of $RMSF_{uv}$. Regions of proteins with higher B-factors, such as loops, have previously been found to exhibit a greater degree of backbone flexibility in solution [34, 326, 461]. Failure to account for backbone flexibility could potentially overstate the intrinsic dynamics of the spin label and decrease the precision of the models generated using the pseudo-rotamers obtained this way. We did not normalize the experimental B-factors to account for differences in experimental crystallographic resolution, since such differences may reflect variations in the backbone disorder of different proteins.

4.3.4 Evaluating coordinate models obtained from raw DEER traces

The data D consist of N decay traces (V_{exp}), e.g., $D = V_{exp,1}, V_{exp,2}, ..., V_{exp,N}$, with the ith decay trace consisting of n_i time points for a total of n_{total} experimental time points among all experimental

traces. In this case, the likelihood of the model was evaluated by the noise-normalized sum-of-squared residuals to the experimental data:

$$\ln\left(L\left(\vartheta|D\right)\right) = -\frac{n_{\text{total}}}{2} * \ln\left(\frac{1}{n_{\text{total}}}\sum_{\text{i}=1}^{N}\sum_{\text{i}_t}^{n_{\text{i}}}\left(\frac{V_{\text{exp,i}}\left(t_{\text{i}_\text{t}}\right) - V_{\text{intra,i}}\left(t_{\text{i}_\text{t}}|\vartheta\right)}{\sigma_{\text{i}}}\right)^2\right) \tag{4.5}$$

Here σ_{i} is the standard deviation of the noise corresponding to the ith decay trace, $V_{\text{exp,i}}\left(t_{\text{i}}\right)$ refers to the experimental data at the i_{t}th time point of decay trace i, and $V_{\text{intra,i}}\left(t_{\text{i}_t}|\vartheta\right)$ refers to the value of the simulated data in decay trace i at time point it given the model parameters ϑ. The values of σ_{i} were calculated from the imaginary component of each DEER trace. Normalizing the data to the noise was necessary to satisfy the assumption that the sum of squared residuals is independently and identically distributed. Forgoing this correction led to overfitting of noisier DEER traces and underfitting of less noisy traces.

Simulation of DEER traces occurred in three steps. First, the distance distributions were obtained from the model coordinates as described above. Second, the intramolecular form factor was calculated for each time point t_{i_t}:

$$V_{\text{intra,t}}\left(t_{\text{i}_\text{t}}|\vartheta\right) = \sum_{\text{j}=1}^{m} P_{\text{sim,i}}\left(r_{\text{j}}|\vartheta\right)\int_0^{\frac{\pi}{2}} \sin\left(\theta\right) * \cos\left(\frac{\left(1-3\cos^2\theta\right)\mu_0\mu_{\text{B}}^2 g^2 t_{\text{i}_\text{t}}}{4\pi\hbar r_{\text{j}}^3}\right)\text{d}\theta \tag{4.6}$$

Here, g is the electron g-factor, μ_0 is the vacuum permeability constant, μ_{B} is the Bohr magneton, t_{i_t} is the i_{t}th time point in microseconds, r is the bin distance in nanometers, and θ is the angle between the bulk magnetic field and the interspin vector.

In the third step, the modulation depth, background slope, and dimensionality (in the case of PfMATE) were determined using nonlinear least-squares minimization. This background was modeled using the stretched exponential function $B\left(t\right) = exp\left(-\left(kt\right)^{\frac{d}{3}}\right)$, where d refers to the dimensionality of the background coupling and was constrained to 3.0 for T4 Lysozyme and to between 2.0 and 3.5 for PfMATE. In the latter case, we generally obtained values ranging from 2.0 to 2.5. These parameters were determined using an initial search as previously described and were fine-tuned throughout the duration of the algorithm using the Levenberg-Marquardt algorithm.

4.3.5 Determination of distance distributions

We used GLADDvu [173] and DeerAnalysis2019b [192] to fit the data and obtain distance distributions. Each DEER trace was truncated by 500 ns to avoid fitting artifacts. Sum-of-Gaussian distributions were obtained with GLADDvu using the interior point method. The distribution with the lowest Bayesian Information Criterion was selected. Distributions were also obtained using Tikhonov regularization with an L-curve criterion with default settings, as well as the generic DeerNet neural network ensemble, using DeerAnalysis2019b. Confidence bands and/or error margins were obtained using the delta method for GLADDvu, the Validation tool for Tikhonov regularization, and built-in ensemble statistics for DeerNet.

4.3.6 Application to T4 Lysozyme and PfMATE

The algorithm as described above was applied to T4 Lysozyme (PDB: 2LZM) and OF PfMATE structure (PDB: 6GWH). For PfMATE, the data were further separated into the extracellular restraints and the intracellular restraints. The algorithm was executed one thousand times for each of these three datasets. Each of the one thousand coordinate models were scored using the AICc (Equation 4.1).

4.3.7 Modeling the OF-to-IF conformational change of PfMATE

Modeling the outward-to-inward conformational change of PfMATE was achieved using an MC fragment insertion approach implemented in RosettaScripts. This protocol randomly changes the backbone dihedral angles of certain residues chosen at random to match those of a similar stretch of residues found in protein structures deposited in the PDB. Only residues 1–50 and 241–268 were perturbed. Peptide fragments were obtained from a July 2011 version of the PDB using the Robetta web server [219] with homologous protein structures removed. The fragment insertion protocol was executed 1000 times in RosettaScripts using the *score3* scoring function and was repeated for 5000 cycles. The Boltzmann temperature was set to 1.0. The following scoring function was then used to quantify the similarity between the experimental and simulated DEER distributions:

$$S_{\mathrm{DEER}} = \sum_{i=1}^{N} \ln\left(\sum_{j=1} p_{\mathrm{sim,ij}} p_{\mathrm{exp,ij}}\right) \tag{4.7}$$

In the event that an experimental and simulated distribution did not overlap, the inner term resolves to $\ln(0)$. Under these circumstances, this value was automatically set to -87.0, which is equivalent to the natural logarithm of the lowest non-negative value that can be described by a single-precision floating point number. The choice of this scoring function is discussed in Appendix B.

4.4 Acknowledgements

We thank Dr. Derek P. Claxton and Dr. Richard A. Stein for critical reading of the manuscript, and Dr. Eric Hustedt for both fruitful discussions regarding the AICc and model-based fitting as well as critical reading of the manuscript.

CHAPTER 5

Structural dynamics of the glutamate-GABA antiporter GadC

This Chapter is based on unpublished data.

5.1 Introduction

Transporters belonging to the APC family shuttle amino acids and their derivatives such as hormones and polyamines through lipid bilayers in organisms across all domains of life [203, 423]. Some APC transporters mediate cation-independent substrate exchange, or antiport, across cell membranes [23, 111, 367], and in humans their upregulation correlates with poor prognosis in a wide variety of cancers [203]. Homologous antiporters allow bacteria to withstand extreme acid stress by importing and exporting the precursors and products, respectively, of proton-consuming amino acid decarboxylases [137, 162, 204]. Of these four "virtual proton pumps" found in the pathogenic *Escherichia coli* strain O157:H7, the Glu/GABA antiporter GadC operates at the lowest pH range [141, 204, 254]. Unlike the others, its knockout sensitizes cells to extremely acidic conditions (pH 1.5-4.0) and sharply decreases host infection and mortality [251]. Experimental studies of GadC may thus reveal both how deadly pathogens involved in food-borne illness survive at low pH and how eukaryotic homologs with disease relevance transport their substrates.

Functional characterization of GadC revealed a stringent dependence of activity on pH, with little to no detectable transport under neutral or weakly alkaline conditions [253, 254, 416]. Its structure, determined by X-ray crystallography in detergent micelles at pH 8.0, was putatively assigned to an inactivated state incapable of substrate translocation. Interestingly, a C-terminal domain unique to GadC was found embedded in the intracellular cavity, suggesting that pH-dependent inactivation appeared to be in part facilitated by autoinhibition. Although mild transport activity was observed under neutral conditions in deletion mutants lacking this domain, detachment of the C-terminus has never been directly detected.

Perhaps more importantly, the structure and dynamics of GadC as it undergoes amino acid ex-

change remain unknown. It has been assumed that the conformational cycles of APC transporters are expected to follow in outline those of well-studied LeuT-fold transporters such as neurotransmitter-sodium symporters and sodium-solute symporters, which couple the otherwise unfavorable import of substrates to inward electrochemical gradients of sodium ions [213]. However, studies of these transporters using solution-state methods such as EPR spectroscopy [76, 212, 213, 310] and HDX/MS [6, 281, 292, 298] have revealed striking divergences in both their elements of alternating access as well as its and their ligand dependence. Moreover, as symporters, NSSs, SSSs, and others undergo fundamentally different transport cycles than GadC [134]. In contrast, the conformational dynamics of antiporters with this fold such as GadC remain understudied and unknown.

Here, DEER spectroscopy [90, 186, 274] and integrative modeling [342, 410] are used to investigate and model the pH-dependent structural changes of GadC in a native lipid environment. We directly detect detachment of the C-terminus at low pH and observe increases in conformational heterogeneity among neighboring helices. Unlike homologous symporters, GadC did not undergo large-amplitude substrate-dependent conformational changes at low pH, which may indicate that its antiport mechanism, rather than resulting from ligand-dependent conformational changes observed in unrelated antiporters [180, 267, 270], could instead stems from stabilization of a transition state that is inaccessible in the absence of substrate. Structural models generated from these distance measurements deviate from the published crystal structure in key aspects and indicate that GadC predominantly adopts an inward-facing occluded conformation in lipid bilayers [111, 367].

5.2 Results

The pH-dependent activity profile of GadC was verified by measuring radiolabeled substrate uptake into proteoliposomes. A construct of wildtype GadC, previously cloned from *E. coli* str. O157:H7, was obtained and expressed in *E. coli* C43 (DE3), purified in β-DDM detergent micelles, and reconstituted into proteoliposomes containing 5 mM Glu at pH 5.5. These proteoliposomes were then tested for substrate transport by detection of [^{3}H]-L-glutamic acid uptake as a function of both external pH and substrate concentration. Additionally, time-dependent Glu transport was measured

in proteoliposomes containing 5 mM GABA at pH 5.5 (Figure G.1). Consistent with previous findings [253, 254, 416], we observed a strong dependence of radioligand uptake on pH, with negligible transport detected at pH 6.5 and above (Figures 5.1.B and C).

To characterize the structural changes associated with pH-dependent activation, we used site-directed spin labeling and EPR spectroscopy [186, 274]. After mutating all three endogenous cysteines in the wildtype sequence to chemically inert residues (C60V, C247A, C380V), a panel of 25 single- and double-cysteine mutants were generated. As with previous studies on structural homologs of GadC, double-cysteine pairs were selected based on their ability to report on inter- and intra-domain motions. To evaluate if these measurements were expected to fall within the detectable range for DEER measurements (15 Å to 60 Å) and to test whether the resulting data were consistent with the crystal structure, distance measurements were first simulated between candidate residue pairs using dummy spin labels modeled over the crystal structure [176, 195]. Following purification and spin-labeling, all mutants were reconstituted into proteoliposomes and tested for transport and pH-dependent inactivation at pH 5.5 (Figure G.2) and 7.5 (Figure G.3), respectively. Additionally, all experimental DEER measurements were carried out in nanodiscs with lipid profiles matching those of the proteoliposomes used for transport assays. This ensured that neither the spin labels nor the membrane environment interfered with the protein's ability to traffic substrates at acidic pH or to undergo inactivation at neutral pH.

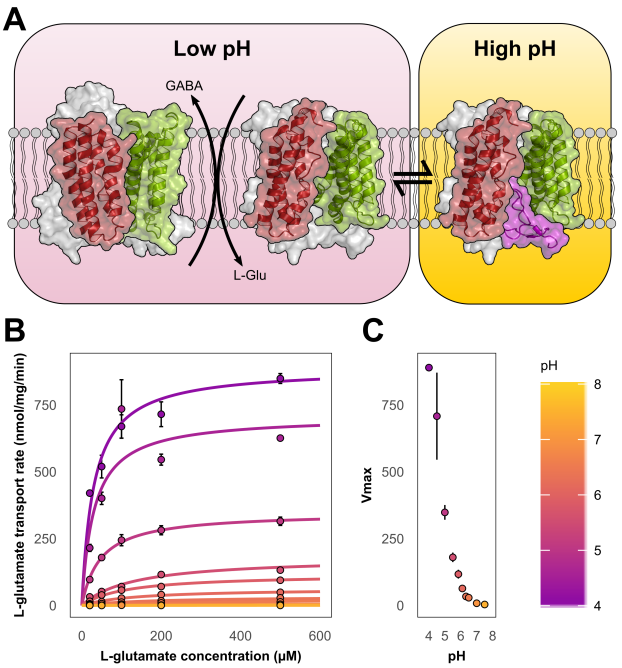

Figure 5.1: Transport activity in the glu/GABA antiporter GadC is dependent on pH. (A) Cartoon depiction of the activation mechanism. (B) Glutamate transport in wildtype GadC reconstituted into proteoliposomes is strongly dependent on pH. (C) Maximal transport rate (Vmax) increases exponentially as a function of pH.

Although GadC was crystallized as an antiparallel homodimer, no experimental evidence of this quaternary assembly was detected in lipid nanodiscs (shown below). Additionally, the addition of substrates induced neither large-scale conformational modulations nor changes in continuous-wave EPR lineshape data (representative pairs shown in Figure G.4). As a result, the following discussion is limited to pH-dependent structural changes.

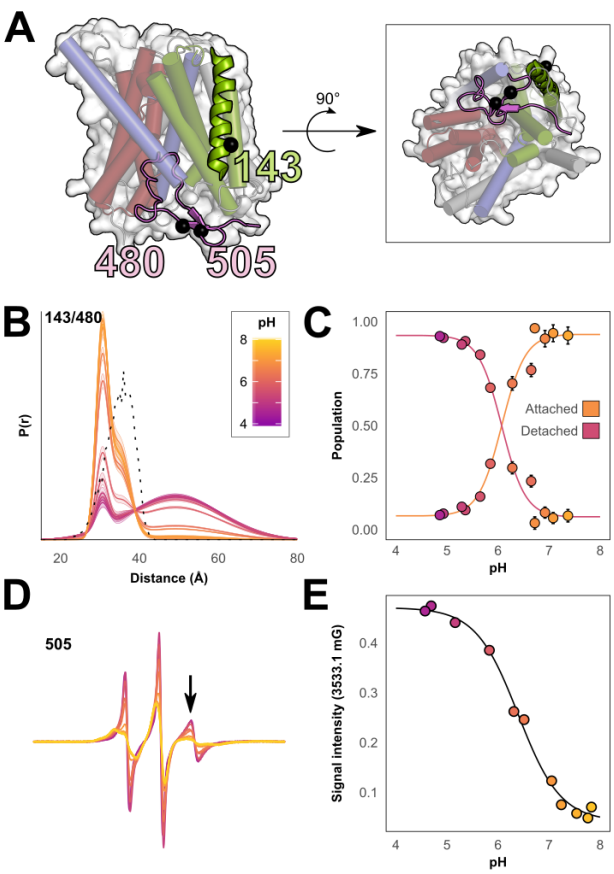

Figure 5.2: Detachment of the C-terminus is triggered by low pH. (A) Position of the C-terminus, shown in pink, relative to the main transmembrane domain of the transporter. Inset: The domain is embedded into the intracellular vestibule. (B) At low pH, a distance component consistent with a distance distribution predicted from the crystal structure (shown in the dashed line) is replaced by a wider, longer-distance component. (C) Titration measurement of the dissociation of the C-terminus. (D) pH-dependent increases in conformational heterogeneity resolved by CW-EPR. (E) Titration measurement of the third moment of the CW spectra reveals a similar profile to the DEER measurements.

5.2.1 Monitoring the detachment of the C-terminus as a function of pH

Abrogation of transport at neutral and alkaline pH has previously been attributed to a coiled domain at the protein's extreme C-terminus (shown in pink in Figures 5.1 and 5.2.A). In the crystal structure of GadC, captured at pH 8.0, this domain is embedded in the intracellular cavity and putatively obstructs closure of the intracellular gate, a prerequisite of alternating access. To test the hypothesis that this domain detaches under acidic conditions, a double-cysteine mutant (143C/480C) was generated and spin-labeled to measure the distance between the C-terminus and the transmembrane domain. Distance distributions of this pairs reported large changes as a function of pH. At neutral pH, the average distance matched that predicted

from the crystal structure, whereas a sharp increase in both the magnitude and the width of the distribution was observed under acidic conditions (Figure 5.2.B, raw data shown in Figure G.5). A nonlinear least-squares fit of a sigmoid function to the amplitudes of the short- and long-distance components revealed that this shift occurred cooperatively with a pKa of 6.07±0.11 (Figure 5.2.C), with both short-distance components diminishing at low pH in a tightly correlated manner (Figure G.6).

To further determine if this cooperative distance change originated from conformational disorder of the C-terminal domain, a second single-cysteine mutant was introduced near the C-terminal extreme of the domain (505C). At neutral pH, the lineshape of this mutant's continuous wave (CW)-EPR spectrum suggested that the domain was relatively structured, consistent with its docked conformation in the crystal structure. By contrast, reducing the pH led to a sharp spectral component that dominated the lineshape at pH 6.0 and below, suggesting increases in the C-terminal domain's disorder and mobility. Nonlinear least squares fit of the signal's intensity of the high field line on as a function of pH yielded a pKa of 6.30±0.04, consistent with the DEER measurements discussed above (Figure 5.2.D and E). Taken together, the data are consistent with the corroborate the hypothesis that the tail detaches from the transmembrane domain and becomes heterogeneous and disordered. Additionally, the pKa of this event closely matched the pH at which transport activity is abrogated, reinforcing this domain's role in regulating substrate exchange under neutral pH conditions.

5.2.2 Characterization of structural changes in the transmembrane domain induced by low pH

To determine if low pH drives additional structural changes following detachment of the C-terminal domain, a systematic analysis of the structure of GadC was undertaken. A variety of motifs defining conformational heterogeneity and alternating access have been observed in structural homologs using both high-resolution methods, such as crystallography and cryo-EM [52, 211, 224, 230, 246, 316, 374, 444, 455, 459], as well as solution-state methods such as EPR spectroscopy [76, 212, 213,

310], FRET [475], and HDX/MS [6, 281, 292, 298]. Depending on the transporter, isomerization between inward- and outward-facing conformations appears to be largely mediated by combinations of rigid-body movements between the bundle and hash domains, movement of EL2 and unwinding of TMH5, and/or independent movement of TMH1, TMH6a, TMH7, and/or EL4. We thus set out to test if movement in these motifs underpin activation of GadC at low pH.

5.2.3 The bundle domain is tilted relative to the crystal structure

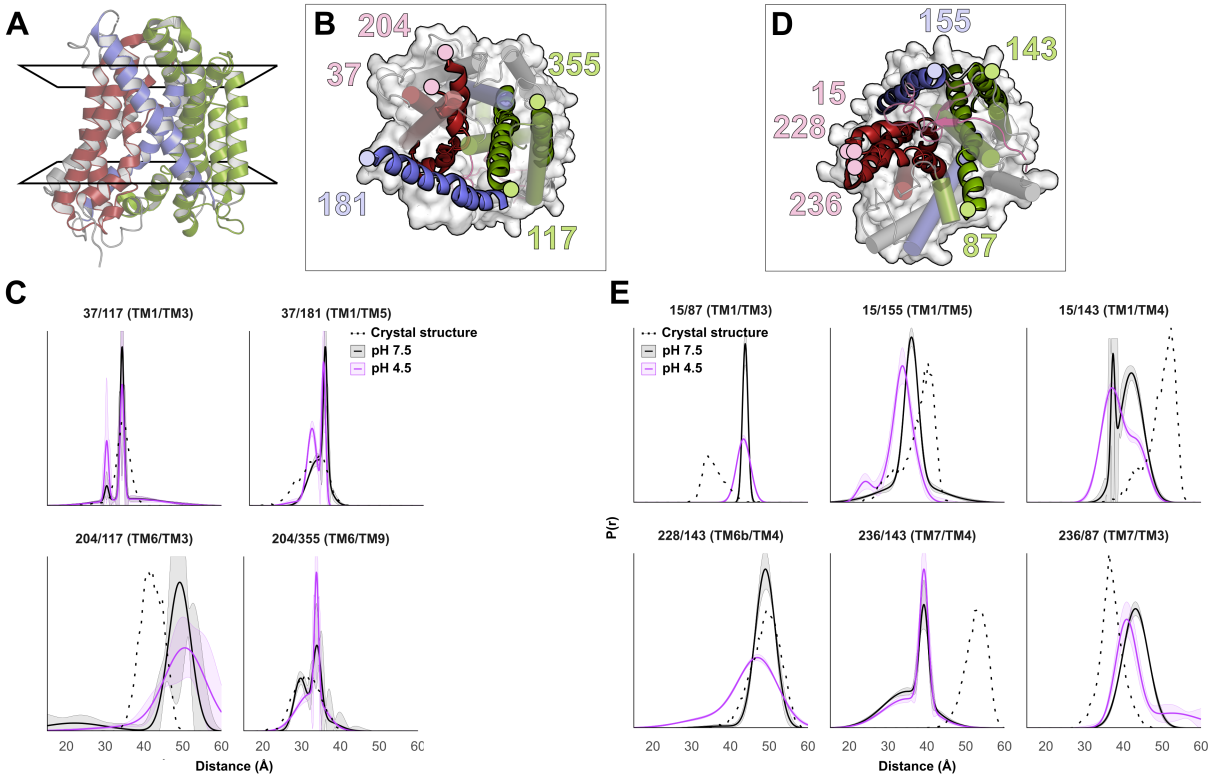

Figure 5.3: Distance measurements between the bundle and scaffold domains deviate from crystal structure. (A) Side view of GadC with the top and bottom slices corresponding to panels B and D, respectively. (C) Distance measurements carried out at pH 4.5 and 7.5 alongside predictions made from the crystal structure. Confidence intervals (95%) shown as shaded regions. (E) Distance measurements on the intracellular side reveal striking inconsistencies with the crystal structure.

A series of measurements were carried out between spin labels attached to the bundle domain and either the hash domain or TMH5 on the extracellular side of the GadC (Figure 5.3.A). Distance distributions collected at pH 7.5 were largely in agreement with predictions made from the crystal structure, in which the extracellular gate was fully closed (Figure 5.3.B). Consistent with this find-

ing, CW profiles involving residue 37, located on the loop connecting helices 1 and 2, were broad and indicative of a constrained environment (Figure G.7). Additionally, cysteine mutants labeled at this residue showed partially abrogated transport at low pH (Figure G.2). Only minor changes were observed in measurements at pH 4.5, suggesting that the extracellular vestibule remained closed even at low pH. Although broadening was observed between TMH3 and TMH6, the considerable overlap between the 95% confidence intervals for both distributions prevented us from definitively concluding that this resulted from conformational rearrangements. The remainder of the data indicate that at both low and high pH, the extracellular gate is consistent with the outward-closed conformation observed in the crystal structure (Figure 5.3.C).

In contrast to these observations, distance measurements carried out on the intracellular side of GadC deviated substantially from predictions made from the crystal structure (Figures 5.3.D, 5.3.E, and G.8). At pH 7.5, we observed that TMH1 and TMH7 in the bundle domain were 10 Å to 15 Å farther from TMH3, and 10 Å to 20 Å closer to TMH4 and TMH5, than the predictions made from the crystal structure. Lowering the pH to 4.5 caused these distance distributions to broaden and, in some distributions, to further shorten by 3 Å to 5 Å. However, the pairwise nature of the data do not immediately indicate which regions of the protein A) deviated from the crystal structure, and B) underwent increases in heterogeneity.

5.2.4 The scaffold domain is largely consistent with the crystal structure

As only minor pH-dependent movement was observed between the bundle and hash domains, subsequent measurements focused on EL4 and IL1, which connected helices 7 and 8 and 2 and 3, respectively. EL4 has been shown to pivot outward and provide access to the extracellular vestibule in LeuT [76, 213], whereas transport activity data suggest that IL1 is involved in mediating pH-dependent activation in the homolog AdiC [434]. In addition to revealing whether the positions of these domains respond to changes in pH, the DEER data could further determine the extent to which the discrepancies observed between our measurements and predictions made from the crystal structure extended to the remainder of the structure.

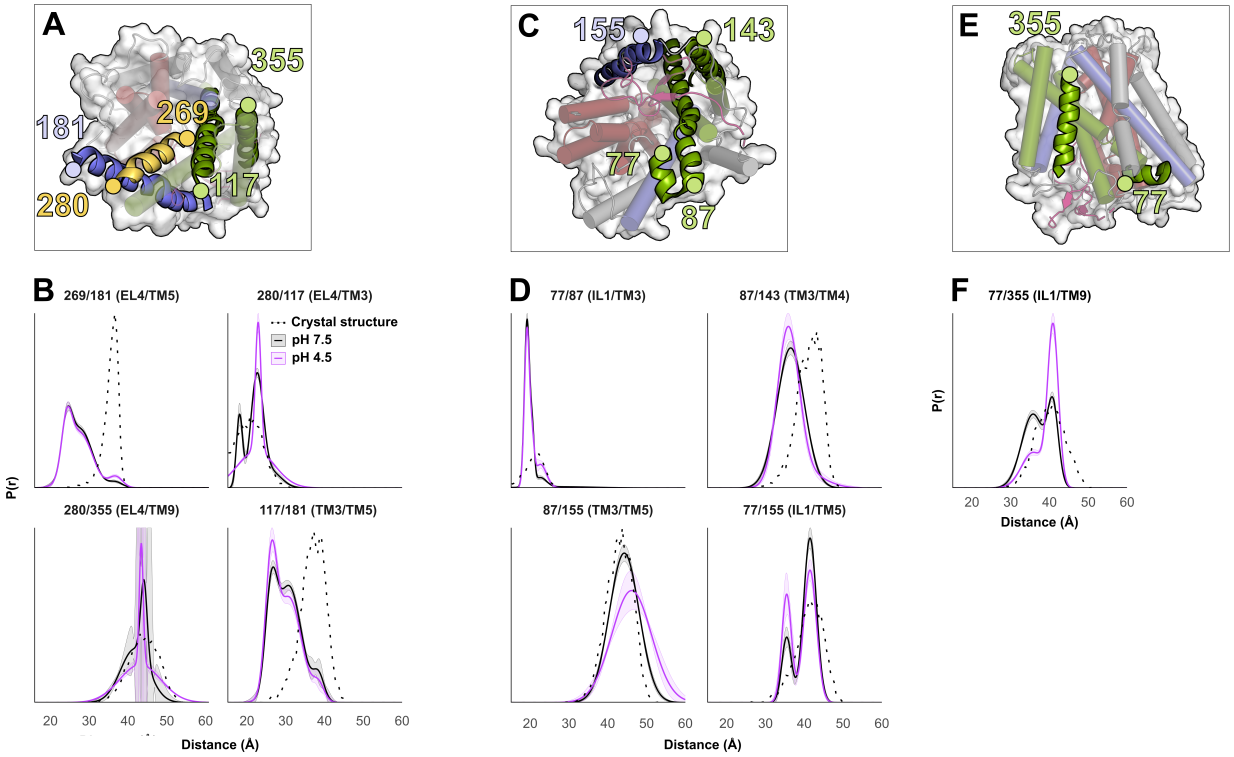

Figure 5.4: No pH-dependent movements are observed in IL1 and EL4. (A) Extracellular view of GadC. (B) Measurements between EL4 and extracellular sites on the scaffold domain suggest that TM5 is closer to the main body of the transporter than in the crystal structure. Confidence intervals (95%) shown as shaded regions. (C) Intracellular view of GadC. (D) Intracellular measurements within the scaffold domain. (D) Intracellular view. (E) Side view. (F) In-to-out measurement suggests a slight pH-dependent movement in TMH9.

On the extracellular side, we collected three distance distributions between EL4 and various points in the structure (Figures 5.4.A, 5.4.B, and G.9). None of the distributions indicated a large-amplitude distance change as would be expected from EL4-mediated opening of the extracellular vestibule. However, whereas two of the distributions involving the C-terminal end of EL4 (residue 280) were in reasonable agreement with the crystal structure, a third pair collected between TMH5 and EL4 showed a shorter-than-expected distance distribution. To ascertain if this discrepancy was due to the position of TMH5, rather than EL4, we collected an additional measurement between TM3 and TM5 and found that it too was also shorter than expected (Figure 5.4.A). A likely explanation is that the extracellular side of TMH5 does not protrude as far as suggested by the crystal structure, and that the position of EL4 is otherwise consistent with the crystal structure.

Distance distributions between IL1 and TMH3 similarly showed little movement and suggest that the two helices are effectively stapled together, consistent with a lack of independent movement observed in other LeuT-fold transporters (Figures 5.4.C, D, and G.10). Further measurements across the hash domain on the intracellular side highlight its structural invariance. However, minor pH-dependent distance changes in IL1/TMH4 are observed and appear to accompany local reconfigurations as evidenced by the CW profile (Figure G.10). Alongside evidence that other distributions involving residue 143 on TMH4 were bimodal, these data suggest that several distinct nitroxide rotamer states may be populated and that their relative proportions shift as a function of pH. Measurements from the hash domain to TMH5, by contrast, do indicate minor changes in their relative positions. This further illustrates how changes in pH and detachment of the C-terminal tail domain coincide with subtle reorganizations of helices on the intracellular side of the protein.

Interestingly, measurements between IL1 and TMH9 on the extracellular side suggested a slight sharpening that may be pH-dependent (Figure 5.4.E and F). The similarity in the distance changes observed between both TMH6/TMH9 (204/355) and EL4/TMH9 (280/355; Figures 5.3.A and 5.4.A, respectively), combined with the lack of distance changes in measurements involving IL1, suggest that these data as slight pH-dependent movements of TMH9. However, this movement is substantially less than what is observed in structural homologs, such as Mhp1 [374, 443, 444], that use the loop connecting TMH9 and TMH10 as a hinge to facilitate entry and exit from the substrate-binding site.

Altogether, the data reveal minor structural rearrangements on the intracellular side but fail to precisely identify which domains are moving and which are stationary.

5.2.5 The bundle domain does not behave like a rigid body

Finally, we evaluated whether the bundle domain acted as a rigid body during this pH transition. The narrowness of this domain, combined with the 15 Å lower limit of DEER, meant that we could only answer this question using distance measurements between the intracellular and extracellular sides of the protein (Figure 5.5.A). Distance changes were observed in every measurement (Figures 5.5.A

and G.11), arguing against the possibility that this domain acts as a rigid body as previously posited. Additionally, they reveal minor discrepancies regarding the position of the extracellular half of TMH6, which was hinted at by measurements described previously involving the hash domain (Figure 5.3.A).

5.2.6 GadC adopts an inward-facing occluded conformation at both pHs

To summarize the experimental DEER data, we observed both substantial deviations between our observations and predictions made from the crystal structure, as well as minor pH-dependent broadening and amplitude changes. To translate these pairwise measurements into fold-level structural models, we generated a series of structural models of GadC using the modeling software suite Rosetta [229, 234]. Unfortunately, high-precision models of the structure of GadC cannot be obtained given the relatively small number of available distance restraints. Therefore, we supplemented the modeling process with a custom-designed statistical potential that quantified each model's similarity to the structures of previously determined homologs. The design of this potential, outlined in Section 5.4, capitalized on the large number of structures of homologs deposited in the Protein Databank. Effectively, this statistical potential served as a form of regularization to penalize the introduction of conformational changes that are not anticipated by the known structures of closely related homologs of GadC.

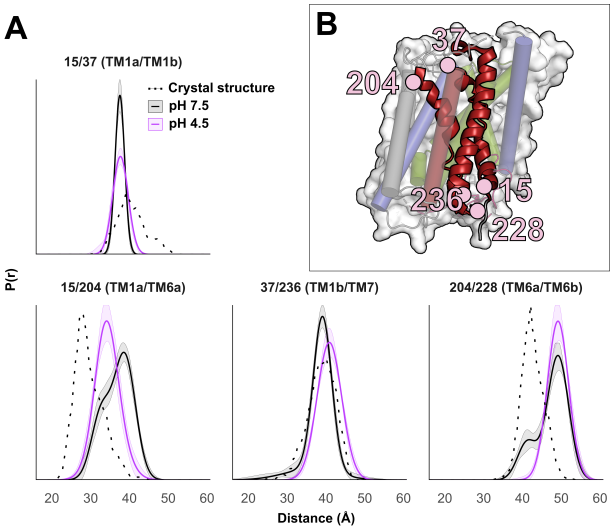

Figure 5.5: Measurements within the bundle domains are inconsistent with a rigid-body pattern of conformational dynamics. (A) In-to-out DEER distance distributions in the bundle domain. Confidence intervals (95%) shown as shaded regions. (B) Side view showing the measurements.

We generated 5,000 Rosetta models for each pH condition using a procedure discussed in detail in section 5.4 and Appendix C. Experimental DEER data was introduced as distance restraints us-

ing the RosettaDEER module (Chapter 3) and the scoring function discussed in Appendix B. Each model generated this way was reweighed based on its structural similarity to homologous transporters (discussed in section 1.4.3). The five best-scoring models generated using data collected at either pH are shown in Figures 5.6.A and G.12. These models principally deviated from the crystal structure in the positions of TMH1 and 7, although we note that the hash domain adopted a slightly more "upright" conformation with TMH4 nearly parallel to the membrane normal. By contrast, the extracellular sides of these models closely resembled that of GadC, with subtle modifications to TMH5 and TMH6a. Initial structural comparisons suggest that these models resemble the inward-facing occluded conformation of the homologs ApcT [367] and MjApcT [201], in which the bundle domain is tilted toward TMH4 and away from TMH3 relative to the crystal structure of GadC as suggested by the data (Figure 5.6). However, unlike ApcT, TMH5 in our models is not bent or puckered and instead adopts a configuration similar to that observed in the GadC crystal structure. In fact, despite differences in the bundle domain, the overall structure of our models closely resembles the structure of GadC, highlighting both the data's limited disagreement with the crystal structure and the effectiveness of our regularization strategy.

These clear deviations from the crystal structure contrasted with far more subtle differences observed between low-pH and high-pH models. In fact, although small pH-dependent distance changes were observed in the data, variation between the top five models at either pH were comparable to differences between models across different pHs (Figure G.12). Therefore, we conclude that the resolution and sparseness of the data prevent the structural basis of pH-dependent activation following release of the C-terminus from being determined with high confidence. It is therefore unclear if increases in heterogeneity observed at low pH are the result of movement in TMH5, as suggested by structures of homologous transporters, or TMH1, which is more strongly supported by the DEER data.

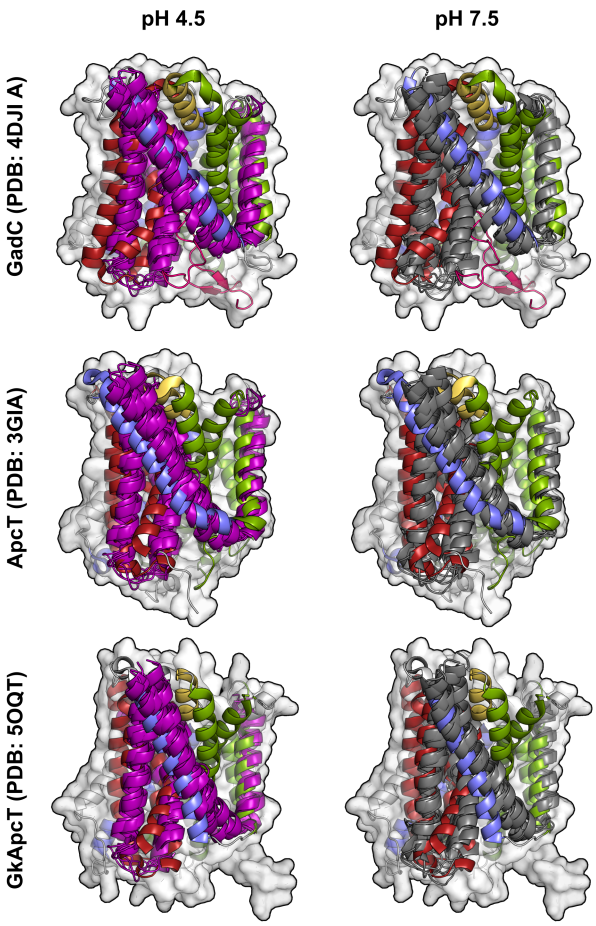

Figure 5.6: Rosetta models of the low- and high-pH conformations generated using the DEER data in purple and dark grey, respectively. Crystal structures of GadC and its homologs ApcT and GkApcT are shown in the colored helices and suggest that GadC adopts an inward-facing occluded conformation.

5.3 Discussion

The DEER data presented here investigates the structure and dynamics of the pH dependent Glu/GABA antiporter GadC at low pH. These measurements, which reflect solution-state backbone dynamics [186, 211], are inconsistent with the conformation stabilized by the crystal lattice, with substantial deviations observed on the intracellular side of the protein. Instead, models generated using Rosetta suggested that GadC adopts a conformation similar to closely related homologs in which its the intracellular vestibule of GadC is partially occluded and resemble the inward-facing occluded conformations observed in the closely related homologs. Indeed, we found that a model of

GadC generated *de novo* using the state-of-the-art method RoseTTAFold [18] generated a similarly IF-occluded model that was more consistent with the data than the crystal structure G.13. This conformation is all the more notable given the crystal structure's similarity to homologs such as BasC [111], Lat-1 [230, 459], Lat-2 [460], and $b^{(0,+)}$AT1 [449, 457]. It should be noted, however, that our measurements reflect the structure and dynamics of GadC in a more physiologically relevant lipid environment. The aforementioned homologs, by contrast, were structurally characterized in detergent micelles and/or when bound to antibodies. It should be noted that lipid/detergent-dependent conformational changes have also been reported in several transporters, including the structural homolog LeuT [330, 383], as well as other membrane proteins.

The absence of substrate-induced conformational changes observed at low pH (Figure G.4) contrasts with results from similar studies of unrelated antiporters, such as those mediating sodium- or proton-dependent drug efflux. However, this discrepancy may be explained by the fact that, unlike the substrates of ion coupled antiporters, neither glutamate nor GABA are hypothesized to move down their concentration gradients under physiological conditions. This owes to bacterial responses to acid stress that trigger rapid decarboxylation and depletion of cytoplasmic amino acids, such as glutamate, and the corresponding spike in cognate polyamines, such as GABA [137, 340]. In fact, all four of the "virtual proton pumps" involved in bacterial acid resistance, including GadC, are co-transcribed with corresponding amino acid decarboxylases [204]. This metabolomic adaptation ensures that neither half of GadC's transport cycle, glutamate import or GABA export, is energetically unfavorable. This, in turn, may would obviate the need for ligand-dependent changes in thermodynamics equivalent to, for example, the pH-dependent conformational rearrangements observed in proton/drug antiporters [89, 180, 270].

On the basis of this observation, we propose a model in which the transport cycle of GadC consists of two half-cycles of uniport in which facilitated diffusion of both Glu and GABA are coupled in opposite directions (Figure 5.7). This posits that substrate binding contributes to the kinetics, rather than the thermodynamics, of the transporter's functional cycle. We propose that this is achieved by stabilization of a high-energy transition state separating the inward-facing and outward-

facing conformations that, under apo conditions, cannot be traversed. Ligand-induced increases in conformational flexibility, which would be consistent with this hypothesis, have previously been reported in the homologous serine/threonine exchanger SteT using single-molecule dynamic force spectroscopy [36]. These findings could extend to homologous amino acid exchangers in humans, such as xCT, which couples the energetically favorable export of glutamine to the import of cystine, which is immediately reduced to two cysteine molecules and thus absent from the cytoplasm [300].

Nevertheless, as these conclusions are derived from data collected under equilibrium conditions, they do not reflect structural changes induced by substrate in the presence of gradients, such as those reported in structural homologs such as SGLT1 [249]. The possibility that the conformational dynamics of APC transporters are responsive to gradients is reinforced by evidence that homologous transporters preferentially bind substrates on specific sides of the membrane [23]. For this reason, further inquiries must determine the structure and function of GadC as it operates in and maintains a pH gradient.

5.4 Materials and Methods

5.4.1 Site-directed mutagenesis

A codon-optimized version of the GadC gene from *Escherichia coli* str. O157:H7 (Genscript) was cloned into a pET19b vector encoding an N-terminal deca-histidine tag. A cysteine-less construct (C60V, C246A, C380V) was generated from this template using site-directed mutagenesis (QuikChange). All single- and double-cysteine mutants were similarly generated from this cysteine-free construct and verified by Sanger sequencing using both T7 forward and reverse primers.

5.4.2 Expression, purification, and spin labeling of GadC

Plasmids encoding either wildtype or mutant GadC were transformed into competent *E. coli* str. C43 (DE3) cells and overexpressed in 1L minimal media A supplemented with ampicillin (Gold Biotechnology) as previously described60. Upon reaching an absorbance (OD600) of 0.7-0.8, GadC expression was induced by adding 1 mM IPTG (Gold Biotechnology) and the temperature was dropped to 20°C. Cells were harvested after 16 hours by centrifugation at 5500 g for 15 minutes,

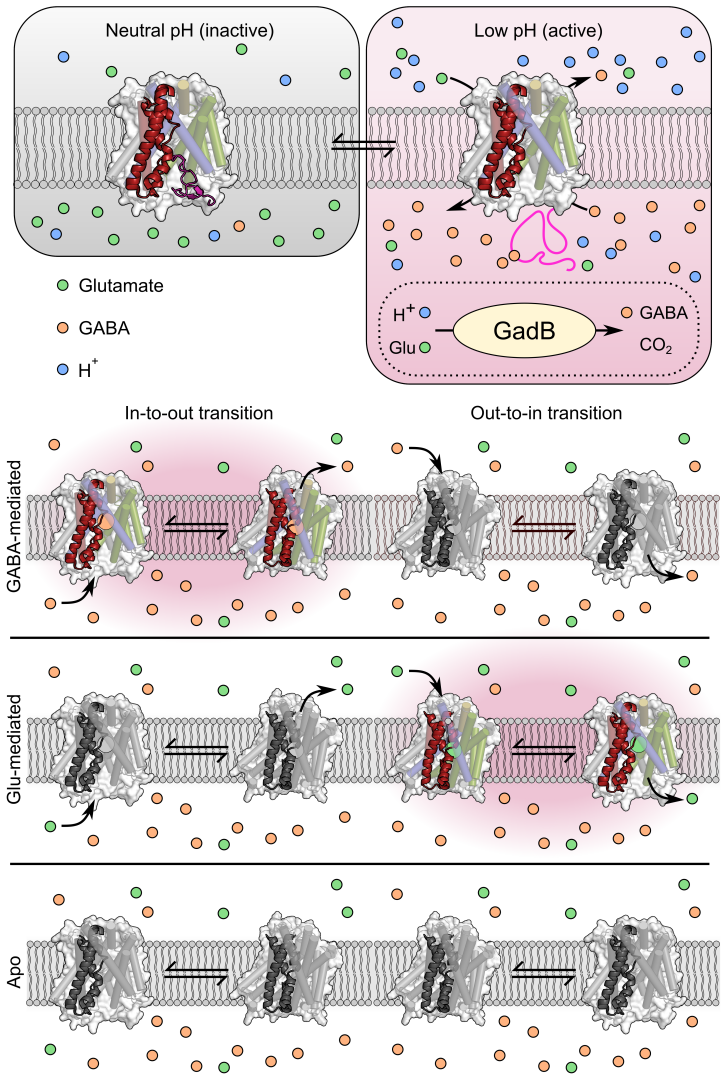

Figure 5.7: Mechanistic model of pH-dependent activation and substrate transport in GadC. (A) Under acidic conditions, the C-terminal domain of GadC detaches. Separately, GadB decarboxylates intracellular glutamate into GABA, consuming protons and increasing intracellular pH. (B) GadC undergoes two half-cycles of energetically downhill uniport. High cytoplasmic concentrations of GABA ensure that its export is the most energetically favorable means of undergoing IF-to-OF isomerization, while low cytoplasmic concentrations of glutamate ensure that its import is the most energetically favorable means of undergoing OF-to-IF isomerization.

resuspended in 22 mL lysis buffer (100 mM KPi, 10 mM DTT, pH 7.5), and lysed by sonication. After centrifugation at 9000 g for 15 minutes, the supernatant was collected and ultracentrifuged at 200 000 g for 90 minutes.

The pelleted membrane fractions were then solubilized in resuspension buffer (50 mM Tris/Mes,

200 mM NaCl, 20% glycerol, 1 mM DTT, pH 7.5) containing 1% β-DDM (Anatrace) and stirred on ice for 60 minutes. Insoluble material was removed by ultracentrifugation at 200.000 g for 30 minutes, and the supernatant was incubated with 1.0 mL Ni-NTA Superflow (Qiagen) resin at 4°C for two hours with 25 mM imidazole. After washing with ten column volumes of resuspension buffer containing 50 mM imidazole and 0.05% B-DDM, purified GadC was eluted from the resin using resuspension buffer with 250 mM imidazole and 0.05% β-DDM.

Following the addition of 60 mM Mes, single- and double-cysteine mutants were labeled with three rounds of 20-fold molar excess MTSSL (Enzo Life Sciences) per cysteine at room temperature and moved to ice overnight after four hours. Samples were then concentrated using Amicon Ultra 50.000 MWCO filter concentrators (Millipore) to a final concentration no greater than 3 mg/ml, as reported by absorbance at 280 nm (ϵ=67.840 $M^{-1}\,cm^{-1}$), and purified into 200 mM Tris/Mes, pH 7.2, 20% glycerol, 0.05% β-DDM by size exclusion chromatography using a Shodex KW-803 column with guard column. Peak fractions were isolated for further studies.

5.4.3 Reconstitution of GadC into proteoliposomes

A 3:1 ratio (weight/weight) of *E. coli* polar lipids and L-α-phosphocholine (Avanti Polar Lipids) were dissolved in chloroform and evaporated with a rotary evaporator. After overnight desiccation in a vacuum chamber, lipids were resuspended in the appropriate buffer, homogenized by ten cycles of freeze-thawing, and stored in small aliquots at -80°C.

Lipids prepared for liposomes were resuspended in 25 mM KPi, 150 mM KCl pH 5.5, and either 5 mM L-Glu or 5 mM GABA to a final concentration of 20 mg/ml (16.4 mM). Before reconstitution, lipids were diluted and destabilized with the addition of 1.25% octyl-β-D-glucopyranoside (β-OG) (Anatrace) and extruded through a 400 nm membrane filter (Whatman). Purified GadC was added to the sample at a 1:200 ratio (weight/weight), bringing the final lipid concentration to 5 mg/mL. Following a thirty-minute incubation at room temperature, detergent was removed from the sample by the gradual addition of 400 mg/mL SM-2 polystyrene Bio-Beads (Bio-Rad) over the course of four hours. After rocking overnight in the dark, the proteoliposome solution was cleared

of biobeads and ultracentrifuged at 150.000 g for 60 minutes. Proteoliposomes were then resuspended in external buffer (25 mM KPi, 150 mM KCl, pH 5.5) and ultracentrifuged to remove external substrates. After repeating this ultracentrifugation step a total of three times, proteoliposomes were suspended in external buffer at a final lipid concentration of 100 mg/ml. GadC concentration was then quantified using SDS/PAGE and densitometry (ImageJ v. 1.53g), with purified GadC in β-DDM serving as a standard curve.

5.4.4 Transport assays

In vitro transport assays were carried out either in triplicate (concentration-dependent) or in duplicate (time-dependent) as previously described [254]. An additional baseline measurement was performed on ice. Glutamic acid (between 25 μM and 1 mM) was added to external buffer and checked for pH immediately prior to all transport experiments. For the time-dependent transport Glu/GABA exchange assays shown in Figure G.1, an fixed external Glu concentration of 50 μM at pH 5.5 was used. In both experiments, proteoliposomes (2 μL) were added to external buffer (98 μL) containing 1 μCi [^{3}H]-L-glutamic acid (approximately 200 nM) and gently agitated. For titration experiments on wildtype GadC, proteoliposomes (1 μL) were added to external buffer (99 μL) containing 1 μCi [^{3}H]-L-glutamic acid. Substrate uptake proceeded for two minutes at 25°C and was quenched by adding ice-cold stop buffer (25 mM glycine, 150 mM KCl, pH 9.5) and vacuum-filtering the solution through a 0.22 μm GSTF filter (Millipore) pre-soaked in stop buffer. The filter was then washed with an additional 6 mL stop buffer, removed, and added to 5 mL Ecoscint H scintillation solution (National Diagnostics). Following quantitation, data were analyzed using Michaelis-Menten kinetics using the *curve_fit* function implemented in SciPy [427]. Baseline measurements were subtracted from the 25°C measurements.

5.4.5 Reconstitution of GadC into lipid nanodiscs

Lipids for nanodisc reconstitution were prepared as described above and resuspended in 50 mM Tris/Mes pH 7.5 to a final concentration of 20 mM. MSP1D1E3 was purified as previously described [180]. Nanodisc reconstitution proceeded using a molar ratio of 1:8 GadC:MSP1D1E3,

1:50 MSP1D1E3:lipid, and 1:5 lipid:cholate. Detergents were gradually removed from the solution using SM-2 Bio-Beads as previously described [180]. After overnight incubation, biobeads were removed from the solution using a 0.20 μm filter. Nanodisc-reconstituted GadC was then isolated from empty nanodiscs by size-exclusion chromatography using a Superdex 200 Increase 10/300 GL column into 50 mM Tris/Mes, pH 7.5, 10% glycerol and concentrated using an Amicon Ultra 100.000 MWCO filter concentrator (Millipore). The pH of all protein samples was carefully determined using a microelectrode and adjusted using 1 M citrate and 1 M Tris. Protein concentration was then evaluated using CW EPR spectroscopy as previously described [483]. Glycerol was added to all DEER samples to a final concentration of 23% vol/vol, which were then flash-frozen in liquid nitrogen prior to DEER spectroscopy.

5.4.6 CW-EPR and DEER spectroscopy and data analysis

Spin-labeled GadC was characterized using CW-EPR at 25°C using a Bruker EMX spectrometer operating at a frequency of 9.5 GHz, a 10 mW incident power, and a modulation amplitude of 1.6 G. DEER measurements were carried out using a dead-time free four-pulse protocol [304] at either 50 K (for 143C/480C) or 83 K (all other double-cysteine mutants). Pulse lengths were as follows: 10 ns to 14 ns (first $\frac{\pi}{2}$ pulse), 20 ns (second and fourth π pulse), and 40 ns (third π pulse). The pump and observation frequencies were separated by 62.26 MHz. Echo decay data were analyzed into distance distributions using GLADDvu with the last 500 ns of the signal truncated [173]. Fitting model parameters were chosen using the Bayesian Information Criterion. To analyze the pH titration distance data collected using GadC 143C/480C, the long-distance component was isolated from the two short-distance components, and was fitted with a sigmoid function using the *curve_fit* function as implemented in SciPy [427]. For all DEER pairs, the distance distributions were compared to predictions generated by MDDS, which was accessed using the CHARMM-GUI web server [195].

5.4.7 Generation of a structure-based statistical potential

After manually isolating the ten core transmembrane helices defining the LeuT-fold, each protein was aligned to each other using TM-Align version 20180426 [471]. For each model, the highest TMscore [470] to a transporter in the APC family was obtained. TMscore quantifies structural similarity and ranges from 0 to 1, with 1 indicating perfect structural overlap [453]; values between different structures of the same transporter were not considered. To reflect the probability of a model belonging to the APC family given its TMscore to structures of other proteins in that family, we used the following exponential function:

$$p(M) = \exp\left(-\frac{1-TMscore}{\beta}\right) \tag{5.1}$$

This function was parametrized by minimization of the total cross-entropy $-\sum \ln(q_i)$, where $q_i = p(M_i)$ for all APC transporters and $q_i = 1 - p(M_i)$ for all non-APC LeuT-fold transporters. Using the *minimize* function implemented in SciPy [427], a value of $\beta = 0.072185$ was obtained.

5.4.8 Initial homology modeling

Structural alignments were obtained using TMAlign. The Rosetta application *partial_thread* was then used to thread the first ten transmembrane helices (residues 1-359) of the GadC sequence over each structure. The three remaining helices were then directly grafted onto the structure by structural alignment; the C-terminal tail (residues 471-511) was omitted. Homology models were constructed using HybridizeMover [384] with five randomly selected templates as well as the hash domain from GadC (PDB: 4DJI chain A). The weight of the DEER data was set to 10.0 during the first two stages and to 0.0 during the final full-atom minimization stage. Sequence fragments used during this step were obtained from the Robetta web server [216]. Each of these models were further refined using ConfChangeMover as described below.

5.4.9 Conformational change modeling using ConfChangeMover

The structural modeling method ConfChangeMover, described in detail in Appendix C, was implemented in Rosetta 3 [229, 234] and consisted of three stages. In the first stage, rigid-body segments consisting of either beta-sheets or individual helices were identified, and cutpoints were introduced in the loops connecting them. This allowed secondary structural elements to be manipulated in three-dimensional space while avoiding the "lever-arm effect" described previously [419]. Sampling consisted of either rigid-body movements or fragment insertions. The former consisted of random rotations and translations, with an average rotation of 15° and an average translation of 2.0 Å, and the latter consisted of modifications to the backbone dihedral angles of the model. During testing, it was found that 50,000 total rounds, consisting of an even mixture of fragment insertions and rigid-body movements, was sufficient to sample a reasonable variety of conformations.

In the second stage, loops were closed using a fragment-based protocol described in detail elsewhere [384]. During this stage, sequence fragments were superimposed over regions of the protein with chainbreaks that were introduced during rigid-body movement, and Cartesian minimization was used to both minimize bond lengths and correct bond angles [341]. Additionally, contiguous regions up to fifteen residues in length were periodically taken from the starting structure and superimposed as fragments this way. We found that 1000 sampling rounds were sufficient to resolve the chainbreaks caused by the first stage.

Lastly, explicit full-atom side chains were added to the model, which was then minimized using FastRelax. An implicit membrane was introduced using RosettaMembrane [466], with membrane-spanning regions determined using OCTOPUS [426].

Several types of restraints were used to drive the model toward conformations consistent with the DEER data while maintaining the LeuT-fold topology of the starting model. During the first two stages, models were restrained using the experimental DEER restraints as implemented in RosettaDEER. A weight of 10.0 was given to this score term. Agreement with the experimental distribution was quantified using the probability function:

$$S_{\text{DEER}} = \sum_{i=1}^{N} \ln\left(\sum_{j=1} p_{\text{sim,ij}} p_{\text{exp,ij}}\right) \tag{5.2}$$

This function reflects the overlap between the experimental and simulated distributions. In the event that any individual simulated distribution does not overlap with its experimental counterpart, the innermost term resolves to $\ln(0)$. To avoid arbitrarily large scores, we automatically set this number to -87.0, which is approximately the negative logarithm of the smallest non-negative value that can be represented by a single-precision floating point number.

To account for the relative invariance of backbone dihedral angles in experimentally observed conformational changes, the model's backbone dihedral angles ϕ_{sim} and ψ_{sim} in radians were restrained using the following circular sigmoid functions:

$$S_{\phi}(x) = \left(1 + \exp\left(|\phi_{\text{sim}} - \phi_{\text{exp}}| - \frac{\pi}{2}\right)\right)^{-1} + \left(1 + \exp\left(|\phi_{\text{sim}} - \phi_{\text{exp}}| + \frac{\pi}{2}\right)\right)^{-1} \tag{5.3}$$

$$S_{\psi}(x) = \left(1 + \exp\left(|\psi_{\text{sim}} - \psi_{\text{exp}}| - \frac{\pi}{2}\right)\right)^{-1} + \left(1 + \exp\left(|\psi_{\text{sim}} - \psi_{\text{exp}}| + \frac{\pi}{2}\right)\right)^{-1} \tag{5.4}$$

Here ϕ_{exp} and ψ_{exp} refer to the backbone dihedral angles in the starting model. This potential minimized the introduction of unnecessary changes to backbone dihedral angles resulting from fragment insertion. These potentials were further limited to regions of the protein with secondary structure during the first stage and to loops during the second stage.

Finally, between the first two stages of modeling, coordinate constraints were placed on the C_{α} backbone atoms belonging to secondary structures. This minimized the probability of reversion to the initial starting pose. Both the dihedral and coordinate constraints were maintained during the full-atom minimization and were given weights of 1.0 throughout conformational change modeling.

CHAPTER 6

Perspectives and future directions

6.1 Synopsis of experimental findings

This document presents experimental research into the pH-dependent activation mechanism and conformational dynamics of GadC, a ”virtual proton pump” found in the APC transporter family. Additionally, it describes a series of structural models generated using these experimental data that attempt to recapitulate the conformation adopted by GadC in solution. While the structural dynamics of more distant homologs, such as prokaryotic and eukaryotic NSSs and SSSs, continue to be extensively studied, prior to this research only a single investigation in a member of the APC family, the homologous serine/threonine antiporter SteT, had been conducted. It is noteworthy that although the work discussed in this dissertation coincided with several breakthroughs in membrane protein structural biology (discussed below), our body of knowledge with respect to structural dynamics of LeuT-fold antiporters barely grew. High-resolution structures of eukaryotic exchangers with disease relevance, including Lat1 and Lat2, were unaccompanied by detailed descriptions of how or even whether they isomerize when bound to ligands.

Given the sensitivity of structural homologs LeuT, Mhp1, and vSGLT to ligands, it came as a surprise that no ligand-dependent conformational dynamics were observed in GadC. Sodium-coupled symporters rely on ion gradients to drive the energetically unfavorable uptake of amino acids and other nutrients into the cell. By contrast, in the low-pH conditions under which GadC is hypothesized to be active, depletion and buildup of intracellular glutamate and GABA, respectively, lead both substrates to be transported down their concentration gradients. The environmental concentrations of both substrates on either side of the cell may thus be sufficient to enforce productive substrate movement required by the bacterial cell under low-pH conditions. Thus, the data presented in Chapter 5 and Appendix G of this dissertation reinforce the structural and conformational diversity of transporters with the LeuT-fold, further highlighting the extent to which the energy

landscapes of proteins with the same fold can diverge. A key takeaway from this work is that experimental evidence of ligand-dependent conformational dynamics in one transporter may not fully correspond with those of structural homologs; in other words, conservation of transporter dynamics at the family-level may not be guaranteed. An important implication of this conclusion is that it may suggest that the conformational dynamics of transporters in eukaryotic organisms in general and humans in particular may not match those of bacterial model systems (see section 1.5). For this reason, it remains unclear whether these results extend to human amino acid exchangers with disease relevance such as Lat1 [230, 459] or xCT [300]. Ultimately this hypothesis will be tested as investigations into human proteins become more widespread.

6.1.1 Perspectives on the effect of substrates on the conformational dynamics of GadC

The DEER data presented here suggests that neither substrate biases the conformational dynamics of GadC, a finding which led us to postulate that glutamate and GABA affect the protein's kinetics, rather than its thermodynamics - such a mechanism would almost certainly be missed by measurements carried out using the DEER technique. Under the proposed mechanism, which is outlined in section 5.3, the contribution of substrate binding to conformational dynamics is limited to stabilization of a hypothetical high-energy transition state that is not traversable under apo conditions. This model, although simple, would explain both the antiport mechanism forbidden substrate-free isomerization as well as the absence of any substrate-mediated changes in the conformational dynamics of GadC observed using DEER. Nevertheless, as the DEER technique only interrogates protein thermodynamics, rather than kinetics, other experiments would be required to test this hypothesis. Direct evaluation of changes in protein kinetics can be achieved using HDX/MS [301], single-molecule FRET [361], or fluorine NMR [261].

6.1.2 Perspectives on the pH-dependent activation mechanism of GadC

This directly ties into outstanding questions regarding the mechanism of pH-dependent substrate transport. Data collected in radioligand transport assays show how activity spikes at low pH and does not appear to plateau in the pH range under experimental observation (Figure 5.1.B and C).

By contrast, detachment of the C-terminus appears to occur with a pKa of 6.0 to 6.25 (Figure 5.2.B, C, and D), which likely rules out the contribution of this domain to the increase in transport rate observed at pH 4.0-5.0. In fact, the change in glutamate exchange observed at pH 6.0-6.5 is negligible compared to the high rates of transport observed at lower pHs, which calls into question the role of this domain in regulating transport at neutral pH.

Two questions naturally follow this line of thinking. First, what other mechanism could explain the pH-dependent activity observed in GadC? Although protonation of the substrates' γ-carboxylate (pKa: 4.25), a prerequisite for transport, may partially explain this phenomenon, abrogated transport of glutamine, which mimics protonated glutamate, at neutral and alkaline pHs is inconsistent with this hypothesis [253]. In the homologous arginine/agmatine antiporter AdiC, activation has been attributed to the proton sensing residue tyrosine Y74, located on the intracellular amphipathic helix connecting TMHs 2 and 3 [434]. Whereas the wildtype similarly undergoes inactivation at neutral pH, AdiC-Y74A maintained high transport activity regardless of pH. This research is relevant because AdiC-Y74F maintained the same pH-dependent inactivation profile of the WT, and in GadC a phenylalanine is found at the equivalent position (residue 76). However, we note that spin-labeled cysteine mutants at residue 77 showed little to no change in either DEER distance measurements or CW spectra, and was inactive at neutral pH. Nevertheless, this hypothesis can be directly tested on F76A background mutants using radioligand transport assays such as those outlined in section 5.4.4. A second hypothesis in AdiC, proposed following MD simulations, suggests that protonation of E218 drives dissociation of the substrate from the active site [481]. This residue is strictly conserved in the pH-dependent "virtual proton pumps" but not the neutral-pH homologous transporters such as ApcT and Lat1 [254]. A simple test of this hypothesis would be to measure the dissociation constants of radiolabeled glutamate, GABA, and glutamine as a function of pH using a scintillation proximity assay [328], which measures substrate binding affinity in detergent-solubilized transporters, in both wildtype and E218Q mutants of GadC. The role of this proposed proton would be equivalent to that of potassium in LeuT, which serves to displace sodium from the substrate-binding site but is otherwise uninvolved in the transport cycle [33].

Second, if the C-terminal domain is negligibly involved in regulation of pH-dependent transport, then what is its primary purpose? We speculate that this domain binds the glutamate decarboxylase GadB, which is cotranscribed with GadC and has been shown to partition to the membrane fraction via an unknown mechanism at acidic pH, but not at neutral pH. In fact, the 2003 publication presenting the crystal structure of GadB proposed this exact hypothesis in passing [64]. While the structure of GadC, determined and published a decade later, was entirely consistent with this mechanism [254], no experimental evidence supporting or refuting this possible protein-protein interaction has, to our knowledge, been published. Moreover, since the publication of GadC's structure, an equivalent mechanism was observed and demonstrated in the structural homolog DrSLC38A9, which has a similar N-terminal domain embedded in its intracellular cavity that, when released into the cytoplasm, binds and recruits the regulatory complex mTORC1 to the lysosomal membrane. A pH-dependent GadB/GadC interaction could easily be tested by spin labeling the C-terminal domain and observing mobility changes in the CW spectrum as a function of both pH and GadB concentration (see Figure 5.2.D for an example of this experiment), as slower tumbling times would be expected following binding of a 330 kDa soluble protein. Follow-up experiments include visualization of GFP-labeled GadB using fluorescence microscopy *in vivo* in cells expressing either full-length or truncated GadC at neutral and acidic pH. Alternatively, the structural basis of this interaction can be determined by crystallography of GadB at low pH with a peptide fragment whose sequence matches that of the C-terminal domain of GadC.

6.1.3 Perspectives on the IF-occluded conformation observed using DEER

We now turn our attention to the IF-occluded conformation modeled using the experimental DEER data. Under physiological conditions, antiporters belonging to the APC transporter family exchange substrates present at micromolar concentrations outside the cell, but millimolar concentrations inside the cell. In some proteins, such as the alanine/serine/cysteine antiporter BasC, this leads to apparent Km values in the micromolar range during import (OF-to-IF) but in the millimolar range during export (IF-to-OF). Technical limitations prevented the DEER experiments presented in this

dissertation from being carried out in a gradient, leading to substrate concentrations identical on both sides of the membrane. We therefore speculate that if GadC interacts with substrates with a similar sidedness as BasC, then under the experimental conditions discussed in this dissertation, this would be expected to lead to more OF-to-IF isomerization than IF-to-OF, consistent with a preponderance of IF GadC observed in our experiments. Stabilization of OF GadC may be achieved by introducing a gradient, which would require reconstitution of GadC into proteoliposomes. To address the possibility that not all GadC molecules are correctly oriented in the proteoliposome, the membrane-impermeable reducing agent TCEP may need to be introduced following reconstitution to reduce any spin-labeled cysteine residues on the wrong side of the membrane (e.g. intracellular cysteine residues on the outside of the proteoliposome).

Other conformations in the transport cycle could conceivably be visualized using DEER by introducing a pH gradient. Maintenance of a pH gradient for extended periods of time has previously been shown to require specific lipid profiles; previous experiments in AdiC and GadC have relied on liposomes comprised of 3:1 POPE:POPG to maintain an outer pH of 2.2 and an inner pH of 5.0 [415, 416, 417]. When executed in conjunction with the preparatory steps outlined above, this experiment could reveal how pH gradients contribute to stabilization of OF GadC and sampling of discrete conformational intermediates in the presence or absence of substrates.

To summarize the experimental findings, we found that 1) the structural basis of pH-dependent activation is partially, but not fully, mediated by detachment of the C-terminal domain as previously hypothesized [254], and 2) the transporter predominantly adopts an inward-facing occluded conformation regardless of whether substrates are presence or not. These findings advance our understanding of transporters with the LeuT-fold by reinforcing the divergent energy landscapes underpinning function. Further research into the breadth of transport mechanisms mediated by symporters, antiporters, and permeases will be necessary to determine whether the observations made in GadC are restricted to amino acid exchangers or transporters in the APC family.

6.2 Synopsis of methodological advancements

While this dissertation nominally focused on studies of GadC, the overwhelming majority of the results presented focus on the development of methods for modeling protein structures using sparse experimental data, particularly DEER data. These methods were designed to tackle the sparse and imprecise nature of the data being collected. Thus, in contrast with equivalent methods designed to model proteins using cryo-EM density or residue coevolutionary restraints, acknowledging the uncertainty and uneven distribution inherent to the experimental data presented in Chapter 5 was fundamental to minimizing the risk of overfitting, which could lead to spurious conclusions (results obtained in *de novo* folding benchmarks of Bax and ExoU in Chapter 3 illustrate how incorrectly folded models can satisfy the experimental data, which is exactly the outcome to be avoided). Therefore, the work presented in this dissertation took a multi-pronged approach to maximize the contribution of these data during modeling:

- Uncertainty was minimized by explicitly modeling the spin label ensemble (Chapter 3). Compared to the previous implementation of DEER restraints in Rosetta, the CONE model, improvements in modeling precision were observed in every protein.

- An effective scoring function was then determined by comparing several candidate functions in Appendix B. This included the possibility that multiple conformers were present in the data, which the distributions in Chapter 5 could not rule out.

- Further improvements in modeling precision were possible in cases where the starting structure was consistent with a set of DEER data. Multilateration of the spin labels using the algorithm discussed in Chapter 4 could conceivably be used to more precisely determine the conformation of interest. Unfortunately, GadC did not adopt the conformation observed in the crystal structure, precluding the use of this method.

- Effective sampling methods discussed in Appendix C allowed modeling to be focused on the immediate conformational vicinity of the starting structure, which prevented precious computational resources from being wasted on sampling unrealistic conformers.

- Finally, a statistical potential capturing our expectation that the structure of GadC resembles those of its homologs served as an additional source of regularization that prevented outrageous structures from being considered (see section 5.4.7).

The combined approach allowed models of IF-occluded GadC to be modeled using only 23 experimental restraints. The resulting models at both pHs closely resembled the structures of two homologs, ApcT and GkApcT, as well as a model generated using RoseTTAFold (see section 6.3 below and Figure G.13). Nevertheless, uncertainty in modeling prevented differences between conformations generated using low- or neutral-pH data from being observed. While that may hint at additional computational innovations that have yet to be realized, it may also suggest that imprecise and/or sparse experimental measurements can only go so far in resolving small-scale conformational changes. Mitigation or elimination of experimental uncertainty could be achieved using more rigid spin labels, such as bifunctional labels or imidazole-derived spin labels, which sample fewer conformers.

In summary, the computational work presented here describes a means to directly integrate DEER data for protein modeling. The results we obtained when using the raw data as experimental restraints, rather than as a means of checking the correctness of structures after the fact, suggests that sparse experimental DEER data can be integrated with computational modeling to complement a wide variety of tasks. In this dissertation the use of DEER data is limited to *de novo* protein fold prediction (Chapter 3, homology modeling (Chapter 5), and conformational change modeling (Chapters 4 and 5), but the data can conceivably be adapted to achieve other tasks, such as to predict the conformations of flexible loop regions that might be unresolved in experimental structures (see Appendix D), or determine the location of paramagnetic ligands (see reference [140]). However, integrating DEER data with other forms of experimentally collected information, such as SAXS data or low-resolution cryo-EM density, will require further fine-tuning.

6.3 Final thoughts: perspectives on integrative modeling using sparse data

As was discussed in section 2.4.1, the overarching goal of integrative modeling is to either explain data that has already been collected or predict future observations [167]. While the former was the predominant goal of the methods development projects discussed here, both approaches contributed to the research presented in Chapter 5 of this dissertation. Although not discussed, experimental design of spin label pairs in GadC was initially guided by an OF homology model of GadC generated using RosettaCM [384] with various structures of AdiC serving as templates [122, 368]. A challenge when using homology models to predict conformational changes between states with RMSD values of 3 Å to 5 Å is that physiologically relevant structural movements are difficult to distinguish from modeling artifacts. By contrast, pairs of experimental structures of a single homolog in both OF and IF conformations can reveal more precisely which regions of a protein move and which stay fixed. Unfortunately, structural characterization of an APC transporter in both conformations did not occur until March 2021, after data collection was concluded. Comparison of these two structures, shown in Figures 1.5 and 1.6, reveal movement in virtually every helix in the core LeuT-fold transmembrane domain. Importantly, they showed helical movements that were initially predicted by the OF homology model of GadC, but misconstrued as modeling artifacts, particularly in the hash domain and TMH10. As a result, no spin pairs were designed to specifically interrogate these movements.

The shortcomings in using integrative modeling to design experiments of this nature hint at a larger problem regarding the use of EPR spectroscopy as an exploratory tool for structural studies. Fortunately, this research coincided with two methodological advancements with major implications for the field of integrative structural biology. First, steady improvements in both software and hardware allowed single-particle cryo-EM to mature from a technique capable of viewing the topologies of large proteins and complexes (>150 kDa) at low-to-medium resolution to one capable of resolving the structures and dynamics of even medium-sized proteins, such as SERT (70 kDa), to atomic detail [78, 478]. Second, state-of-the-art *de novo* protein structure prediction algorithms were recently developed that are capable of routinely achieving sub-angstrom modeling accuracy

from sequence alone [18, 200]. Appendix A presents anecdotal evidence that multiple discrete conformers may be modeled to high accuracy this way, solving a major outstanding problem in *de novo* structure prediction [297]. Therefore, access to atomic-detail structures and structural models is expected to be less of a barrier to high-impact structural biology research in the coming years.

What will be the effect of these developments on integrative modeling? The research outlined in this dissertation followed a playbook used in previous integrative structural biology investigations [213, 310] that focuses on two types of models: structural models and mechanistic models. These technological advancements point to a future in which the integration of sparse data, such as distance data collected using EPR, for the purposes of structure prediction may soon be unnecessary. By contrast, these technological improvements have the potential to facilitate the design of enormously informative experiments seeking to describe how and when different structures of a protein, either obtained experimentally or predicted *de novo*, interconvert in response to mutagenesis, ligand binding, or environmental changes such as pH or lipid composition. As stated above, it was unclear during data collection if the absence of conformational dynamics in the DEER data resulted from structural uniformity or from uninformative spin pair design. Having multiple protein structures and/or high-accuracy structural models in different conformations can mitigate the possibility of the latter, which would be a welcome change when designing experiments that reports on the characteristics of a protein's energy landscape. Thus, experimental design that aims to track population changes, rather than precise conformational details, plays to the strength of the DEER technique and will ensure its relevance for years to come.

References

[1] Dinar Abdullin, Nicole Florin, Gregor Hagelueken, and Olav Schiemann. EPR-based approach for the localization of paramagnetic metal Ions in biomolecules. *Angewandte Chemie - International Edition*, 54(6):1827–1831, 2015.

[2] Dinar Abdullin, Gregor Hagelueken, and Olav Schiemann. Determination of nitroxide spin label conformations via PELDOR and X-ray crystallography. *Physical Chemistry Chemical Physics*, 18(15):10428–10437, apr 2016.

[3] Dinar Abdullin and Olav Schiemann. Localization of metal ions in biomolecules by means of pulsed dipolar EPR spectroscopy. *Dalton Transactions*, 50(3):808–815, 2021.

[4] Jeff Abramson and Ernest M. Wright. Structure and function of Na+-symporters with inverted repeats. *Current Opinion in Structural Biology*, 19(4):425–432, 2009.

[5] Paul D. Adams, Navraj S. Pannu, Randy J. Read, and Axel T. Brünger. Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proceedings of the National Academy of Sciences of the United States of America*, 94(10):5018–5023, 1997.

[6] Suraj Adhikary, Daniel J Deredge, Anu Nagarajan, Lucy R Forrest, Patrick L Wintrode, and Satinder K Singh. Conformational dynamics of a neurotransmitter: Sodium symporter in a lipid bilayer. *Proceedings of the National Academy of Sciences of the United States of America*, 114(10):E1786—-E1795, mar 2017.

[7] Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory*, pages 199–213, 1998.

[8] Nathan Alexander, Ahmad Al-Mestarihi, Marco Bortolus, Hassane Mchaourab, and Jens Meiler. De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data. *Structure*, 16(2):181–195, 2008.

[9] Nathan S Alexander, Anita M Preininger, Ali I Kaya, Richard A Stein, Heidi E Hamm, and Jens Meiler. Energetic analysis of the rhodopsin-G-protein complex links the α5 helix to GDP release. *Nature Structural and Molecular Biology*, 21(1):56–63, 2014.

[10] Nathan S Alexander, Richard A Stein, Hanane A Koteiche, Kristian W Kaufmann, Hassane S Mchaourab, and Jens Meiler. RosettaEPR: Rotamer Library for Spin Label Structure and Dynamics. *PLoS ONE*, 8(9), 2013.

[11] Rebecca F Alford, Patrick J Fleming, Karen G Fleming, and Jeffrey J Gray. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. *Biophysical Journal*, 118(8):2042–2055, mar 2020.

[12] Rebecca F Alford, Julia Koehler Leman, Brian D Weitzner, Amanda M Duran, Drew C Tilley, Assaf Elazar, and Jeffrey J Gray. An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Computational Biology*, 11(9):1–23, 2015.

[13] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'meara, Frank P. Dimaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13:3031–3048, 2017.

[14] Noelia Alonso-García, Inés García-Rubio, José A. Manso, Rubén M. Buey, Hector Urien, Arnoud Sonnenberg, Gunnar Jeschke, and José M. De Pereda. Combination of X-ray crystallography, SAXS and DEER to obtain the structure of the FnIII-3,4 domains of integrin $\alpha 6\beta 4$. *Acta Crystallographica Section D: Biological Crystallography*, 71:969–985, 2015.

[15] Christian Altenbach, Ana Karin Kusnetzow, Oliver P. Ernst, Klaus Peter Hofmann, and Wayne L. Hubbell. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(21):7439–7444, 2008.

[16] Melanie L Aprahamian and Steffen Lindert. Utility of Covalent Labeling Mass Spectrometry Data in Protein Structure Prediction with Rosetta. *Journal of Chemical Theory and Computation*, 15(5):3410–3424, may 2019.

[17] Mykhailo Azarkh, Anna Bieber, Mian Qi, Jörg W A Fischer, Maxim Yulikov, Adelheid Godt, and Malte Drescher. Gd(III)-Gd(III) Relaxation-Induced Dipolar Modulation Enhancement for In-Cell Electron Paramagnetic Resonance Distance Determination. *Journal of Physical Chemistry Letters*, 10(7):1477–1481, apr 2019.

[18] Minkyung Baek, Frank Dimaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R Glassman, Andy Degiovanni, Jose H Pereira, Andria V Rodrigues, Alberdina A Van Dijk, Ana C Ebrecht, Diederik J Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K Rathinaswamy, Udit Dalwadi, Calvin K Yip, John E Burke, K Christopher Garcia, Nick V Grishin, Paul D Adams, Randy J Read, and David Baker. Accurate prediction of protein structures and interactions using a 3-track network. *bioRxiv*, page 2021.06.14.448402, jul 2021.

[19] Aidin R Balo, Hannes Feyrer, and Oliver P Ernst. Toward Precise Interpretation of DEER-Based Distance Distributions: Insights from Structural Characterization of V1 Spin-Labeled Side Chains. *Biochemistry*, 55(37):5256–5263, 2016.

[20] J. E. Banham, G. Jeschke, and C. R. Timmel. Evidence from EPR that nitroxide spin labels attached to human hemoglobin alter their conformation upon freezing. *Molecular Physics*, 105(15-16):2041–2047, aug 2007.

[21] Marta Barrios, María-Isabel Moreno-Carralero, Nuria Cuadrado-Grande, María Baro, José-Luis Vivanco, and María-Josefa Morán-Jiménez. The homozygous mutation G75R in the human SLC11A2 gene leads to microcytic anaemia and iron overload. *British Journal of Haematology*, 157(4):510–514, 2012.

[22] Katja Barth, Susanne Hank, Philipp E Spindler, Thomas F Prisner, Robert Tampé, and Benesh Joseph. Conformational Coupling and trans-Inhibition in the Human Antigen Transporter Ortholog TmrAB Resolved with Dipolar EPR Spectroscopy. *Journal of the American Chemical Society*, 140(13):4527–4533, apr 2018.

[23] Paola Bartoccioni, Joana Fort, Antonio Zorzano, Ekaitz Errasti-Murugarren, and Manuel Palacín. Functional characterization of the alanine-serine-cysteine exchanger of Carnobacterium sp AT7. *Journal of General Physiology*, 151(4):505–517, 2019.

[24] Kathleen N Beasley, Brian T Sutch, Ma'Mon M Hatmal, Ralf Langen, Peter Z Qin, and Ian S Haworth. *Computer Modeling of Spin Labels: NASNOX, PRONOX, and ALLNOX*, volume 563. Elsevier Inc., 1 edition, 2015.

[25] Francois Berenger, Rojan Shrestha, Yong Zhou, David Simoncini, and Kam Y.J. Zhang. Durandal: Fast exact clustering of protein decoys. *Journal of Computational Chemistry*, 33(4):471–474, 2012.

[26] Frances C Bernstein, Thomas F Koetzle, Williams GJ, Edgar F Meyer, Michael D Brice, John R Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The Protein Data Bank : A Computer-based Archival Structures. *Archives of Biochemistry and biophysics*, 185(2):584–591, 1978.

[27] Thijs Beuming, Lei Shi, Jonathan A Javitch, and Harel Weinstein. A comprehensive structure-based alignment of prokaryotic and eukaryotic neurotransmitter/Na+ symporters (NSS) aids in the use of the LeuT structure to probe NSS structure and function. *Molecular Pharmacology*, 70(5):1630–1642, 2006.

[28] Shreyas Bhat, Marco Niello, Klaus Schicker, Christian Pifl, Harald H. Sitte, Michael Freissmuth, and Walter Sandtner. Handling of intracellular k+ determines voltage dependence of plasmalemmal monoamine transporter function. *eLife*, 10:1–23, 2021.

[29] Jaya Bhatnagar, Peter P. Borbat, Abiola M. Pollard, Alexandrine M. Bilwes, Jack H. Freed, and Brian R. Crane. Structure of the ternary complex formed by a chemotaxis receptor signaling domain, the CheA histidine kinase, and the coupling protein CheW As determined by pulsed dipolar ESR spectroscopy. *Biochemistry*, 49(18):3824–3841, 2010.

[30] Jaya Bhatnagar, Jack H Freed, and Brian R Crane. Rigid Body Refinement of Protein Complexes with Long-Range Distance Restraints from Pulsed Dipolar ESR. *Methods in Enzymology*, 423(07):117–133, 2007.

[31] Stefan Bibow, Yevhen Polyhach, Cédric Eichmann, Celestine N. Chi, Julia Kowal, Stefan Albiez, Robert A. McLeod, Henning Stahlberg, Gunnar Jeschke, Peter Güntert, and Roland Riek. Solution structure of discoidal high-density lipoprotein particles with a shortened apolipoprotein A-I. *Nature Structural and Molecular Biology*, 24(2):187–193, feb 2017.

[32] Christian B. Billesbølle, Mie B. Krüger, Lei Shi, Matthias Quick, Zheng Li, Sebastian Stolzenberg, Julie Kniazeff, Kamil Gotfryd, Jonas S. Mortensen, Jonathan A. Javitch, Harel Weinstein, Claus J. Loland, and Ulrik Gether. Substrate-induced unlocking of the inner gate

determines the catalytic efficiency of a neurotransmitter: Sodium symporter. *Journal of Biological Chemistry*, 290(44):26725–26738, 2015.

[33] Christian B. Billesbølle, Jonas S. Mortensen, Azmat Sohail, Solveig G. Schmidt, Lei Shi, Harald H. Sitte, Ulrik Gether, and Claus J. Loland. Transition metal ion FRET uncovers K+ regulation of a neurotransmitter/sodium symporter. *Nature Communications*, 7, sep 2016.

[34] M Billeter, J Vendrell, G Wider, F X Avilés, M Coll, A Guasch, R Huber, and K Wüthrich. Comparison of the NMR solution structure with the X-ray crystal structure of the activation domain from procarboxypeptidase B. Technical Report 1, 1992.

[35] Benjamin P. Binder, Andrew R. Thompson, and David D. Thomas. Atomistic Models from Orientation and Distance Constraints Using EPR of a Bifunctional Spin Label. *Biophysical Journal*, 117(2):319–330, 2019.

[36] Christian A. Bippes, Antra Zeltina, Fabio Casagrande, Merce Ratera, Manuel Palacin, Daniel J. Muller, and Dimitrios Fotiadis. Substrate binding tunes conformational flexibility and kinetic stability of an amino acid antiporter. *Journal of Biological Chemistry*, 284(28):18651–18663, 2009.

[37] Paola Bisignano, Chiara Ghezzi, Hyunil Jo, Nicholas F. Polizzi, Thorsten Althoff, Chakrapani Kalyanaraman, Rosmarie Friemann, Matthew P. Jacobson, Ernest M. Wright, and Michael Grabe. Inhibitor binding mode and allosteric regulation of Na+-glucose symporters. *Nature Communications*, 9(1):1–10, 2018.

[38] Mandy E. Blackburn, Angelo M. Veloro, and Gail E. Fanucci. Monitoring inhibitor-induced conformational population shifts in HIV-1 protease by pulsed EPR spectroscopy. *Biochemistry*, 48(37):8765–8767, sep 2009.

[39] Stephanie Bleicken, Gunnar Jeschke, Carolin Stegmueller, Raquel Salvador-Gallego, Ana J. García-Sáez, and Enrica Bordignon. Structural Model of Active Bax at the Membrane. *Molecular Cell*, 56(4):496–505, 2014.

[40] Wouter Boomsma and Thomas Hamelryck. Full cyclic coordinate descent: Solving the protein loop closure problem in Cα space. *BMC Bioinformatics*, 6, jun 2005.

[41] Peter P Borbat, Elka R Georgieva, and Jack H Freed. Improved sensitivity for long-distance measurements in biomolecules: Five-pulse double electron-electron resonance. *Journal of Physical Chemistry Letters*, 4(1):170–175, jan 2013.

[42] Petr P. Borbat, Hassane S. Mchaourab, and Jack H. Freed. Protein structure determination using long-distance constraints from double-quantum coherence ESR: Study of T4 lysozyme. *Journal of the American Chemical Society*, 124(19):5304–5314, may 2002.

[43] Igor V. Borovykh, Stefano Ceola, Prasad Gajula, Peter Gast, Heinz Jürgen Steinhoff, and Martina Huber. Distance between a native cofactor and a spin label in the reaction centre of Rhodobacter sphaeroides by a two-frequency pulsed electron paramagnetic resonance method and molecular dynamics simulations. *Journal of Magnetic Resonance*, 180(2):178–185, jun 2006.

[44] Patrick D. Bosshart and Dimitrios Fotiadis. Secondary Active Transporters. *Subcellular Biochemistry*, 92:275–299, 2019.

[45] Olga Boudker and Grégory Verdon. Structural perspectives on secondary active transporters. *Trends in Pharmacological Sciences*, 31(9):418–426, 2010.

[46] Evzen Boura, Bartosz Rózycki, Hoi Sung Chung, Dawn Z. Herrick, Bertram Canagarajah, David S. Cafiso, William A. Eaton, Gerhard Hummer, and James H. Hurley. Solution structure of the ESCRT-I and -II supercomplex: Implications for membrane budding and scission. *Structure*, 20(5):874–886, 2012.

[47] Evzen Boura, Bartosz Rózycki, Dawn Z Herrick, Hoi Sung Chung, Jaroslav Vecer, William A Eaton, David S Cafiso, Gerhard Hummer, and James H Hurley. Solution structure of the ESCRT-I complex by smallangle X-ray scattering, EPR, and FRET spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9437–9442, 2011.

[48] Alice M Bowen, Eachan O D Johnson, Francesco Mercuri, Nicola J Hoskins, Ruihong Qiao, James S O McCullagh, Janet E Lovett, Stephen G Bell, Weihong Zhou, Christiane R Timmel, Luet Lok Wong, and Jeffrey R Harmer. A Structural Model of a P450-Ferredoxin Complex from Orientation-Selective Double Electron-Electron Resonance Spectroscopy. *Journal of the American Chemical Society*, 140(7):2514–2527, feb 2018.

[49] Andrew Bowman, Richard Ward, Hassane El-Mkami, Tom Owen-Hughes, and David G. Norman. Probing the (H3-H4)2 histone tetramer structure using pulsed EPR spectroscopy combined with site-directed spin labelling. *Nucleic Acids Research*, 38(2):695–707, 2009.

[50] Aaron T. Bozzi, Lukas B. Bane, Wilhelm A. Weihofen, Abhishek Singharoy, Eduardo R. Guillen, Hidde L. Ploegh, Klaus Schulten, and Rachelle Gaudet. Crystal Structure and Conformational Change Mechanism of a Bacterial Nramp-Family Divalent Metal Transporter. *Structure*, 24(12):2102–2114, 2016.

[51] Aaron T. Bozzi and Rachelle Gaudet. Molecular Mechanism of Nramp-Family Transition Metal Transport. *Journal of Molecular Biology*, (xxxx):166991, 2021.

[52] Aaron T. Bozzi, Christina M. Zimanyi, John M. Nicoludis, Brandon K. Lee, Casey H. Zhang, and Rachelle Gaudet. Structures in multiple conformations reveal distinct transition metal and proton pathways in an nramp transporter. *eLife*, 8:1–63, 2019.

[53] Suzanne Brandon, Albert H Beth, and Eric J Hustedt. The global analysis of DEER data. *Journal of Magnetic Resonance*, 218:93–104, 2012.

[54] Theresa Braun, Malte Drescher, and Daniel Summerer. Expanding the genetic code for site-directed spin-labeling. *International Journal of Molecular Sciences*, 20(2):373, jan 2019.

[55] Frauke D Breitgoff, Yevhen O Polyhach, and Gunnar Jeschke. Reliable nanometre-range distance distributions from 5-pulse double electron electron resonance. *Physical Chemistry Chemical Physics*, 19(24):15754–15765, 2017.

[56] Rasmus Bro and Sijmen De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, 11(5):393–401, 1997.

[57] Louise J. Brown, Ken L. Sale, Ron Hills, Clement Rouviere, Likai Song, Xiaojun Zhang, and Piotr G. Fajer. Structure of the inhibitory region of troponin by site directed spin labeling electron paramagnetic resonance. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12765–12770, 2002.

[58] Axel T Brünger, Paul D Adams, G Marius Clore, Warren L Delano, Piet Gros, Ralf W Grossekunstleve, Jian Sheng Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, Randy J Read, Luke M Rice, Thomas Simonson, and Gregory L Warren. Crystallography NMR system: A new software suite for macromolecular structure determination. Technical Report 5, 1998.

[59] Kenneth P. Burnham and David R. Anderson. *Model selection and multi-model inference*. 2002.

[60] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. Technical Report 5, 1995.

[61] Richard H. Byrd, Robert B. Schnabel, and Gerald A. Shultz. Trust Region Algorithm for Nonlinearly Constrained Optimization. *SIAM Journal on Numerical Analysis*, 24(5):1152–1170, 1987.

[62] Antonio N Calabrese, Scott M Jackson, Lynsey N Jones, Oliver Beckstein, Florian Heinkel, Joerg Gsponer, David Sharples, Marta Sans, Maria Kokkinidou, Arwen R Pearson, Sheena E Radford, Alison E Ashcroft, and Peter J F Henderson. Topological Dissection of the Membrane Transport Protein Mhp1 Derived from Cysteine Accessibility and Mass Spectrometry. *Analytical Chemistry*, 89(17):8844–8852, sep 2017.

[63] Adrian A. Canutescu and Roland L. Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12(5):963–972, may 2003.

[64] Guido Capitani, Daniela De Biase, Caterina Aurizi, Heinz Gut, Francesco Bossa, and Markus G. Grütter. Crystal structure and functional analysis of Escherichia coli glutamate decarboxylase. *EMBO Journal*, 22(16):4027–4037, 2003.

[65] Oliviero Carugo and Sándor Pongor. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein Science*, 10(7):1470–1473, 2008.

[66] Thomas M Casey and Gail E Fanucci. Spin labeling and Double Electron-Electron Resonance (DEER) to Deconstruct Conformational Ensembles of HIV Protease. In *Methods in Enzymology*, volume 564, pages 153–187. Academic Press Inc., 2015.

[67] Mary Hongying Cheng and Ivet Bahar. Monoamine transporters: structure, intrinsic dynamics and allosteric regulation, jul 2019.

[68] Thomas A. Chew, Benjamin J. Orlando, Jinru Zhang, Naomi R. Latorraca, Amy Wang, Scott A. Hollingsworth, Dong Hua Chen, Ron O. Dror, Maofu Liao, and Liang Feng. Structure and mechanism of the cation–chloride cotransporter NKCC1. *Nature*, 572(7770):488–492, aug 2019.

[69] Thomas A. Chew, Jinru Zhang, and Liang Feng. High-resolution views and transport mechanisms of the NKCC1 and KCC transporters. *Journal of Molecular Biology*, 433(16):167056, 2021.

[70] Gamma Chi, Rebecca Ebenhoch, Henry Man, Haiping Tang, Laurence E Tremblay, Gabriella Reggiano, Xingyu Qiu, Tina Bohstedt, Idlir Liko, Fernando G Almeida, Alexandre P Garneau, Dong Wang, Gavin McKinley, Christophe P Moreau, Kiran D Bountra, Patrizia Abrusci, Shubhashish M M Mukhopadhyay, Alejandra Fernandez-Cid, Samira Slimani, Julie L Lavoie, Nicola A Burgess-Brown, Ben Tehan, Frank DiMaio, Ali Jazayeri, Paul Isenring, Carol V Robinson, and Katharina L Dürr. Phospho-regulation, nucleotide binding and ion access control in potassium-chloride cotransporters. *The EMBO Journal*, pages 1–20, 2021.

[71] Ximin Chi, Xiaorong Li, Yun Chen, Yuanyuan Zhang, Qiang Su, and Qiang Zhou. Molecular basis for regulation of human potassium chloride cotransporters. *bioRxiv*, 2020.

[72] Yun Wei Chiang, Peter P. Borbat, and Jack H. Freed. Maximum entropy: A complement to Tikhonov regularization for determination of pair distance distributions by pulsed ESR. *Journal of Magnetic Resonance*, 177(2):184–196, dec 2005.

[73] Yun Wei Chiang, Peter P. Borbat, and Jack H. Freed. The determination of pair distance distributions by pulsed ESR using Tikhonov regularization. *Journal of Magnetic Resonance*, 172(2):279–295, feb 2005.

[74] Yoonjoo Choi and Charlotte M Deane. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins: Structure, Function and Bioinformatics*, 78(6):1431–1440, 2010.

[75] Pieter Chys and Pablo Chacón. Random coordinate descent with spinor-matrices and geometric filters for efficient loop closure. *Journal of Chemical Theory and Computation*, 9(3):1821–1829, mar 2013.

[76] Derek P Claxton, Matthias Quick, Lei Shi, Fernanda Delmondes De Carvalho, Harel Weinstein, Jonathan A Javitch, and Hassane S McHaourab. Ion/substrate-dependent conformational dynamics of a bacterial homolog of neurotransmitter:sodium symporters. *Nature Structural and Molecular Biology*, 17(7):822–829, 2010.

[77] Jonathan A. Coleman, Evan M. Green, and Eric Gouaux. X-ray structures and mechanism of the human serotonin transporter. *Nature*, 532(7599):334–339, apr 2016.

[78] Jonathan A. Coleman, Dongxue Yang, Zhiyu Zhao, Po Chao Wen, Craig Yoshioka, Emad Tajkhorshid, and Eric Gouaux. Serotonin transporter–ibogaine complexes illuminate mechanisms of inhibition and transport. *Nature*, 569(7754):141–145, 2019.

[79] Alberto Collauto, Hannah A Deberg, Royi Kaufmann, William N Zagotta, Stefan Stoll, and Daniella Goldfarb. Rates and equilibrium constants of the ligand-induced conformational transition of an HCN ion channel protein domain determined by DEER spectroscopy. *Physical Chemistry Chemical Physics*, 19(23):15324–15334, 2017.

[80] Philipp Consentius, Ulrich Gohlke, Bernhard Loll, Claudia Alings, Robert Müller, Udo Heinemann, Martin Kaupp, Markus Wahl, and Thomas Risse. Tracking Transient Conformational States of T4 Lysozyme at Room Temperature Combining X-ray Crystallography and Site-Directed Spin Labeling. *Journal of the American Chemical Society*, 138(39):12868–12875, oct 2016.

[81] Patrick Conway, Michael D. Tyka, Frank DiMaio, David E. Konerding, and David Baker. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science*, 23(1):47–55, 2014.

[82] Anthonya Cooper, Donna Woulfe, and Fusun Kilic. Post-translational modifications of serotonin transporter. *Pharmacological Research*, 140(October 2018):7–13, 2019.

[83] Robin A. Corey, Zainab Ahdash, Anokhi Shah, Euan Pyle, William J. Allen, Tomas Fessl, Janet E. Lovett, Argyris Politis, and Ian Collinson. ATP-induced asymmetric pre-protein folding as a driver of protein translocation through the sec machinery. *eLife*, 8:1–25, 2019.

[84] Timothy F. Cunningham, Marshall S. McGoff, Ishita Sengupta, Christopher P. Jaroniec, W. Seth Horne, and Sunil Saxena. High-resolution structure of a protein spin-label in a solvent-exposed β-sheet and comparison with DEER spectroscopy. *Biochemistry*, 51(32):6350–6359, 2012.

[85] Timothy F Cunningham, Soraya Pornsuwan, W Seth Horne, and Sunil Saxena. Rotameric preferences of a protein spin label at edge-strand β-sheet sites. *Protein Science*, 25(5):1049–1060, 2016.

[86] Timothy F Cunningham, Miriam R Putterman, Astha Desai, W Seth Horne, and Sunil Saxena. The double-histidine Cu2+-binding motif: A highly rigid, site-specific spin probe for electron spin resonance distance measurements. *Angewandte Chemie - International Edition*, 54(21):6330–6334, 2015.

[87] Marjorie Damian, Maxime Louet, Antoniel Augusto Severo Gomes, Céline M'Kadmi, Séverine Denoyelle, Sonia Cantel, Sophie Mary, Paulo M Bisch, Jean-Alain Fehrentz, Laurent J Catoire, Nicolas Floquet, and Jean-Louis Banères. Allosteric modulation of ghrelin receptor signaling by lipids. *Nature Communications*, 12(1):3938, dec 2021.

[88] Reza Dastvan, Eva Maria Brouwer, Denise Schuetz, Oliver Mirus, Enrico Schleiff, and Thomas F. Prisner. Relative Orientation of POTRA Domains from Cyanobacterial Omp85 Studied by Pulsed EPR Spectroscopy. *Biophysical Journal*, 110(10):2195–2206, may 2016.

[89] Reza Dastvan, Axel W Fischer, Smriti Mishra, Jens Meiler, and Hassane S McHaourab. Protonation-dependent conformational dynamics of the multidrug transporter EmrE. *Proceedings of the National Academy of Sciences of the United States of America*, 113(5):1220–1225, 2016.

[90] Reza Dastvan, Smriti Mishra, Yelena B. Peskova, Robert K. Nakamoto, and Hassane S. Mchaourab. Mechanism of allosteric modulation of P-glycoprotein by transport substrates and inhibitors. *Science*, 364(6441):689–692, may 2019.

[91] C M Deane. CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Protein Science*, 10(3):599–612, mar 2001.

[92] Charlotte M Deane and Tom L Blundell. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Technical Report 1, 2000.

[93] Vincent Debruycker, Andrew Hutchin, Matthieu Masureel, Emel Ficici, Chloé Martens, Pierre Legrand, Richard A. Stein, Hassane S. Mchaourab, José D. Faraldo-Gómez, Han Remaut, and Cédric Govaerts. An embedded lipid in the multidrug transporter LmrP suggests a mechanism for polyspecificity. *Nature Structural and Molecular Biology*, 27(9):829–835, sep 2020.

[94] Diego Del Alamo, Axel W. Fischer, Rocco Moretti, Nathan S. Alexander, Jeffrey Mendenhall, Nicholas J. Hyman, and Jens Meiler. Efficient Sampling of Protein Loop Regions Using Conformational Hashing Complemented with Random Coordinate Descent. *Journal of Chemical Theory and Computation*, 17(1):560–570, jan 2021.

[95] Diego del Alamo, Cédric Govaerts, and Hassane S. Mchaourab. AlphaFold2 predicts the inward-facing conformation of the multidrug transporter LmrP. *Proteins: Structure, Function and Bioinformatics*, (March):10–12, 2021.

[96] Diego del Alamo, Kevin L. Jagessar, Jens Meiler, and Hassane S. Mchaourab. Methodology for rigorous modeling of protein conformational changes by Rosetta using DEER distance restraints. *PLOS Computational Biology*, 17(6):e1009107, 2021.

[97] Diego del Alamo, Maxx H Tessmer, Richard A Stein, Jimmy B Feix, Hassane S Mchaourab, and Jens Meiler. Rapid Simulation of Unprocessed DEER Decay Data for Protein Fold Prediction. *Biophysical Journal*, 118(2):366–375, 2020.

[98] David Drew and Olga Boudker. Shared Molecular Mechanisms of Membrane Transporters. *Annual Review of Biochemistry*, 85:543–572, 2016.

[99] Olivier Duss, Erich Michel, Maxim Yulikov, Mario Schubert, Gunnar Jeschke, and Frédéric H.T. Allain. Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature*, 509(7502):588–592, may 2014.

[100] Olivier Duss, Maxim Yulikov, Frédéric H T Allain, and Gunnar Jeschke. Combining NMR and EPR to determine structures of large RNAs and protein-RNA complexes in solution. *Methods in Enzymology*, 558(1):279–331, 2015.

[101] Olivier Duss, Maxim Yulikov, Gunnar Jeschke, and Frédéric H T Allain. EPR-aided approach for solution structure determination of large RNAs or protein-RNA complexes. *Nature Communications*, 5(May):3669, 2014.

[102] Sergei A. Dzuba. The determination of pair distance distribution by double electron-electron resonance: Regularization by the length of distance discretization with Monte Carlo calculations. *Journal of Magnetic Resonance*, 269:113–119, aug 2016.

[103] Sarah J. Edwards, Christopher W. Moth, Sunghoon Kim, Suzanne Brandon, Zheng Zhou, Charles E. Cobb, Eric J. Hustedt, Albert H. Beth, Jarrod A. Smith, and Terry P. Lybrand. Automated structure refinement for a protein heterodimer complex using limited EPR spectroscopic data and a rigid-body docking algorithm: A three-dimensional model for an ankyrin-cdb3 complex. *Journal of Physical Chemistry B*, 118(18):4717–4726, 2014.

[104] Thomas H. Edwards and Stefan Stoll. A Bayesian approach to quantifying uncertainty from experimental noise in DEER spectroscopy. *Journal of Magnetic Resonance*, 270:87–97, 2016.

[105] Thomas H. Edwards and Stefan Stoll. Optimal Tikhonov regularization for DEER spectroscopy. *Journal of Magnetic Resonance*, 288:58–68, 2018.

[106] Ines A. Ehrnstorfer, Eric R. Geertsma, Els Pardon, Jan Steyaert, and Raimund Dutzler. Crystal structure of a SLC11 (NRAMP) transporter reveals the basis for transition-metal ion transport. *Nature Structural and Molecular Biology*, 21(11):990–996, 2014.

[107] Ines A. Ehrnstorfer, Cristina Manatschal, Fabian M. Arnold, Juerg Laederach, and Raimund Dutzler. Structural and mechanistic basis of proton-coupled metal ion transport in the SLC11/NRAMP family. *Nature Communications*, 8:1–11, 2017.

[108] Martin Lorenz Eisinger, Aline Ricarda Dörrbaum, Hartmut Michel, Etana Padan, and Julian David Langer. Ligand-induced conformational dynamics of the Escherichia coli Na+/H+ antiporter NhaA revealed by hydrogen/deuterium exchange mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 114(44):11691–11696, 2017.

[109] Hassane El Mkami and David G. Norman. EPR Distance Measurements in Deuterated Proteins. *Methods in Enzymology*, 564:125–152, 2015.

[110] Burkhard Endeward, Joel A Butterwick, Roderick MacKinnon, and Thomas F Prisner. Pulsed electron-electron double-resonance determination of spin-label distances and orientations on the tetrameric potassium ion channel KcsA. *Journal of the American Chemical Society*, 131(42):15246–15250, 2009.

[111] Ekaitz Errasti-Murugarren, Joana Fort, Paola Bartoccioni, Lucía Díaz, Els Pardon, Xavier Carpena, Meritxell Espino-Guarch, Antonio Zorzano, Christine Ziegler, Jan Steyaert, Juan Fernández-Recio, Ignacio Fita, and Manuel Palacín. L amino acid transporter structure and molecular bases for the asymmetry of substrate interaction. *Nature Communications*, 10(1), dec 2019.

[112] Ekaitz Errasti-Murugarren and Manuel Palacín. Heteromeric Amino Acid Transporters in Brain: from Physiology to Pathology. *Neurochemical Research*, (0123456789), 2021.

[113] Narayanan Eswar, Ben Webb, Marc A Marti-Renom, M S Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali. Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics*, 15(1), sep 2006.

[114] Eric G.B. Evans, Jacob L.W. Morgan, Frank DiMaio, William N. Zagotta, and Stefan Stoll. Allosteric conformational change of a cyclic nucleotide-gated ion channel revealed by DEER spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 117(20):10839–10847, may 2020.

[115] Luis Fábregas Ibáñez and Gunnar Jeschke. General regularization framework for DEER spectroscopy. *Journal of Magnetic Resonance*, 300:28–40, mar 2019.

[116] Luis Fábregas Ibáñez and Gunnar Jeschke. Optimal background treatment in dipolar spectroscopy. *Physical Chemistry Chemical Physics*, 22(4):1855–1868, 2020.

[117] Luis Fábregas Ibáñez, Gunnar Jeschke, and Stefan Stoll. DeerLab: a comprehensive software package for analyzing dipolar electron paramagnetic resonance spectroscopy data. *Magnetic Resonance*, 1(2):209–224, 2020.

[118] Salem Faham, Akira Watanabe, Gabriel Mercado Besserer, Duilio Cascio, Alexandre Specht, Bruce A Hirayama, Ernest M Wright, and Jeff Abramson. The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na+/sugar symport. *Science*, 321(5890):810–814, 2008.

[119] Piotr Fajer, Mikolai Fajer, Michael Zawrotny, and Wei Yang. *Full Atom Simulations of Spin Label Conformations*, volume 563. Elsevier Inc., 1 edition, 2015.

[120] Jianjun Fan, Yang Xiao, Matthias Quick, Yuwei Yang, Ziyi Sun, Jonathan A. Javitch, and Xiaoming Zhou. Crystal structures of LeuT reveal conformational dynamics in the outward-facing states. *Journal of Biological Chemistry*, 296:100609, 2021.

[121] Bertrand T. Fang. Trilateration and extension to global positioning system navigation. *Journal of Guidance, Control, and Dynamics*, 9(6):715–717, 1986.

[122] Yiling Fang, Hariharan Jayaram, Tania Shane, Ludmila Kolmakova-Partensky, Fang Wu, Carole Williams, Yong Xiong, and Christopher Miller. Structure of a prokaryotic virtual proton pump at 3.2 Åresolution. *Nature*, 460(7258):1040–1043, 2009.

[123] Marc Fasnacht, Ken Butenhof, Anne Goupil-Lamy, Francisco Hernandez-Guzman, Hongwei Huang, and Lisa Yan. Automated antibody structure prediction using Accelrys tools: Results and best practices. *Proteins: Structure, Function and Bioinformatics*, 82(8):1583–1598, 2014.

[124] Niklas Fehr, Carsten Dietz, Yevhen Polyhach, Tona Von Hagens, Gunnar Jeschke, and Harald Paulsen. Modeling of the N-terminal section and the lumenal loop of trimeric light harvesting complex II (LHCII) by using EPR. *Journal of Biological Chemistry*, 290(43):26007–26020, 2015.

[125] R. M. Fine, H. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal. Predicting antibody hypervariable loop conformations II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins: Structure, Function, and Bioinformatics*, 1(4):342–362, 1986.

[126] Axel W. Fischer, Nathan S. Alexander, Nils Woetzel, Mert Karakas, Brian E. Weiner, and Jens Meiler. BCL::MP-fold: Membrane protein structure prediction guided by EPR restraints. *Proteins: Structure, Function and Bioinformatics*, 83(11):1947–1962, 2015.

[127] Axel W. Fischer, David M. Anderson, Maxx H. Tessmer, Dara W. Frank, Jimmy B. Feix, and Jens Meiler. Structure and Dynamics of Type III Secretion Effector Protein ExoU As determined by SDSL-EPR Spectroscopy in Conjunction with de Novo Protein Folding. *ACS Omega*, 2(6):2977–2984, 2017.

[128] Axel W Fischer, Enrica Bordignon, Stephanie Bleicken, Ana J García-Sáez, Gunnar Jeschke, and Jens Meiler. Pushing the size limit of de novo structure ensemble prediction guided by sparse SDSL-EPR restraints to 200 residues: The monomeric and homodimeric forms of BAX. *Journal of Structural Biology*, 195(1):62–71, 2016.

[129] Axel W. Fischer, Sten Heinze, Daniel K. Putnam, Bian Li, James C. Pino, Yan Xia, Carlos F. Lopez, and Jens Meiler. CASP11 - An evaluation of a modular BCL: Fold-based protein structure prediction pipeline. *PLoS ONE*, 11(4):1–19, 2016.

[130] András Fiser and Andrej Sali. ModLoop: Automated modeling of loops in protein structures. Technical Report 18, 2003.

[131] Gabriel A. Fitzgerald, Daniel S. Terry, Audrey L. Warren, Matthias Quick, Jonathan A. Javitch, and Scott C. Blanchard. Quantifying secondary transport at single-molecule resolution. *Nature*, 575(7783):528–534, 2019.

[132] Sarel J Fleishman, Andrew Leaver-Fay, Jacob E Corn, Eva Maria Strauch, Sagar D Khare, Nobuyasu Koga, Justin Ashworth, Paul Murphy, Florian Richter, Gordon Lemmon, Jens Meiler, and David Baker. Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS ONE*, 6(6), 2011.

[133] Mark R. Fleissner, Duilio Cascio, and Wayne L. Hubbell. Structural origin of weakly ordered nitroxide motion in spin-labeled proteins. *Protein Science*, 18(5):893–908, may 2009.

[134] Lucy R Forrest and Gary Rudnick. The rocking bundle: A mechanism for ion-coupled solute flux by symmetrical transporters. *Physiology*, 24(6):377–386, 2009.

[135] Lucy R Forrest, Sotiria Tavoulari, Yuan Wei Zhang, Gary Rudnick, and Barry Honig. Identification of a chloride ion binding site in Na+/Cl –dependent transporters. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12761–12766, 2007.

[136] Lucy R. Forrest, Yuan Wei Zhang, Miriam T. Jacobs, Joan Gesmonde, Li Xie, Barry H. Honig, and Gary Rudnick. Mechanism for alternating access in neurotransmitter transporters. *Proceedings of the National Academy of Sciences of the United States of America*, 105(30):10338–10343, 2008.

[137] John W. Foster. Escherichia coli acid resistance: Tales of an amateur acidophile. *Nature Reviews Microbiology*, 2(11):898–907, 2004.

[138] Matthew C Franklin, Jimin Wang, and Thomas A Steitz. Structure of the Replicating Complex of a Pol Family DNA Polymerase DNA polymerases, with the exception of DNA polymer-ase (pol), have a common structural core, containing the two universally conserved aspartate residues (De. Technical report, 2001.

[139] Simon A. Fromm, Rosalie E. Lawrence, and James H. Hurley. Structural mechanism for amino acid-dependent Rag GTPase nucleotide state switching by SLC38A9. *Nature Structural and Molecular Biology*, 27(11):1017–1023, 2020.

[140] Betty J. Gaffney, Miles D. Bradshaw, Stephen D. Frausto, Fayi Wu, Jack H. Freed, and Peter Borbat. Locating a lipid at the portal to the lipoxygenase active site. *Biophysical Journal*, 103(10):2134–2144, nov 2012.

[141] Xiang Gao, Lijun Zhou, Xuyao Jiao, Feiran Lu, Chuangye Yan, Xin Zeng, Jiawei Wang, and Yigong Shi. Mechanism of substrate recognition and transport by an amino acid antiporter. *Nature*, 463(7282):828–832, 2010.

[142] Ethan G. Geier, Avner Schlessinger, Hao Fan, Jonathan E. Gable, John J. Irwin, Andrej Sali, and Kathleen M. Giacomini. Structure-based ligand discovery for the Large-neutral Amino Acid Transporter 1, LAT-1. *Proceedings of the National Academy of Sciences of the United States of America*, 110(14):5480–5485, 2013.

[143] Elka R Georgieva, Peter P Borbat, Christopher Ginter, Jack H Freed, and Olga Boudker. Conformational ensemble of the sodium-coupled aspartate transporter. *Nature Structural and Molecular Biology*, 20(2):215–221, 2013.

[144] Elka R. Georgieva, Aritro S. Roy, Vladimir M. Grigoryants, Petr P. Borbat, Keith A. Earle, Charles P. Scholes, and Jack H. Freed. Effect of freezing conditions on distances and their distributions derived from Double Electron Electron Resonance (DEER): A study of doubly-spin-labeled T4 lysozyme. *Journal of Magnetic Resonance*, 216:69–77, mar 2012.

[145] Angeliki Giannoulis, Yin Yang, Yan Jun Gong, Xiaoli Tan, Akiva Feintuch, Raanan Carmieli, Thorsten Bahrenberg, Yangping Liu, Xun Cheng Su, and Daniella Goldfarb. DEER distance measurements on trityl/trityl and Gd(iii)/trityl labelled proteins. *Physical Chemistry Chemical Physics*, 21(20):10217–10227, 2019.

[146] Lucia Gigli, Witold Andrałojć, Arina Dalaloyan, Giacomo Parigi, Enrico Ravera, Daniella Goldfarb, and Claudio Luchinat. Assessing protein conformational landscapes: Integration of DEER data in Maximum Occurrence analysis. *Physical Chemistry Chemical Physics*, 20(43):27429–27438, 2018.

[147] Kamil Gotfryd, Thomas Boesen, Jonas S Mortensen, George Khelashvili, Matthias Quick, Daniel S Terry, Julie W Missel, Michael V LeVine, Pontus Gourdon, Scott C Blanchard, Jonathan A Javitch, Harel Weinstein, Claus J Loland, Poul Nissen, and Ulrik Gether. X-ray structure of LeuT in an inward-facing occluded conformation reveals mechanism of substrate release. *Nature Communications*, 11(1), dec 2020.

[148] Markus Gränz, Nicole Erlenbach, Philipp Spindler, Dnyaneshwar B Gophane, Lukas S Stelzl, Snorri Th. Sigurdsson, and Thomas F Prisner. Dynamics of Nucleic Acids at Room Temperature Revealed by Pulsed EPR Spectroscopy. *Angewandte Chemie*, 130(33):10700–10703, aug 2018.

[149] Günnur Güler, Rebecca M. Gärtner, Christine Ziegler, and Werner Mäntele. Lipid-protein interactions in the regulated betaine symporter BetP probed by infrared spectroscopy. *Journal of Biological Chemistry*, 291(9):4295–4307, 2016.

[150] Peter Güntert. Automated NMR structure calculation with CYANA. In *Methods in molecular biology (Clifton, N.J.)*, volume 278, pages 353–378. 2004.

[151] Zhefeng Guo, Duilio Cascio, Kálmán Hideg, and Wayne L Hubbell. Structural determinants of nitroxide motion in spin-labeled proteins: Solvent-exposed sites in helix B of T4 lysozyme. *Protein Science*, 17(2):228–239, 2008.

[152] Gregor Hagelueken, Dinar Abdullin, Richard Ward, and Olav Schiemann. MtsslSuite: In silico spin labelling, trilateration and distance-constrained rigid body docking in PyMOL. *Molecular Physics*, 111(18-19):2757–2766, oct 2013.

[153] Gregor Hagelueken, Richard Ward, James H Naismith, and Olav Schiemann. MtsslWizard: In Silico Spin-Labeling and Generation of Distance Distributions in PyMOL. *Applied Magnetic Resonance*, 42(3):377–391, 2012.

[154] Nina Hansra, Shruti Arya, and Michael W. Quick. Intracellular Domains of a Rat Brain GABA Transporter that Govern Transport. *Journal of Neuroscience*, 24(16):4082–4087, 2004.

[155] M Haridas, B F Anderson, and E N Baker. Structure of human diferrric lactoferrin refined at 2.2 Angstrom resolution. *Acta Crystallographica - Section D Biological Crystallography*, 51(5):629–646, 1995.

[156] Ma'Mon M Hatmal, Yiyu Li, Balachandra G Hegde, Prabhavati B Hegde, Christine C Jao, Ralf Langen, and Ian S Haworth. Computer modeling of nitroxide spin labels on proteins. *Biopolymers*, 97(1):35–44, 2012.

[157] Jennifer M. Hays, David S. Cafiso, and Peter M. Kasson. Hybrid Refinement of Heterogeneous Conformational Ensembles Using Spectroscopic Data. *Journal of Physical Chemistry Letters*, 10(12):3410–3414, 2019.

[158] Jennifer M. Hays, Marissa K. Kieber, Jason Z. Li, Ji In Han, Linda Columbus, and Peter M. Kasson. Refinement of Highly Flexible Protein Structures using Simulation-Guided Spectroscopy. *Angewandte Chemie - International Edition*, 57(52):17110–17114, 2018.

[159] Tania Henriquez, Larissa Wirtz, Dan Su, and Heinrich Jung. Prokaryotic solute/sodium symporters: Versatile functions and mechanisms of a transporter family †. *International Journal of Molecular Sciences*, 22(4):1–21, 2021.

[160] Lim Heo, Collin F. Arbour, and Michael Feig. Driven to near-experimental accuracy by refinement via molecular dynamics simulations. *Proteins: Structure, Function and Bioinformatics*, 87(12):1263–1275, 2019.

[161] Dawn Z. Herrick, Weiwei Kuo, Hao Huang, Charles D. Schwieters, Jeffrey F. Ellena, and David S. Cafiso. Solution and Membrane-Bound Conformations of the Tandem C2A and C2B Domains of Synaptotagmin 1: Evidence for Bilayer Bridging. *Journal of Molecular Biology*, 390(5):913–923, jul 2009.

[162] Bradley M. Hersh, Farees T. Farooq, Danielle N. Barstad, Darcy L. Blankenhorn, and Joan L. Slonczewski. A glutamate-dependent acid resistance gene in Escherichia coli. *Journal of Bacteriology*, 178(13):3978–3981, 1996.

[163] Peter W Hildebrand, Andrean Goede, Raphael A Bauer, Bjoern Gruening, Jochen Ismer, Elke Michalsky, and Robert Preissner. SuperLooper - A prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Research*, 37(SUPPL. 2), 2009.

[164] D Hilger, Y Polyhach, E Padan, H Jung, and G Jeschke. High-resolution structure of a Na+/H+ antiporter dimer obtained by pulsed electron paramagnetic resonance distance measurements. *Biophysical Journal*, 93(10):3675–3683, 2007.

[165] Stephanie J Hirst, Nathan Alexander, Hassane S Mchaourab, and Jens Meiler. RosettaEPR: An integrated tool for protein structure determination from sparse EPR data. *Journal of Structural Biology*, 173(3):506–514, 2011.

[166] Matthew D Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. Technical report, 2014.

[167] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, Alessandro Vespignani, and Tal Yarkoni. Integrating explanation and prediction in computational social science. *Nature*, jun 2021.

[168] Susanne Hofmann, Dovile Januliene, Ahmad R. Mehdipour, Christoph Thomas, Erich Stefan, Stefan Brüchert, Benedikt T. Kuhn, Eric R. Geertsma, Gerhard Hummer, Robert Tampé, and Arne Moeller. Conformation space of a heterodimeric ABC exporter under turnover conditions. *Nature*, 571(7766):580–583, 2019.

[169] H. J. Hogben, M. Krzystyniak, G. T.P. Charnock, P. J. Hore, and Ilya Kuprov. Spinach - A software library for simulation of spin dynamics in large spin systems. *Journal of Magnetic Resonance*, 208(2):179–194, 2011.

[170] Chwan Deng Hsiao, Yuh Ju Sun, John Rose, and Bi Cheng Wang. The crystal structure of glutamine-binding protein from Escherichia coli. Technical Report 2, 1996.

[171] Wayne L. Hubbell, David S. Cafiso, and Christian Altenbach. Identifying conformational changes with site-directed spin labeling. *Nature Structural Biology*, 7(9):735–739, 2000.

[172] Wayne L. Hubbell, Carlos J. López, Christian Altenbach, and Zhongyu Yang. Technological advances in site-directed spin labeling of proteins. *Current Opinion in Structural Biology*, 23(5):725–733, oct 2013.

[173] Eric Hustedt, Fabrizio Marinelli, Richard Stein, José Faraldo-Gómez, and Hassane Mchaourab. Confidence Analysis of DEER Data and its Structural Interpretation with Ensemble-Biased Metadynamics. *Confidence Analysis of DEER Data and Its Structural Interpretation with Ensemble-Biased Metadynamics*, page 299941, 2018.

[174] Hüseyin Ilgü, Jean Marc Jeckelmanna, Vytautas Gapsys, Zöhre Ucurum, Bert L. De Grootc, and Dimitrios Fotiadis. Insights into the molecular basis for substrate binding and specificity of the wild-type L-arginine/agmatine antiporter AdiC. *Proceedings of the National Academy of Sciences of the United States of America*, 113(37):10358–10363, 2016.

[175] Shahidul M. Islam and Benoît Roux. Simulating the distance distribution between spin-labels attached to proteins. *Journal of Physical Chemistry B*, 119(10):3901–3911, 2015.

[176] Shahidul M. Islam, Richard A. Stein, Hassane S. McHaourab, and Benoiît Roux. Structural refinement from restrained-ensemble simulations based on EPR/DEER data: Application to T4 lysozyme. *Journal of Physical Chemistry B*, 117(17):4740–4754, 2013.

[177] Junji Iwahara, Charles D. Schwieters, and G. Marius Clore. Ensemble Approach for NMR Structure Refinement against 1H Paramagnetic Relaxation Enhancement Data Arising from a Flexible Paramagnetic Group Attached to a Macromolecule. *Journal of the American Chemical Society*, 126(18):5879–5896, 2004.

[178] A. Jack and M. Levitt. Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Crystallographica Section A*, 34(6):931–935, 1978.

[179] Matthew P Jacobson, David L Pincus, Chaya S Rapp, Tyler J F Day, Barry Honig, David E Shaw, and Richard A Friesner. A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins: Structure, Function and Genetics*, 55(2):351–367, may 2004.

[180] Kevin L Jagessar, Derek P Claxton, Richard A Stein, and Hassane S Mchaourab. Sequence and structural determinants of ligand-dependent alternating access of a MATE transporter. *Proceedings of the National Academy of Sciences of the United States of America*, 117(9):4732–4740, mar 2020.

[181] Pavel Janoš and Alessandra Magistrato. All-atom Simulations Uncover the Molecular Terms of NKCC1 Transport Mechanism. *Journal of Chemical Information and Modeling*, pages 1–16, jul 2021.

[182] Oleg Jardetzky. Simple allosteric model for membrane pumps [27]. *Nature*, 211(5052):969–970, 1966.

[183] Jean Marc Jeckelmann and Dimitrios Fotiadis. Sub-nanometer cryo-EM density map of the human heterodimeric amino acid transporter 4F2HC-LAT2. *International Journal of Molecular Sciences*, 21(19):1–13, 2020.

[184] G Jeschke. Characterization of protein conformational changes with sparse spin-label distance constraints. *Journal of Chemical Theory and Computation*, 8(10):3854–3863, oct 2012.

[185] G Jeschke, G Panek, A Godt, A Bender, and H Paulsen. Data analysis procedures for pulse ELDOR measurements of broad distance distributions. Technical Report 1-2, 2004.

[186] Gunnar Jeschke. DEER distance measurements on proteins. *Annual Review of Physical Chemistry*, 63(January):419–446, 2012.

[187] Gunnar Jeschke. Conformational dynamics and distribution of nitroxide spin labels. *Progress in nuclear magnetic resonance spectroscopy*, 72:42–60, 2013.

[188] Gunnar Jeschke. Ensemble models of proteins and protein domains based on distance distribution restraints. *Proteins: Structure, Function and Bioinformatics*, 84(4):544–560, 2016.

[189] Gunnar Jeschke. MMM: A toolbox for integrative structure modeling. *Protein Science*, 27(1):76–85, 2018.

[190] Gunnar Jeschke. The contribution of modern EPR to structural biology. *Emerging Topics in Life Sciences*, 2(1):9–18, 2018.

[191] Gunnar Jeschke. MMM: Integrative ensemble modeling and ensemble analysis. *Protein Science*, 30(1):125–135, oct 2021.

[192] Gunnar Jeschke, V. Chechik, P. Ionita, A. Godt, H. Zimmermann, J. Banham, C. R. Timmel, D. Hilger, and H. Jung. DeerAnalysis2006 - A comprehensive software package for analyzing pulsed ELDOR data. *Applied Magnetic Resonance*, 30(3-4):473–498, 2006.

[193] Gunnar Jeschke, Achim Koch, Ulrich Jonas, and Adelheid Godt. Direct conversion of EPR dipolar time evolution data to distance distributions. *Journal of Magnetic Resonance*, 155(1):72–82, 2002.

[194] Gunnar Jeschke, Christoph Wegener, Monika Nietschke, Heinrich Jung, and Heinz Jürgen Steinhoff. Interresidual Distance Determination by Four-Pulse Double Electron-Electron Resonance in an Integral Membrane Protein: The Na +/Proline Transporter PutP of Escherichia coli. *Biophysical Journal*, 86(4):2551–2557, 2004.

[195] Sunhwan Jo, Xi Cheng, Shahidul M. Islam, Lei Huang, Huan Rui, Allen Zhu, Hui Sun Lee, Yifei Qi, Wei Han, Kenno Vanommeslaeghe, Alexander D. MacKerell, Benoît Roux, and Wonpil Im. CHARMM-GUI PDB manipulator for advanced modeling and simulations of proteins containing nonstandard residues. *Advances in Protein Chemistry and Structural Biology*, 96:235–265, 2014.

[196] Jumper John and Richard Evans. High Accuracy Protein Structure Prediction Using Deep Learning. In *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*, pages 22–24, 2020.

[197] David T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.

[198] Benesh Joseph, Arthur Sikora, and David S Cafiso. Ligand Induced Conformational Changes of a Membrane Transporter in E. coli Cells Observed with DEER/PELDOR. *Journal of the American Chemical Society*, 138(6):1844–1847, feb 2016.

[199] Manuel F. Juette, Daniel S. Terry, Michael R. Wasserman, Zhou Zhou, Roger B. Altman, Qinsi Zheng, and Scott C. Blanchard. The bright future of single-molecule fluorescence imaging. *Current Opinion in Chemical Biology*, 20(1):103–111, 2014.

[200] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, jul 2021.

[201] Katharina E J Jungnickel, Joanne L Parker, and Simon Newstead. Structural basis for amino acid transport by the CAT family of SLC7 transporters. *Nature Communications*, 9(1):1–12, 2018.

[202] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[203] Palanivel Kandasamy, Gergely Gyimesi, Yoshikatsu Kanai, and Matthias A Hediger. Amino acid transporters revisited: New views in health and disease, oct 2018.

[204] Usheer Kanjee and Walid A. Houry. Mechanisms of Acid Resistance in Escherichia coli. *Annual Review of Microbiology*, 67(1):65–81, 2013.

[205] Mert Karakaş, Nils Woetzel, Rene Staritzbichler, Nathan Alexander, Brian E. Weiner, and Jens Meiler. BCL::Fold - De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements. *PLoS ONE*, 7(11), nov 2012.

[206] Caline S. Karam and Jonathan A. Javitch. Phosphorylation of the Amino Terminus of the Dopamine Transporter: Regulatory Mechanisms and Implications for Amphetamine Action. *Advances in Pharmacology*, 82:205–234, 2018.

[207] Yasaman Karami, Frédéric Guyon, Sjoerd De Vries, and Pierre Tufféry. DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins. *Scientific Reports*, 8(1), dec 2018.

[208] Yasaman Karami, Julien Rey, Guillaume Postic, Samuel Murail, Pierre Tufféry, and Sjoerd J De Vries. DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Research*, 47(W1):W423—-W428, jul 2019.

[209] Daniel R Kattnig, Jörg Reichenwallner, and Dariush Hinderberger. Modeling excluded volume effects for the faithful description of the background signal in double electron-electron resonance. *Journal of Physical Chemistry B*, 117(51):16542–16557, dec 2013.

[210] Kelli Kazmier, Nathan S Alexander, Jens Meiler, and Hassane S Mchaourab. Algorithm for selection of optimized EPR distance restraints for de novo protein structure determination. *Journal of Structural Biology*, 173(3):549–557, mar 2011.

[211] Kelli Kazmier, Derek P. Claxton, and Hassane S. Mchaourab. Alternating access mechanisms of LeuT-fold transporters: trailblazing towards the promised energy landscapes. *Current Opinion in Structural Biology*, 45(Figure 1):100–108, 2017.

[212] Kelli Kazmier, Shruti Sharma, Shahidul M. Islam, Benoît Roux, Hassane S. Mchaourab, and Ernest M. Wright. Conformational cycle and ion-coupling mechanism of the Na+/hydantoin transporter Mhp1. *Proceedings of the National Academy of Sciences of the United States of America*, 111(41):14752–14757, 2014.

[213] Kelli Kazmier, Shruti Sharma, Matthias Quick, Shahidul M. Islam, Benoît Roux, Harel Weinstein, Jonathan A. Javitch, and Hassane S. McHaourab. Conformational dynamics of ligand-dependent alternating access in LeuT. *Nature Structural and Molecular Biology*, 21(5):472–479, 2014.

[214] Kamil Khafizov, Camilo Perez, Caroline Koshy, Matthias Quick, Klaus Fendler, Christine Ziegler, and Lucy R Forrest. Investigation of the sodium-binding sites in the sodium-coupled betaine transporter BetP. *Proceedings of the National Academy of Sciences of the United States of America*, 109(44), oct 2012.

[215] George Khelashvili, Nathaniel Stanley, Michelle A. Sahai, Jaime Medina, Michael V. LeVine, Lei Shi, Gianni De Fabritiis, and Harel Weinstein. Spontaneous Inward Opening of the Dopamine Transporter Is Triggered by PIP2-Regulated Dynamics of the N-Terminus. *ACS Chemical Neuroscience*, 6(11):1825–1837, 2015.

[216] David E. Kim, Dylan Chivian, and David Baker. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 32(WEB SERVER ISS.):526–531, 2004.

[217] Miyeon Kim, Sergey A. Vishnivetskiy, Ned Van Eps, Nathan S. Alexander, Whitney M. Cleghorn, Xuanzhi Zhan, Susan M. Hanson, Takefumi Morizumi, Oliver P. Ernst, Jens Meiler, Vsevolod V. Gurevich, and Wayne L. Hubbell. Conformation of receptor-bound visual arrestin. *Proceedings of the National Academy of Sciences of the United States of America*, 109(45):18407–18412, 2012.

[218] Seung Joong Kim, Javier Fernandez-Martinez, Ilona Nudelman, Yi Shi, Wenzhu Zhang, Barak Raveh, Thurston Herricks, Brian D Slaughter, Joanna A Hogan, Paula Upla, Ilan E

Chemmama, Riccardo Pellarin, Ignacia Echeverria, Manjunatha Shivaraju, Azraa S Chaudhury, Junjie Wang, Rosemary Williams, Jay R Unruh, Charles H Greenberg, Erica Y Jacobs, Zhiheng Yu, M Jason De La Cruz, Roxana Mironska, David L Stokes, John D Aitchison, Martin F Jarrold, Jennifer L Gerton, Steven J Ludtke, Christopher W Akey, Brian T Chait, Andrej Sali, and Michael P Rout. Integrative structure and functional anatomy of a nuclear pore complex. *Nature*, 555(7697):475–482, mar 2018.

[219] Sunghoon Kim, Suzanne Brandon, Zheng Zhou, Charles E Cobb, Sarah J Edwards, Christopher W Moth, Christian S Parry, Jarrod A Smith, Terry P Lybrand, Eric J Hustedt, and Albert H Beth. Determination of structural models of the complex between the cytoplasmic domain of erythrocyte band 3 and ankyrin-R repeats 13-24. *Journal of Biological Chemistry*, 286(23):20746–20757, jun 2011.

[220] Daniel Klose, Johann P Klare, Dina Grohmann, Christopher W M Kay, Finn Werner, and Heinz Jürgen Steinhoff. Simulation vs. reality: A comparison of in silico distance predictions with DEER and FRET measurements. *PLoS ONE*, 7(6), 2012.

[221] Lukasz Kowalczyk, Mercè Ratera, Antonella Paladino, Paola Bartoccioni, Ekaitz Errasti-Murugarren, Eva Valencia, Guillem Portella, Susanna Bial, Antonio Zorzano, Ignacio Fita, Modesto Orozco, Xavier Carpena, José Luis Vázquez-Ibar, and Manuel Palacín. Molecular basis of substrate-induced permeation by an amino acid antiporter. *Proceedings of the National Academy of Sciences of the United States of America*, 108(10):3935–3940, 2011.

[222] Reinhard Krämer and Christine Ziegler. Regulative interactions of the osmosensing C-terminal domain in the trimeric glycine betaine transporter BetP from Corynebacterium glutamicum. *Biological Chemistry*, 390(8):685–691, aug 2009.

[223] Eva Maria Krammer and Martine Prévost. Function and Regulation of Acid Resistance Antiporters. *Journal of Membrane Biology*, 252(4-5):465–481, oct 2019.

[224] Harini Krishnamurthy and Eric Gouaux. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature*, 481(7382):469–474, 2012.

[225] Georgii G. Krivov, Maxim V. Shapovalov, and Roland L. Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function and Bioinformatics*, 77(4):778–795, 2009.

[226] Ulrike Krug, Nathan S Alexander, Richard A Stein, Antje Keim, Hassane S McHaourab, Norbert Sträter, and Jens Meiler. Characterization of the Domain Orientations of E. coli 5 -Nucleotidase by Fitting an Ensemble of Conformers to DEER Distance Distributions. *Structure*, 24(1):43–56, 2016.

[227] Andrey A Kuzhelev, Rodion K Strizhakov, Olesya A Krumkacheva, Yuliya F Polienko, Denis A Morozov, Georgiy Yu Shevelev, Dmitrii V Pyshnyi, Igor A Kirilyuk, Matvey V Fedin, and Elena G Bagryanskaya. Room-temperature electron spin relaxation of nitroxides immobilized in trehalose: Effect of substituents adjacent to NO-group. *Journal of Magnetic Resonance*, 266:1–7, may 2016.

[228] Alex L. Lai, Hao Huang, Dawn Z. Herrick, Natalie Epp, and David S. Cafiso. Synaptotagmin 1 and SNAREs form a complex that is structurally heterogeneous. *Journal of Molecular Biology*, 405(3):696–706, jan 2011.

[229] Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487(C):545–574, 2011.

[230] Yongchan Lee, Pattama Wiriyasermkul, Chunhuan Jin, Lili Quan, Ryuichi Ohgaki, Suguru Okuda, Tsukasa Kusakizako, Tomohiro Nishizawa, Kazumasa Oda, Ryuichiro Ishitani, Takeshi Yokoyama, Takanori Nakane, Mikako Shirouzu, Hitoshi Endou, Shushi Nagamori, Yoshikatsu Kanai, and Osamu Nureki. Cryo-EM structure of the human L-type amino acid transporter 1 in complex with glycoprotein CD98hc. *Nature Structural and Molecular Biology*, 26(6):510–517, jun 2019.

[231] Hsiang Ting Lei, Jinming Ma, Silvia Sanchez Martinez, and Tamir Gonen. Crystal structure of arginine-bound lysosomal transporter SLC38A9 in the cytosol-open state. *Nature Structural and Molecular Biology*, 25(6):522–527, 2018.

[232] Hsiang Ting Lei, Xuelang Mu, Johan Hattne, and Tamir Gonen. A conformational change in the N terminus of SLC38A9 signals mTORC1 activation. *Structure*, 29(5):426–432.e8, 2021.

[233] Julia Koehler Leman, Ralf Mueller, Mert Karakas, Nils Woetzel, and Jens Meiler. Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins: Structure, Function and Bioinformatics*, 81(7):1127–1140, 2013.

[234] Julia Koehler Leman, Brian D. Weitzner, Steven M. Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F. Alford, Melanie Aprahamian, David Baker, Kyle A. Barlow, Patrick Barth, Benjamin Basanta, Brian J. Bender, Kristin Blacklock, Jaume Bonet, Scott E. Boyken, Phil Bradley, Chris Bystroff, Patrick Conway, Seth Cooper, Bruno E. Correia, Brian Coventry, Rhiju Das, René M. De Jong, Frank DiMaio, Lorna Dsilva, Roland Dunbrack, Alexander S. Ford, Brandon Frenz, Darwin Y. Fu, Caleb Geniesse, Lukasz Goldschmidt, Ragul Gowthaman, Jeffrey J. Gray, Dominik Gront, Sharon Guffy, Scott Horowitz, Po Ssu Huang, Thomas Huber, Tim M. Jacobs, Jeliazko R. Jeliazkov, David K. Johnson, Kalli Kappel, John Karanicolas, Hamed Khakzad, Karen R. Khar, Sagar D. Khare, Firas Khatib, Alisa Khramushin, Indigo C. King, Robert Kleffner, Brian Koepnick, Tanja Kortemme, Georg Kuenze, Brian Kuhlman, Daisuke Kuroda, Jason W. Labonte, Jason K. Lai, Gideon Lapidoth, Andrew Leaver-Fay, Steffen Lindert, Thomas Linsky, Nir London, Joseph H. Lubin, Sergey Lyskov, Jack Maguire, Lars Malmström, Enrique Marcos, Orly Marcu, Nicholas A. Marze, Jens Meiler, Rocco Moretti, Vikram Khipple Mulligan, Santrupti Nerli, Christoffer Norn, Shane Ó'Conchúir, Noah Ollikainen, Sergey Ovchinnikov, Michael S. Pacella, Xingjie Pan,

Hahnbeom Park, Ryan E. Pavlovicz, Manasi Pethe, Brian G. Pierce, Kala Bharath Pilla, Barak Raveh, P. Douglas Renfrew, Shourya S.Roy Burman, Aliza Rubenstein, Marion F. Sauer, Andreas Scheck, William Schief, Ora Schueler-Furman, Yuval Sedan, Alexander M. Sevy, Nikolaos G. Sgourakis, Lei Shi, Justin B. Siegel, Daniel Adriano Silva, Shannon Smith, Yifan Song, Amelie Stein, Maria Szegedy, Frank D. Teets, Summer B. Thyme, Ray Yu Ruei Wang, Andrew Watkins, Lior Zimmerman, and Richard Bonneau. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods*, 17(7):665–680, 2020.

[235] Vanessa Leone, Izabela Waclawska, Katharina Kossmann, Caroline Koshy, Monika Sharma, Thomas F. Prisner, Christine Ziegler, Burkhard Endeward, and Lucy R. Forrest. Interpretation of spectroscopic data using molecular simulations for the secondary active transporter BetP. *Journal of General Physiology*, 151(3):381–394, mar 2019.

[236] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. Technical Report 2, 1944.

[237] Michael V. LeVine, Daniel S. Terry, George Khelashvili, Zarek S. Siegel, Matthias Quick, Jonathan A. Javitch, Scott C. Blanchard, and Harel Weinstein. The allosteric mechanism of substrate-specific transport in SLC6 is mediated by a volumetric sensor. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15947–15956, 2019.

[238] Jing Li, Saher A Shaikh, Giray Enkavi, Po Chao Wen, Zhijian Huang, and Emad Tajkhorshid. Transient formation of water-conducting states in membrane transporters. *Proceedings of the National Academy of Sciences of the United States of America*, 110(19):7696–7701, may 2013.

[239] Ying Li, Sergey Korolev, and Gabriel Waksman. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: Structural basis for nucleotide incorporation. Technical Report 24, 1998.

[240] Shide Liang, Chi Zhang, and Yaoqi Zhou. LEAP: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *Journal of Computational Chemistry*, 35(4):335–341, feb 2014.

[241] George Liapakis, Merrill M. Simpson, and Jonathan A. Javitch. The Substituted-Cysteine Accessibility Method (SCAM) to Elucidate Membrane Protein Structure . *Current Protocols in Neuroscience*, 8(1):1–10, 1999.

[242] James E.D. Lillington, Janet E. Lovett, Steven Johnson, Pietro Roversi, Christiane R. Timmel, and Susan M. Lea. Shigella flexneri Spa15 crystal structure verified in solution by double electron electron resonance. *Journal of Molecular Biology*, 405(2):427–435, 2011.

[243] Sunghyuk Lim, Graham Roseman, Igor Peshenko, Grace Manchala, Diana Cudia, Alexander M Dizhoor, Glenn Millhauser, and James B Ames. Retinal guanylyl cyclase activating protein 1 forms a functional dimer. *PLoS ONE*, 13(3), mar 2018.

[244] Robert F. Ling, Charles L. Lawson, and Richard J. Hanson. *Solving Least Squares Problems.*, volume 72. 1977.

[245] Shenglong Ling, Wei Wang, Lu Yu, Junhui Peng, Xiaoying Cai, Ying Xiong, Zahra Hayati, Longhua Zhang, Zhiyong Zhang, Likai Song, and Changlin Tian. Structure of an E. coli integral membrane sulfurtransferase and its structural transition upon SCN- binding defined by EPR-based hybrid method. *Scientific Reports*, 6(October):1–12, 2016.

[246] Si Liu, Shenghai Chang, Binming Han, Lingyi Xu, Mingfeng Zhang, Cheng Zhao, Wei Yang, Feng Wang, Jingyuan Li, Eric Delpire, Sheng Ye, Xiao Chen Bai, and Jiangtao Guo. Cryo-EM structures of the human cation-chloride cotransporter KCC1. *Science*, 366(6464):505–508, 2019.

[247] Claus J. Loland. The use of LeuT as a model in elucidating binding sites for substrates and inhibitors in neurotransmitter transporters. *Biochimica et Biophysica Acta - General Subjects*, 1850(3):500–510, 2015.

[248] Claus Juul Loland, Lene Norregaard, Thomas Litman, and Ulrik Gether. Generation of an activating Zn2+ switch in the dopamine transporter: Mutation of an intracellular tyrosine constitutively alters the conformational equilibrium of the transport cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1683–1688, 2002.

[249] Donald D.F. Loo, Bruce A. Hirayama, Elsa M. Gallardo, Jason T. Lam, Eric Turk, and Ernest M. Wright. Conformational changes couple Na+ and glucose transport. *Proceedings of the National Academy of Sciences of the United States of America*, 95(13):7789–7794, 1998.

[250] Donald D.F. Loo, Bruce A. Hirayama, Movses H. Karakossian, Anne Kristine Meinild, and Ernest M. Wright. Conformational dynamics of hSGLT1 during Na +/glucose cotransport. *Journal of General Physiology*, 128(6):701–720, 2006.

[251] Peilong Lu, Dan Ma, Yuling Chen, Yingying Guo, Guo Qiang Chen, Haiteng Deng, and Yigong Shi. L-glutamine provides acid resistance for Escherichia coli through enzymatic release of ammonia. *Cell Research*, 23(5):635–644, 2013.

[252] Mark A. Lukas. Robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems*, 22(5):1883–1902, sep 2006.

[253] Dan Ma, Peilong Lu, and Yigong Shi. Substrate selectivity of the acid-activated glutamate/γ-aminobutyric acid (GABA) antiporter GadC from Escherichia coli. *Journal of Biological Chemistry*, 288(21):15148–15153, 2013.

[254] Dan Ma, Peilong Lu, Chuangye Yan, Chao Fan, Ping Yin, Jiawei Wang, and Yigong Shi. Structure and mechanism of a glutamate-GABA antiporter. *Nature*, 483(7391):632–636, 2012.

[255] Jinming Ma, Hsiang Ting Lei, Francis E Reyes, Silvia Sanchez-Martinez, Maen F Sarhan, Johan Hattne, and Tamir Gonen. Structural basis for substrate binding and specificity of a

sodium-alanine symporter AgcS. *Proceedings of the National Academy of Sciences of the United States of America*, 116(6):2086–2090, 2019.

[256] Justin L. MacCallum, Alberto Perez, and Ken A. Dill. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22):6985–6990, 2015.

[257] Ulrika Magnusson, Branka Salopek-Sondi, Linda A Luck, and Sherry L Mowbray. X-ray Structures of the Leucine-binding Protein Illustrate Conformational Changes and the Basis of Ligand Specificity. *Journal of Biological Chemistry*, 279(10):8747–8752, mar 2004.

[258] Lina Malinauskaite, Matthias Quick, Linda Reinhard, Joseph A Lyons, Hideaki Yano, Jonathan A Javitch, and Poul Nissen. A mechanism for intracellular release of Na+ by neurotransmitter/sodium symporters. *Nature Structural and Molecular Biology*, 21(11):1006–1012, 2014.

[259] Lina Malinauskaite, Saida Said, Caglanur Sahin, Julie Grouleff, Azadeh Shahsavar, Henriette Bjerregaard, Pernille Noer, Kasper Severinsen, Thomas Boesen, Birgit Schiøtt, Steffen Sinning, and Poul Nissen. A conserved leucine occupies the empty substrate site of LeuT in the Na+-free return state. *Nature Communications*, 7(1):1–11, 2016.

[260] Daniel J. Mandell, Evangelos A. Coutsias, and Tanja Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, 6(8):551–552, 2009.

[261] Aashish Manglik, Tae Hun Kim, Matthieu Masureel, Christian Altenbach, Zhongyu Yang, Daniel Hilger, Michael T. Lerch, Tong Sun Kobilka, Foon Sun Thian, Wayne L. Hubbell, R. Scott Prosser, and Brian K. Kobilka. Structural insights into the dynamic process of β2-adrenergic receptor signaling. *Cell*, 161(5):1101–1111, 2015.

[262] Fabrizio Marinelli and José D. Faraldo-Gómez. Ensemble-Biased Metadynamics: A Molecular Simulation Method to Sample Experimental Distributions. *Biophysical Journal*, 108(12):2779–2782, 2015.

[263] Fabrizio Marinelli and Giacomo Fiorin. Structural Characterization of Biomolecules through Atomistic Simulations Guided by DEER Measurements. *Structure*, 27(2):359–370.e12, 2019.

[264] Andriy Marko and Thomas F Prisner. An algorithm to analyze PELDOR data of rigid spin label pairs. *Physical Chemistry Chemical Physics*, 15(2):619–627, 2013.

[265] Claire Marks, Jaroslaw Nowak, Stefan Klostermann, Guy Georges, James Dunbar, Jiye Shi, Sebastian Kelm, and Charlotte M Deane. Sphinx: Merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*, 33(9):1346–1353, may 2017.

[266] Donald W. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[267] Chloé Martens, Richard A. Stein, Matthieu Masureel, Aurélie Roth, Smriti Mishra, Rosie Dawaliby, Albert Konijnenberg, Frank Sobott, Cédric Govaerts, and Hassane S. McHaourab. Lipids modulate the conformational dynamics of a secondary multidrug transporter. *Nature Structural and Molecular Biology*, 23(8):744–751, 2016.

[268] A C R Martin, J C Cheetham, and A R Rees. Modeling antibody hypervariable loops: A combined algorithm. *Proceedings of the National Academy of Sciences of the United States of America*, 86(23):9268–9272, 1989.

[269] Maria T Mas, Kenneth C Smith, David L Yarmush, Kazuo Aisaka, and Richard M Fine. Modeling the anti-CEA antibody combining site by homology and conformational search. Technical Report 4, 1992.

[270] Matthieu Masureel, Chloé Martens, Richard A Stein, Smriti Mishra, Jean-Marie Ruysschaert, Hassane S Mchaourab, and Cédric Govaerts. Protonation drives the conformational switch in the multidrug transporter LmrP. *Nature chemical biology*, 10(2):149–155, feb 2014.

[271] Erez Matalon, Thomas Huber, Gregor Hagelueken, Bim Graham, Veronica Frydman, Akiva Feintuch, Gottfried Otting, and Daniella Goldfarb. Gadolinium(III) spin labels for high-sensitivity distance measurements in transmembrane helices. *Angewandte Chemie - International Edition*, 52(45):11831–11834, nov 2013.

[272] Anna G. Matveeva, Vyacheslav M. Nekrasov, and Alexander G. Maryasov. Analytical solution of the PELDOR inverse problem using the integral Mellin transform. *Physical Chemistry Chemical Physics*, 19(48):32381–32388, dec 2017.

[273] Tatiana Maximova, Ryan Moffatt, Buyong Ma, Ruth Nussinov, and Amarda Shehu. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics, apr 2016.

[274] Hassane S McHaourab, P Ryan Steed, and Kelli Kazmier. Toward the fourth dimension of membrane protein structure: Insight into dynamics from spin-labeling EPR spectroscopy. *Structure*, 19(11):1549–1561, 2011.

[275] Catherine A McPhalen, Michael G Vincent, and Johan N Jansonius. X-ray structure refinement and comparison of three forms of mitochondrial aspartate aminotransferase. *Journal of Molecular Biology*, 225(2):495–517, 1992.

[276] Catherine A McPhalen, Michael G Vincent, Daniel Picot, Johan N Jansonius, Arthur M Lesk, and Cyrus Chothia. Domain closure in mitochondrial aspartate aminotransferase. *Journal of Molecular Biology*, 227(1):197–213, 1992.

[277] Jens Meiler and David Baker. Rapid protein fold determination using unassigned NMR data. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15404–15409, 2003.

[278] Anne Kristine Meinild, Bruce A. Hirayama, Ernest M. Wright, and Donald D.F. Loo. Fluorescence studies of ligand-induced conformational changes of the Na+/glucose cotransporter. *Biochemistry*, 41(4):1250–1258, 2002.

[279] Jeffrey Mendenhall and Jens Meiler. Prediction of Transmembrane Proteins and Regions using Fourier Spectral Analysis and Advancements in Machine Learning. *Sermacs 2014*, (April 2015):2–3, 2014.

[280] F Mentink-Vigier, A Collauto, A Feintuch, I Kaminker, V Tarle, and D Goldfarb. Increasing sensitivity of pulse EPR experiments using echo train detection schemes. *Journal of Magnetic Resonance*, 236:117–125, 2013.

[281] Patrick S Merkle, Kamil Gotfryd, Michel A Cuendet, Katrine Z Leth-Espensen, Ulrik Gether, Claus J Loland, and Kasper D Rand. Substrate-modulated unwinding of transmembrane helices in the NSS transporter LeuT. Technical Report 5, 2018.

[282] Gregory E Merz, Peter P Borbat, Ashley J Pratt, Elizabeth D Getzoff, Jack H Freed, and Brian R Crane. Copper-based pulsed dipolar ESR spectroscopy as a probe of protein conformation linked to disease states. *Biophysical Journal*, 107(7):1669–1674, oct 2014.

[283] Virginia Meyer, Michael A Swanson, Laura J Clouston, Przemysław J Boratyński, Richard A Stein, Hassane S McHaourab, Andrzej Rajca, Sandra S Eaton, and Gareth R Eaton. Room-temperature distance measurements of immobilized Spin-labeled Protein by DEER/PELDOR. *Biophysical Journal*, 108(5):1213–1219, mar 2015.

[284] E Michalsky, A Goede, and R Preissner. Loops In Proteins (LIP) - A comprehensive loop database for homology modelling. *Protein Engineering*, 16(12):979–985, dec 2003.

[285] Sergey Milikisiyants, Shenlin Wang, Rachel A. Munro, Matthew Donohue, Meaghan E. Ward, David Bolton, Leonid S. Brown, Tatyana I. Smirnova, Vladimir Ladizhansky, and Alex I. Smirnov. Oligomeric Structure of Anabaena Sensory Rhodopsin in a Lipid Bilayer Environment by Combining Solid-State NMR and Long-range DEER Constraints. *Journal of Molecular Biology*, 429(12):1903–1920, jun 2017.

[286] A D Milov, A G Maryasov, and Y D Tsvetkov. Pulsed electron double resonance (PELDOR) and its applications in free-radicals research. Technical Report 1, 1998.

[287] A D Milov, A B Ponomarev, and Yu D Tsvetkov. Electron-electron double resonance in electron spin echo: Model biradical systems and the sensitized photolysis of decalin. Technical Report 1, 1984.

[288] Smriti Mishra, Brandy Verhalen, Richard A. Stein, Po Chao Wen, Emad Tajkhorshid, and Hassane S. Mchaourab. Conformational dynamics of the nucleotide binding domains and the power stroke of a heterodimeric ABC transporter. *eLife*, 2014(3):e02740, 2014.

[289] Peter Mitchell. A general theory of membrane transport from studies of bacteria. *Nature*, 180(4577):134–136, jul 1957.

[290] Shriyaa Mittal and Diwakar Shukla. Predicting Optimal DEER Label Positions to Study Protein Conformational Heterogeneity. *Journal of Physical Chemistry B*, 121(42):9761–9770, oct 2017.

[291] Ingvar R. Möller, Patrick S. Merkle, Dionisie Calugareanu, Gerard Comamala, Solveig Gaarde Schmidt, Claus J. Loland, and Kasper D. Rand. Probing the conformational impact of detergents on the integral membrane protein LeuT by global HDX-MS. *Journal of Proteomics*, 225(May):103845, 2020.

[292] Ingvar R. Möller, Marika Slivacka, Anne Kathrine Nielsen, Søren G.F. Rasmussen, Ulrik Gether, Claus J. Loland, and Kasper D. Rand. Conformational dynamics of the human serotonin transporter during substrate and drug binding. *Nature Communications*, 10(1), dec 2019.

[293] Michelle Y. Monette, Suma Somasekharan, and Biff Forbush. Molecular motions involved in NA-K-Cl cotransporter-mediated ion transport and transporter activation revealed by internal cross-linking between transmembrane domains 10 and 11/12. *Journal of Biological Chemistry*, 289(11):7569–7579, 2014.

[294] Anna Mullen, Jenny Hall, Janika Diegel, Isa Hassan, Adam Fey, and Fraser MacMillan. Membrane transporters studied by EPR spectroscopy: Structure determination and elucidation of functional dynamics. *Biochemical Society Transactions*, 44(3):905–915, 2016.

[295] Lara Napolitano, Michele Galluccio, Mariafrancesca Scalise, Chiara Parravicini, Luca Palazzolo, Ivano Eberini, and Cesare Indiveri. Novel insights into the transport mechanism of the human amino acid transporter LAT1 (SLC7A5). Probing critical residues for substrate translocation. *Biochimica et Biophysica Acta - General Subjects*, 1861(4):727–736, apr 2017.

[296] Radford M Neal. Probabilistic Inference Using Markov Chain Monte Carlo Methods Acknowledgements. Technical Report CRG-TR-93-1. Technical Report September, 1993.

[297] John M. Nicoludis and Rachelle Gaudet. Applications of sequence coevolution in membrane protein biochemistry. *Biochimica et Biophysica Acta - Biomembranes*, 1860(4):895–908, 2018.

[298] Anne Kathrine Nielsen, Ingvar R. Möller, Yong Wang, Søren G.F. Rasmussen, Kresten Lindorff-Larsen, Kasper D. Rand, and Claus J. Loland. Substrate-induced conformational dynamics of the dopamine transporter. *Nature Communications*, 10(1):1–14, 2019.

[299] G E Norris, B F Anderson, and E N Baker. Molecular replacement solution of the structure of apolactoferrin, a protein displaying large-scale conformational change. *Acta Crystallographica Section B*, 47(6):998–1004, 1991.

[300] Kazumasa Oda, Yongchan Lee, Pattama Wiriyasermkul, Yoko Tanaka, Mizuki Takemoto, Keitaro Yamashita, Shushi Nagamori, Tomohiro Nishizawa, and Osamu Nureki. Consensus mutagenesis approach improves the thermal stability of system xc− transporter, xCT, and enables cryo-EM analyses. *Protein Science*, 29(12):2398–2407, dec 2020.

[301] Irina Oganesyan, Cristina Lento, and Derek J. Wilson. Contemporary hydrogen deuterium exchange mass spectrometry. *Methods*, 144(April):27–42, 2018.

[302] Sergey Ovchinnikov, Lisa Kinch, Hahnbeom Park, Yuxing Liao, Jimin Pei, David E. Kim, Hetunandan Kamisetty, Nick V. Grishin, and David Baker. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife*, 4(September):1–25, 2015.

[303] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298, 2017.

[304] M Pannier, S Veit, A Godt, G Jeschke, and H W Spiess. Dead-Time Free Measurement of Dipole–Dipole Interactions between Electron Spins. *Journal of Magnetic Resonance*, 142(2):331–340, 2000.

[305] Hahnbeom Park, Gyu Rie Lee, Lim Heo, and Chaok Seok. Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments. *PLoS ONE*, 9(11), nov 2014.

[306] Hahnbeom Park, Gyu Rie Lee, David E. Kim, Ivan Anishchenko, Qian Cong, and David Baker. High-accuracy refinement using Rosetta in CASP13. *Proteins: Structure, Function and Bioinformatics*, 87(12):1276–1282, 2019.

[307] Hahnbeom Park, Sergey Ovchinnikov, David E. Kim, Frank DiMaio, and David Baker. Protein homology model refinement by large-scale energy optimization. *Proceedings of the National Academy of Sciences of the United States of America*, 115(12):3054–3059, 2018.

[308] Sang Youn Park, Peter P. Borbat, Gabriela Gonzalez-Bonet, Jaya Bhatnagar, Abiola M. Pollard, Jack H. Freed, Alexandrine M. Bilwes, and Brian R. Crane. Reconstruction of the chemotaxis receptor-kinase assembly. *Nature Structural and Molecular Biology*, 13(5):400–407, may 2006.

[309] Simon G. Patching. Recent developments in nucleobase cation symporter-1 (NCS1) family transport proteins from bacteria, archaea, fungi and plants. *Journal of Biosciences*, 43(4):797–815, 2018.

[310] Aviv Paz, Derek P. Claxton, Jay Prakash Kumar, Kelli Kazmier, Paola Bisignano, Shruti Sharma, Shannon A. Nolte, Terrin M. Liwag, Vinod Nayak, Ernest M. Wright, Michael Grabe, Hassane S. McHaourab, and Jeff Abramson. Conformational transitions of the sodium-dependent sugar transporter, vSGLT. *Proceedings of the National Academy of Sciences of the United States of America*, 115(12):E2742–E2751, 2018.

[311] Aravind Penmatsa, Kevin H Wang, and Eric Gouaux. X-ray structure of dopamine transporter elucidates antidepressant mechanism. *Nature*, 503(7474):85–90, 2013.

[312] Aravind Penmatsa, Kevin H Wang, and Eric Gouaux. X-ray structures of Drosophila dopamine transporter in complex with nisoxetine and reboxetine. *Nature Structural and Molecular Biology*, 22(6):506–508, jun 2015.

[313] Alberto Perez, Joseph A. Morrone, Emiliano Brini, Justin L. MacCallum, and Ken A. Dill. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances*, 2(11):1–7, 2016.

[314] Camilo Perez, Belinda Faust, Ahmad Reza Mehdipour, Kevin A. Francesconi, Lucy R. Forrest, and Christine Ziegler. Substrate-bound outward-open state of the betaine transporter BetP provides insights into Na+ coupling. *Nature Communications*, 5, jul 2014.

[315] Camilo Perez, Kamil Khafizov, Lucy R Forrest, Reinhard Krämer, and Christine Ziegler. The role of trimerization in the osmoregulated betaine transporter BetP. *EMBO Reports*, 12(8):804–810, aug 2011.

[316] Camilo Perez, Caroline Koshy, Özkan Yildiz, and Christine Ziegler. Alternating-access mechanism in conformationally asymmetric trimers of the betaine transporter BetP. *Nature*, 490(7418):126–130, 2012.

[317] Martin F. Peter, Anne T. Tuukkanen, Caspar A. Heubach, Alexander Selsam, Fraser G. Duthie, Dmitri I. Svergun, Olav Schiemann, and Gregor Hagelueken. Studying Conformational Changes of the Yersinia Type-III-Secretion Effector YopO in Solution by Integrative Structural Biology. *Structure*, 27(9):1416–1426.e3, sep 2019.

[318] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[319] Richard L. Pio. Euler Angle Transformations. *IEEE Transactions on Automatic Control*, 11(4):707–715, 1966.

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

[321] Chayne L Piscitelli, Harini Krishnamurthy, and Eric Gouaux. Neurotransmitter/sodium symporter orthologue LeuT has a single high-affinity substrate site. *Nature*, 468(7327):1129–1133, 2010.

[322] Yevhen Polyhach, Enrica Bordignon, and Gunnar Jeschke. Rotamer libraries of spin labelled cysteines for protein studies. *Physical Chemistry Chemical Physics*, 13(6):2356–2366, 2011.

[323] Luca Ponzoni, She Zhang, Mary Hongying Cheng, and Ivet Bahar. Shared dynamics of LeuT superfamily members and allosteric differentiation by structural irregularities and multimerization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1749), 2018.

[324] Alexey Potapov, Hiromasa Yagi, Thomas Huber, Slobodan Jergic, Nicholas E. Dixon, Gottfried Otting, and Daniella Goldfarb. Nanometer-scale distance measurements in proteins using Gd3+ spin labeling. *Journal of the American Chemical Society*, 132(26):9040–9048, 2010.

[325] Florian A Potra and Stephen J Wright. Interior-point methods. Technical report, 2000.

[326] R Powers, G M Clore, D S Garrett, and A M Gronenborn. Relationships Between the Precision of High-Resolution Protein NMR Structures, Solution-Order Parameters, and Crystallographic B Factors. *Journal of Magnetic Resonance, Series B*, 101(3):325–327, 1993.

[327] Eric A Price, Brian T Sutch, Qi Cai, Peter Z Qin, and Ian S Haworth. Computation of nitroxide-nitroxide distances in spin-labeled DNA duplexes. *Biopolymers*, 87(1):40–50, sep 2007.

[328] Matthias Quick and Jonathan A. Javitch. Monitoring the function of membrane transport proteins in detergent-solubilized form. *Proceedings of the National Academy of Sciences of the United States of America*, 104(9):3603–3608, 2007.

[329] Matthias Quick, Lei Shi, Britta Zehnpfennig, Harel Weinstein, and Jonathan A Javitch. Experimental conditions can obscure the second high-affinity site in LeuT. *Nature Structural and Molecular Biology*, 19(2):207–212, 2012.

[330] Matthias Quick, Anne Marie Lund Winther, Lei Shi, Poul Nissen, Harel Weinstein, and Jonathan A. Javitch. Binding of an octylglucoside detergent molecule in the second substrate (S2) site of LeuT establishes an inhibitor-bound conformation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5563–5568, 2009.

[331] Michael Raba, Sabrina Dunkel, Daniel Hilger, Kamila Lipiszko, Yevhen Polyhach, Gunnar Jeschke, Susanne Bracher, Johann P. Klare, Matthias Quick, Heinrich Jung, and Heinz Jürgen Steinhoff. Extracellular loop 4 of the proline transporter PutP controls the periplasmic entrance to ligand binding sites. *Structure*, 22(5):769–780, may 2014.

[332] H Raghuraman, Shahidul M Islam, Soumi Mukherjee, Benoit Roux, and Eduardo Perozo. Dynamics transitions at the outer vestibule of the KcsA potassium channel during gating. *Proceedings of the National Academy of Sciences of the United States of America*, 111(5):1831–1836, 2014.

[333] Asghar M. Razavi, George Khelashvili, and Harel Weinstein. How structural elements evolving from bacterial to human SLC6 transporters enabled new functional properties. *BMC Biology*, 16(1):1–14, 2018.

[334] Randy J. Read, Massimo D. Sammito, Andriy Kryshtafovych, and Tristan I. Croll. Evaluation of model refinement in CASP13. *Proteins: Structure, Function and Bioinformatics*, 87(12):1249–1262, 2019.

[335] Manuele Rebsamen, Lorena Pochini, Taras Stasyk, Mariana E.G. De Araújo, Michele Galluccio, Richard K. Kandasamy, Berend Snijder, Astrid Fauster, Elena L. Rudashevskaya,

Manuela Bruckner, Stefania Scorzoni, Przemyslaw A. Filipek, Kilian V.M. Huber, Johannes W. Bigenzahn, Leonhard X. Heinz, Claudine Kraft, Keiryn L. Bennett, Cesare Indiveri, Lukas A. Huber, and Giulio Superti-Furga. SLC38A9 is a component of the lysosomal amino acid sensing machinery that controls mTORC1. *Nature*, 519(7544):477–481, 2015.

[336] Katrin Reichel, Lukas S. Stelzl, Jürgen Köfinger, and Gerhard Hummer. Precision DEER Distances from Spin-Label Ensemble Refinement. *Journal of Physical Chemistry Letters*, 9(19):5748–5752, 2018.

[337] Michelle S. Reid, David M. Kern, and Stephen Graf Brohawn. Cryo-EM structure of the potassium-chloride cotransporter KCC4 in lipid nanodiscs. *eLife*, 9, apr 2020.

[338] Núria Reig, César Del Rio, Fabio Casagrande, Mercè Ratera, Josep Lluís Gelpí, David Torrents, Peter J.F. Henderson, Hao Xie, Stephen A. Baldwin, Antonio Zorzano, Dimitrios Fotiadis, and Manuel Palacín. Functional and structural characterization of the first prokaryotic member of the L-amino acid transporter (LAT) family: A model for APC transporters. *Journal of Biological Chemistry*, 282(18):13270–13281, may 2007.

[339] Susanne Ressl, Anke C. Terwisscha Van Scheltinga, Clemens Vonrhein, Vera Ott, and Christine Ziegler. Molecular basis of transport and regulation in the Na+/betaine symporter BetP. *Nature*, 458(7234):47–52, 2009.

[340] Hope Richard and John W. Foster. Escherichia coli glutamate- and arginine-dependent acid resistance systems increase internal pH and reverse transmembrane potential. *Journal of Bacteriology*, 186(18):6032–6041, 2004.

[341] Carol A. Rohl, Charlie E.M. Strauss, Dylan Chivian, and David Baker. Modeling Structurally Variable Regions in Homologous Proteins with Rosetta. *Proteins: Structure, Function and Genetics*, 55(3):656–677, apr 2004.

[342] Michael P. Rout and Andrej Sali. Principles for Integrative Structural Biology Studies. *Cell*, 177(6):1384–1403, may 2019.

[343] Benoiît Roux and Shahidul M Islam. Restrained-ensemble molecular dynamics simulations based on distance histograms from double electron-electron resonance spectroscopy. *Journal of Physical Chemistry B*, 117(17):4733–4739, 2013.

[344] Gary Rudnick. Serotonin transporters - Structure and function. *Journal of Membrane Biology*, 213(2):101–110, 2006.

[345] Gary Rudnick and Walter Sandtner. Serotonin transport in the 21st century. *Journal of General Physiology*, 151(11):1248–1264, 2019.

[346] John M Russell. Sodium-Potassium-Chloride Cotransport. *Physiology Reviews*, pages 211–276, 2000.

[347] Indra D. Sahu, Andrew F. Craig, Megan M. Dunagum, Robert M. McCarrick, and Gary A. Lorigan. Characterization of Bifunctional Spin Labels for Investigating the Structural and Dynamic Properties of Membrane Proteins Using EPR Spectroscopy. *Journal of Physical Chemistry B*, 121(39):9185–9195, oct 2017.

[348] Indra D. Sahu and Gary A. Lorigan. Site-Directed Spin Labeling EPR for Studying Membrane Proteins. *BioMed Research International*, 2018:1–13, 2018.

[349] Milton H. Saier, Vamsee S. Reddy, Brian V. Tsu, Muhammad Saad Ahmed, Chun Li, and Gabriel Moreno-Hagelsieb. The Transporter Classification Database (TCDB): Recent advances. *Nucleic Acids Research*, 44(D1):D372–D379, 2016.

[350] Tomohide Saio, Soya Hiramatsu, Mizue Asada, Hiroshi Nakagawa, Kazumi Shimizu, Hiroyuki Kumeta, Toshikazu Nakamura, and Koichiro Ishimori. Conformational ensemble of a multidomain protein explored by Gd3+ electron paramagnetic resonance. *Biophysical Journal*, jul 2021.

[351] Monica Sala-Rabanal, Bruce A. Hirayama, Donald D.F. Loo, Vincent Chaptal, Jeff Abramson, and Ernest M. Wright. Bridging the gap between structure and kinetics of human SGLT1. *American Journal of Physiology - Cell Physiology*, 302(9):1293–1305, 2012.

[352] Ken Sale, Jean-Loup Faulon, Genetha A. Gray, Joseph S. Schoeniger, and Malin M. Young. Optimal bundling of transmembrane helices using sparse distance constraints. *Protein Science*, 13(10):2613–2627, 2009.

[353] Ken Sale, Likai Song, Yi Shiuan Liu, Eduardo Perozo, and Piotr Fajer. Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER. *Journal of the American Chemical Society*, 127(26):9334–9335, 2005.

[354] Andrej Sali. From integrative structural biology to cell biology, jan 2021.

[355] Jessica L. Sarver, Michael Zhang, Lishan Liu, David Nyenhuis, and David S. Cafiso. A Dynamic Protein-Protein Coupling between the TonB-Dependent Transporter FhuA and TonB. *Biochemistry*, 57(6):1045–1053, feb 2018.

[356] Marion F. Sauer, Alexander M. Sevy, James E. Crowe, and Jens Meiler. Multi-state design of flexible proteins predicts sequences optimal for conformational change. *PLoS Computational Biology*, 16(2):1–29, 2020.

[357] M J Schmidt, A Fedoseev, D Summerer, and M Drescher. Genetically Encoded Spin Labels for in Vitro and In-Cell EPR Studies of Native Proteins. In *Methods in Enzymology*, volume 563, pages 483–502. Academic Press Inc., 2015.

[358] Moritz J Schmidt, Artem Fedoseev, Dennis Bücker, Julia Borbas, Christine Peter, Malte Drescher, and Daniel Summerer. EPR Distance Measurements in Native Proteins with Genetically Encoded Spin Labels. *ACS Chemical Biology*, 10(12):2764–2771, dec 2015.

[359] Thomas Schmidt, Jaekyun Jeon, Yusuke Okuno, Sai C Chiliveri, and G Marius Clore. Submillisecond Freezing Permits Cryoprotectant-Free EPR Double Electron–Electron Resonance Spectroscopy. *ChemPhysChem*, 21(12):1224–1229, jun 2020.

[360] Thomas Schmidt, Marielle A. Wälti, James L. Baber, Eric J. Hustedt, and G. Marius Clore. Long Distance Measurements up to 160 Å in the GroEL Tetradecamer Using Q-Band DEER EPR Spectroscopy. *Angewandte Chemie - International Edition*, 55(51):15905–15909, 2016.

[361] Benjamin Schuler. Single-molecule FRET of protein structure and dynamics - a primer. *Journal of nanobiotechnology*, 11 Suppl 1, 2013.

[362] Sabrina Schulze, Stefan Köster, Ulrike Geldmacher, Anke C. Terwisscha Van Scheltinga, and Werner Kühlbrandt. Structural basis of Na+-independent and cooperative substrate/product antiport in CaiT. *Nature*, 467(7312):233–236, 2010.

[363] Charles D Schwieters, John J Kuszewski, Nico Tjandra, and G Marius Clore. The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance*, 160(1):65–73, 2003.

[364] Emanuela Screpanti and Carola Hunte. Discontinuous membrane helices in transport proteins and their correlation with function. *Journal of Structural Biology*, 159(2 SPEC. ISS.):261–267, 2007.

[365] K. Ilker Sen, Timothy M. Logan, and Piotr G. Fajer. Protein dynamics and monomer-monomer interactions in AntR activation by electron paramagnetic resonance and double electron-electron resonance. *Biochemistry*, 46(41):11639–11649, 2007.

[366] Pedro Sfriso, Miquel Duran-Frigola, Roberto Mosca, Agustí Emperador, Patrick Aloy, and Modesto Orozco. Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins. *Structure*, 24(1):116–126, 2016.

[367] Paul L. Shaffer, April Goehring, Aruna Shankaranarayanan, and Eric Gouaux. Structure and mechanism of a Na+-independent amino acid transporter. *Science*, 325(5943):1010–1014, 2009.

[368] Paul L. Shaffer, April Goehring, Aruna Shankaranarayanan, and Eric Gouaux. Structure and mechanism of a Na+-independent amino acid transporter. *Science*, 325(5943):1010–1014, 2009.

[369] Azadeh Shahsavar, Peter Stohler, Gleb Bourenkov, Iwan Zimmermann, Martin Siegrist, Wolfgang Guba, Emmanuel Pinard, Steffen Sinning, Markus A Seeger, Thomas R Schneider, Roger J P Dawson, and Poul Nissen. Structural insights into the inhibition of glycine reuptake. *Nature*, 591(7851):677–681, mar 2021.

[370] Peter S. Shenkin, David L. Yarmush, Richard M. Fine, Huajun Wang, and Cyrus Levinthal. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers*, 26(12):2053–2085, 1987.

[371] Lei Shi and Weinstein Harel. Conformational rearrangements to the intracellular open states of the LeuT and ApcT transporters are modulated by common mechanisms. *Biophysical Journal*, 99(12):L103–L105, 2010.

[372] Lei Shi, Matthias Quick, Yongfang Zhao, Harel Weinstein, and Jonathan A Javitch. The Mechanism of a Neurotransmitter:Sodium Symporter-Inward Release of Na+ and Substrate Is Triggered by Substrate in a Second Binding Site. *Molecular Cell*, 30(6):667–677, 2008.

[373] Yigong Shi. Common folds and transport mechanisms of secondary active transporters. *Annual Review of Biophysics*, 42(1):51–72, 2013.

[374] Tatsuro Shimamura, Simone Weyand, Oliver Beckstein, Nicholas G. Rutherford, Jonathan M. Hadden, David Sharpies, Mark S.P. Sansom, So Iwata, Peter J.F. Henderson, and Alexander D. Cameron. Molecular basis of alternating access membrane transport by the sodium-hydantoin transporter Mhp1. *Science*, 328(5977):470–473, 2010.

[375] D. Shortle, K. T. Simons, and D. Baker. Clustering of low-energy conformations near the native structures of small proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 95(19):11158–11162, 1998.

[376] Fabian Sievers, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(539), 2011.

[377] Katie J Simmons, Scott M Jackson, Florian Brueckner, Simon G Patching, Oliver Beckstein, Ekaterina Ivanova, Tian Geng, Simone Weyand, David Drew, Joseph Lanigan, David J Sharples, Mark SP Sansom, So Iwata, Colin WG Fishwick, A Peter Johnson, Alexander D Cameron, and Peter JF Henderson. Molecular mechanism of ligand recognition by membrane transport protein, Mhp1. *The EMBO Journal*, 33(16):1831–1844, 2014.

[378] Kim T. Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Technical Report 1, 1997.

[379] Kevin Singewald, Matthew J Lawless, and Sunil Saxena. Increasing nitroxide lifetime in cells to enable in-cell protein structure and dynamics measurements by electron spin resonance spectroscopy. *Journal of Magnetic Resonance*, 299:21–27, feb 2019.

[380] Natesh Singh and Gerhard F. Ecker. Insights into the structure, function, and ligand discovery of the large neutral amino acid transporter 1, lat1. *International Journal of Molecular Sciences*, 19(5):1–32, 2018.

[381] Satinder K Singh, Chayne L Piscitelli, Atsuko Yamashita, and Eric Gouaux. A competitive inhibitor traps LeuT in an open-to-out conformation. *Science*, 322(5908):1655–1661, 2008.

[382] Gregory G Slabaugh. Computing Euler angles from a rotation matrix, 1999.

[383] Azmat Sohail, Kumaresan Jayaraman, Santhoshkannan Venkatesan, Kamil Gotfryd, Markus Daerr, Ulrik Gether, Claus J Loland, Klaus T Wanner, Michael Freissmuth, Harald H Sitte, Walter Sandtner, and Thomas Stockner. The Environment Shapes the Inner Vestibule of LeuT. *PLoS Computational Biology*, 12(11), nov 2016.

[384] Yifan Song, Frank Dimaio, Ray Yu Ruei Wang, David Kim, Chris Miles, Tj Brunette, James Thompson, and David Baker. High-resolution comparative modeling with RosettaCM. *Structure*, 21(10):1735–1742, 2013.

[385] Sebastian Spicher, Dinar Abdullin, Stefan Grimme, and Olav Schiemann. Modeling of spin-spin distance distributions for nitroxide labeled biomacromolecules. *Physical Chemistry Chemical Physics*, 22(42):24282–24290, 2020.

[386] Philipp E Spindler, Izabela Waclawska, Burkhard Endeward, Jörn Plackmeyer, Christine Ziegler, and Thomas F Prisner. Carr-Purcell Pulsed Electron Double Resonance with Shaped Inversion Pulses. *Journal of Physical Chemistry Letters*, 6(21):4331–4335, oct 2015.

[387] Ashutosh Srivastava, Sandhya Premnath Tiwari, Osamu Miyashita, and Florence Tama. Integrative/Hybrid Modeling Approaches for Studying Biomolecules, 2020.

[388] Madhur Srivastava, C. Lindsay Anderson, and Jack H. Freed. A New Wavelet Denoising Method for Selecting Decomposition Levels and Noise Thresholds. *IEEE Access*, 4:3862–3877, 2016.

[389] Madhur Srivastava and Jack H. Freed. Singular Value Decomposition Method to Determine Distance Distributions in Pulsed Dipolar Electron Spin Resonance: II. Estimating Uncertainty. *Journal of Physical Chemistry A*, 123(1):359–370, 2019.

[390] Amelie Stein and Tanja Kortemme. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS ONE*, 8(5), may 2013.

[391] Richard A Stein, Albert H Beth, and Eric J Hustedt. *A straightforward approach to the analysis of double electron-electron resonance data*, volume 563. Elsevier Inc., 1 edition, 2015.

[392] Lukas S. Stelzl, Philip W. Fowler, Mark S.P. Sansom, and Oliver Beckstein. Flexible gates generate occluded intermediates in the transport cycle of LacY. *Journal of Molecular Biology*, 426(3):735–751, 2014.

[393] Alasdair C. Steven and Wolfgang Baumeister. The future is hybrid. *Journal of Structural Biology*, 163(3):186–195, 2008.

[394] Sebastian Stolzenberg, Zheng Li, Matthias Quick, Lina Malinauskaite, Poul Nissen, Harel Weinstein, Jonathan A. Javitch, and Lei Shi. The role of transmembrane segment 5 (TM5) in Na2 release and the conformational transition of neurotransmitter:sodium symporters toward the inward-open state. *Journal of Biological Chemistry*, 292(18):7372–7384, 2017.

[395] Nariaki Sugiura. Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26, 1978.

[396] Yuh Ju Sun, John Rose, Bi Cheng Wang, and Chwan Deng Hsiao. The structure of glutamine-binding protein complexed with glutamine at 1.94 Åresolution: Comparisons with other amino acid binding proteins. *Journal of Molecular Biology*, 278(1):219–229, 1998.

[397] Zhoutong Sun, Qian Liu, Ge Qu, Yan Feng, and Manfred T Reetz. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chemical Reviews*, 2019.

[398] Ramasubramanian Sundaramoorthy, Amanda L. Hughes, Vijender Singh, Nicola Wiechens, Daniel P. Ryan, Hassane El-Mkami, Maxim Petoukhov, Dmitri I. Svergun, Barbara Treutlein, Salina Quack, Monika Fischer, Jens Michaelis, Bettina Böttcher, David G. Norman, and Tom Owen-Hughes. Structural reorganization of the chromatin remodeling enzyme Chd1 upon engagement with nucleosomes. *eLife*, 6:1–28, 2017.

[399] Tai Ching Sung, Ching Yu Li, Yei Chen Lai, Chien Lun Hung, Orion Shih, Yi Qi Yeh, U. Ser Jeng, and Yun Wei Chiang. Solution structure of apoptotic BAX oligomer: Oligomerization likely precedes membrane insertion. *Structure*, 23(10):1878–1888, oct 2015.

[400] Sarah R Sweger, Stephan Pribitzer, and Stefan Stoll. Bayesian Probabilistic Analysis of DEER Spectroscopy Data Using Parametric Distance Distribution Models. *Journal of Physical Chemistry A*, 124(30):6193–6202, jul 2020.

[401] Yoshiki Tanaka, Christopher J. Hipolito, Andrés D. Maturana, Koichi Ito, Teruo Kuroda, Takashi Higuchi, Takayuki Katoh, Hideaki E. Kato, Motoyuki Hattori, Kaoru Kumazaki, Tomoya Tsukazaki, Ryuichiro Ishitani, Hiroaki Suga, and Osamu Nureki. Structural basis for the drug extrusion mechanism by a MATE multidrug transporter. *Nature*, 496(7444):247–251, 2013.

[402] Lin Tang, Lin Bai, Wen Hua Wang, and Tao Jiang. Crystal structure of the carnitine transporter and insights into the antiport mechanism. *Nature Structural and Molecular Biology*, 17(4):492–496, 2010.

[403] Narin S. Tangprasertchai, Xiaojun Zhang, Yuan Ding, Kenneth Tham, Remo Rohs, Ian S. Haworth, and Peter Z. Qin. An Integrated Spin-Labeling/Computational-Modeling Approach for Mapping Global Structures of Nucleic Acids. *Methods in Enzymology*, 564:427–453, 2015.

[404] Igor Tascón, Joana S Sousa, Robin A Corey, Deryck J Mills, David Griwatz, Nadine Aumüller, Vedrana Mikusevic, Phillip J Stansfeld, Janet Vonck, and Inga Hänelt. Structural basis of proton-coupled potassium transport in the KUP family. *Nature Communications*, 11(1), dec 2020.

[405] Sotiria Tavoulari, Ahsan N Rizwan, Lucy R Forrest, and Gary Rudnick. Reconstructing a chloride-binding site in a bacterial neurotransmitter transporter homologue. *Journal of Biological Chemistry*, 286(4):2834–2842, 2011.

[406] Pedro L Teixeira, Jeff L Mendenhall, Sten Heinze, Brian Weiner, Marcin J Skwark, and Jens Meiler. Membrane protein contact and structure prediction using co-evolution in conjunction with machine learning. *PLoS ONE*, 12(5), may 2017.

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

[408] Giulio Tesei, João M. Martins, Micha B.A. Kunze, Yong Wang, Ramon Crehuet, and Kresten Lindorff-Larsen. DEER-PREdict: Software for efficient calculation of spin-labeling EPR and NMR data from conformational ensembles. *PLoS Computational Biology*, 17(1):2020.08.09.243030, 2021.

[409] Maxx H. Tessmer, David M. Anderson, Adam Buchaklian, Dara W. Frank, and Jimmy B. Feix. Cooperative substrate-cofactor interactions and membrane localization of the bacterial phospholipase A2 (PLA2) enzyme, ExoU. *Journal of Biological Chemistry*, 292(8):3411–3419, 2017.

[410] Maxx H. Tessmer, David M. Anderson, Adam M. Pickrum, Molly O. Riegert, Rocco Moretti, Jens Meiler, Jimmy B. Feix, and Dara W. Frank. Identification of a ubiquitin-binding interface using Rosetta and DEER. *Proceedings of the National Academy of Sciences of the United States of America*, 115(3):525–530, 2018.

[411] Thomas B Thompson, Michael G Thomas, Jorge C Escalante-Semerena, and Ivan Rayment. Three-dimensional structure of adenosylcobinamide kinase/adenosylcobinamide phosphate guanylyltransferase from Salmonella typhimurium determined to 2.3 Åresolution. Technical Report 21, 1998.

[412] Thomas B Thompson, Michael G Thomas, Jorge C Escalante-Semerena, and Ivan Rayment. Three-dimensional structure of adenosylcobinamide kinase/adenosylcobinamide phosphate guanylyltransferase (CobU) complexed with GMP: Evidence for a substrate-induced transferase active site. *Biochemistry*, 38(40):12995–13005, oct 1999.

[413] Andrey Nikolaevich Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Doklady Akademii Nauk SSSR*, 151:501–504, 1963.

[414] Ching Ju Tsai and Christine Ziegler. Coupling electron cryomicroscopy and X-ray crystallography to understand secondary active transport. *Current Opinion in Structural Biology*, 20(4):448–455, 2010.

[415] Ming Feng Tsai, Yiling Fang, and Christopher Miller. Sided functions of an arginine-agmatine antiporter oriented in liposomes. *Biochemistry*, 51(8):1577–1585, 2012.

[416] Ming Feng Tsai, Patrick McCarthy, and Christopher Miller. Substrate selectivity in glutamate-dependent acid resistance in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15):5898–5902, 2013.

[417] Ming Feng Tsai and Christopher Miller. Substrate selectivity in arginine-dependent acid resistance in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15):5893–5897, apr 2013.

[418] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A A Kohl, Anna Potapenko, Andrew J Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman,

Stig Petersen, Andrew W Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, jul 2021.

[419] Michael D. Tyka, Kenneth Jung, and David Baker. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. *Journal of Computational Chemistry*, 33(31):2483–2491, 2012.

[420] Pedro M. Valero-Mora. *ggplot2: Elegant Graphics for Data Analysis* , volume 35. 2010.

[421] Ned Van Eps, Christian Altenbach, Lydia N Caro, Naomi R Latorraca, Scott A Hollingsworth, Ron O Dror, Oliver P Ernst, and Wayne L Hubbell. Gi- and Gs-coupled GPCRs show different modes of G-protein binding. *Proceedings of the National Academy of Sciences of the United States of America*, 115(10):2383–2388, mar 2018.

[422] Ned Van Eps, Anita M. Preininger, Nathan Alexander, Ali I. Kaya, Scott Meier, Jens Meiler, Heidi E. Hamm, and Wayne L. Hubbell. Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9420–9424, 2011.

[423] Ake Vastermark, Simon Wollwage, Michael E. Houle, Rita Rio, and Milton H. Saier. Expansion of the APC superfamily of secondary carriers. *Proteins: Structure, Function and Bioinformatics*, 82(10):2797–2811, 2014.

[424] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, sep 2017.

[425] Brandy Verhalen, Reza Dastvan, Sundarapandian Thangapandian, Yelena Peskova, Hanane A. Koteiche, Robert K. Nakamoto, Emad Tajkhorshid, and Hassane S. McHaourab. Energy transduction and alternating access of the mammalian ABC transporter P-glycoprotein. *Nature*, 543(7647):738–741, 2017.

[426] Håkan Viklund and Arne Elofsson. OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, 24(15):1662–1668, 2008.

[427] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A Nicholson, David R Hagen, Dmitrii V Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G Young, Gavin A Price, Gert Ludwig Ingold,

Gregory E Allen, Gregory R Lee, Hervé Audren, Irvin Probst, Jörg P Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A Brodtkorb, Perry Lee, Robert T McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J Pingel, Thomas P Robitaille, Thomas Spura, Thouis R Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[428] Weixiao Y. Wahlgren, Elin Dunevall, Rachel A. North, Aviv Paz, Mariafrancesca Scalise, Paola Bisignano, Johan Bengtsson-Palme, Parveen Goyal, Elin Claesson, Rhawnie Caing-Carlsson, Rebecka Andersson, Konstantinos Beis, Ulf J. Nilsson, Anne Farewell, Lorena Pochini, Cesare Indiveri, Michael Grabe, Renwick C.J. Dobson, Jeff Abramson, S. Ramaswamy, and Rosmarie Friemann. Substrate-bound outward-open structure of a Na+-coupled sialic acid symporter reveals a new Na+ site. *Nature Communications*, 9(1):1–14, 2018.

[429] Guoli Wang and Roland L. Dunbrack. PISCES: A protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[430] Guoli Wang and Roland L. Dunbrack. PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33(SUPPL. 2):94–98, 2005.

[431] Hui Wang, April Goehring, Kevin H Wang, Aravind Penmatsa, Ryan Ressler, and Eric Gouaux. Structural basis for action by diverse antidepressants on biogenic amine transporters. *Nature*, 503(7474):141–145, 2013.

[432] Kevin H Wang, Aravind Penmatsa, and Eric Gouaux. Neurotransmitter and psychostimulant recognition by the dopamine transporter. *Nature*, 521(7552):322–327, may 2015.

[433] Ray Yu Ruei Wang, Yifan Song, Benjamin A. Barad, Yifan Cheng, James S. Fraser, and Frank DiMaio. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife*, 5(September2016):1–22, 2016.

[434] Sheng Wang, Renhong Yan, Xi Zhang, Qi Chu, and Yigong Shi. Molecular mechanism of pH-dependent substrate transport by an arginine-agmatine antiporter. *Proceedings of the National Academy of Sciences of the United States of America*, 111(35):12734–12739, 2014.

[435] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13):1650–1655, 2003.

[436] R Ward, M Zoltner, L Beer, H El Mkami, I R Henderson, T Palmer, and D G Norman. The Orientation of a Tandem POTRA Domain Pair, of the Beta-Barrel Assembly Protein BamA, Determined by PELDOR Spectroscopy. *Structure*, 17(9):1187–1194, 2009.

[437] Dora Toledo Warshaviak, Valery V Khramtsov, Duilio Cascio, Christian Altenbach, and Wayne L Hubbell. Structure and dynamics of an imidazoline nitroxide side chain with strongly hindered internal motion in proteins. *Journal of Magnetic Resonance*, 232:53–61, 2013.

[438] Akira Watanabe, Seungho Choe, Vincent Chaptal, John M Rosenberg, Ernest M Wright, Michael Grabe, and Jeff Abramson. The mechanism of sodium and substrate release from the binding pocket of vSGLT. *Nature*, 468(7326):988–991, 2010.

[439] L H Weaver and B W Matthews. Structure of bacteriophage T4 lysozyme refined at 1.7 Åresolution. *Journal of Molecular Biology*, 193(1):189–199, 1987.

[440] Christoph Wegener, Sandra Tebbe, Heinz Jürgen Steinhoff, and Heinrich Jung. Spin labeling analysis of structure and dynamics of the Na+/proline transporter of Escherichia coli. *Biochemistry*, 39(16):4831–4837, 2000.

[441] Brian E. Weiner, Nathan S. Alexander, Louesa R. Akin, Nils Woetzel, Mert Karakas, and Jens Meiler. BCL::Fold – Protein topology determination from limited NMRrestraints. *Proteins*, 82(4):587–595, 2014.

[442] Jonathan Yaacov Weinstein, Assaf Elazar, and Sarel Jacob Fleishman. A lipophilicity-based energy function for membrane-protein modelling and design. *PLoS Computational Biology*, 15(8), 2019.

[443] Simone Weyand, Tatsuro Shimamura, Oliver Beckstein, Mark S.P. Sansom, So Iwata, Peter J.F. Henderson, and Alexander D. Cameron. The alternating access mechanism of transport as observed in the sodium-hydantoin transporter Mhp1. *Journal of Synchrotron Radiation*, 18(1):20–23, 2011.

[444] Simone Weyand, Tatsuro Shimamura, Shunsuke Yajima, Shu N.Ichi Suzuki, Osman Mirza, Kuakarun Krusong, Elisabeth P. Carpenter, Nicholas G. Rutherford, Jonathan M. Hadden, John O'Reilly, Pikyee Ma, Massoud Saidijam, Simon G. Patching, Ryan J. Hope, Halina T. Norbertczak, Peter C.J. Roach, So Iwata, Peter J.F. Henderson, and Alexander D. Cameron. Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science*, 322(5902):709–713, 2008.

[445] Nicholas R J Whitelegg and Anthony R Rees. WAM: An improved algorithm for modelling antibodies on the WEB. Technical Report 12, 2000.

[446] Laura M Wingler, Matthias Elgeti, Daniel Hilger, Naomi R Latorraca, Michael T Lerch, Dean P Staus, Ron O Dror, Brian K Kobilka, Wayne L Hubbell, and Robert J Lefkowitz. Angiotensin Analogs with Divergent Bias Stabilize Distinct Receptor Conformations. *Cell*, 176(3):468—-478.e11, 2019.

[447] Nils Woetzel, Mert Karakaş, Rene Staritzbichler, Ralf Müller, Brian E. Weiner, and Jens Meiler. BCL::Score-Knowledge Based Energy Potentials for Ranking Protein Models Represented by Idealized Secondary Structure Elements. *PLoS ONE*, 7(11), nov 2012.

[448] Steven G. Worswick, James A. Spencer, Gunnar Jeschke, and Ilya Kuprov. Deep neural network processing of DEER data. *Science Advances*, 4(8):eaat5218, 2018.

[449] Di Wu, Tamara N Grund, Sonja Welsch, Deryck J Mills, Max Michel, Schara Safarian, and Hartmut Michel. Structural basis for amino acid exchange by a human heteromeric amino acid transporter. *Proceedings of the National Academy of Sciences of the United States of America*, 117(35):21281–21287, sep 2020.

[450] Kaiqi Wu, Chaowei Shi, Juan Li, Haipeng Wang, Pan Shi, Liu Chen, Fangming Wu, Ying Xiong, and Changlin Tian. Efficient long-distance NMR-PRE and EPR-DEER restraints for two-domain protein structure determination. *Protein and Cell*, 4(12):893–896, dec 2013.

[451] Yan Xia, Axel W. Fischer, Pedro Teixeira, Brian Weiner, and Jens Meiler. Integrated Structural Biology for α-Helical Membrane Protein Structure Determination. *Structure*, 26(4):657–666.e2, 2018.

[452] Yuan Xie, Shenghai Chang, Cheng Zhao, Feng Wang, Si Liu, Jin Wang, Eric Delpire, Sheng Ye, and Jiangtao Guo. Structures and an activation mechanism of human potassium-chloride cotransporters. *Science Advances*, 6(50), 2020.

[453] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[454] Hiromasa Yagi, Debamalya Banerjee, Bim Graham, Thomas Huber, Daniella Goldfarb, and Gottfried Otting. Gadolinium tagging for high-precision measurements of 6 nm distances in protein assemblies by EPR. *Journal of the American Chemical Society*, 133(27):10418–10421, jul 2011.

[455] Atsuko Yamashita, Satinder K. Singh, Toshimitsu Kawate, Yan Jin, and Eric Gouaux. Crystal structure of a bacterial homologue of Na+/Cl –dependent neurotransmitter transporters. *Nature*, 437(7056):215–223, 2005.

[456] Renhong Yan, Yaning Li, Jennifer Müller, Yuanyuan Zhang, Simon Singer, Lu Xia, Xinyue Zhong, Jürg Gertsch, Karl Heinz Altmann, and Qiang Zhou. Mechanism of substrate transport and inhibition of the human LAT1-4F2hc amino acid transporter. *Cell Discovery*, 7(1), dec 2021.

[457] Renhong Yan, Yaning Li, Yi Shi, Jiayao Zhou, Jianlin Lei, Jing Huang, and Qiang Zhou. Cryo-EM structure of the human heteromeric amino acid transporter b0,+AT-rBAT. Technical Report 16, 2020.

[458] Renhong Yan, Yuanyuan Zhang, Yaning Li, Lu Xia, Yingying Guo, and Qiang Zhou. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Technical Report 6485, 2020.

[459] Renhong Yan, Xin Zhao, Jianlin Lei, and Qiang Zhou. Structure of the human LAT1–4F2hc heteromeric amino acid transporter complex. *Nature*, 568(7750):127–130, apr 2019.

[460] Renhong Yan, Jiayao Zhou, Yaning Li, Jianlin Lei, and Qiang Zhou. Structural insight into the substrate recognition and transport mechanism of the human LAT2–4F2hc complex. *Cell Discovery*, 6(1):20–23, 2020.

[461] Lee Wei Yang, Eran Eyal, Chakra Chennubhotla, Jun Goo Jee, Angela M. Gronenborn, and Ivet Bahar. Insights into Equilibrium Dynamics of Proteins from Comparison of NMR and X-Ray Data with Computational Predictions. *Structure*, 15(6):741–749, jun 2007.

[462] Yin Yang, Feng Yang, Xia Yan Li, Xun Cheng Su, and Daniella Goldfarb. In-Cell EPR Distance Measurements on Ubiquitin Labeled with a Rigid PyMTA-Gd(III) Tag. *Journal of Physical Chemistry B*, 123(5):1050–1059, feb 2019.

[463] Yunhuang Yang, Theresa A Ramelot, Robert M McCarrick, Shuisong Ni, Erik A Feldmann, John R Cort, Huang Wang, Colleen Ciccosanti, Mei Jiang, Haleema Janjua, Thomas B Acton, Rong Xiao, John K Everett, Gaetano T Montelione, and Michael A Kennedy. Combining NMR and EPR methods for homodimer protein structure determination. *Journal of the American Chemical Society*, 132(34):11910–11913, 2010.

[464] Zhimin Yang, Richard A Stein, Thacien Ngendahimana, Maren Pink, Suchada Rajca, Gunnar Jeschke, Sandra S Eaton, Gareth R Eaton, Hassane S McHaourab, and Andrzej Rajca. Supramolecular Approach to Electron Paramagnetic Resonance Distance Measurement of Spin-Labeled Proteins. *Journal of Physical Chemistry B*, 124(16):3291–3299, apr 2020.

[465] Zhongyu Yang, Michael R Kurpiewski, Ming Ji, Jacque E Townsend, Preeti Mehta, Linda Jen-Jacobson, and Sunil Saxena. ESR spectroscopy identifies inhibitory Cu 2+ sites in a DNA-modifying enzyme to reveal determinants of catalytic specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 109(17):3–10, 2012.

[466] Vladimir Yarov-Yarovoy, Jack Schonbrun, and David Baker. Multipass membrane protein structure prediction using Rosetta. *Proteins: Structure, Function and Genetics*, 62(4):1010–1025, 2006.

[467] Henna Ylikangas, Kalle Malmioja, Lauri Peura, Mikko Gynther, Emmanuel O. Nwachukwu, Jukka Leppänen, Krista Laine, Jarkko Rautio, Maija Lahtela-Kakkonen, Kristiina M. Huttunen, and Antti Poso. Quantitative insight into the design of compounds recognized by the L-type amino acid transporter 1 (LAT1). *ChemMedChem*, 9(12):2699–2707, 2014.

[468] Sandra Zakrzewska, Ahmad Reza Mehdipour, Viveka Nand Malviya, Tsuyoshi Nonaka, Juergen Koepke, Cornelia Muenke, Winfried Hausner, Gerhard Hummer, Schara Safarian, and Hartmut Michel. Inward-facing conformation of a multidrug resistance MATE family transporter. *Proceedings of the National Academy of Sciences of the United States of America*, 116(25):12275–12284, 2019.

[469] Sensen Zhang, Jun Zhou, Yuebin Zhang, Tianya Liu, Perrine Friedel, Wei Zhuo, Suma Somasekharan, Kasturi Roy, Laixing Zhang, Yang Liu, Xianbin Meng, Haiteng Deng, Wenwen Zeng, Guohui Li, Biff Forbush, and Maojun Yang. The structural basis of function and regulation of neuronal cotransporters NKCC1 and KCC2. *Communications Biology*, 4(1):1–15, 2021.

[470] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics*, 57(4):702–710, 2004.

[471] Yang Zhang and Jeffrey Skolnick. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.

[472] Yuan Wei Zhang, Sotiria Tavoulari, Steffen Sinning, Antoniya A Aleksandrova, Lucy R Forrest, and Gary Rudnick. Structural elements required for coupling ion and substrate transport in the neurotransmitter transporter homolog LeuT. *Proceedings of the National Academy of Sciences of the United States of America*, 115(38):E8854—-E8862, sep 2018.

[473] Yuan Wei Zhang, Stacy Uchendu, Vanessa Leone, Richard T. Bradshaw, Ntumba Sangwa, Lucy R. Forrest, and Gary Rudnick. Chloride-dependent conformational changes in the GlyT1 glycine transporter. *Proceedings of the National Academy of Sciences of the United States of America*, 118(10):1–10, 2021.

[474] Chunfeng Zhao and Sergei Yu Noskov. The role of local hydration and hydrogen-bonding dynamics in ion and solute release from ion-coupled secondary transporters. *Biochemistry*, 50(11):1848–1856, 2011.

[475] Yongfang Zhao, Matthias Quick, Lei Shi, Ernest L Mehler, Harel Weinstein, and Jonathan A Javitch. Substrate-dependent proton antiport in neurotransmitter:sodium symporters. *Nature Chemical Biology*, 6(2):109–116, 2010.

[476] Yongxiang Zhao, Jiemin Shen, Qinzhe Wang, Ming Zhou, and Erhu Cao. Inhibitory and Transport Mechanisms of the Human Cation-Chloride Cotransport KCC1. *bioRxiv*, pages 1–9, 2020.

[477] Wenjun Zheng and Bernard R Brooks. Normal-modes-based prediction of protein conformational changes guided by distance constraints. *Biophysical Journal*, 88(5):3109–3117, 2005.

[478] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods*, 18(2):176–185, feb 2021.

[479] Zheng Zhou, Susan C. DeSensi, Richard A. Stein, Suzanne Brandon, Mrinalini Dixit, Erin J. McArdle, Eric M. Warren, Heather K. Kroh, Likai Song, Charles E. Cobb, Eric J. Hustedt, and Albert H. Beth. Solution structure of the cytoplasmic domain of erythrocyte membrane band 3 determined by site-directed spin labeling. *Biochemistry*, 44(46):15115–15128, 2005.

[480] Christine Ziegler, Erhard Bremer, and Reinhard Krämer. The BCCT family of carriers: From physiology to crystal structure, 2010.

[481] Elia Zomot and Ivet Bahar. Protonation of glutamate 208 induces the release of agmatine in an outward-facing conformation of an arginine/agmatine antiporter. *Journal of Biological Chemistry*, 286(22):19693–19701, 2011.

[482] Elia Zomot, Annie Bendahan, Matthias Quick, Yongfang Zhao, Jonathan A Javitch, and Baruch I Kanner. Mechanism of chloride interaction with neurotransmitter:sodium symporters. *Nature*, 449(7163):726–730, 2007.

[483] Ping Zou and Hassane S. Mchaourab. Increased sensitivity and extended range of distance measurements in Spin-labeled membrane proteins: Q-band double electron-electron resonance and nanoscale bilayers. *Biophysical Journal*, 98(6):L18–L20, 2010.

Appendix A

AlphaFold2 predicts the inward-facing conformation of the multidrug transporter LmrP

The contents of this Appendix have been previously published [95].

As part of the 14th annual Critical Assessment of Structure Prediction (CASP), the protein structure prediction algorithm AlphaFold2 generated multiple models of the proton/drug antiporter LmrP. Previous experimental data from DEER spectroscopy, a technique which reports distance distributions between spin labels attached to proteins, suggest that one of the lower-ranked models may have captured a conformation that has so far eluded experimental structure determination.

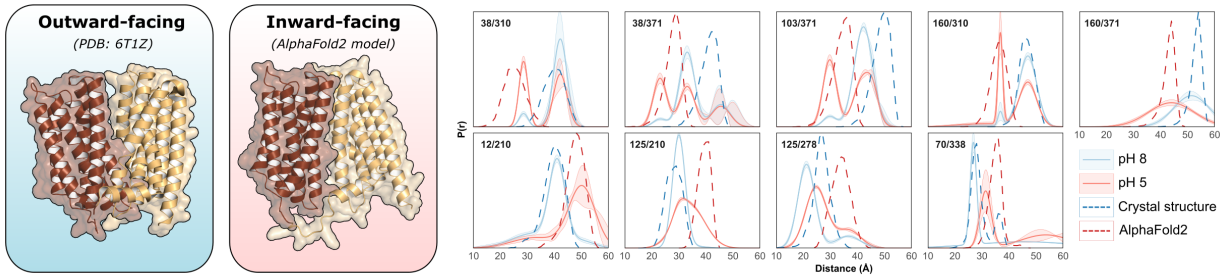

Figure A.1: An IF model of LmrP generated by AlphaFold2 is consistent with experimental data. Left: OF and IF conformations determined using X-ray crystallography and modeled by AlphaFold2, respectively. Right: Experimental DEER distance distributions on the extracellular and intracellular sides of the protein, respectively, overlap with distances predicted by AlphaFold2 model 1. Dashed lines are distance distributions predicted by either the crystal structure (blue) or the model (red). These data have been previously published.

A.1 Main Text

Active transporters such as LmrP alternate between OF and IF conformations during their transport cycles [45, 270, 267]. Whereas the crystal structure captures LmrP in the former [93], AlphaFold2 modeled LmrP in the latter [196] (Figure A.1). Because LmrP is a proton/drug antiporter, we carried out DEER distance measurements [186, 90, 274] at low and neutral pH to stabilize the IF and OF conformation, respectively (shown in red and blue in Figure A.1). To evaluate the IF model's consistency with the low pH DEER data, we modeled the predicted distances *in silico* using

MDDS [176], a program hosted on the CHARMM-GUI web server [195]. Not only do the predicted distances overlap remarkably well with our experimental data (Figure A.1, dashed and solid lines, respectively), but importantly the magnitudes of the experimental distance changes agree with those predicted between the OF crystal structure and AlphaFold2's IF model. These results suggest that the AlphaFold2 model depicts a functionally relevant intermediate of LmrP.

The significance of this breakthrough in modeling transporter conformations is reinforced by comparison of this model to those submitted by other contestants, which overwhelmingly depicted LmrP in an occluded conformation (Figure A.2). Occluded models result from methodological biases that favor compactness [297]. Therefore, the success of AlphaFold2 in modeling IF LmrP suggests that these biases may finally have been overcome. Additionally, it sets the stage for the structural characterization of transporters and their functional intermediates by integrating computational modeling with experimental spectroscopy.

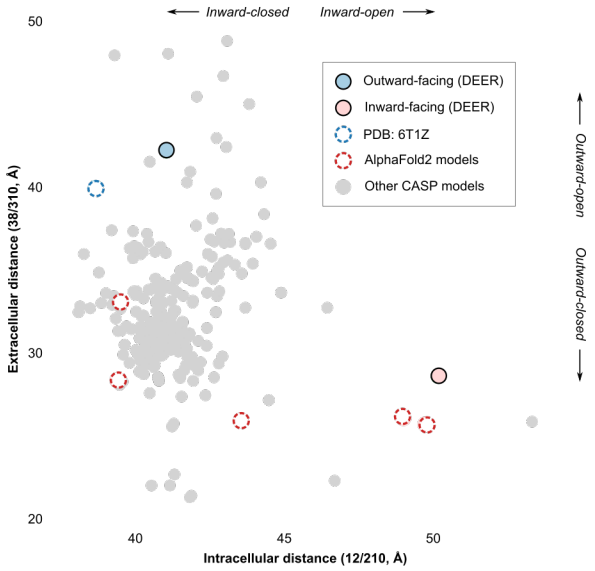

Figure A.2: Predicted DEER distances of all CASP14 LmrP models. X and Y axes reflect the average predicted DEER distances of all CASP LmrP models on the intracellular side and extracellular sides, respectively. Solid blue and red circles represent components from the experimental DEER data corresponding to outward- and inward-facing conformations, respectively. The inward-facing AlphaFold2 model shown in panel A is located on the bottom-right.

A.2 Acknowledgments

The work presented here was supported by the National Institutes of Health (GM077659).

Appendix B

Evaluation of scoring approaches for integrative modeling using DEER distance data

This Appendix is based on unpublished data.

Conformational changes define the functional cycles of many proteins. The complete characterization of functional intermediates, such as those that are infrequently sampled or are transiently populated, continues to elude established structural biology techniques. Integrating experimental distance data collected using DEER spectroscopy with computational modeling methods promises to overcome these barriers and provide a glimpse at these states. However, experimental DEER distance distributions cannot be reliably predicted or reproduced *in silico*, preventing the identification of correctly folded models to high accuracy. In part because of this fact, the scoring functions used to evaluate protein structural models using experimental DEER data in the literature vary wildly and are highly non-standard. Here we use experimental data collected in the model system PfMATE to evaluate a panel of scoring functions with the goal of determining the most effective metric for identifying correctly folded protein structures. In general, most metrics were comparable in performance when scoring unimodal distributions consistent with a single structure. However, when the data indicated a bimodal distribution representing two populations in equilibrium, only a small subset of these metrics could effectively identify native-like models. We conclude that methods comparing the overlap between the simulated and experimental distributions, rather than their average distances as is commonly performed, are more effective at identifying native-like models when structural uniformity cannot be guaranteed by the data. Nevertheless, our results indicate that optimal results can only be achieved by *a priori* determination of individual conformations in the data.

B.1 Introduction

The integration of sparse experimental data collected using cryo-EM, X-ray crystallography, or NMR with Rosetta protein structure prediction allows protein models to be generated at atomic-detail accuracy [16, 47, 387, 393, 451]. Recently, distance data collected using EPR spectroscopy in conjunction with SDSL have been a new focus for analyzing protein structure and dynamics [171, 172, 190, 274]. Even a small number of distance measurements obtained using DEER spectroscopy, which range from 15 Å to 80 Å, effectively complement short-range distance data obtained using nuclear magnetic resonance spectroscopy [245, 463]. As a consequence, the number of computational tools that integrate these data into modeling continues to grow [157, 165, 263].

The recent interest in the DEER technique has been accompanied by methodological improvements in the way the primary spectroscopic data are converted in the distance restraints for computation. A number of paradigms exist for fitting these data [389, 448] with model-free fitting continuing to be the most widely-used strategy [192, 193]. Recent studies have identified improvements in background correction methods [116] and have detailed the optimal balance between fitting the data and regularization [105, 115]. By contrast, among methods that model distributions as sums of Gaussian functions, substantial work has gone into identifying more effective fitting algorithms and model selection criteria [117, 173, 365, 391, 400]. In both cases, these advancements have coincided with the more widespread use of confidence bands in distance distributions to visualize experimental uncertainty [104, 117, 173, 400].

Less research has been completed into determining the best approach for converting these distance data into accurate protein structural models. As a result, there is no standard approach for integrating these data for protein structural modeling. For example, it is still common to use the average or peak distance as a single restraint while disregarding the width or shape of the distance distribution [88, 243, 421]. In other cases, distribution widths have been used as function bounds [8, 89, 114, 128, 245]. When spin labels are explicitly modeled as flexible side chains or pseudo-rotamers, restraints have been introduced that attempt to maximize the overlap between the entire simulated and experimental distributions [157, 213, 262, 332, 343]. A functionally similar

approach is to minimize the area between the integrals [226]. Finally, the primary DEER data has been used directly for model-building and refinement [48, 164, 263, 336]. Although direct fitting has the potential to sidestep analytical artifacts, it is unclear whether there is any quantifiable improvement in model quality as a result of using this approach. Therefore, despite the variation in these approaches, to the best of our knowledge, the optimal scoring function for protein modeling has not been determined.

Here we employ the modeling suite Rosetta to compare several different scoring approaches using the proton-coupled multidrug transporter PfMATE as a model system. Different scoring functions are found to vary substantially in their ability to identify native-like structural models using sparse DEER data. Importantly, several commonly used metrics, such as naive integration of average distance values, fail to identify native-like models when the data are multimodal, which is commonly observed in conformationally heterogeneous proteins. Using scoring functions that compute the overlap between the experimental and simulated distributions could facilitate more meaningful interpretation of protein structures from limited DEER distance data.

B.2 Results and Discussion

B.2.1 Overview of the hybrid energy function

Protein structural models are commonly evaluated using a hybrid score representing the sum of its energy and its agreement with the data. This can be written as $E_{\text{total}} = E_{\text{model}} + w_{\text{data}}E_{\text{data}}$, where E_{model} is a model's score derived from an energy or scoring function, E_{data} is the function evaluating a model's goodness-of-fit to the experimental data, and w_{data} is the weight assigned to the experimental data relative to the native energy function [5, 178]. Ideally, both E_{model} and E_{data} would increase monotonically as a function of a model's deviation from the target conformation of interest. In practice, neither term alone is sufficient to unequivocally identify the experimentally determined structure of a protein. The former crudely approximates the physical forces acting upon biomacromolecules in solution, while the latter reflects agreement with measurements that may be sparse, ambiguous, and unevenly distributed. Both functions must contribute to the calculation of

physiologically meaningful models.

The challenge when applying experimental DEER data as modeling constraints is the fact that the distance distributions simulated *in silico* tend to overstate the dynamics of the spin label while neglecting the dynamics of the protein backbone [153, 156, 176, 322]. Any single structural model can by definition only depict a subset of the backbone conformations sampled in solution. By contrast, because clash evaluation is predominantly used to remove rotamers from structural models, the conformations of the spin labels sampled in solution are likely a subset of the rotamer libraries used to simulate these distributions. As such, it is highly unlikely that any individual structural model can exactly reproduce the experimental data.

B.2.2 Overview of the benchmark

We thus sought to determine the most effective function of E_{data} that could most effectively identify native-like models despite these factors. Several reasons guided the choice to use the multidrug transporter PfMATE as a model system to quantify the effectiveness of various scoring methods (Table B.1). First, several crystal structures in different conformations have previously been published [401, 468]. Second, a comprehensive panel of experimental DEER data has been previously collected and found to be largely consistent with two of these structures that face either outward (to the periplasm) or inward (to the cytoplasm) [180]. Third, the Rosetta suite is well equipped to model intermediate states between these two conformations. We thus generated a library of "decoy" structural models of PfMATE using Rosetta by perturbing the dihedral angles of this helix, leading to approximately 3000 alternative conformations that ranged from OF to IF, fully occluded, and fully open (see Chapter 4 for details on how these were modeled). We then calculated the C_α RMSD of each model to the OF state. Finally, since this conformation is sampled at neutral pH, we curated four sets of distance distributions collected at pH 7.5 that interrogated this inter-lobe distance on both the intracellular and extracellular sides of the membrane (see Table 4.1 in Chapter 4).

In addition to evaluating the effect of each function on scoring, we also focused on the con-

tribution of experimental noise and uncertainty to scoring. The data preparation pipeline, which is discussed in detail in B.3, proceeded as follows. First, the experimental distance data at pH 7.5 were converted into raw DEER traces in the time domain with a step size of 8 ns. Background intermolecular coupling was modeled as a stretched exponential function with background slope values ranging from 10^{-6} to 10^{-1} and modulation depth values ranging uniformly from 0.05 to 0.40, which roughly corresponds to values that would be obtained with Q-band DEER without the use of an arbitrary waveform generator. Second, random Gaussian noise was added to these time traces to simulate the effect of SNR that were either high (noise comprises 0.5% of the signal on average), medium (2%), or low (10%). Third, these data were truncated at time window durations corresponding to the number of oscillations observed, which ranged from 0.2 to 3.0 with step sizes of 0.2. Previous research suggests that at least one complete oscillation is required to accurately resolve features in the distribution beyond the mean, such as the width and multimodality. Fourth, these data were converted into distance distributions with 95% confidence intervals in

Table B.1: Scoring metrics used for the benchmark. Symbols: μ: average distance; σ: standard deviation; *cdf*: cumulative density function; $p_X(r)$: probability of distance r in distribution X.

Method	**Formula**
Average distance only	$\left(\mu_{\text{sim}} - \mu_{\text{exp}}\right)^2$
Bounded range	$max\left(0.0, \lvert\mu_{\text{sim}} - \mu_{\text{exp}} - \sigma\rvert\right)$
Overlap	$\sum_{i=1} \lvert p_{\text{sim}}(r_i) - p_{\text{exp}}(r_i)\rvert$
Wasserstein	$\sum_{i=1} \lvert cdf_{\text{sim}}(r_i) - cdf_{\text{exp}}(r_i)\rvert$
Discrepancy	$max\left(\lvert p_{\text{sim}}(r) - p_{\text{exp}}(r)\right)$
Kolmogorov-Smirnov	$max\left(\lvert cdf_{\text{sim}}(r) - cdf_{\text{exp}}(r)\right)$
Chi-squared	$\sum_{i=1} \frac{(p_{\text{sim}}(r_i)) - p_{\text{exp}}(r_i)^2}{p_{\text{exp}}(r_i)}$
Reverse Chi-squared	$\sum_{i=1} \frac{(p_{\text{exp}}(r_i)) - p_{\text{sim}}(r_i)^2}{p_{\text{sim}}(r_i)}$
Cross entropy	$-\sum_{i=1} p_{\text{exp}}(r_i) \ln\left(p_{\text{sim}}(r_i)\right)$
Jensen-Shannon distance	$\sqrt{\sum_{i=1}\left(\frac{p_{\text{sim}}(r_i)}{2} \ln\left(\frac{2 * p_{\text{sim}}(r_i)}{p_{\text{sim}}(r_i) + p_{\text{exp}}(r_i)}\right)\right) + \frac{p_{\text{exp}}(r_i)}{2} \ln\left(\frac{2 * p_{\text{exp}}(r_i)}{p_{\text{sim}}(r_i) + p_{\text{exp}}(r_i)}\right)}$
Bhattacharyya	$-\ln\left(\sum_{i=1} \sqrt{p_{\text{sim}}(r_i)\, p_{\text{exp}}(r_i)}\right)$
Hellinger	$1 - \sum_{i=1} \sqrt{p_{\text{sim}}(r_i)\, p_{\text{exp}}(r_i)}$
Jaccard index	$\sum_{i=1} \frac{min\left(p_{\text{sim}}(r_i), p_{\text{exp}}(r_i)\right)}{max\left(p_{\text{sim}}(r_i), p_{\text{exp}}(r_i)\right)}$
Joint probability	$\sum_{i=1} p_{\text{sim}}(r_i)\, p_{\text{exp}}(r_i)$

an automated fashion using the analytical software DeerLab [117]. Finally, to evaluate the effect of restraint quantity on scoring, between one and ten such distributions were used as restraints to evaluate each structural model of PfMATE using each of the metrics listed in Table B.1.

B.2.3 Unimodal distance distribution benchmark

The results are presented in Figure B.1.A and show that most metrics achieve Spearman correlation coefficients of approximately 0.7 under ideal conditions (three oscillations, ten restraints, high SNR). Interestingly, experimental SNR, but not duration in the time domain, appeared to have an outsized impact on correlation quality, which may be indicative of recent advancements in background correction methods during the analysis of DEER data [116]. The least effective metrics for model quality by Spearman correlation were the average, cross-entropy, maximum discrepancy, and Chi-squared, each of which we discuss in turn. The average distance may have been overly sensitive to long-distance components that were occasionally added when time domain data were transformed to the distance domain, but which may be absent when more deliberate data analyses are performed. Indeed, simply using a distance range appeared to ameliorate this problem. The cross-entropy metric may have been sensitive to differences in the width of the experimental and simulated distributions, since it requires perfect overlap between the domains of the two distributions (e.g., the X-axes). This may be challenging for this specific model system, as the distance data are characterized by wide distributions likely resulting from backbone disorder. The discrepancy metric captures the largest difference in amplitudes at any point in the distance distribution; given the difficulty of exactly simulating DEER distance distributions *in silico*, it is unsurprising that this a poor reporter for model quality. By contrast, we are unable to rationalize why the Chi-squared metric, but not the reverse Chi-squared metric, is ineffective of distinguishing native-like from non-native-like models. None of these results were substantially affected by the inclusion of 95% confidence intervals during calculation (data not shown).

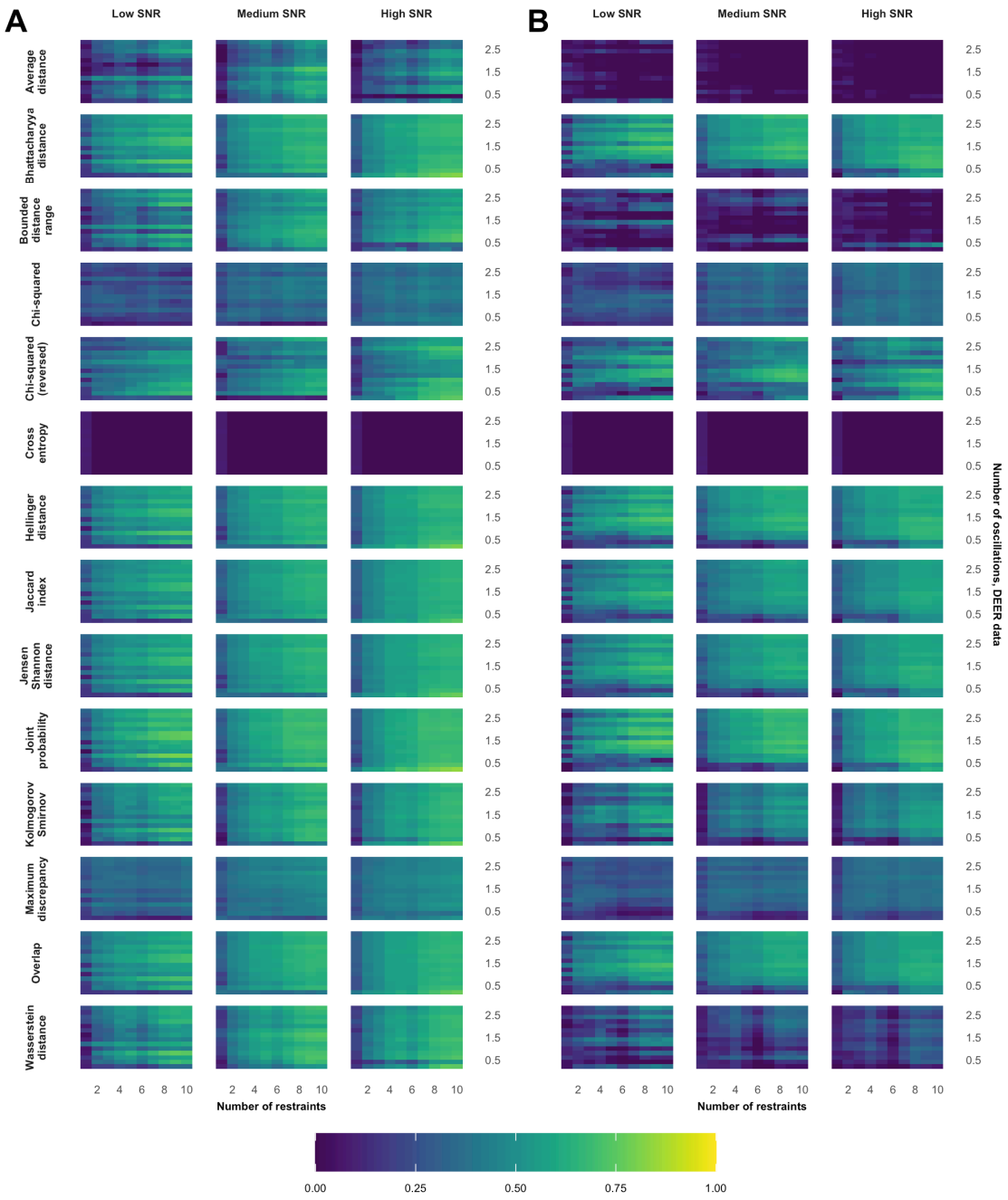

Figure B.1: Spearman correlation coefficients between model RMSD and score as a function of number of restraints, number of oscillations in the data, and scoring function. Data consist of distance distributions that are either A) unimodal or B) bimodal. Only a small number of metrics, namely those that evaluate the overlap between experimental and simulated DEER distance distributions, generate scores that correlate with RMSD when using bimodal distributions.

B.2.4 Multimodal distance distribution benchmark

The results presented above are qualified by the fact that they were obtained using experimental data that is ideal for modeling purposes, insofar as they represent a single population or component consistent with the target crystal structure from which RMSD values were measured. In practice, experimental data often features multiple populations with distinct distance components, some of which may be analytical artifacts or "ghost" peaks. While such data can be cleaned prior to modeling using strategies such as non-negative matrix factorization and/or global analysis [173, 446], it is not always possible to exactly distinguish which distance components belong to the conformations of interest. Under circumstances where multiple components are present in the data, we would expect metrics that consider agreement with entire distributions, such as the average distance, or the Wasserstein (also called the earth mover's) distance quantifying the area between the integrals of two distributions, to be overly sensitive, and thus respond poorly, to multimodality in the data.

To test this hypothesis, we repeated the simulation pipeline outlined above, but added a distance component consistent with the IF conformation of PfMATE prior to simulation of the data in the time domain. The remaining steps of the analytical pipeline were unchanged. Each distribution in this second set of data thus consisted of an even mixture of distance components belonging to outward- and inward-facing PfMATE. The results, plotted in Figures B.1.B and B.2, reveal that the majority of metrics considered in this study are incapable of correctly identifying models using these data, with many of their respective correlation coefficients plummeting even under ideal conditions (three oscillations, high SNR, ten restraints). Because performance was irrespective of SNR or duration in the time domain, we believe these reflect fundamental shortcomings in the ways scores are computed, rather than how the data were analyzed. We note that results obtained using these distributions, unlike the unimodal distributions discussed above, appeared to be more sensitive to time domain truncation, which may reflect the increased difficulty of resolving the shape of these distribution when there is insufficient data in the time domain.

Correlation coefficients calculated using several methods, such as the Jaccard index and the Joint probability, appear to be minimally affected by the presence of new components in the distance

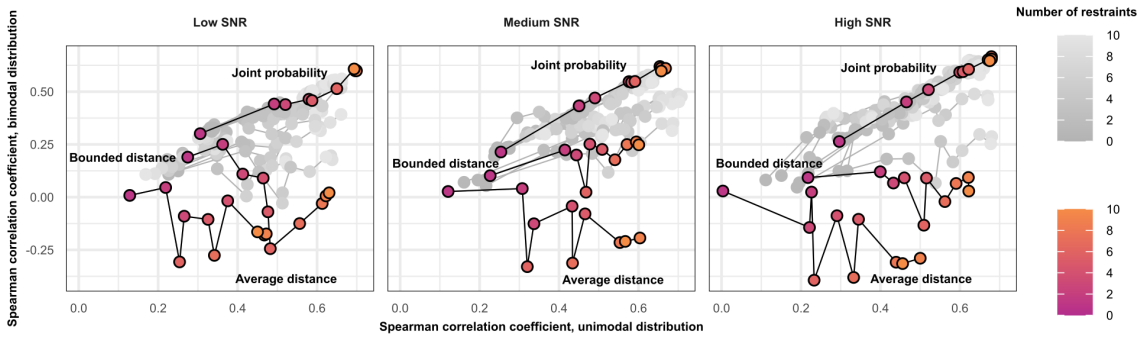

Figure B.2: Comparison of scoring methods when evaluating unimodal and bimodal distributions. All data used for scoring results shown here consisted of three oscillations in the time domain.

data. Interestingly, these metrics all favorably score partial overlap between distance components, while penalizing distributions that fall between the two components. This may allow specific conformations consistent with the data to be pinpointed, including the outward-open conformation from which RMSD values were calculated. It should be noted that these methods were more sensitive to time domain truncation than the unimodal distributions described above, which may demonstrate the extent to which low-quality data interferes with high-precision modeling of protein structures. Nonetheless, all of them return values that correlate slightly worse than when scoring models using unimodal distributions (Figure B.1.A), again demonstrating the value of preprocessing the data prior to modeling.

B.2.5 Concluding remarks

This study evaluated several scoring functions using experimental DEER data in the conformationally heterogeneous model system PfMATE. We elected to use this model system, rather than a more static system such as T4 Lysozyme, due to its nonnegligible backbone dynamics that are also observed in other proteins of biophysical interest [79, 90, 114, 267, 425]. Indeed, while these experimental data may be broadly classified as consisting of components belonging to either outward-facing or inward-facing conformations, they also indicate substantial heterogeneity and disorder. Thus, reducing the distribution to a single value, such as an average distance, appears in this case to be detrimental to high-precision modeling. In general, the majority of scoring functions studied

here return values that correlate well with model quality when the data consist of only one component. By contrast, while none of these scoring functions perfectly maintain this performance when the data consist of multiple components, a minority of these scoring functions can nonetheless return score values for models that correlate with RMSD. Nevertheless, in all cases the Spearman correlation coefficients decreased, indicating the importance of preprocessing the data and, if possible, isolating individual components in the data prior to modeling.

B.3 Materials and Methods

B.3.1 Preparation of PfMATE decoy models

A library of 2,855 structural models of PfMATE was generated using the software suite Rosetta 3.10 [229, 234], with both its OF (PDB: 6GWH) and IF (PDB: 6FHZ) conformations serving as template models. These models were generated using fragment insertion, in which the backbone dihedral angles of transmembrane helices 1 (residues 1-50) and 7 (residues 240-268) were perturbed using sequence fragments obtained from the Protein Databank. Sequence fragments were obtained from the Robetta web server as previously described [216]. A total of 5000 rounds of fragment insertion was executed using the scoring function *score3*. Each model's structural similarity to the outward-facing state (PDB: 6GWH) was then calculated using the *score_jd2* application with residues 1-50 omitted. Finally, we binned these models by this RMSD value and balanced the dataset to avoid the overrepresentation of models that were either fully occluded or highly inaccurate.

B.3.2 Simulation and analysis of DEER data

All experimental DEER distributions used in this study have been previously published [180]. Unimodal distance distributions were simulated using the data at pH 7.5 as follows. First, each distance distribution was converted into a DEER trace in the time domain with 8 ns time steps as previously described. Each trace was normalized such that the signal intensity at 0 µs was equal to 1, and their duration was set to three oscillations, which we calculated from the average distance r_{avg} in angstroms such that one oscillation had a duration of $\frac{r_{\text{avg}}^3}{5.2*10^4}$ microseconds. Second, background coupling was simulated using the function $B(t) = \exp(-kt)$, where k represents the contribution of

intermolecular background coupling to the signal and t is the time in microseconds. Then, normally distributed noise was added to simulate the contribution of experimental noise in the data, with the standard deviation equal to either 0.005, 0.02, or 0.1 for the high, medium, and low SNR datasets, respectively. Finally, these DEER traces were truncated to specific durations ranging from 0.2 to 3.0 oscillations and were analyzed using the *fitmodel* function implemented in DeerLab v0.13.1 with default settings [117].

Simulation of bimodal distributions was identical, except the experimental DEER data collected at pH 4.0 were added to the experimental distributions collected at pH 7.5 prior to transformation to the time domain.

B.3.3 Scoring of PfMATE models using simulated DEER distributions

We used the *score_jd2* application to score each PfMATE model using each of the metrics listed in Table B.1. All scoring methods were implemented in RosettaDEER. Scores were correlated with DEER values using the *spearmanr* function as implemented in SciPy [427].

B.4 Acknowledgements

We would like to thank Luis Fábregas Ibáñez for helpful assistance with DeerLab and Karan Bhardwaj for contributing exploratory data analysis to this study.

Appendix C

ConfChangeMover: Integrative modeling of conformational changes in LeuT-fold transporters using sparse spectroscopy data

This Appendix is based on unpublished data.

C.1 Introduction

Active transporters, such as those with the LeuT-fold, undergo conformational changes to import and export substrates into and out of the cell [211]. Canonical models of symport and antiport mandate that these proteins adopt several different conformers. However, their structures are often experimentally resolved in one specific state, leaving scientists guessing at the the molecular drivers of substrate translocation. Sparse experimental data, collected using techniques such as EPR [186, 213] and/or HDX/MS [281], can report on the structural basis of alternating access when higher-resolution methods such as cryo-EM and X-ray crystallography fail. In conjunction with computational modeling, these data can provide a glimpse of the molecular details of these unknown conformers. Nonetheless, the sparseness and ambiguity of these experimental restraints make it challenging to model structures that are consistent with the data provided while retaining atomic-detail information provided by the starting X-ray or cryo-EM structure.

We reasoned that the general problem surrounding the accurate modeling of conformational changes is similar in principle to that of structure refinement, in which the atomic details of structural models are deduced from low-quality models [160, 306, 334]. Recent advancements in MC refinement methods have relied on conservative sampling described as ”broken-chain kinematics” [307], in which rigid-body segments of structural models are manipulated in isolation and rejoined using loop closure (see section D.1 for a more detailed discussion on protein loop closure methods). This strategy has been instrumental to recent advances in homology modeling and is the foundation of widely-used methods including RosettaCM [384] and Modeller [113]. However, MC methods

are rarely employed for conformational change modeling problems, which have instead generally been tackled using either gradient minimization [114] or MD simulations [263, 366]. Both methods are computationally expensive and can potentially understate the extent to which protein backbones reconfigure.

Here we describe a general-purpose MC sampling method, ConfChangeMover, which we implemented in the macromolecular modeling program Rosetta [229, 234] and designed to model conformational changes using sparse experimental data (Figure C.1). The method combines recent conceptual advances borrowed from cutting-edge homology modeling methods with novel sampling approaches designed to conservatively manipulate a starting structure. Whereas homology modeling is concerned with identifying the dihedral angles of a protein that satisfy spatial restraints provided by one or more template models, our approach instead samples spatial rearrangements consistent with dihedral angles observed in the starting structure. We first demonstrate ConfChangeMover on a panel of soluble proteins using simulated C_α-C_α distance restraints. Then, using experimental EPR data that has been previously published, we apply the method to three transporters with the LeuT-fold fold (described in detail in Chapter 1).

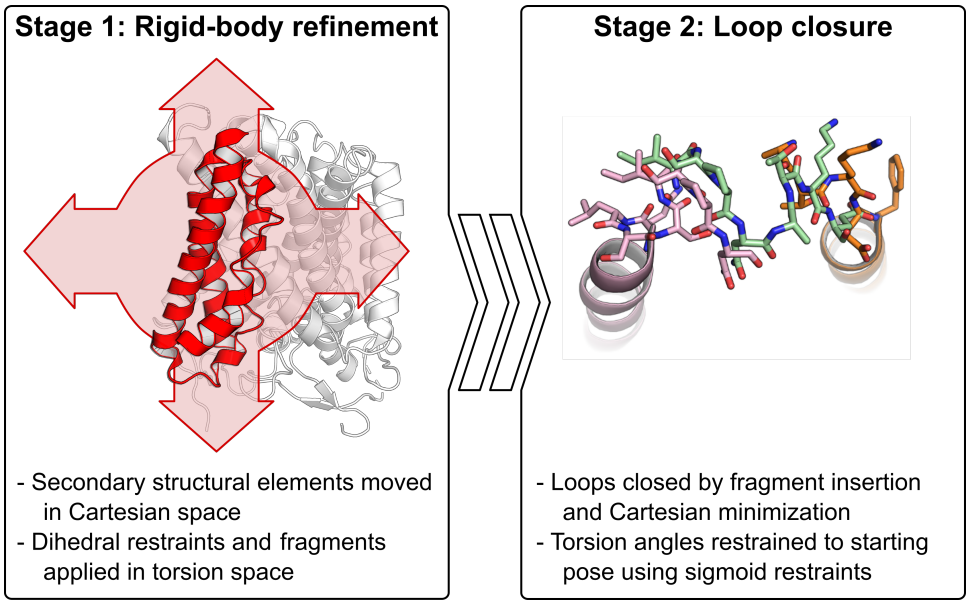

Figure C.1: Overview of the ConfChangeMover sampler.

C.2 Materials and Methods

C.2.1 Overview of the sampling approach

ConfChangeMover was implemented in Rosetta and executed in RosettaScripts [132]. As with RosettaCM, it samples candidate structural models using a two-stage strategy (Figure C.1). Prior to the first stage, the input structure is converted to a coarse-grained model with side chains replaced by immobile centroid pseudo-atoms. Cutpoints are introduced at residues located on loops connecting pairs of rigid bodies, or segments, which consist of either α-helices or β-sheets identified using the Dictionary of Secondary Structure of Proteins (DSSP) [202]. These cutpoints allow the positions and conformations of these segments to be perturbed either in isolation or in concert in the first stage without downstream propagation to rest of the protein via the "lever-arm effect" [419]. A series of sigmoid dihedral constraints, discussed below in section C.2.2, are added to model.

During the first stage, several types of perturbations are randomly introduced in the structural model. These include:

- **Rigid-body movements of one secondary structural element (SSE).** Rotation angles and translation vectors are randomly drawn from normal distributions with standard deviations of 15° and 2.0 Å, respectively.

- **Rigid-body movements of multiple spatially adjacent SSEs.** The number of SSEs randomly ranges from 2 to $N-1$, where N is the number of segments in the model, and the movement parameters match those used for a single SSE.

- **Helical twists.** Helices are twisted by a randomly chosen angle drawn from a normal distribution with a standard deviation of 15°.

- **Fragment insertion.** The dihedral angles of a three-residue stretch of the protein are modified to match those of a randomly chosen sequence fragment obtained from the PDB. Fragments were obtained using the Robetta web server as previously described [216].

Throughout this stage, SSEs that are adjacent in sequence are constrained such that their N- and C-terminal ends can be plausibly bridged by the loop connecting them. We achieved this by

automatically rejecting moves that separate the termini of consecutive SSEs by distances greater than $2.65n_{\text{res}} + 2.11$ Å, where n_{res} is the number of residues in the loop [447].

During the second stage, loops are closed using a method similar to one described previously [341] that is used by RosettaCM [384]. However, ConfChangeMover supplements this procedure using stretches of the starting model ranging from three to fifteen residues in length. The perturbations available include:

- **Fragment superimposition over gap regions.** Nine-residue fragments obtained from the PDB are superimposed over unresolved gaps in the structure.
- **Fragment superimposition over randomly chosen regions.** Nine-residue fragments are superimposed over randomly-chosen regions, which can include those that do not contain chainbreaks.
- **Template superimposition.** Stretches of residues with lengths ranging from nine to fifteen residues are copied from the starting conformation and superimposed over the model.

As with RosettaCM, during the final 25% of the second stage, each move is followed by a brief Cartesian minimization [81] using the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm [60]. At the end of stage 2, the entire model is minimized using the same approach.

During an optional third stage, all-atom side chains replace the centroid pseudo-atoms, and the entire model is iteratively minimized as previously described [81, 384].

C.2.2 Application of constraints during modeling

Several types of constraints are applied to these models throughout the algorithm. To account for the relative invariance of protein dihedral angles during conformational change modeling (see section C.3.1), circular sigmoidal restraints are added to the ϕ and ψ angles of the model based on either the starting conformation or a separate model provided by the user:

$$S_{\phi}(x) = \left(1 + \exp\left(|\phi_{\text{sim}} - \phi_{\text{exp}}| - \frac{\pi}{2}\right)\right)^{-1} + \left(1 + \exp\left(|\phi_{\text{sim}} - \phi_{\text{exp}}| + \frac{\pi}{2}\right)\right)^{-1} \quad \text{(C.1)}$$

$$S_{\psi}(x) = \left(1 + \exp\left(|\psi_{\text{sim}} - \psi_{\text{exp}}| - \frac{\pi}{2}\right)\right)^{-1} + \left(1 + \exp\left(|\psi_{\text{sim}} - \psi_{\text{exp}}| + \frac{\pi}{2}\right)\right)^{-1} \quad \text{(C.2)}$$

Additionally, following stage 1, coordinate constraints were applied to the C_{α} atoms of all residues belonging to SSEs. This allowed fragment insertions during stage 2 to close loops while minimally affecting the dihedral angles obtained following stage 1.

C.2.3 Benchmark on soluble proteins using simulated distance restraints

We first tested ConfChangeMover on seven soluble proteins (Table C.1). These topologically dissimilar proteins were selected from previous benchmarks [186, 366]. We first modeled all missing residues using RosettaCM [384]. Distance restraints between C_{α} atoms were selected from the starting structure using a modification of the Zheng-Brooks algorithm as implemented in the program MMM [186, 190, 322, 477], with one restraint per twenty residues in the total protein sequence, rounding up. For the benchmark on soluble proteins, we compared ConfChangeMover to fragment insertion, in which the dihedral angles of the starting structures were directly modified. For both methods, regions that did not undergo conformational transitions between the two states were not permitted to move and were not used for RMSD calculations.

C.2.4 Benchmark on LeuT-fold transporters proteins using experimental EPR restraints

For the benchmark using experimental data, ConfChangeMover was tested using three LeuT-fold transporter proteins: LeuT [224, 455], Mhp1 [374, 444], and vSGLT [438] (Table C.1). These three proteins have previously been studied using EPR, and all undergo ligand-dependent conformational transitions between IF and OF conformations [76, 212, 213, 310].

Table C.1: Protein structures used in the benchmark of ConfChangeMover. † A model of OF vSGLT was generated from the X-ray structure of the homolog SiaT. This model was not used as a target model in this benchmark. ‡ Insufficient experimental restraints were available to model the OF-to-IF transition in LeuT.

Protein *(Organism)*	**PDB A (Resolution)**	**PDB B (Resolution)**	**Length**	**References**
Soluble proteins (simulated data)				
Adenosylcobinamide kinase (*Salmonella enterica*)	1CBU (2.30 Å)	1C9K (2.20 Å)	181	[411, 412]
DNA polymerase I (*Thermophilus aquaticus*)	2KTQ (2.30 Å)	3KTQ (2.30 Å)	832	[239]
Glutamine-binding protein (*Escherichia coli*)	1GGG (2.30 Å)	1WDN (1.94 Å)	248	[170, 396]
Lactoferrin (*Homo sapiens*)	1LFH (2.80 Å)	1LFG (2.20 Å)	710	[155, 299]
Leucine-binding protein (*E. coli*)	1USI (1.80 Å)	1USG (1.53 Å)	369	[257]
Mitochondrial aspartate aminotransferase (*Gallus gallus*)	1AMA (2.30 Å)	9AAT (2.20 Å)	423	[275, 276]
Pol alpha DNA polymerase (*Escherichia* phage RB69)	1IG9 (2.60 Å)	1IH7 (2.21 Å)	903	[138]
LeuT-fold proteins (experimental data)				
LeuT (*Aquifex aeolicus*)	2A65 (1.65 Å)	3TT3‡ (3.22 Å)	513	[224, 455]
Mhp1 (*Microbacterium tumefaciens*)	2JLN (2.85 Å)	2X79 (3.80 Å)	489	[374, 444]
vSGLT (*Vibrio parahaemolyticus*)	2XQ2 (2.73 Å)	5NV9† (1.95 Å)	543	[428, 438]

Four transitions of interest were used to benchmark ConfChangeMover (Table C.1). In each case, experimental data was provided using the RosettaDEER module (see Chapter 3). For LeuT, the IF-to-OF transition took advantage of distance restraints collected on LeuT in the presence of the detergent β-OG [213]. In Mhp1, the OF-to-IF and IF-to-OF transitions were simulated using DEER data collected either without substrates or with both sodium and benzylhydantoin, respectively [212]. Finally, because vSGLT has only been crystallized in IF conformations [118, 438], we modeled the OF-to-IF transition by starting from a previously published OF homology model generated from the homologous sialic acid transporter SiaT [310, 428]. The IF-to-OF transition of vSGLT and the OF-to-IF transition of LeuT were not modeled.

For both benchmarks, we set stages 1 and 2 to consist of 50,000 rounds and 1,000 rounds, respectively. The weights of the score terms *dihedral_constraint*, *atompair_constraint*, and *coordinate_constraint* were set to 1.0, 10.0, and 1.0, respectively. The Rosetta scoring functions *score3* and *score4_smooth_cart* were used for stages 1 and 2, respectively. We generated 100 models using each method for the soluble protein benchmark and 1000 models for the LeuT-fold benchmark.

C.3 Results and Discussion

C.3.1 Most dihedral angles do not change during conformational isomerization

To determine the extent to which dihedral ϕ and ψ angles rotate during conformational change, we compared pairs of structures from a panel of conformational changes used in a recent benchmark [184]. Comparison of proteins in multiple states revealed that the overwhelming majority of backbone dihedral angles in proteins remain unchanged when undergoing conformational transitions, suggesting that interconversion between two states may be facilitated by a small fraction of residues as previously suggested [356] (Figure C.2). This expectation was implemented as a modeling restraint using circular sigmoidal functions on the backbone ϕ and ψ angles (see section C.2.2). Sigmoidal restraints have previously been used for modeling proteins using restraints with a small number of false positives, such as those obtained using residue coevolution [302, 303, 406], because they favor structural models that satisfy as many restraints as possible while minimizing the contribution of noisy data to the resulting model.

C.3.2 Benchmark on soluble proteins

We first benchmarked ConfChangeMover on a panel of seven soluble proteins using simulated distance restraints (Table C.1). Both directions were used as part of the benchmark, leading to fourteen total transitions. These proteins were specifically chosen for their complex modes of conformational isomerization, wherein loops and SSEs moves in ways that are unlikely to be easily recapitulated by simple rotation of backbone dihedral angles [89, 226]. To test that this was the case, the performance of ConfChangeMover was compared to simple fragment insertion. For both methods, regions that do not move in between these two states were not manipulated. Distance restraints between C_α atoms were chosen using a previously published restraint-picking algorithm [184, 477] and simulated from the target structure. Following the simulation, C_α RMSD values were calculated exclusively from the mobile regions of the protein.

The results are plotted in Figure C.3. In the absence of restraints, the majority of models sampled using ConfChangeMover were nearly identical to the starting model, indicated by the dashed

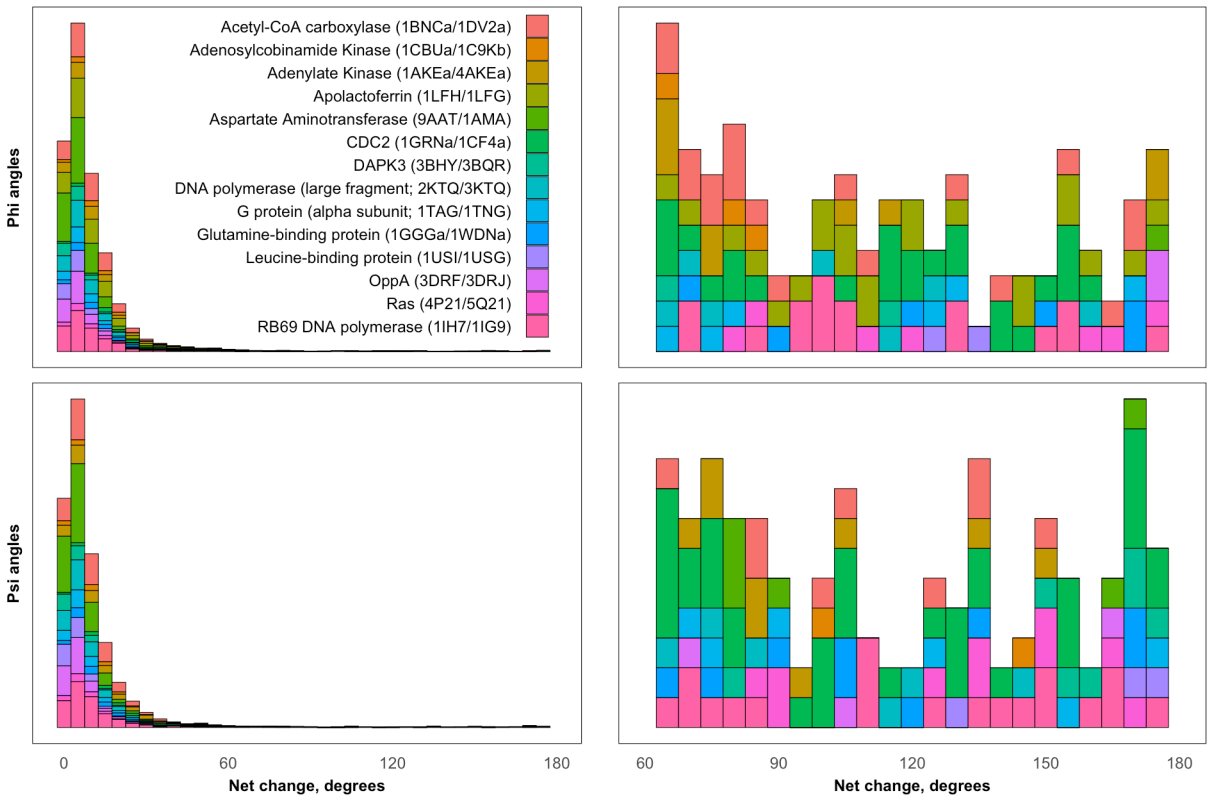

Figure C.2: Rotational changes observed in the dihedral angles observed in various proteins undergoing conformational changes. Left: the overwhelming majority of dihedral angles in the dataset rotate less than 30° during isomerization. Right: Close-up of cases that rotate more than 60°.

line, suggesting that the starting structure generally occupied an energy minimum that may be difficult to escape. Introducing simulated C_α-C_α restraints led to improvements in RMSD in every protein except lactoferrin. In contrast, using restraints with fragment insertion generally led to unfolding of the models and worsening of RMSD relative to the starting structure (see Figure C.3.A), despite the fact that fragments were only inserted in stretches of mobile residues. In general, the breadth of models sampled this way was both far larger and universally poorer in quality than those obtained ConfChangeMover. The exception to this second point, lactoferrin, appeared to be modeled relatively easily by simple changes in the dihedral angles connecting two domains. Additionally, the majority of simulated restraints obtained using the modified Zheng-Brookes restraint-

picking algorithm discussed above did not cover mobile portions of the protein. As a result, we believe the aggressive sampling available to fragment insertion may have been beneficial to escaping the local minimum and sampling conformers similar to the target. Nevertheless, the majority of models did worsen in RMSD relative to the starting structure.

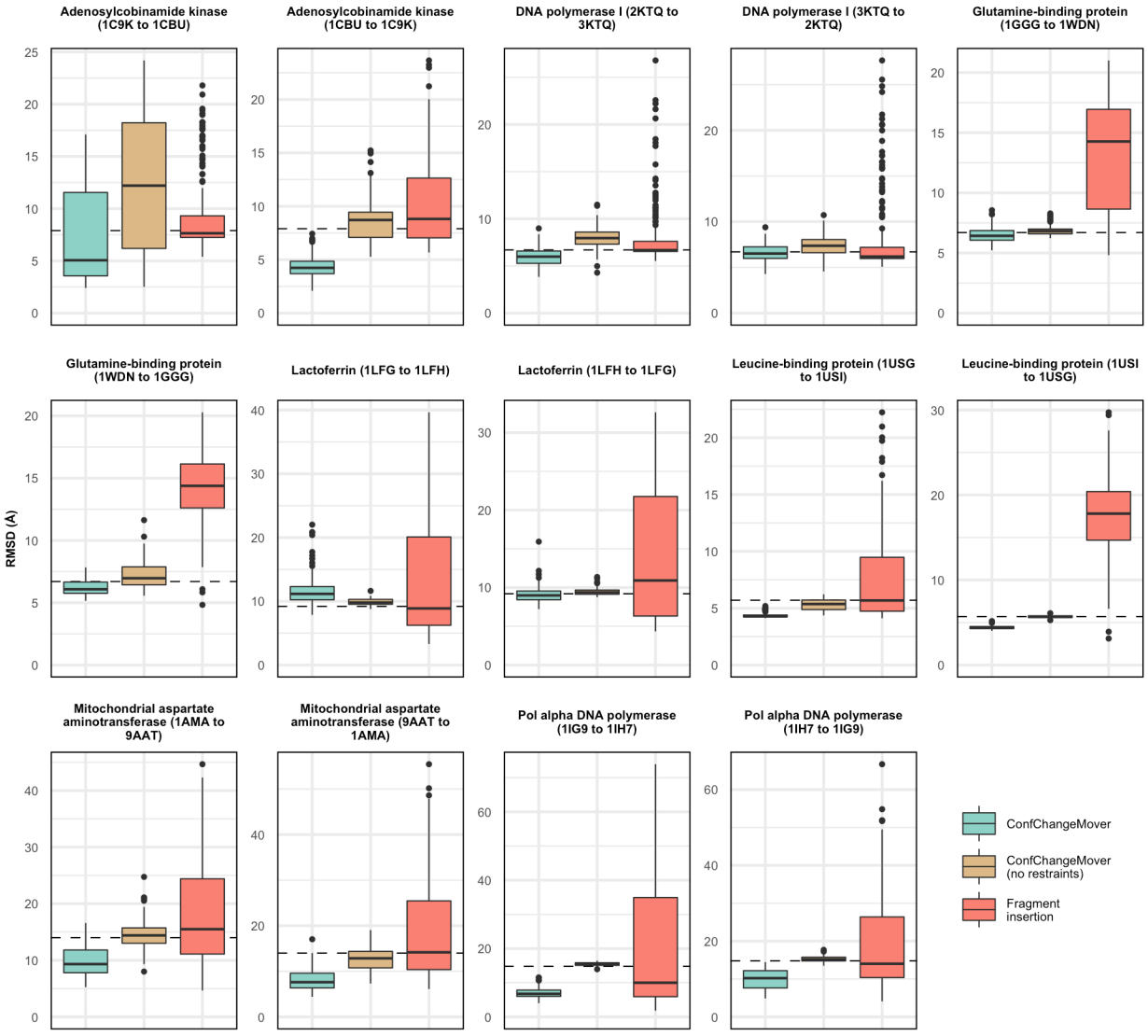

Figure C.3: ConfChangeMover outperforms fragment insertion in Rosetta when modeling conformational changes in soluble proteins using simulated C_α restraints. Starting models indicated by the dashed black line.

The remaining proteins in the benchmark set generally show a consistent pattern in which sampling using ConfChangeMover is focused and worsening of models is generally avoided. This is the

principal challenge of structure refinement [307], as there are many more ways to ruin a structural model than improve it.

C.3.3 Benchmark on LeuT-fold transporters using experimental data

With these results in hand, we then applied this method to LeuT-fold transporters that have been experimentally studied using EPR. The transmembrane domains of these transporters are entirely α-helical and undergo various divergent modes of isomerization to facilitate translocation of substrates across the membrane (see Chapter 1 for details). Libraries of DEER measurements collected in LeuT, Mhp1, and vSGLT are generally consistent with experimental structures. For LeuT, due to outstanding controversies surrounding the conformational details of the IF state [213, 383], we exclusively simulated the IF-to-OF transition using experimental measurements collected in the detergent β-OG. Helices in the bundle domain, as well as TMH5 and EL4, were permitted to move, while the positions of the hash domain and TMH10-12 were fixed. For Mhp1, both the IF and OF conformations were largely found to be consistent with EPR data collected in the apo and substrate-bound state and indicated rigid-body movement of the hash domain and bending of TMH5. For vSGLT, which has only been crystallized in IF conformations, we modeled the OF-to-IF conformational change starting from an OF homology model generated from the structure of the homolog SiaT [428] that was generated for a previous study [310]. The conformational changes suggested by the EPR data are largely consistent with those expected from this model and suggest a mechanism of alternating access that combines elements of both LeuT and Mhp1 that involve helices across the whole protein. We repeated the conformational change modeling procedure and compared it to both fragment insertion and to the homology modeling program RosettaCM [384]. Additionally, to evaluate the impact of experimental restraints on modeling, ConfChangeMover was also run with simulated DEER restraints generated using MDDS [176], as well as simulated C_α-C_α distance restraints generated from the target model.

The results are shown in Figure C.4 and repeat the pattern observed in those obtained in the soluble protein benchmark. ConfChangeMover largely samples structures similar to the starting

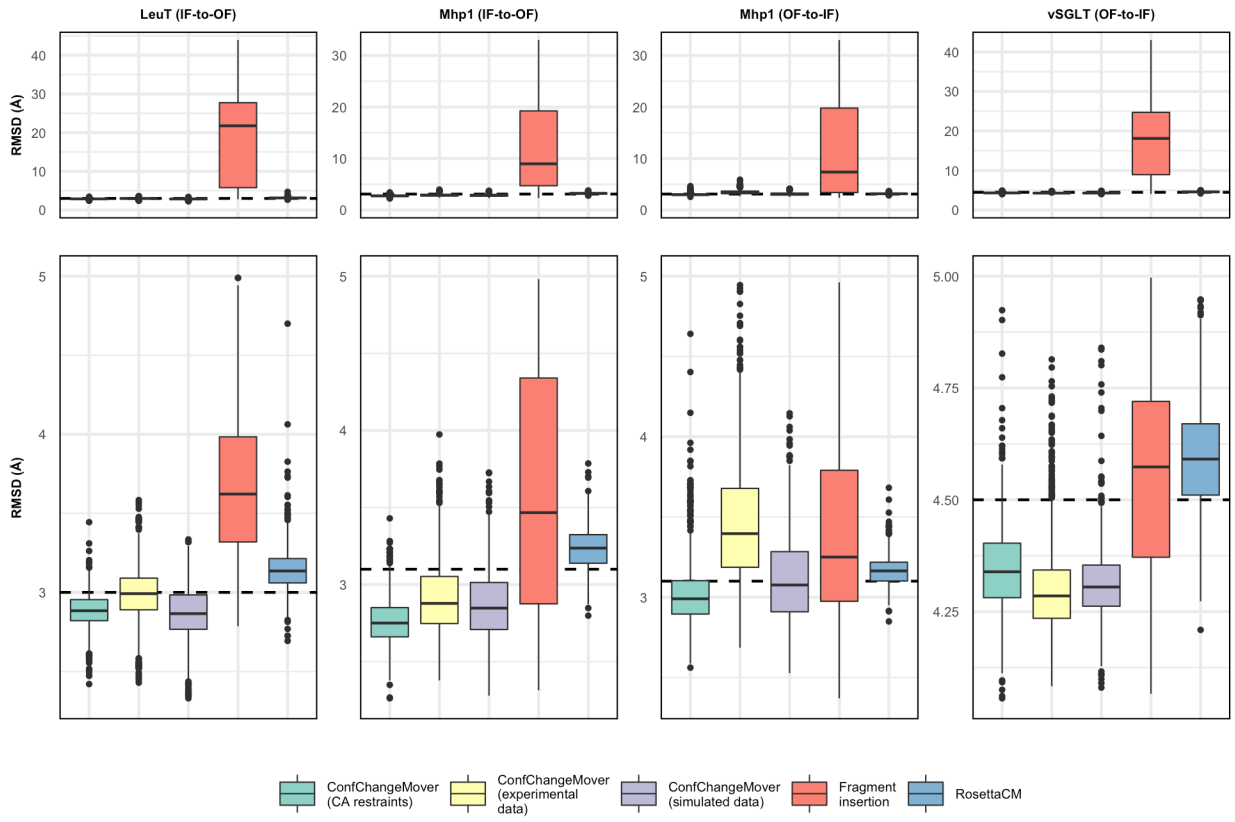

Figure C.4: ConfChangeMover outperforms available methods in Rosetta when modeling conformational changes in LeuT-fold transporters using EPR data. Starting models indicated by the dashed black line.

model, with virtually all models improving in RMSD relative to the starting structure. The exception, OF-to-IF Mhp1, saw most models get worse in quality. The striking different between models generated using simulated and experimental data suggested that poor model quality could be attributed to differences between predicted and experimental spin label distances. Indeed, a distance restraint between TMH1a and TMH9 on the intracellular side of the protein (30/338) differed from predictions obtained using MDDS by 14 Å [212]. Additionally, visual examination of representative models obtained using simulated DEER restraints revealed that the kinked TMH5 facilitating substrate egress into the cytoplasm could not be recapitulated using this method (Figure C.5).

Related to this observation, we found that in three cases the C_α-C_α distance restraints provided the greatest benefit, suggesting that the precision obtained when directly restraining the backbone is

generally superior to those obtained from spin probes measured using DEER. The results from the fourth case, vSGLT, were not statistically different from either experimental or simulated DEER data (unpaired two-tailed t-test), highlighting the consistency of that crystal structure with the experimental DEER data.

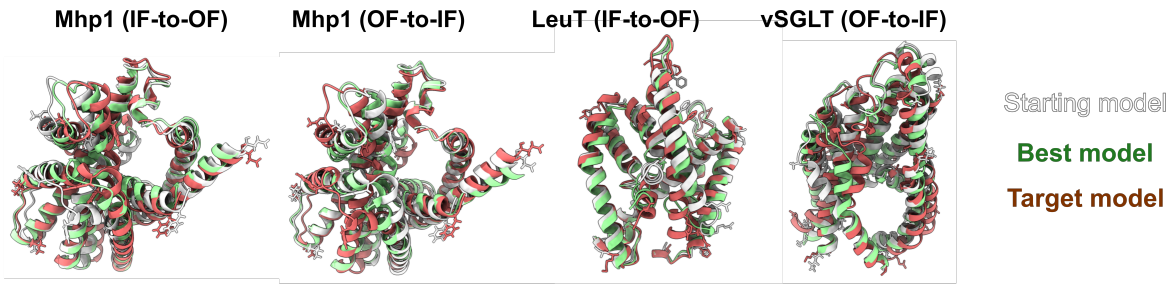

Figure C.5: Lowest RMSD models obtained using ConfChangeMover with experimental EPR restraints.

In marked contrast, and as was seen in the soluble protein benchmark, we found that fragment insertion similarly unfolded these models, with average RMSD values of 10 Å to 25 Å (Figure C.3). Unlike the soluble protein benchmark, in no cases did it outperform ConfChangeMover among proteins with this topology. Additionally, the homology modeling program RosettaCM sampled too conservatively and did not allow models to be modified away from the starting structure.

C.3.4 Concluding remarks

The modeling protocol ConfChangeMover is presented and discussed. Benchmarks in both soluble proteins using simulated distance restraints and membrane transporter proteins using experimental restraints show how this method can tackle various modes of conformational interconversion while generating models that nonetheless resemble the starting conformation. Its implementation as a Mover in Rosetta allows it to be used in combination with countless other modeling methods. Further development should focus on combining ConfChangeMover with other protocols, or embedding it in more complex modeling approaches [307].

C.3.5 Acknowledgements

We would like to thank Dr. Davide Sala for collaborating on this project and Dr. Marion F. Sauer for fruitful discussions during its early stages. The research in this Appendix was supported by the National Institutes of Health (R01 GM080403 and R01 GM073151), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB1423, project number 421152132, subproject A07 and Z04.

Appendix D

Efficient sampling of loop conformations using conformational hashing and random coordinate descent

The contents of this Appendix have been previously published [94].

De novo construction of loop regions is an important problem in computational structural biology. Compared to regions with well-defined secondary structure, loops tend to exhibit significant conformational heterogeneity. As a result, their structures are often ambiguous when determined using experimental data obtained by crystallography, cryo-EM, or NMR. Although structurally diverse models could provide a more relevant representation of proteins in their native states, obtaining large numbers of biophysically realistic and physiologically relevant loop conformations is a resource-consuming task. To address this need, we developed a novel loop construction algorithm, Hash/RCD, that combines knowledge-based conformational hashing with RCD. This hybrid approach achieved a closure rate of 100% on a benchmark set of 195 loops in 29 proteins that range from three to thirty-one residues. More importantly, the use of templates allows Hash/RCD to maintain the accuracy of state-of-the-art coordinate descent methods while reducing sampling time from over 400 ms to 141 ms. These results highlight how the integration of coordinate descent with knowledge-based sampling overcomes barriers inherent to either approach in isolation. This method may facilitate the identification of native-like loop conformations using experimental data or full-atom scoring functions by allowing rapid sampling of large numbers of loops. In this manuscript, we investigate and discuss the advantages, bottlenecks, and limitations of combining conformational hashing with RCD. By providing a detailed technical description of the Hash/RCD algorithm, we hope to facilitate its implementation by other researchers.

D.1 Introduction

Despite its importance to computational structural biology, the prediction of protein loops remains a challenge [238]. Without the periodic backbone hydrogen bonds defining regular secondary structure, a large conformational space needs to be searched. Moreover, loops can interconvert between many isoenergetic conformations, complicating efforts to identify a single conformation at a global energy minimum. Perturbation of loops in structures determined using experimental techniques such as crystallography further complicates the development of loop modeling methods [179], as the conformations observed in a crystal lattice may be artifacts of experimental design and/or data collection.

Algorithms that predict loop regions in proteins generally use one of several strategies. Template-based methods rely on experimentally determined loop conformations deposited in the Protein Databank to build missing loop regions. For example, in the Loophash algorithm [419], which is implemented in Rosetta [229, 234], the sequence of the loop target is threaded onto a template selected from a loop library. Other examples of template-based methods include Superlooper [163], which searches the Loops-in-proteins database [284]; FREAD [74], which uses several criteria to identify experimentally determined loops of interest; and DaReUs-Loop [207, 208], which uses loop flanking regions to identify suitable candidate loops for modeling. In general, template-based loop prediction has the advantage of being fast, since the loop dihedral angles have been experimentally observed. However, they are limited by the underrepresentation of long loops in the PDB (11 or more residues), which leads to fewer templates. Consequently, this approach is generally suitable for short and medium-sized loops (10 or fewer residues).

An alternative approach for modeling missing loops is to do so *de novo*. This approach achieves loop closure by relying on an energy function to optimize the dihedral angles of the residues comprising the loop. Various methods have been described that use this strategy, such as GalaxyLoop-PS2 [305], ModLoop [130], LEAP [240], PETRA [92], and Rosetta-KIC [260] and NGK [390]. A widely-used method is the cyclic coordinate descent (CCD) algorithm [63, 40], which was inspired by the random tweak algorithm used in robotics. Both CCD and the closely-related RCD algorithm

[75] rotate the loop's backbone torsion angles to place the a "virtual" terminal residue over the loop's anchor point. Coordinate descent methods have the advantage of high closure rates, even among longer loops. However, as with all *de novo* methods, they are limited by their time complexity, which depends on the loop's sequence length. Moreover, they can introduce distortions in the loop's dihedral angles.

Finally, a series of "hybrid" approaches hold promise to find middle ground between low time complexity of template-based modeling with the high closure rate of *de novo*-based methods. For example, both CODA [91] and Sphinx [265] arrive at consensus predictions by combining the template-based predictions from FREAD [74] with loops modeled *de novo*. The comparative modeling protocol RosettaCM [384] predicts loop conformations by assembling three- and nine-residue fragments from the PDB and closing loop regions using Cartesian minimization. Similar protocols have used a hybrid approach to model long hypervariable loops in antibodies [123, 268, 269, 445]. In summary, these methods allow long loops to be predicted in a reasonable amount of time without being restrained by the lack of experimentally determined templates of a given length.

An optimal loop construction algorithm would find the middle ground between low time complexity and high closure rate. Here we introduce and discuss an algorithm, Hash/RCD, that combines conformational hashing with RCD. Hash/RCD circumvents the lack of templates for long loops by constructing them from shorter fragments using a MC framework. The resulting hybrid algorithm, which is implemented as part of the BioChemical Library (BCL) [205, 447], combines the advantages of conformational hashing and coordinate descent while mitigating their respective limitations.

This Appendix discusses the implementation and performance of conformational hashing with and without RCD and is organized as follows. First, Materials and Methods (section D.2) describes the general methodology that combines conformational hashing and RCD within an MC framework. Next, generation of the loop template library is discussed. This is followed by the mathematical details regarding the parametrization of 1) loop conformations for conformational hashing and 2) fragments for recombination into longer loop templates. We then provide a technical description

of the RCD method used in this study and a summary of the benchmark set used to quantify the performance of Hash/RCD. Finally, this section is concluded with a description of the method's performance compared to the Rosetta Loophash algorithm [419], RCD in isolation, and the orthogonal sampling approach RosettaCM [384]. In the Results section we discuss the performance of Hash/RCD, which we evaluated using several metrics, including closure rate and central processing unit (CPU) time consumption. We also explored the limits of conformational hashing and compare the loops generated by Hash/RCD to those generated by Rosetta Loophash, RCD in isolation, and RosettaCM.

D.2 Materials and Methods

D.2.1 General methodology and generation of the loop template library

The sampling approach used by Hash/RCD is described in Figure D.1 and consists of two stages and a post-processing step. The first stage uses conformational hashing to construct loop regions from precomputed templates (the term "template" refers to any experimentally determined loop structure used for modeling purposes). The second stage identifies and closes loops that could not be closed during the first stage using RCD. Finally, a post-processing step constructs loops that may be missing from the protein's N- and C-termini.

We compiled the initial template library used by the first stage from a nonredundant subset of experimentally determined structures in the PDB [26]. The Dunbrack lab's protein sequence culling server (PISCES) server was used to filter out structures whose resolutions exceeded 3.0 Å and did not fully or partially consist of C_α-traces [429, 430]. This avoided the overrepresentation of protein structures that have been determined in multiple nearly identical conformations and resulted in about 87,000 structures. From there, we discarded the SSE definitions provided in the PDB files, instead using the SSE definition program DSSP [202]. Loops were defined as contiguous regions in sequence space that could not assigned any regular secondary structure. This approach both ensured reproducibility and standardized the assignment of secondary structure on the basis of backbone hydrogen bonding geometry. Finally, we removed loops containing unresolved backbone

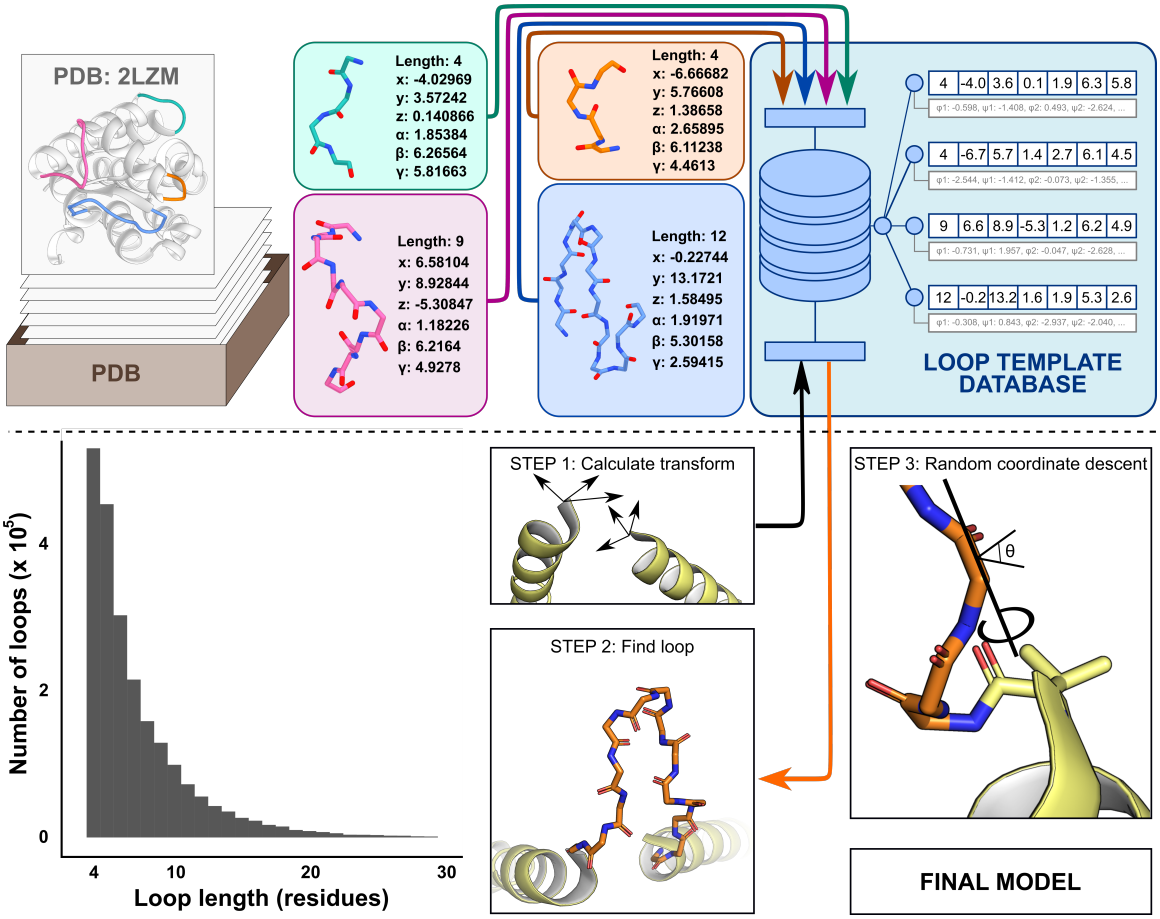

Figure D.1: Overview of the Hash/RCD algorithm. Top: Loop parameterization using a hash key computed from the length and relative orientation of its anchor points. A hash look-up identifies and selects suitable template conformations. Bottom left: The initial template library consisted of about 3.7 million loop conformations with different sequence lengths. These loops were collected from a set of about 87,000 protein structures deposited in the PDB. Bottom right: Depiction of the conformational hashing and RCD stages of this algorithm. An MC framework embeds both stages and allows loop templates to be added, replaced, and removed.

coordinates. Using this approach, we collected about 3.7 million loop conformations (Figure D.1).

The templates were then translated into hash keys describing the geometric aspects of the anchor residues flanking the loop (see next section for details). Lists of loops corresponding to specific geometries were stored together in a hash table. As we discuss below, Hash/RCD draws from this hash table to sample the addition and replacement of loops for given protein models by computing the parametrization of the missing loop, calculating the associated hash key, and inserting a suitable template chosen at random. The loop's sequence is then threaded over this template and inserted

into the protein model. By relying on a hash table to identify templates, this approach has the advantage of $O(1)$ computation time and generally requires CPU time on the order of microseconds. Since long loops have fewer templates in the loop library, they could be assembled from shorter fragments using a procedure discussed below.

Following conformational hashing, Hash/RCD uses random coordinate descent, previously described by Canutescu et al. [63] and Chys and Chacon [75], to minimize the distance between a moving (loop end) and target (anchor) set by calculating the rotation that must occur around a given axis (ϕ, ψ). We took several steps to diversify the loops generated this way. First, we found that randomly choosing which dihedral angle (either ϕ or ψ) at every step allowed Hash/RCD to avoid getting stuck in non-closable conformations. Additionally, only a random fraction of the rotation is applied. Further, we modified the original protocol to bend the terminal regions of the SSEs flanking the loop (Figure D.1). Finally, supplementing this protocol with scoring functions allows it to identify and reject rotations that cause the loop to clash with the rest of the protein model.

The final step of the protocol constructs the terminal loop regions. These were initialized with dihedral angles that are randomly chosen from a (ϕ, ψ) distribution derived from experimentally determined protein structures. The coordinate descent algorithm was then executed to resolve steric interferences and/or energetically unfavorable configurations.

D.2.2 Parametrization of loop conformations and selection of suitable conformations

In addition to their length, loop conformations can be defined by the relative rotational and translational orientation of the anchor residues flanking them. Therefore, we defined a local orthonormal coordinate system for the anchor points of each loop as (e_x, e_y, e_z) based on their backbone coordinates. Here e_x is the normalized $C_\alpha - C$ vector, e_y is the normalized component of the $C_\alpha - O$ vector orthogonal to e_x, and e_z is computed from e_x and e_y such that $e_z = e_x \times e_y$. Accordingly, the translation vector resides within this coordinate system and is defined as follows:

$$\vec{t} = (t_x, t_y, t_z) = (\alpha_{c,x} - \alpha_{n,x}, \alpha_{c,y} - \alpha_{n,y}, \alpha_{c,z} - \alpha_{n,z}) \tag{D.1}$$

Here $\alpha_{c,x}$ and $\alpha_{n,x}$ are the x-coordinates of the C_α-atom of the N-terminal and C-terminal anchors, respectively. The relative rotational orientation of the two anchor points was quantified using Euler angles (α,β,γ) following the extrinsic x-y-z convention [319]. These can be readily extracted from the matrix $\mathbf{M}_r$ describing the rotation between both coordinate systems that can be computed as $\mathbf{M}_r = \mathbf{M}_n^{-1} \cdot \mathbf{M}_c$, where $\mathbf{M}_n$ and $\mathbf{M}_c$ are the transformation matrices of the local coordinate systems at the N- and C-terminal anchor points [382].

Thus, each loop was parametrized into seven parameters: the loop's sequence length (d), translation vector (t_x,t_y,t_z), and Euler angles (α,β,γ). Each parameter is discretized into bins and translated into a one-dimensional hash key k using the hash function f:

$$f : d \times (t_x,t_y,t_z) \times (\alpha,\beta,\gamma) \rightarrow k \tag{D.2}$$

By grouping structurally similar loops into the same bins using this function, sparse populations within the hash map are avoided. We evaluated several different bin sizes in this study and found bin sizes of 1 Å for the translation vector and 60° for the Euler angles provided the optimal balance between closure rates and accuracy. Additionally, the hash map only stores the dihedral angles (ϕ,ψ) of each residue in the loop conformation. As a result, each loop conformation c can be described using only $2d+2$ parameters, e.g. $c = (\psi_N,\phi_1,\psi_1,...,\phi_d,\psi_d,\psi_C)$. Here d is the length of the loop in amino acid residues, ψ_N is the ψ-angle of the N-terminal anchor point, (ϕ_i,ψ_i) are the dihedral angles of the ith residue of the loop, and ϕ_C is the ϕ-angle of the C-terminal anchor point. The key-value pair (k,c) of each conformation is stored in the hash map accordingly.

These steps are all carried out during the generation of the loop template prior to loop prediction. During modeling, loop look-up proceeds as follows. First, the coordinate systems for the anchor points are computed and converted into a hash function (Equation D.2). Second, a range of suitable conformations capable of closing this loop are returned in $O(1)$ time, and one is chosen at random. Third, the sequence of the loop being modeled is threaded onto this randomly chosen conformation, a process that happens in $O(n)$ time. Therefore, the overall time complexity of the algorithm is limited by the linear-time computation of this last step, which is in turn determined by the length

of the loop being modeled.

D.2.3 Integration of conformational hashing within a Monte Carlo Metropolis framework

In a protein model containing multiple loop regions, closing a certain loop with a certain conformation might hinder the closure of other loops. Consequently, our algorithm must be able to sample different combinations of loops without needlessly increasing computational complexity. Owing to the previously demonstrated success of MC algorithms [229, 234, 205, 129], we embedded the conformational hashing step in an MC framework. Effectively, the loop construction algorithm consists of two sub-algorithms, conformational hashing and RCD, which are executed back-to-back. Each sub-algorithm places a "pseudo-residue" at the terminal end of a loop and tries to perfectly superimpose it over the corresponding anchor residue. Loop closure is calculated using the RMSD of this pseudo-residue and the anchor residue; we use an RMSD cutoff of 0.08 Å to account to allow for minor inaccuracies in bond lengths and angles.

In the MC implementation of the conformational hashing algorithm (Figure D.1), a loop is randomly selected, perturbed, and evaluated using a scoring function. Several perturbations can be sampled. First, missing loops can be added directly from the hash map. Second, subregions of a loop can be replaced with loops added from the hash map. Third, short stretches of up to three residues can be cut back at the anchor residues and replaced with loop conformations obtained from the hash map. Fourth, when suitable loops with sequence length $d+2$ are absent from the template library, stepwise construction of loops can instead be achieved by randomly selecting loop conformations with sequence length less than $d-2$ from the template library and applying them to the selected template.

Depending on the impact of these perturbations on the score of the model, the new model is either accepted or rejected [205]. Knowledge-based score terms used to evaluate loop conformations include clash evaluation, consistency with Ramachandran potentials, and residue-residue interactions. We also added score terms to evaluate a model's consistency with secondary structure prediction algorithms such as PSI-blast based secondary structure Prediction [197, 435], Jufo9D

[233], and MASP [279]. Finally, a score term consisting of 30% of the scoring function pushes the algorithm toward loop closure by linearly penalizing stretches of missing residues.

D.2.4 Construction of missing loop regions using random coordinate descent

The RCD algorithm closed loops only when they could not be closed using the conformational hashing algorithm. Its implementation is modeled on the CCD algorithm described by Canutescu et al. [63], which is in turn based on the random tweak algorithm [125, 370]; additionally, we included several modifications discussed by Chys and Chacon [75]. This portion of the algorithm can be divided into a pre-stage and a main-stage component.

During the pre-stage, missing residues are dynamically added to the anchor residues. The backbone dihedral angles of these residues are initialized with (ϕ, ψ) angles derived from a probability distribution of experimentally observed backbone dihedral angles. Then, using a knowledge-based potential, these (ϕ, ψ) angles are subsequently perturbed and evaluated. Potential inaccuracies in the secondary structure prediction are accounted for by adding and/or removing residues from the anchor SSEs. Throughout this pre-stage, sampling is guided by scoring terms evaluating the completeness of the amino acid sequence, steric interference between residues, residue-residue interactions, and the loop trajectory towards its anchor point. This module also constructed the terminal loop regions of the protein models. The main-stage portion of the algorithm iteratively uses RCD to calculate the rotation that must occur around a given axis (ϕ of ψ) to minimize the distance between the end of the loop and the target coordinates over many iterations to close a chain break. Throughout this step, residue-residue interactions and steric interferences between residues were evaluated using scoring functions.

We found that running the conformational hashing algorithm for 500 iterations appeared to offer the best balance between performance (loop closure) and time complexity. The algorithm could be terminated early if no score improvements were identified after 50 iterations. For the RCD algorithm, we obtained the best results when the algorithm was run for 2000 total iterations with the option of terminating early after 500 iterations without any improvements.

D.2.5 Compensation for lack of templates for long loops

The initial set of about 87,000 protein structures contained about 3.7 million loop templates. Of those, the majority of these templates (about 2.2 million) were four or more residues in length (Figure D.1). By contrast, only 12% of the templates were ten or more residues in length. Moreover, longer loops cover a greater conformational space, making the conformational hashing algorithm less likely to close these loops. We developed a two-pronged approach to overcome this challenge.

First, we supplemented our algorithm with a method that combines two short loop templates into a larger loop template by superimposing the backbone coordinates of one loop's C-terminal anchor point with the backbone coordinates of the other loop's N-terminal anchor point. The resulting template has a sequence length of $d = d_1 + d_2 + 1$, with d_n being the sequence length of the nth template. The translation vector $\vec{t}$ and the Euler angles (α, β, γ) are then computed in a straightforward manner. The local coordinate system of the N-terminal anchor point of the second template is first transformed into the local coordinate system of the N-terminal anchor point of the first template. This is achieved by multiplying the first coordinate system's rotation matrix the inverse of the second coordinate system's rotation matrix, i.e., $\mathbf{M} = \mathbf{M}_2^{-1} \cdot \mathbf{M}_1$. By multiplying the translation vector t_2 of the second template with this matrix, the resulting translation vector can be computed by simple vector addition, i.e., $\vec{t} = t_1 + (t_2 \cdot \mathbf{M})$.

Nevertheless, although this approach is theoretically sound, we found that small inaccuracies, likely due the binning strategy used in the original hash map, could be propagated when templates generated this way are in turn combined into new templates. Therefore, the stored dihedral angles for both templates were recombined into a sequence consisting of $2(d_1 + d_2 + 2)$ dihedral angles: $(\psi_{n,1}, \phi_1, 1, \psi_1, 1, ..., \phi_{1,d}, \psi_{1,d}, \phi_{1,c}, \psi_{\mathrm{n},2}, \phi_2, 1, \psi_2, 1, ..., \phi_{2,d}, \psi_{2,d}, \phi_{2,c})$. An artificial amino acid sequence of length $d + 2$ was generated within the algorithm and fitted against the combined sequence of dihedral angles, permitting the accurate computation of this template's parameterization.

D.2.6 The benchmark set used to evaluate the algorithm

We evaluated the performance of this algorithm using a benchmark set consisting of twenty-nine soluble and membrane proteins (Table H.1). This included the set of soluble proteins previously used by Tyka et al. to benchmark the Rosetta Loophash algorithm [419], as well as the eleven membrane proteins previously used to benchmark the protein structure prediction algorithm BCL::MP-Fold [126]. These proteins ranged in size from 57 to 1,560 residues with varying α-helical and β-strand secondary structure content. The 195 loops in this benchmark set with four or more residues had lengths between three and 31 residues. As we described earlier, secondary structure definitions were obtained using DSSP [202]; loops identified this way were then removed from each of the PDB structures. Loops with stretches of missing coordinates were excluded from the benchmark set.

For the purposes of this benchmark, we used a modified loop template library in which we removed templates belonging to homologs of proteins in the benchmark set (for heterooligomers, we kept chains that were not homologous, but removed those that were). For these purposes, a cutoff of 25% sequence identity was used when defining homology. Homologs were identified by pairwise alignment of all 87,000 proteins to each protein in the benchmark set using Clustal Omega [376]. This reduced the size of the loop template library by 3.4%.

D.2.7 Comparison with RCD, Rosetta Loophash, and RosettaCM

Three loop prediction protocols were compared to Hash/RCD. We used the Rosetta scoring function *score4_smooth_cart* to score all models generated this way. The first method, RCD, is simply the Hash/RCD algorithm without any conformational hashing. The second method, Rosetta Loophash [419], was modified slightly to change the focus from loop diversification to loop closure. The "relax" stage of the procedure was skipped, leaving only resampling- and minimization-based loop closure stages. Because Loophash assumes "ideal" bond length and angles, we ran the Rosetta idealize application on each structural model in the benchmark set prior to use. The same set of structures obtained in the previous section were used to create a database for Loophash (unlike our

loop database, however, even regions of the protein with secondary structure were included in the Loophash database). The 195 non-terminal loops present in the benchmark set were each tested individually, in the context of the full experimentally determined structure of the remaining loops, with parameters set to skip RMSD-based filtering. A total of 100 output structures were sampled for each loop in the benchmark set, and each output structure represents the result of 100 randomly selected database loops matching the required geometry. Although Loophash always produced models, we found that its substitution and minimization approach frequently perturbed the protein structure, even in regions outside the loop. For this reason, whenever at least one of the 100 output structures were within 1 Å C_α RMSD from the input structure, Loophash's ability to close that loop was defined as successful. Correspondingly, if none of the output structures met this criterion, Loophash's ability to close the loop was defined as a failure. Runtimes for Loophash are reported as an average for a single output structure and include only the time spent actively sampling the loop.

The third method, RosettaCM, is a homology modeling protocol that fills loops using a hybrid approach combining fragment insertion with full-atom Cartesian minimization [384]. We chose this protocol as an orthogonal approach for the purposes of benchmarking Hash/RCD for two reasons. First, it is widely used for homology modeling, a task that often involving loop prediction and closure. Second, the strategy for loop closure, which involves the insertion of three- and nine-residue fragments followed by Cartesian minimization, is superficially similar to the strategy used by Hash/RCD. The fragment templates used by RosettaCM were obtained from the Robetta web server with homologs excluded [216]. During the coarse-grained stages of the protocol, the *atom_pair_constraint* score term was set to 5.0 and the *frag_weight_aligned* option was set to 0.0 to avoid modifying non-loop regions. The full-atom Cartesian minimization step was then carried out with coordinate constraints from the starting model. Models were then scored using the scoring function *ref2015_cart*. Because RosettaCM explicitly models all sidechains and has an execution time ranging from minutes to hours, we do not report runtimes for this protocol and use it only to compare the RMSD values of the resulting models.

D.3 Results

This section describes the distribution of templates collected from structures in the PDB and rationalizes the need for template recombination. Following this, the performance of the conformational hashing algorithm is reported using loop closure rate and compute time. Additionally, we discuss the algorithm's performance as a function of different optimization parameters. This section concludes with a comparison of models generated by Hash/RCD, RCD alone, Rosetta Loophash, and RosettaCM.

D.3.1 Effect of parameter bin size and loop length on loop closure by conformational hashing

Before computing the hash key for a loop template, the loop's parametrization needs to be discretized, which is achieved through binning (see Section D.2). The hash map's granularity is heavily influenced by the bin width, which in turn significantly influences the closure rate and physical reasonableness of the loop regions it generates. For example, whereas larger bin widths are more densely populated, the resulting loop regions can be physically unreasonable. Smaller bin widths, by contrast, result in a sparser population of the hash map and come at the cost of a lower loop closure rate. Therefore, we quantified the influence of the bin width on the loop closure rate by repeating loop construction for the benchmark set with different bin widths. Specifically, the rotation bin width was increased in 30° steps from 30° to 120° (the translation bin width was kept at 1 Å because larger translation bin widths would require post-processing of the fitted loop to avoid overextension of the peptide bonds). This evaluation did not use a post-processing step, such as minimization, to reduce computational complexity.

The results are shown in Figure D.2 and demonstrate how loop closure rate decreased approximately linearly with loop length. For example, among loops no more than five residues long, a bin width of 60° led to a loop closure rate of 94%. Among loops between six and ten residues long, however, it dropped to 61%, and for loops greater than ten residues long, it fell further to 33%, for an average total loop closure rate of 70%. We observed a similar, almost linear relation between loop length and loop closure rates when using the other three evaluated angle bin widths. More

generally, we found that loop closure strongly depended on angle bin width. For the evaluated angle bin widths of 30°, 60°, 90°, and 120°, the total loop closure rate of the conformational hashing algorithm arrived at 58%, 70%, 78% and 89% (Figure D.2). Although these results suggest that larger angle bin widths are preferable, upon closer examination we found that many of the resulting models revealed had unnatural angles within the peptide bonds connecting the anchor SSEs and the loop. Rather than increase the computational complexity of the algorithm by using minimization or other post-processing steps, we opted for an angle bin width of 60°.

D.3.2 Conformational hashing achieves a high loop closure rate for short loops but RCD is required for long loops

We evaluated the percentage of missing loop regions among our benchmark set (Table H.1) that could be successfully constructed by first removing all non-terminal loop regions from each protein structure. We then constructed the missing loop regions using our algorithm and computed the loop closure rate per benchmark protein. This was performed with the original template library either prior to or following extension by template recombination and fragment-based loop construction (see section D.2). In both cases, the loop libraries were binned at 60°.

When additional templates through fragment combination are excluded, the algorithm achieved a loop closure rate of 54% over all twenty-nine benchmark proteins. As demonstrated above, a loop's length strongly influenced whether it could be successfully closed. For example, loops up to five residues long were successfully closed 87% of the time, whereas loops ten residues long or greater were closed only 21% of the time. By contrast, when the template library was extended to include loops generated using fragment-based construction, the overall loop closure rate could be improved by nearly 30% to an overall closure rate of 70% (Figure D.2). Further increases in either the size of this library through additional fragment recombination did not appear to improve loop closure rate (data not shown). Nevertheless, as we mentioned above, following this step with RCD achieved a closure rate of 100%. We found that Rosetta Loophash achieves a 99% loop closure rate on the loops with a length of ten residues or less. However, this plummeted to 39% when

loop length exceeded ten residues. We note that this high loop closure rate is achieved in part by using a 2 Å bin width, which is wider than the 1 Å width used by Hash/RCD. A consequence of this approach is its reliance on optimization and refinement following the fitting, which increases the computational effort required (discussed below).

D.3.3 CPU time requirement is dominated by the evaluation of steric interference

Although the time complexity of the template look-up is O(1), the task of threading the target sequence against the template is O(n), which leads to linear time complexity for the overall loop construction algorithm. We therefore studied the effect this had on the CPU time required to execute the algorithm (specifically, the time between entering the MC algorithm and leaving the MC algorithm divided by the number of successfully constructed loop regions). To evaluate the contribution of the scoring term evaluating steric interference between residues relative to the overall CPU time requirement, the computation time needed by this term was evaluated separately. This was repeated for both RCD alone and Hash/RCD.

When steric interference was not included during sampling, conformational hashing required on average 27±4 ms CPU time per loop, whereas RCD required 159±11 ms. By contrast, Hash/RCD required only 68±7 ms (Figure D.2). When steric interference is calculated during sampling, the required CPU time increased to 59±5 ms for conformational hashing, 468±41 ms for RCD, and 161±13 ms for Hash/RCD. The evaluation of steric interferences dominated the computational burden by accounting for 54%, 66%, and 58% of the total CPU time requirement among conformational hashing, RCD, and Hash/RCD, respectively. Overall, these results demonstrate that using templates prior to coordinate descent can lead to a threefold decrease in the computation time of loop closure.

In contrast, the Rosetta Loophash protocol has a CPU runtime of 160 s to sample each loop. This runtime is highly correlated with total protein size ($R^2 = 0.8$). We found that it is overwhelmingly devoted to minimization, as only 5% of the algorithm's runtime involved conformational sampling.

D.3.4 Hash/RCD samples experimentally observed conformations

Hash/RCD was developed to efficiently sample both major and minor populations that a loop might adopt. We assume that the experimentally determined structures deposited in the PDB correctly and accurately represent one of the proteins' major populations (minor populations, by contrast, can only rarely be verified experimentally and are not considered for this benchmark). To test whether Hash/RCD correctly samples those conformations, we generated 100 models with constructed loop regions for each protein in the benchmark set. These conformations were subsequently compared to the experimentally determined structures using the RMSD of the C_α-atoms. For the membrane protein structure 3P5N, two non-terminal loops were not resolved in the X-ray-derived model and consequently excluded from this comparison. Moreover, in the case of homo-oligomeric proteins in the benchmark set, we excluded each instance of a given loop beyond the first to avoid their overrepresentation in the benchmark set. This led to 195 non-terminal loops comprised of three or more residues.

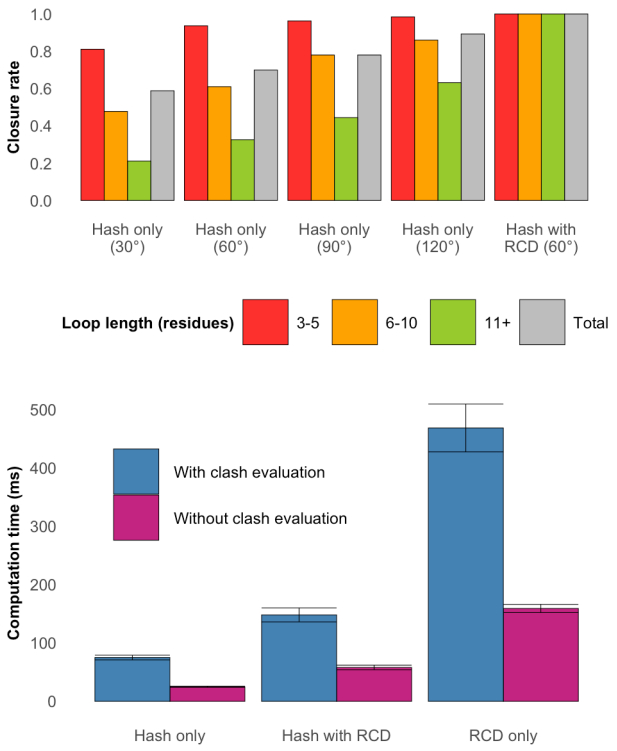

Figure D.2: Evaluating conformational hashing and RCD algorithms for loop construction. Top: The rotation angle bin width of the hash map influenced the loop closure rate of longer loops when using conformational hashing alone. When modeling short loops, we found high loop closure rates regardless of bin width. When conformational hashing was combined with RCD, the loop closure rate increased to 100% in our benchmark set. Bottom: The calculation of steric interference constituted the overwhelming majority of this algorithm's CPU time requirement. Combining conformational hashing with RCD was found to improve CPU time efficiency.

To focus on how effectively Hash/RCD could sample the major loop population, we determined the loop with the lowest C_α RMSD among each of the 100 conformers sampled for each of the

195 loops in the benchmark set (Figure D.3, top left). As expected, the lowest RMSD among the conformers we sampled depended on the length of the loop, which in turn dictated the size of the sampling space. However, in the majority of loops, at least one conformer was within 2 Å of the experimentally observed structure, and all of the longer loops sampled at least one conformer within 5 Å C_α RMSD. We found that these results were comparable to those obtained using RCD alone, as well as the orthogonal hybrid method RosettaCM. In sharp contrast, Rosetta Loophash was far less capable of sampling native-like loop conformers with lengths exceeding ten residues. These results suggest that the lack of long templates prevents these loops from being closed using physiologically meaningful structures, and that sampling becomes the predominant obstacle for modeling long loops using conformational hashing.

When native loop conformations are unavailable and RMSD values cannot be calculated, scoring functions must be used to infer which loop conformers are structurally relevant. We therefore focused on the lowest-scoring loops obtained using each method. Distributions are shown in Figure D.3 (right panels), and pairwise comparisons between Hash/RCD and other methods are shown in Figure D.4. The results reinforce the findings discussed above. Notably, the lowest-scoring loop conformer in most cases had RMSD values only slightly higher than the lowest-RMSD loop. We interpret this to suggest that the sampling space of the Hash/RCD method is focused on native-like conformers and ignores portions of the conformational space that are unlikely to be physiologically relevant. Moreover, these RMSD values are comparable to those obtained from the lowest-scoring loops obtained using RCD alone and contrast with the lowest-scoring conformers of long loops obtained by Rosetta Loophash. Finally, they improve upon loop modeling using RosettaCM, which interestingly was less effective at predicting short loops than we expected but reproduced the model quality of long loops that was observed using either Hash/RCD or RCD alone.

Our results highlight the shortcomings of an exclusively template-based strategy when attempting to model long loops. When taken alongside the improvements in computation time, the results suggest that Hash/RCD arrives at conformations similar in quality to RCD alone, but skips over a large number of intermediate conformations that would otherwise be expensive to sample. By

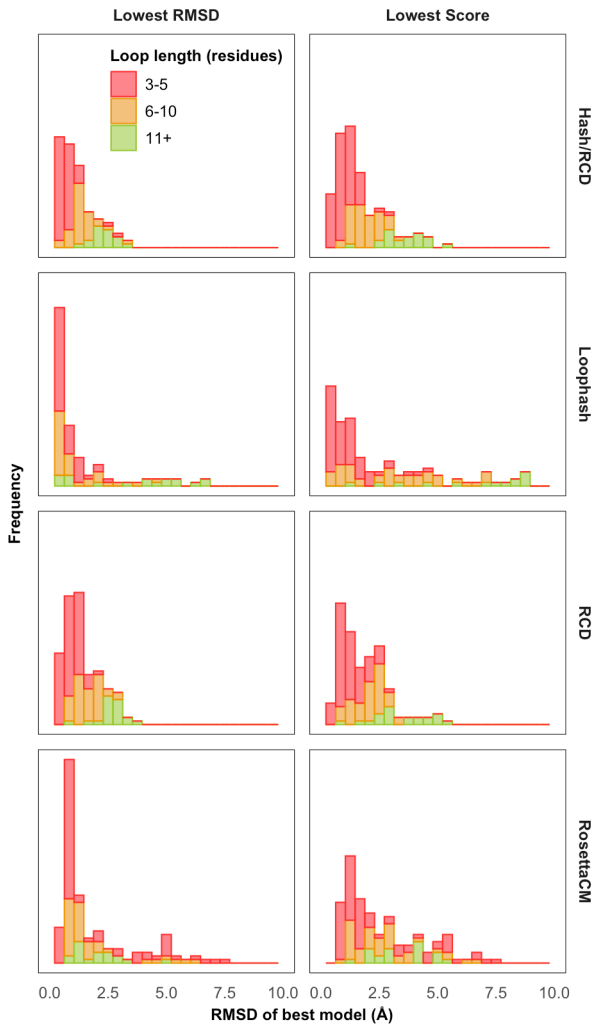

Figure D.3: Loops generated using Hash/RCD are comparable in quality to those using RCD alone. Left: Histograms of the lowest RMSD values among loop conformers sampled by a variety of methods. Although Rosetta Loophash slightly outperformed Hash/RCD, RCD alone, and RosettaCM when modeling short loops, it was unable to model native-like conformations for long loops. Right: Histograms of the RMSD values of the lowest-scoring loop conformers.

further improving conformations obtained using conformational hashing, Hash/RCD achieves results comparable to RCD alone on computational timescales comparable to those of template-based methods.

D.4 Discussion

D.4.1 Complementing conformational hashing with template-independent modeling

Loop modeling algorithms that employ conformational hashing methods must address the fact that most loops found in structures deposited in the PDB have sequence lengths of less than ten residues. This is evidenced by our initial loop library, which was derived from about 87,000 protein structures and disproportionately consists of short loops (Figure D.1). Additionally, longer loops can cover a larger conformational space, further impeding construction of long loops using conformational hashing. When we designed Hash/RCD, we found that this led to discrepancies between the loop closure rates for different loop lengths. For example, whereas loops between three to five residues long were closed 96% of the time, the loop closure rate dropped to 61% and 33% among loops with lengths of either six to ten residues or eleven or more residues, respectively. (Figure 2). We mitigated this problem by using template recombination and fragment-based loop construction (see section D.2), which improved the loop closure rate from 54% to 70%. Nonetheless, these results reinforced the need for template-independent conformational sampling. In this study, we integrated and applied an implementation of random coordinate descent to portions of loops that could not be constructed by conformational hashing, which compensated for the latter's inability to close long

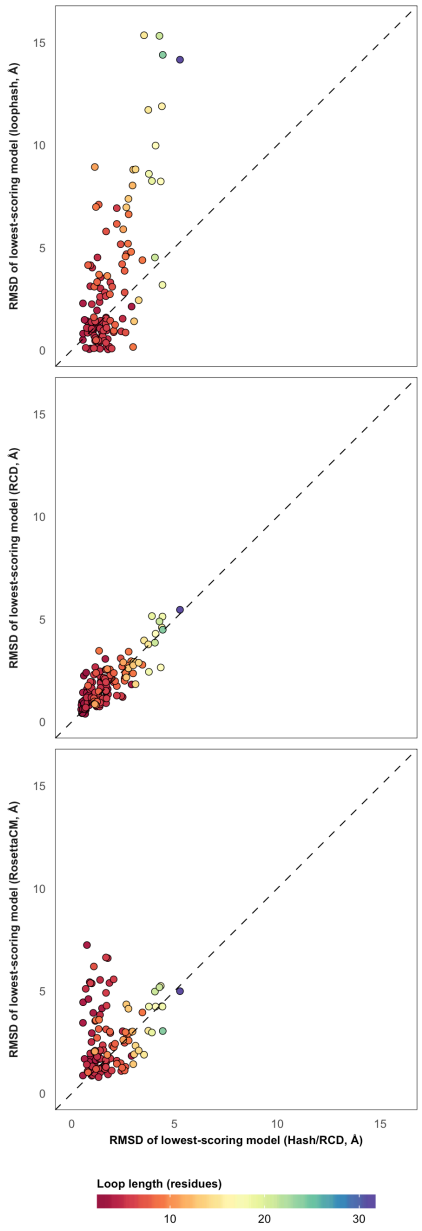

Figure D.4: Pairwise comparison of RMSD values among the best-scoring loop conformations obtained using Hash/RCD, Rosetta Loophash, RCD alone, or RosettaCM.

loops. As a result, loop closure improved to 100%. Consequently, whereas a stand-alone conformational hashing approach might suffice when the goal is to construct very short loop regions, loops found in most proteins will require a hybrid template-based/template-independent loop construction algorithm.

D.4.2 Hash/RCD efficiently samples structurally diverse loop conformations

Prediction of structural heterogeneity in proteins can be achieved by sampling diverse conformations (Figure D.5), which could capture major and minor populations of the protein in equilibrium. Random coordinate descent, although computationally demanding, achieves a high loop closure rate that cannot be replaced by more CPU-efficient template-based methods. Relying on RCD only in cases where conformational hashing was unsuccessful led to a significant reduction in CPU time; this was demonstrated by a drop from 468ms using RCD alone to 161 ms using Hash/RCD (Figure D.2). Thus, integrating conformational hashing with RCD combines the efficiency of conformational hashing with the high loop closure rate of RCD. The reduction in CPU time allows a wider range of possible loop conformations to be considered (Figure D.4). However, it needs to be noted that for longer loops in the benchmark set, the conformation of the experimentally determined structure was not sampled in any of the models to within 2 Å (Figure D.3). Since this problem was exclusively faced by loops longer than ten residues in length, it can be solved by incorporating experimental data from technique such as electron paramagnetic resonance spectroscopy or fluorescence resonance energy transfer, to reduce the size of the conformational space. Alternatively, models obtained using Hash/RCD can be further refined in molecular dynamics simulations, which may save considerable sampling time and assist in identifying more physiologically relevant loop conformers.

D.5 Conclusion

The hybrid loop modeling method Hash/RCD provides an efficient way to sample structurally diverse loop conformations and is significantly faster than using the template-independent approach RCD alone. We found that the constructed loop regions largely exhibit naturally occurring di-

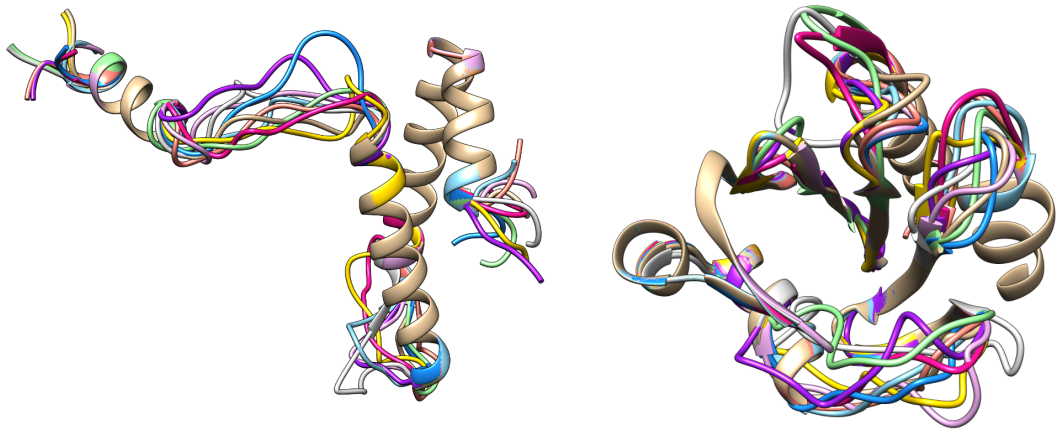

Figure D.5: Representative loop predictions obtained using Hash/RCD.

hedral angles due to their construction from experimentally observed conformations. While this algorithm's millisecond-timescale computation time was only quantified when implemented in the BCL, we believe similar performance can be theoretically achieved in any protein structural modeling program.

Two applications of this approach are proposed. First, it could be used for the prediction of conformational ensembles in loop regions. For example, one could conceivably fit these loops against experimental data to determine a weighted distribution of conformers that represents the protein in question under equilibrium conditions. Second, the simultaneous prediction of adjacent loops could reveal a protein's topological details, for example by capturing to what extent two loops may be intertwined. This may be relevant to multipass integral membrane proteins, as their solvent-exposed loop regions are least likely to be resolved.

D.6 Acknowledgments

We want to thank Dr. Marion F. Sauer for thorough proofreading of this manuscript. Work in the Meiler laboratory is supported through NIH (R01 GM080403 and R01 GM073151). The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB1423, project number 421152132, subproject A07 and Z04. Parts of the data analysis were performed using R in conjunction with the ggplot2 package [420]. Figures depict-

ing protein models were created using Chimera [318] and composite figures were created using Inkscape. To determine sequence identities between the template set and the benchmark set, the sequences were aligned using Clustal Omega [376].

Appendix E

Supplement to "Rapid simulation of unprocessed DEER decay data for protein fold prediction"

This Appendix contains supplementary information for Chapter 3.

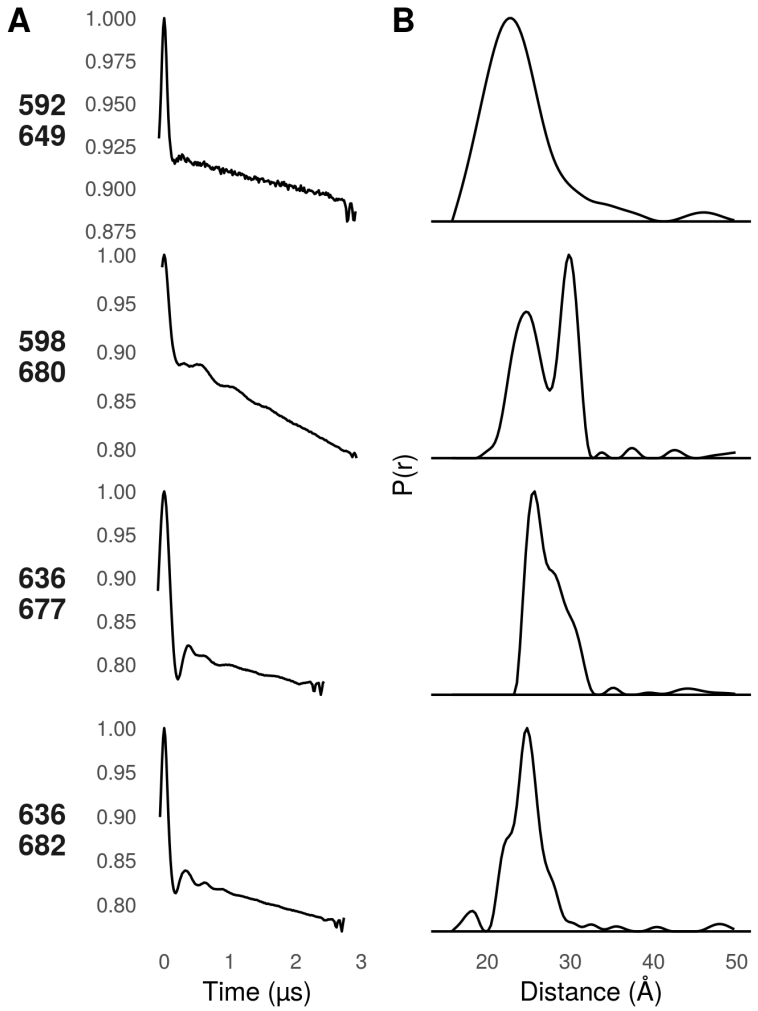

Figure E.1: Data gathered in the ExoU C-terminus for this study.

Table E.1: List of spin-labeled proteins in the Protein Databank.

Protein	PDB	Angle	Distance	Ref.
T4 Lysozyme	1ZYTa	5.23	0.83	[133]
T4 Lysozyme	2CUUa	7.24	0.84	[133]
T4 Lysozyme	2CUUa	3.39	0.78	[133]
T4 Lysozyme	2IGCa	11.49	0.81	[151]
T4 Lysozyme	2NTHa	3.06	0.8	[151]
T4 Lysozyme	2OU8a	7.13	0.82	[151]
T4 Lysozyme	2OU8a	9.08	0.86	[151]
T4 Lysozyme	2OU9a	15.13	0.84	[151]
T4 Lysozyme	2Q9Da	10.99	0.8	[151]
T4 Lysozyme	2Q9Ea	14.55	0.86	[151]
T4 Lysozyme	2Q9Eb	14.55	0.86	[151]
T4 Lysozyme	2Q9Ec	8.08	0.84	[151]
GB1	3V3Xb	4.78	0.83	[84]
GB1	3V3Xb	8.53	0.82	[84]
GB1	3V3Xb	1.13	1.02	[84]
GB1	3V3Xc	2	0.8	[84]
GB1	3V3Xd	5.65	0.78	[84]
GB1	3V3Xd	1.39	0.8	[84]
GB1	3V3Xd	2.71	0.81	[84]
GB1	5BMGa	10.97	1.08	[85]
GB1	5BMGb	4.57	0.79	[85]
GB1	5BMGc	13.52	1.13	[85]
GB1	5BMGd	5.89	0.8	[85]
GB1	5BMGe	1.45	0.99	[85]
GB1	5BMGf	7.7	0.84	[85]
GB1	5BMGg	8.47	0.87	[85]
GB1	5BMHa	8.29	0.77	[85]
GB1	5BMHa	9.38	0.77	[85]
GB1	5BMIa	7.42	0.83	[85]
Azurin	5I26a	4.8	0.81	[80]
Azurin	5I26b	0.55	0.99	[80]
Azurin	5I26c	0.92	1.01	[80]
Azurin	5I26d	0.89	1	[80]
Azurin	5I28a	0.5	1	[80]
Azurin	5I28b	0.71	1	[80]
Azurin	5I28c	0.32	1.01	[80]
Azurin	5I28d	0.25	1.01	[80]
Azurin	5I28e	4.49	0.96	[80]
Azurin	5I28f	1.81	1.01	[80]
Azurin	5I28f	2.15	0.98	[80]
Azurin	5I28h	2.1	0.89	[80]
Azurin	5I28i	5.44	0.8	[80]
Azurin	5I28j	6.8	0.85	[80]
T4 Lysozyme	5JDTa	2.52	0.78	[80]
T4 Lysozyme	5JDTa	0.85	0.76	[80]

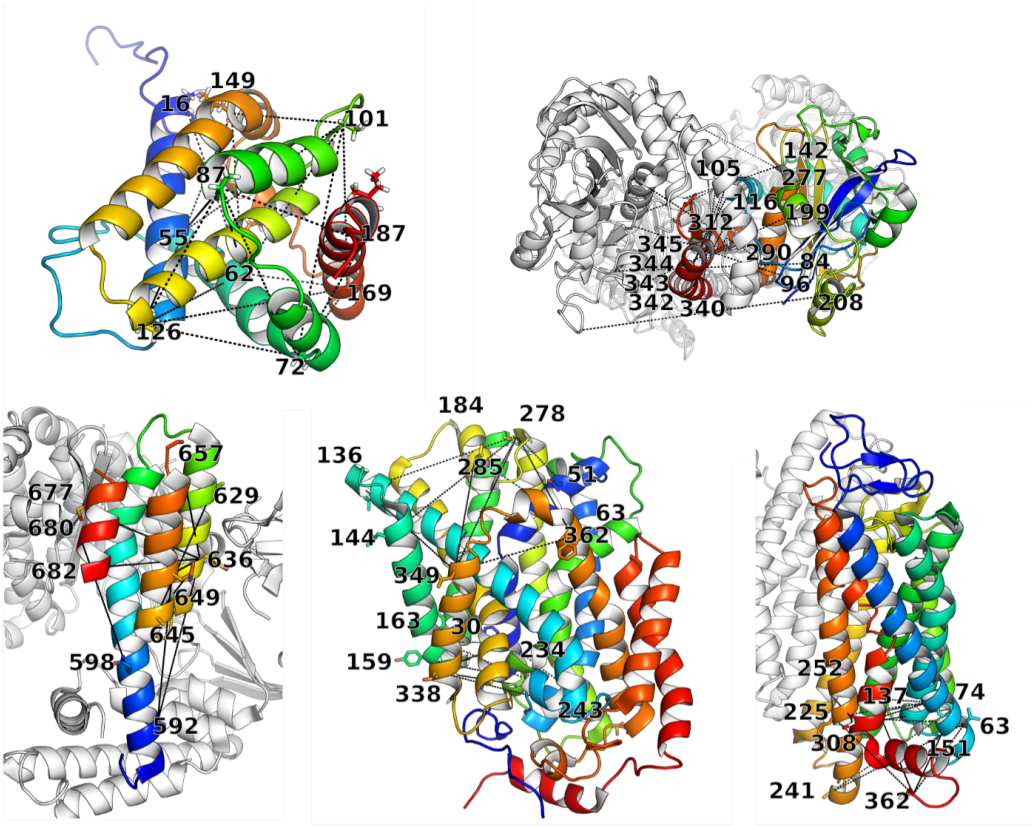

Figure E.2: Placement of experimental DEER restraints on protein structures used in this study. Clockwise from top left: Bax (PDB: 1F16 model 8), CDB3 (PDB: 1HYN chains R/S), Rhodopsin (1GZM chain A), Mhp1 (PDB: 2JLN), and ExoU (PDB: 3TU3, C-terminus only).

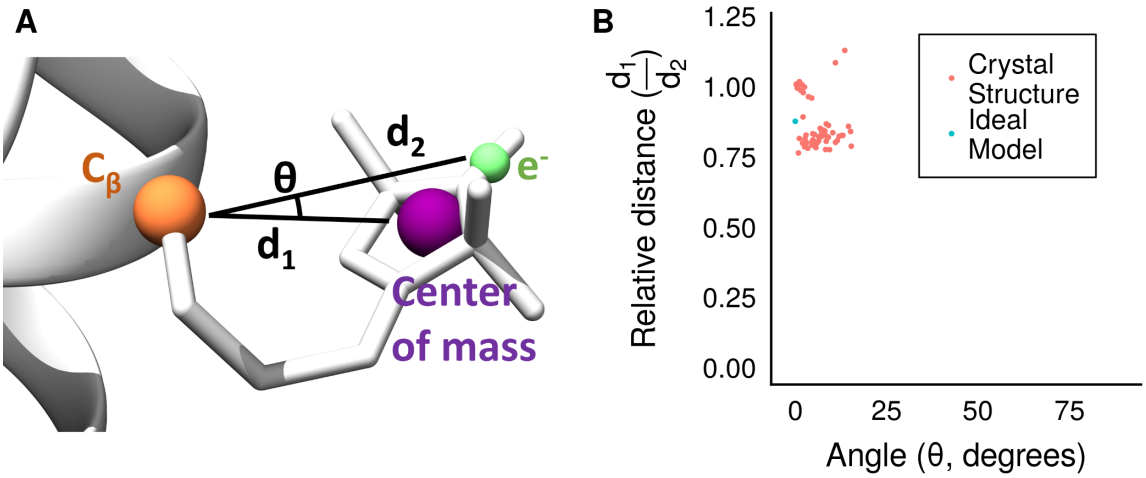

Figure E.3: Nitroxide centers of mass fall along the C_β-electron vector. A) Depiction of spin label from PDB: 2Q9D showing the nitroxide center of mass (purple) and the nitroxide bond midpoint (green). B). Angle and relative distance of nitroxide center of mass along the C_β-electron vector.

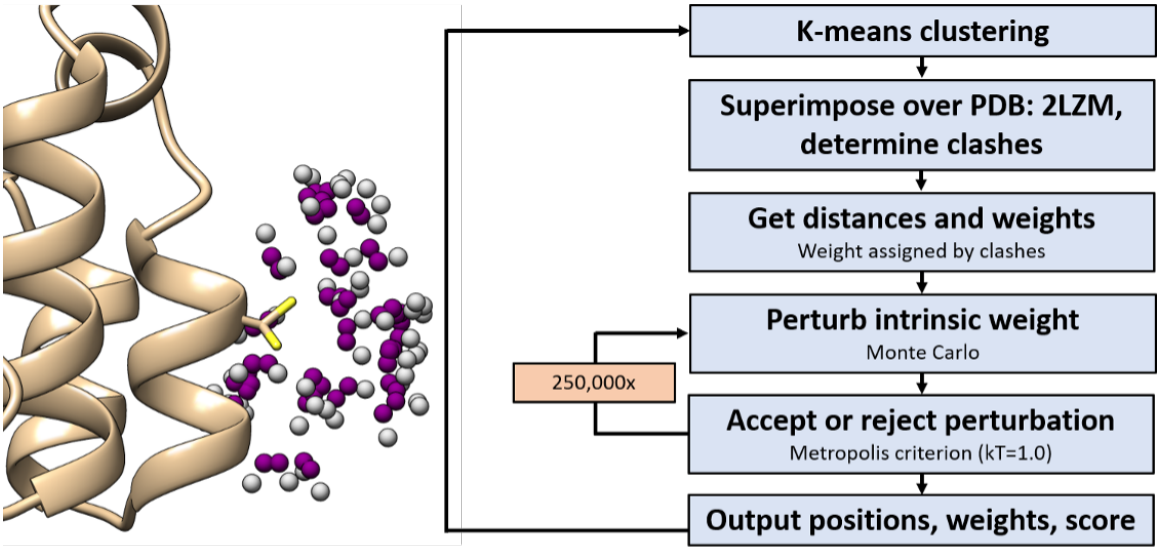

Figure E.4: Optimization of RosettaDEER measurement coordinates. Left: Each of the rotamers in Rosetta's MTSSL rotamer library was converted into two coordinates: one representing the nitroxide ring center of mass (purple), which was used to evaluate clashes; and one representing the nitroxide bond midpoint (silver), from which distances were measured. Shown over PDB 2CUU residue 131. Right: Optimization scheme for reducing the number of measurement coordinates using experimental T4 Lysozyme distance data. One thousand replicates were performed for each of N clusters, with N ranging from 3 to 53 coordinates.

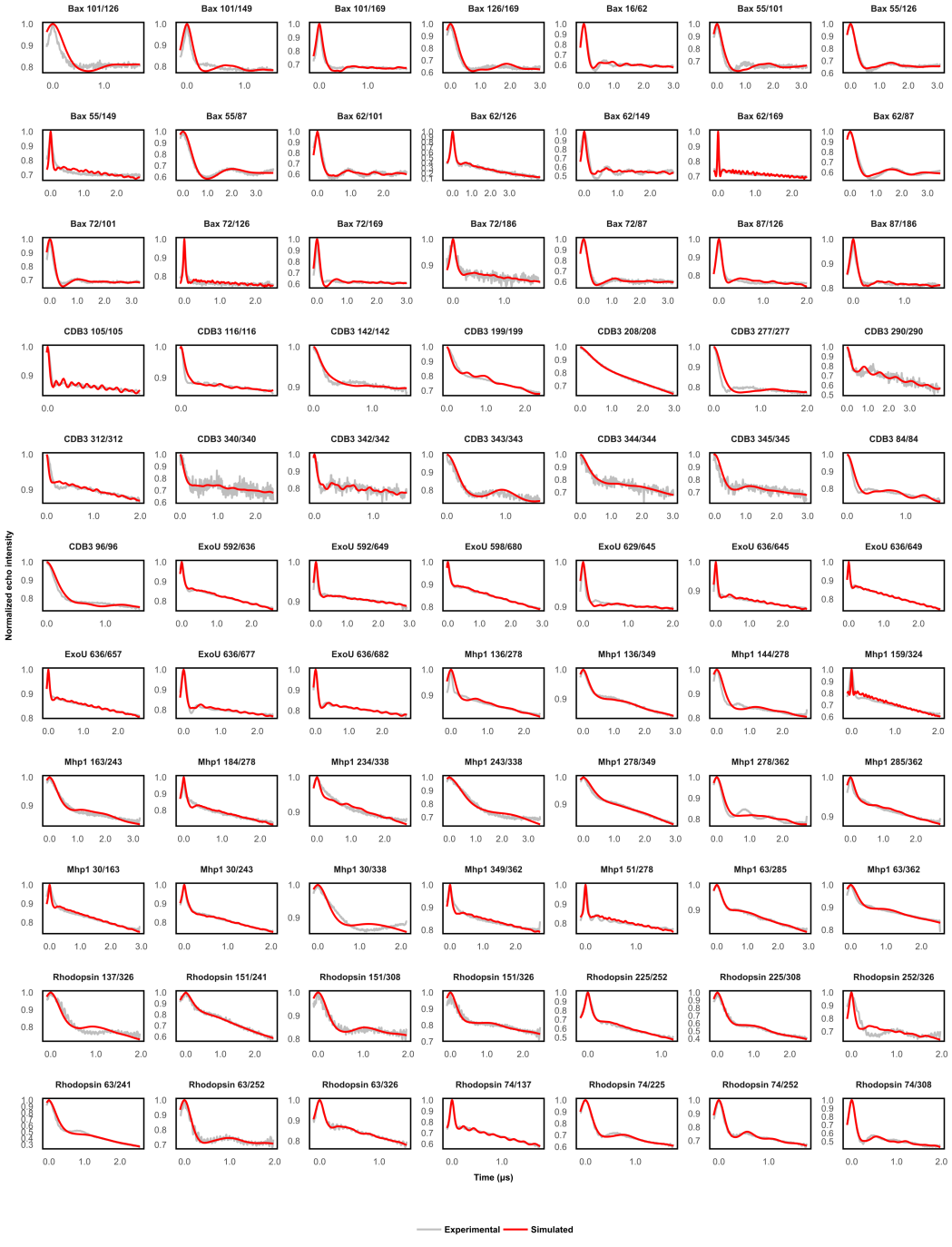

Figure E.5: All simulated and experimental DEER decay data used in this study between experimentally resolved residues. RosettaDEER could generally, but not always, simulate DEER traces from native-like models that are comparable to the experimental data.

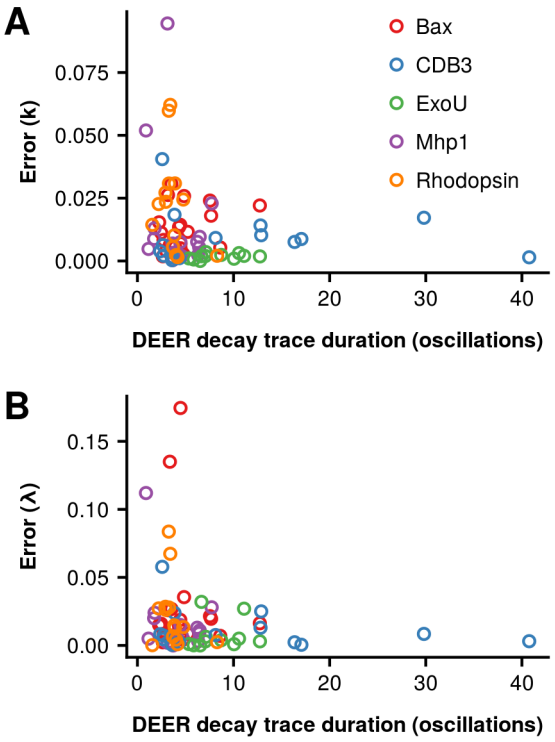

Figure E.6: Deviation between experimental and simulated background decay (k) and modulation depths (λ).

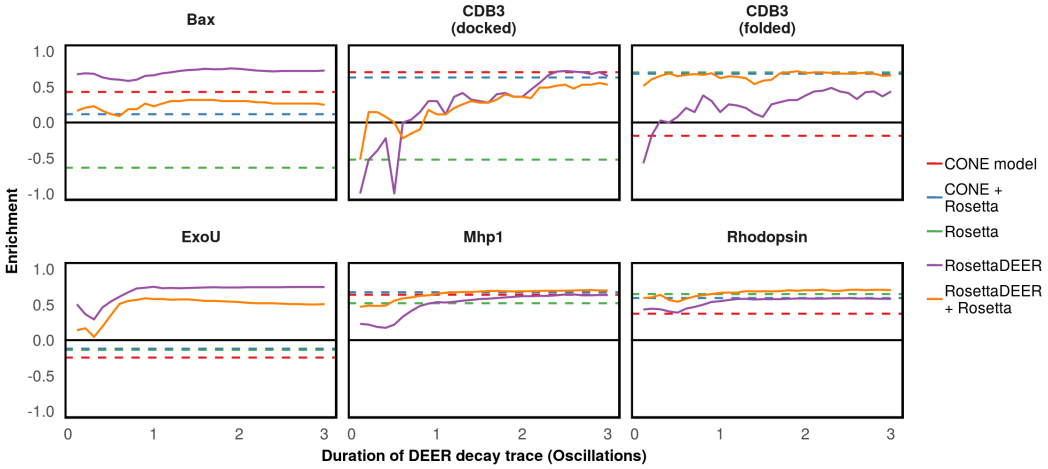

Figure E.7: Enrichment of misfolded and misdocked decoys as a function of DEER decay trace duration. Enrichment was quantified as the logarithm of the percentage of native-like models (top 10% by $\text{RMSD}_{100\text{SSE}}$) that were also in the top 10% by score. An enrichment of 1 indicates that the set of models constituting the top 10% by RosettaDEER score was identical to the set of models constituting the top 10% by Rosetta score.

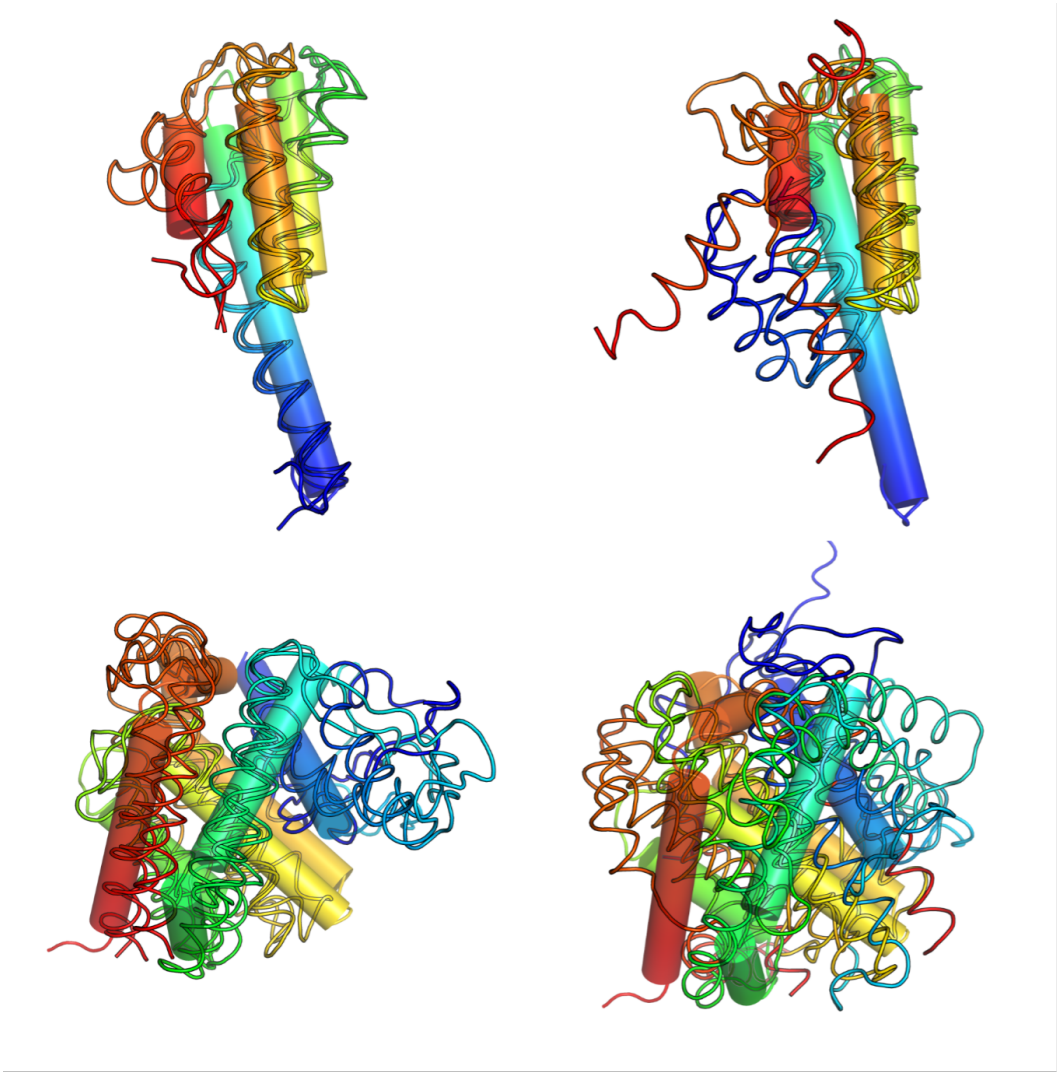

Figure E.8: Effect of DEER restraints on structure prediction of Bax and ExoU. Top 3 best-scoring models of ExoU (top) and Bax (bottom) folded either with (left) or without (right) experimental DEER restraints. The native models are shown as cylinders for comparison.

Appendix F

Supplement to "Methodology for rigorous modeling of protein conformational changes by Rosetta using DEER distance restraints"

This Appendix contains supplementary information for Chapter 4.

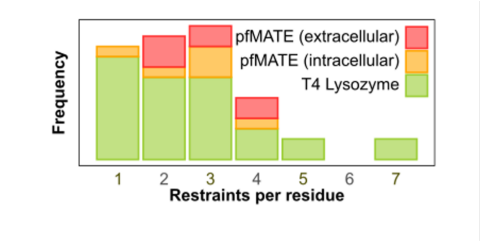

Figure F.1: Number of DEER restraints per spin-labeled residue across T4 Lysozyme and PfMATE.

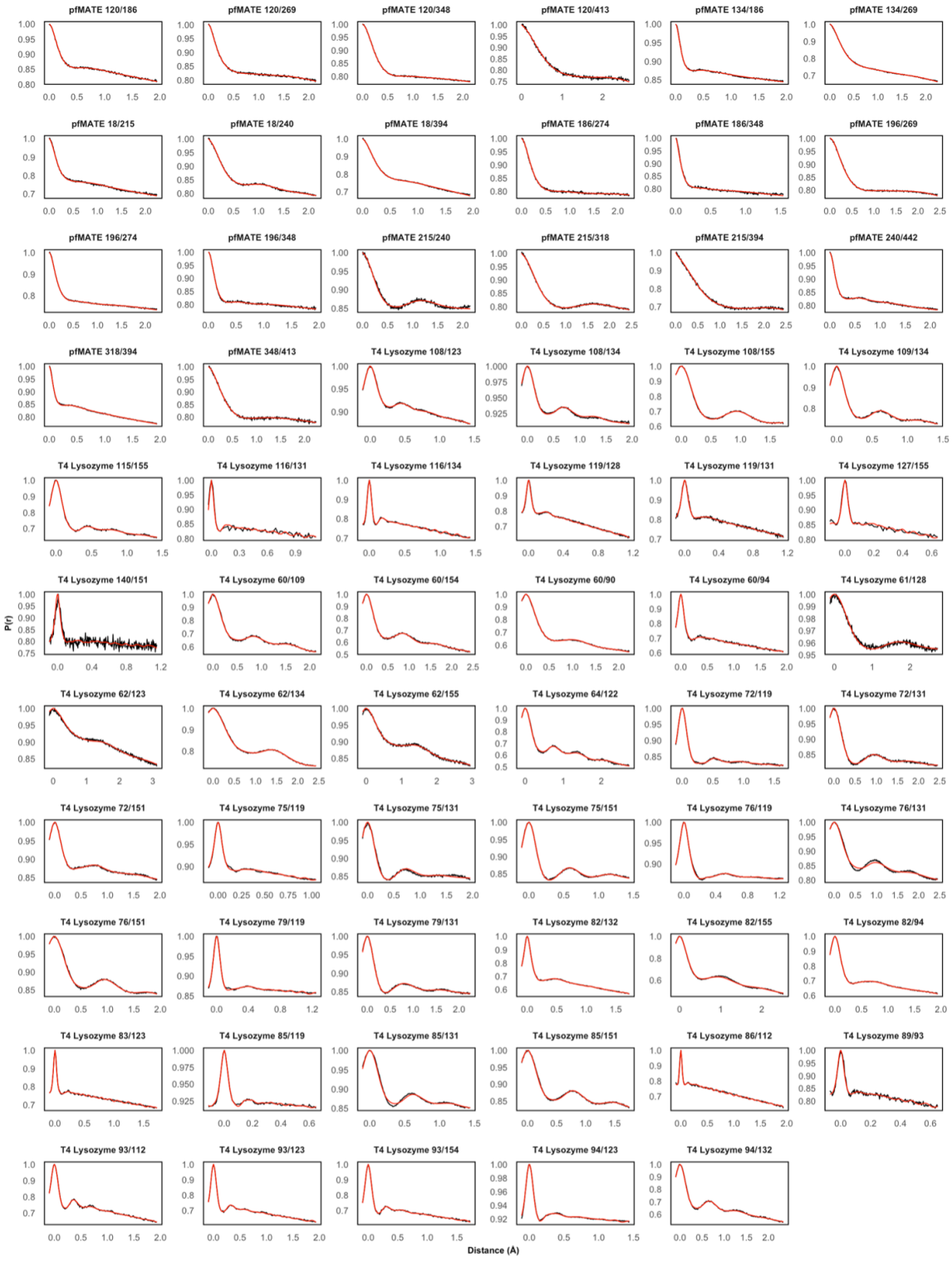

Figure F.2: All simulated and experimental DEER decay data used in this study between experimentally resolved residues. All DEER traces determined by multilateration are shown in red. Experimental DEER traces are shown in black.

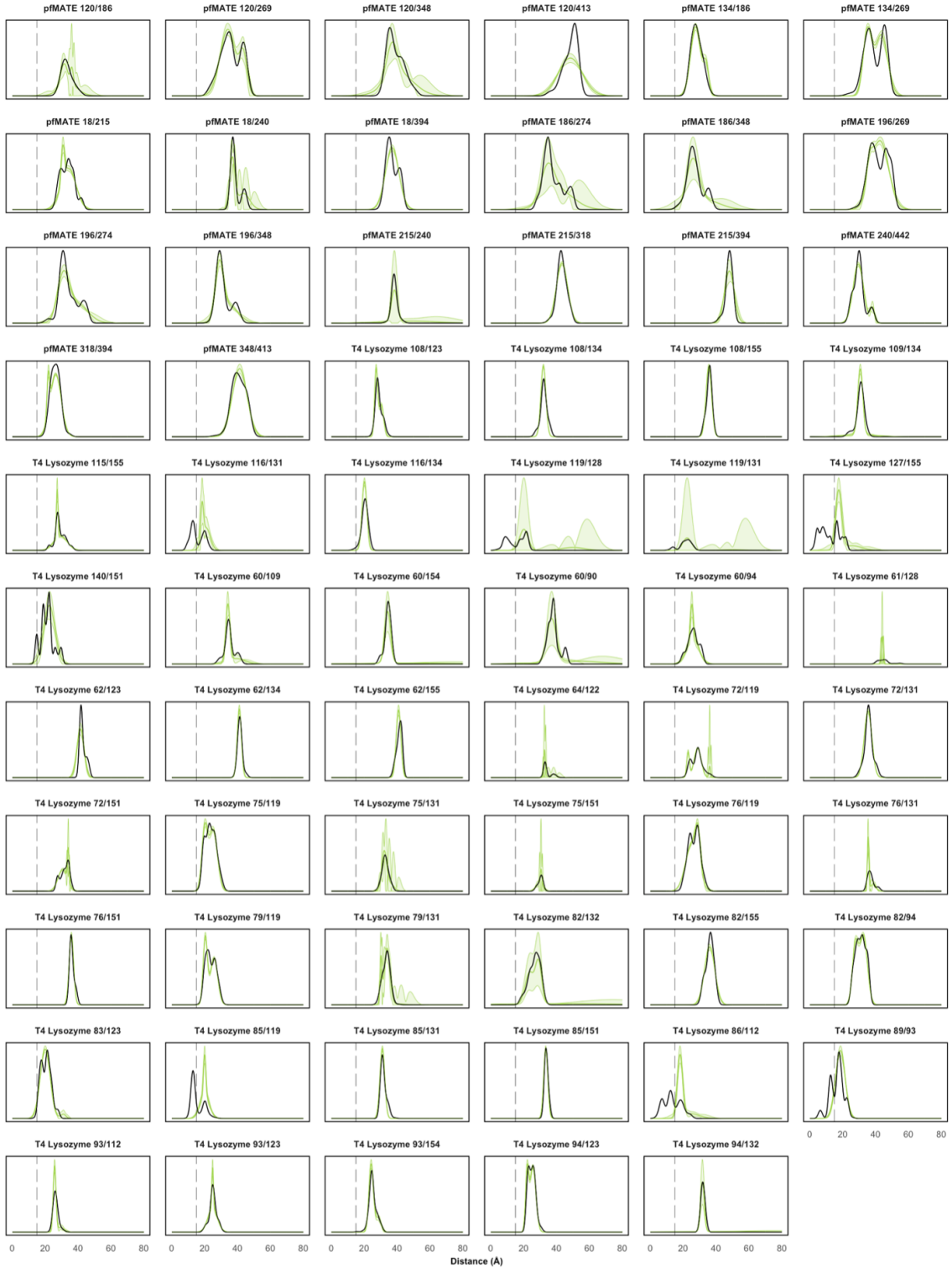

Figure F.3: Comparison of distributions obtained using GLADDvu and those using the RosettaDEER multilateration algorithm. All DEER distance distributions determined by multilateration are shown in black. DEER distributions calculated using GladdVU are shown in green, with the shaded regions indicating 95% confidence intervals. Distance values shorter than 15 Å (indicated by the dashed line) were not used to simulate DEER traces.

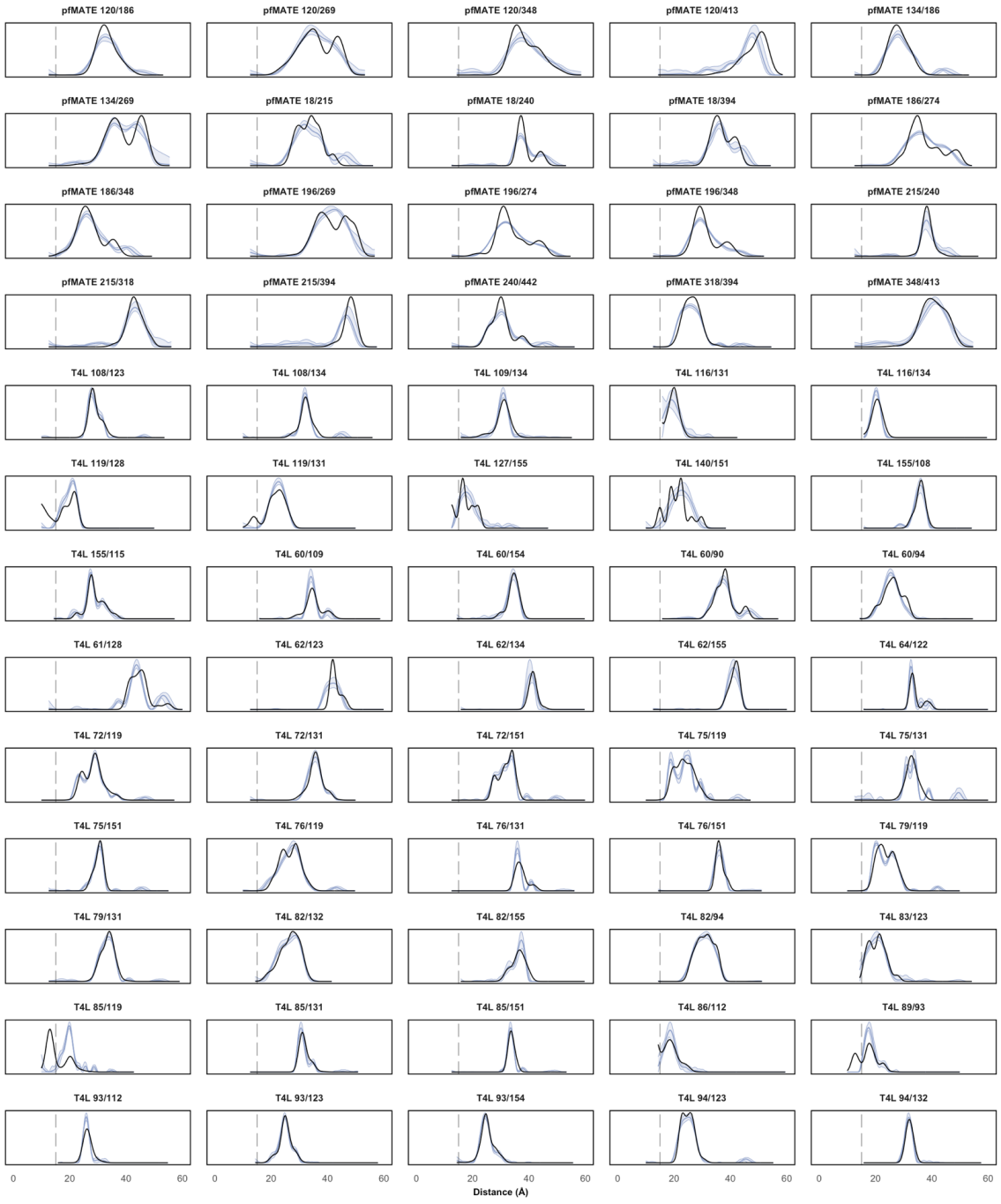

Figure F.4: Comparison of distributions obtained using DeerAnalysis and those using the RosettaDEER multilateration algorithm. All DEER distance distributions determined by multilateration are shown in black. DEER distributions calculated using DeerAnalysis are shown in green, with the shaded regions obtained using the validation tool.

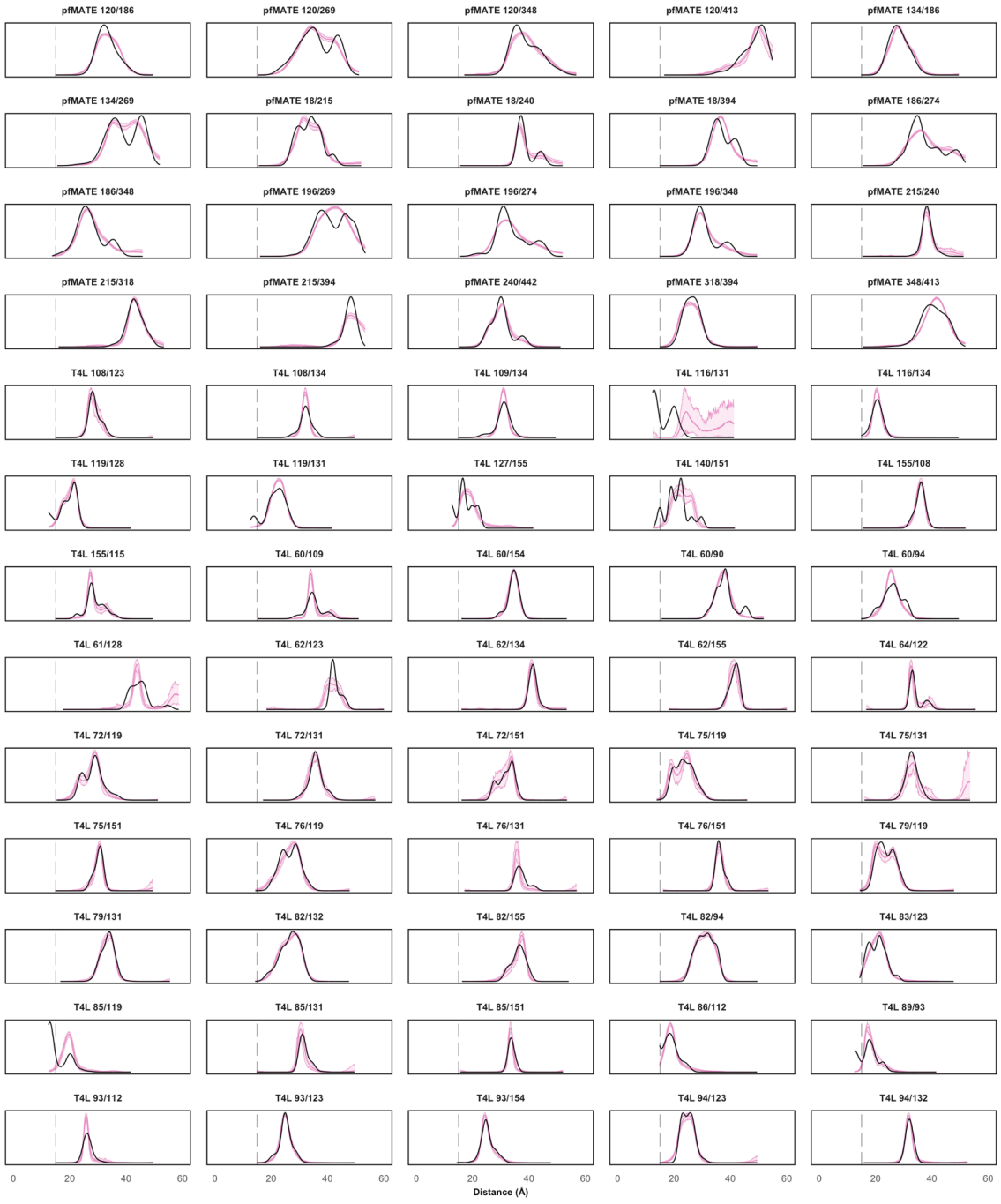

Figure F.5: Comparison of distributions obtained using DeerNet and those using the RosettaDEER multilateration algorithm. All DEER distance distributions determined by multilateration are shown in black. DEER distributions calculated using DeerNet are shown in pink, with the shaded regions obtained using ensemble statistics.

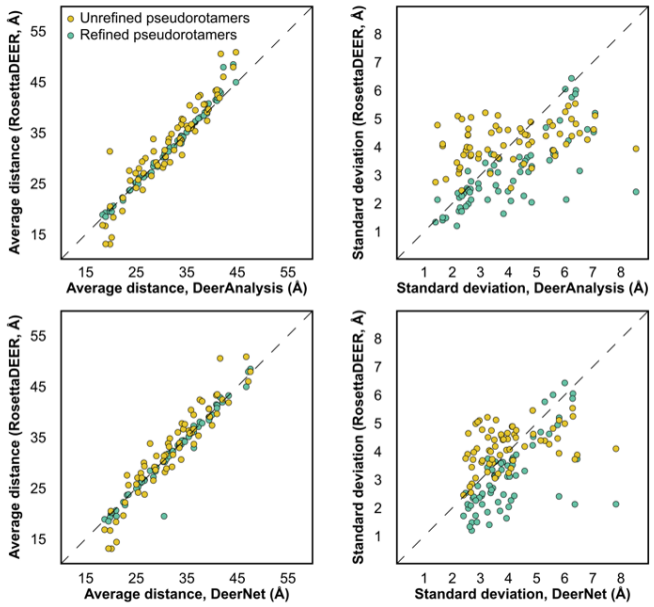

Figure F.6: Comparison of average and standard deviation values obtained when fitting DEER data collected in pfMATE and T4 Lysozyme to values obtained using DeerAnalysis and DeerNet. Long-distance fitting artifacts were removed from fits obtained using DeerAnalysis. These fits appeared to overstate the standard deviation values relative to GLADDvu, whereas those obtained using DeerNet appeared to be biased toward certain width values.

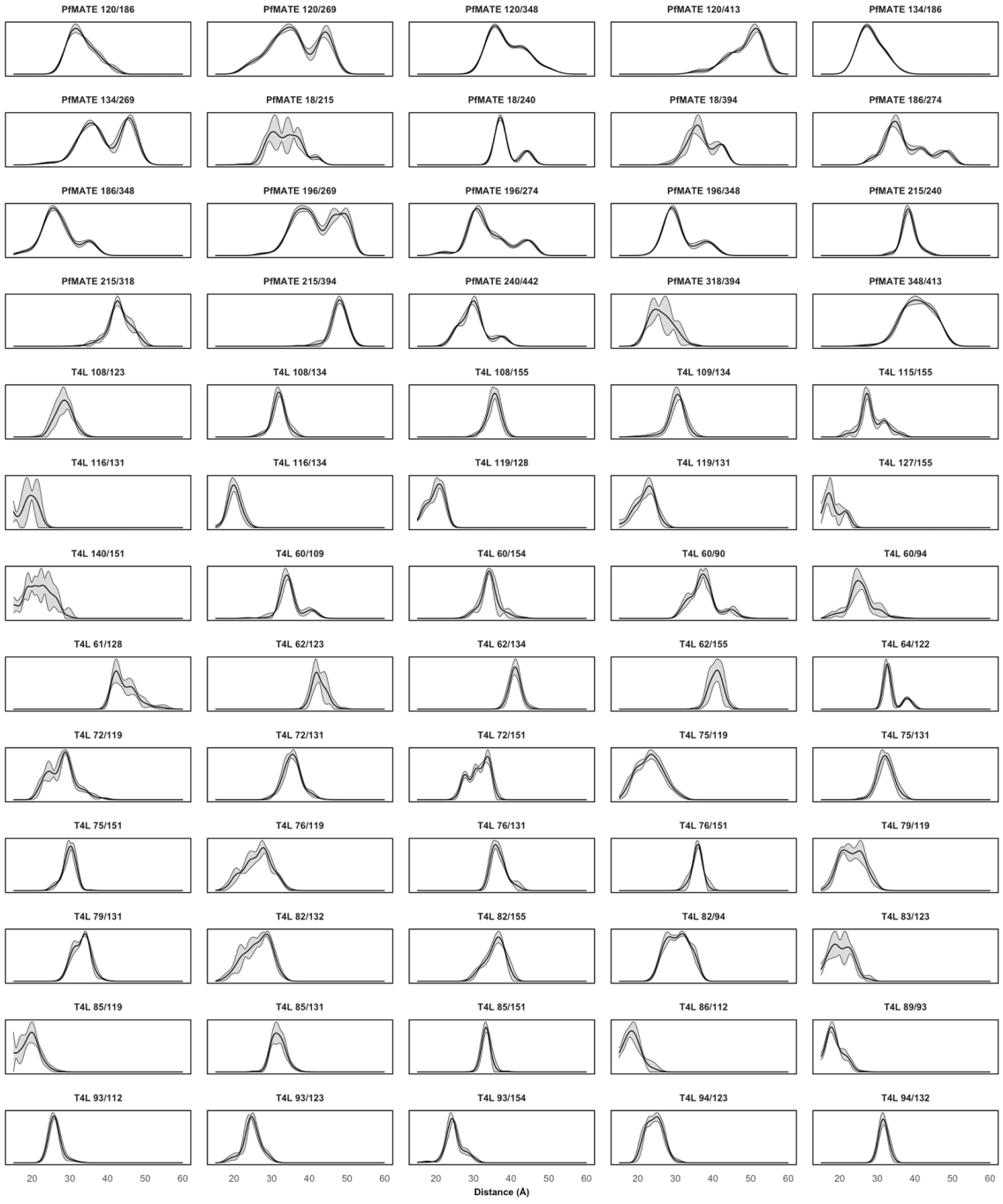

Figure F.7: Confidence analysis among the five best-scoring rotamer ensembles generated using the RosettaDEER multilateration algorithm. Shaded regions depict 95% confidence intervals, and line represents the mean distribution. Ensembles were selected using the AICc.

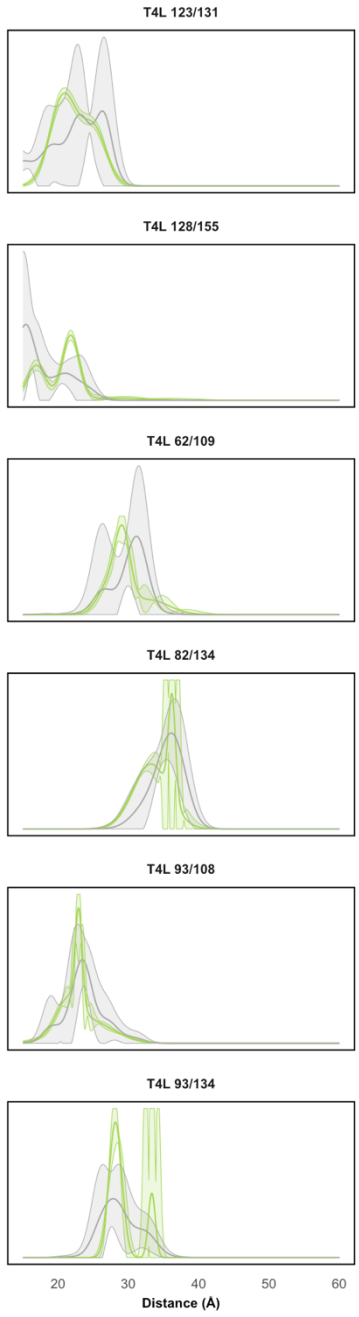

Figure F.8: Comparison of DEER distance distributions used to validate pseudo-rotamers obtained using the RosettaDEER multilateration algorithm. Distributions obtained using GLADDvu and RosettaDEER are shown in green and grey, respectively. Confidence bands for RosettaDEER depict the five best sets of pseudo-rotamers.

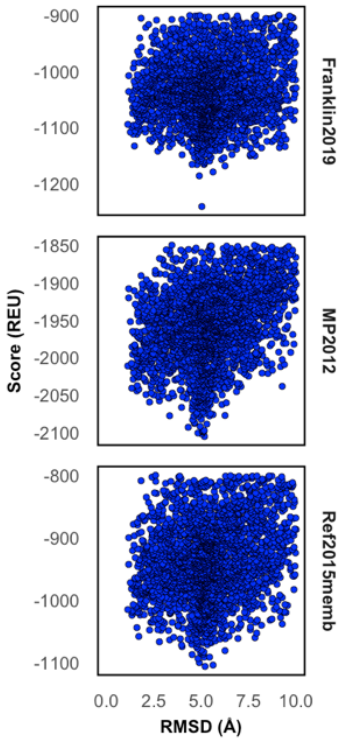

Figure F.9: Rosetta energy functions for membrane proteins cannot identify the inward-facing conformation of PfMATE. In all three cases, the lowest-energy models are fully occluded from both sides of the membrane. RMSD is measured from the inward-facing crystal structure (PDB: 6FHZ); the first 50 residues were omitted.

Appendix G

Supplement to ”pH-dependent structural dynamics of the glutamate-GABA antiporter GadC”

This Appendix contains supplementary information for Chapter 5.

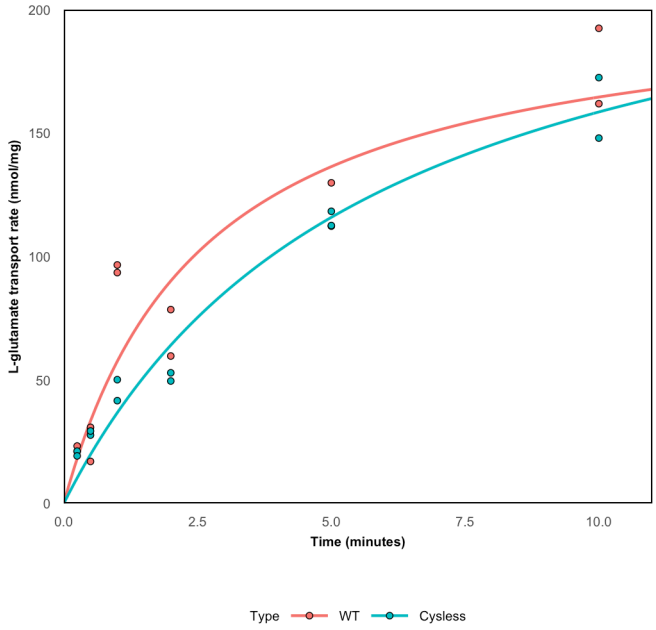

Figure G.1: Time-dependent glutamate transport by wildtype and cysless GadC reconstituted into proteoliposomes filled with 5 mM GABA.

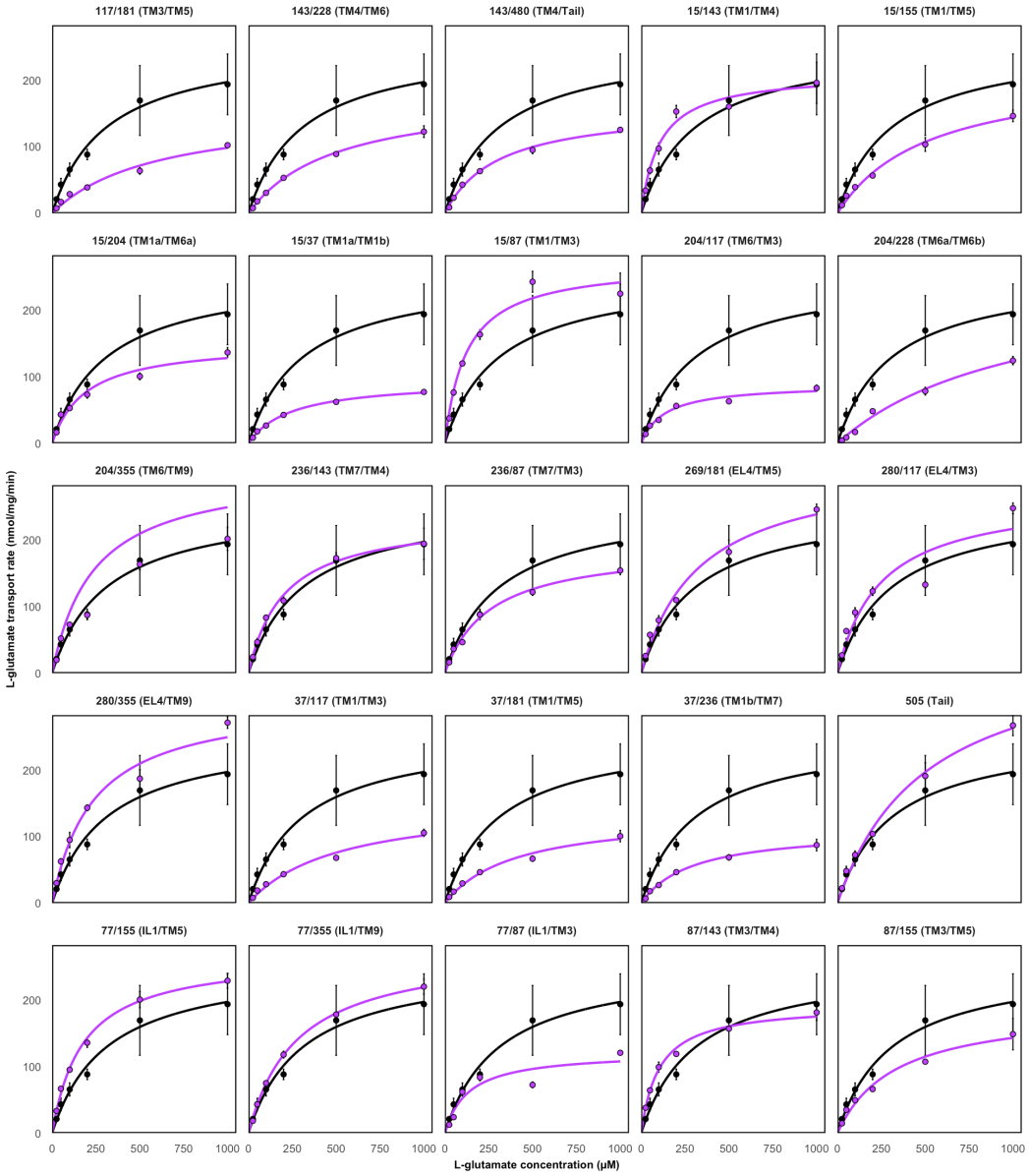

Figure G.2: Glutamate transport activity by GadC cysteine mutants. All experiments executed in triplicate and baseline-normalized. Wildtype transport rate is shown in black. Error bars show the standard error of the mean.

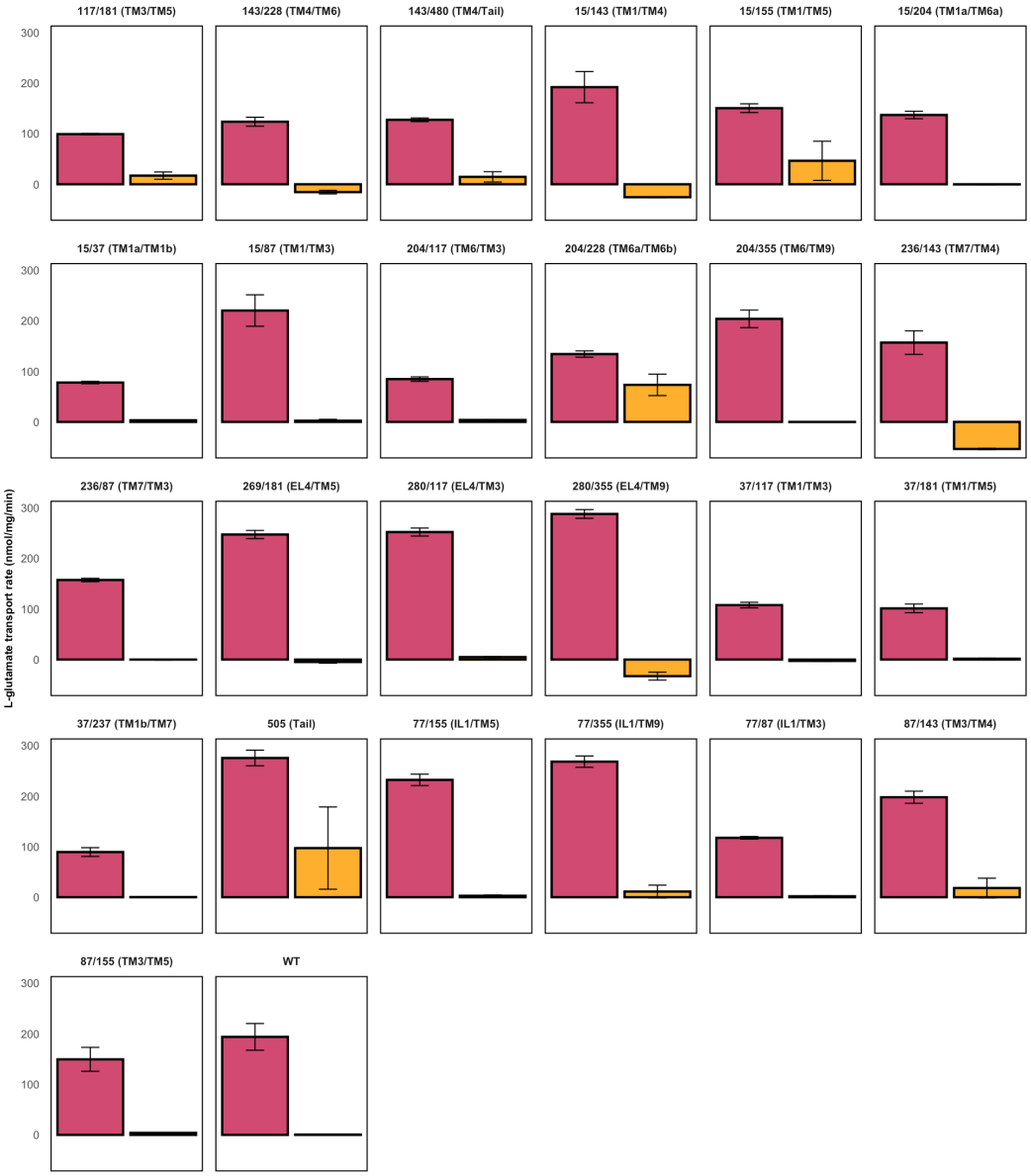

Figure G.3: pH-dependent inactivation of glutamate transport activity by GadC cysteine mutants. Transport activity at pH 4.5 and 7.5 are shown in mauve and orange, respectively. All experiments executed in triplicate and baseline-normalized.

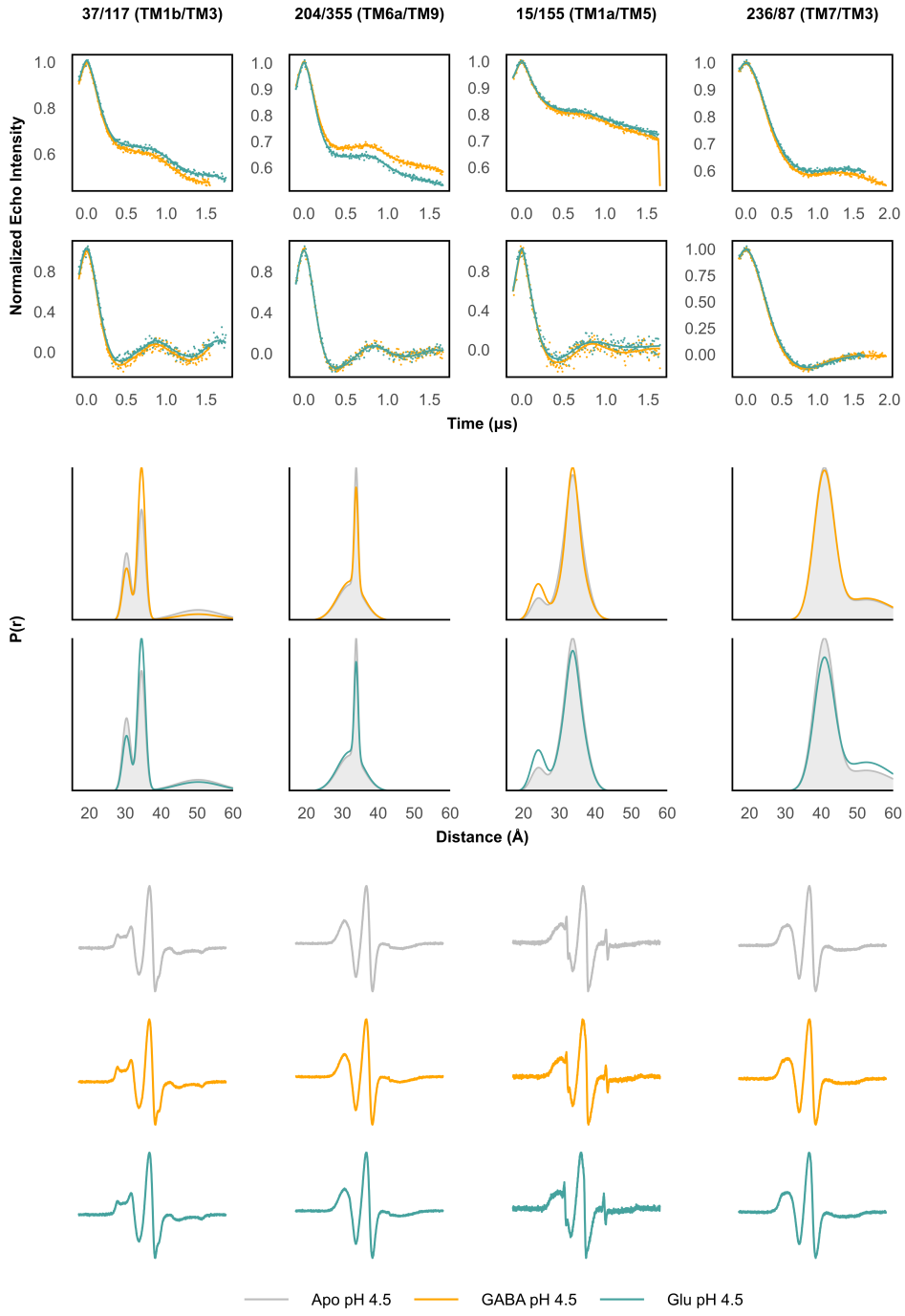

Figure G.4: Representative EPR pairs do not show evidence of large-scale substrate-dependent conformational changes in GadC. Top: DEER traces prior to and following background-correction. Middle: DEER distance distributions at pH 4.5 with 1 mM GABA (orange) or glutamate (teal). Apo distributions shown in grey. Bottom: continuous-wave EPR lineshapes with or without substrates.

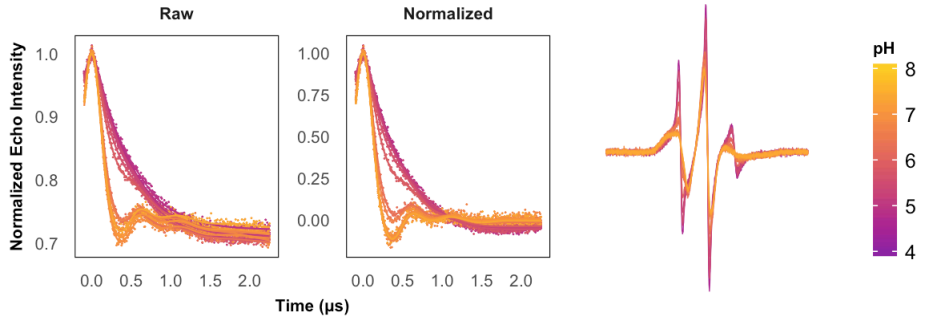

Figure G.5: pH-dependent DEER data and continuous-wave EPR spectra of GadC 143/480.

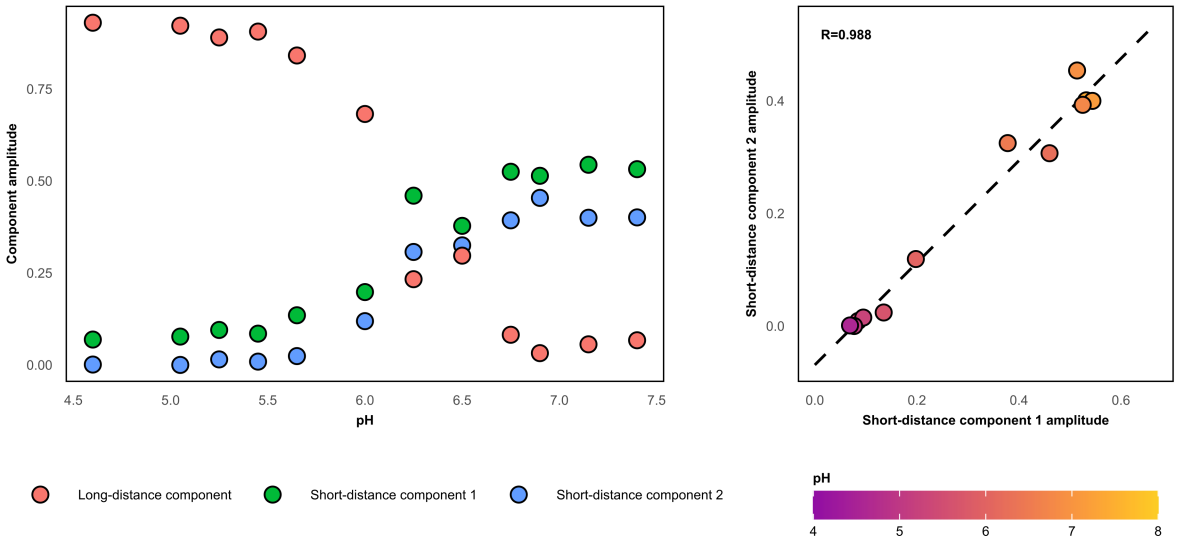

Figure G.6: Correlation between short-distance DEER components in spin pair 143/480 during detachment of the tail.

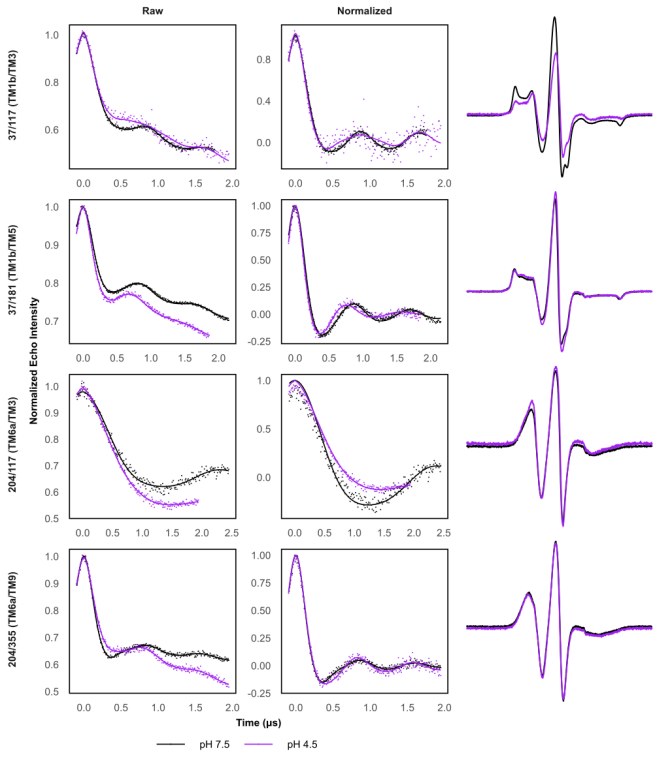

Figure G.7: DEER data and CW profiles of double-cysteine mutants labeled on both the bundle and scaffold domains on the extracellular side.

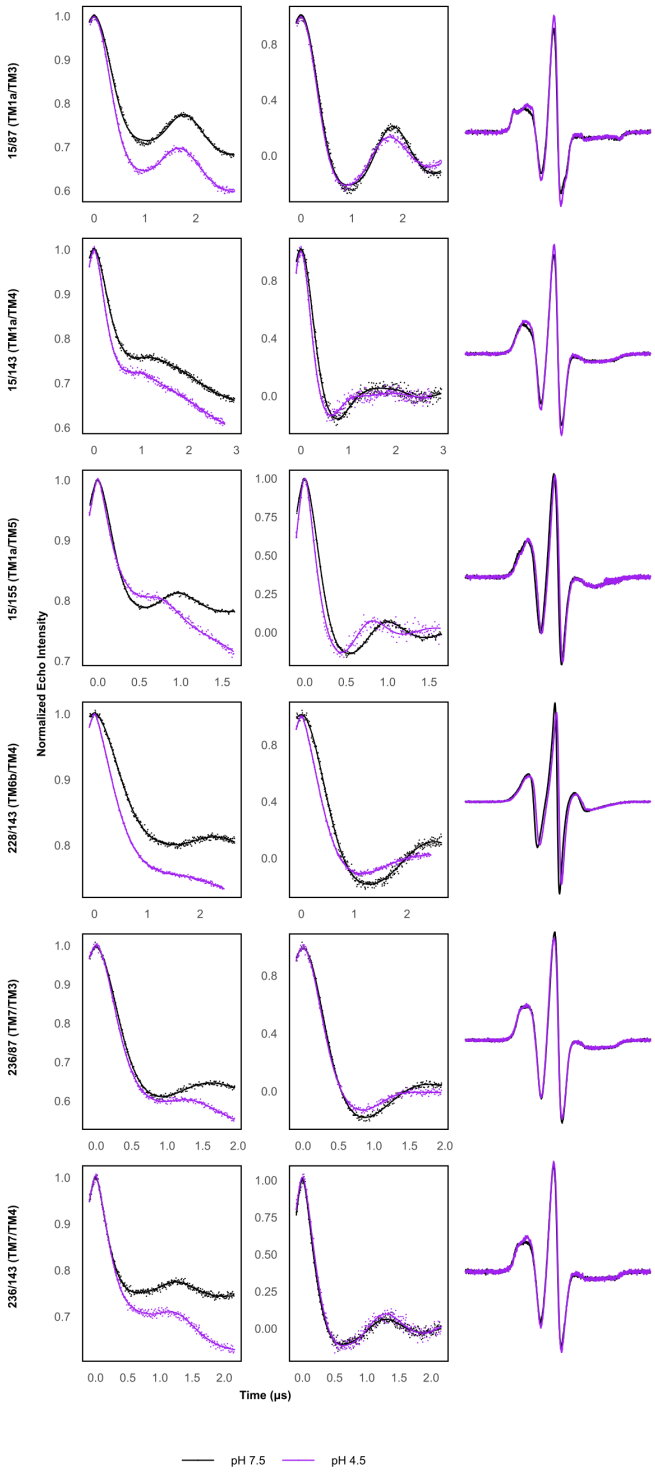

Figure G.8: DEER data and CW profiles of double-cysteine mutants labeled on both the bundle and scaffold domains on the intracellular side.

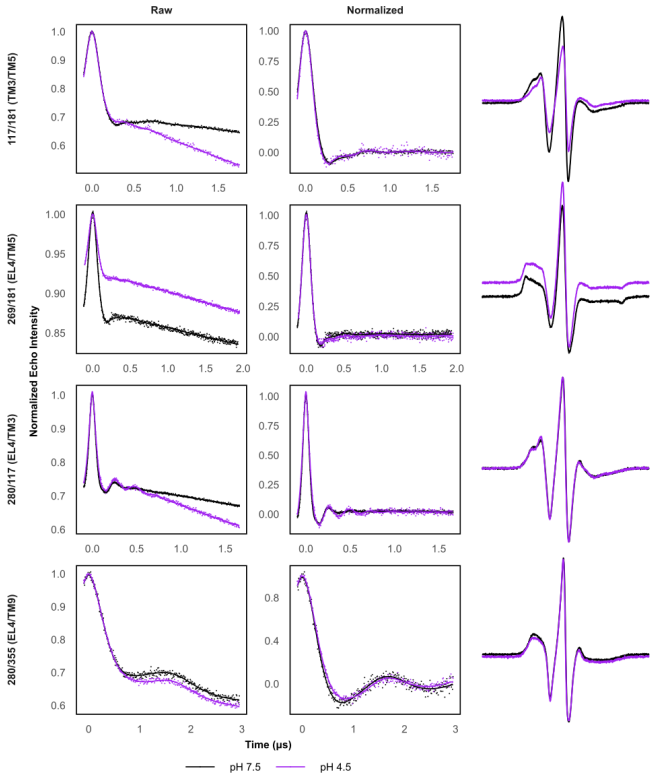

Figure G.9: DEER data and CW profiles of double-cysteine mutants labeled in EL4 and the scaffold domain on the extracellular side.

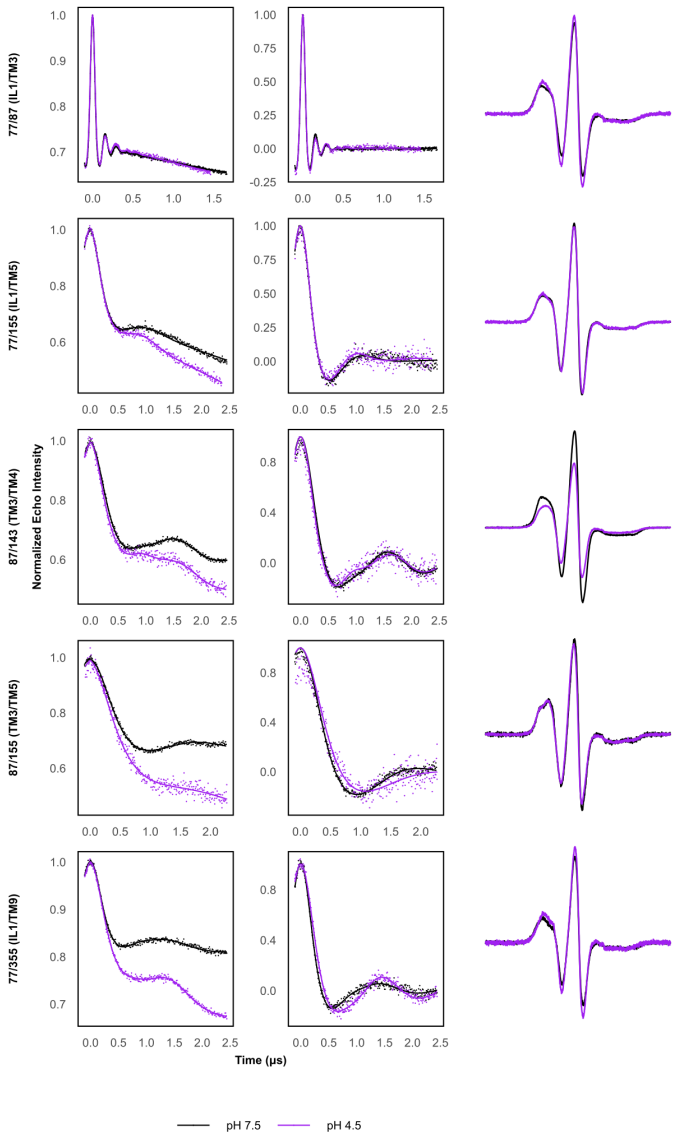

Figure G.10: DEER data and CW profiles of double-cysteine mutants labeled in IL1 and the scaffold domain on the extracellular side.

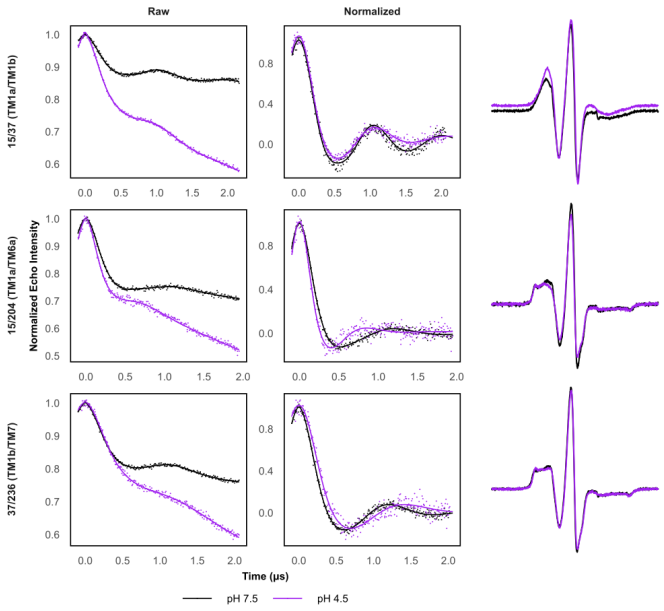

Figure G.11: DEER data and CW profiles of double-cysteine mutants labeled in the bundle domain.

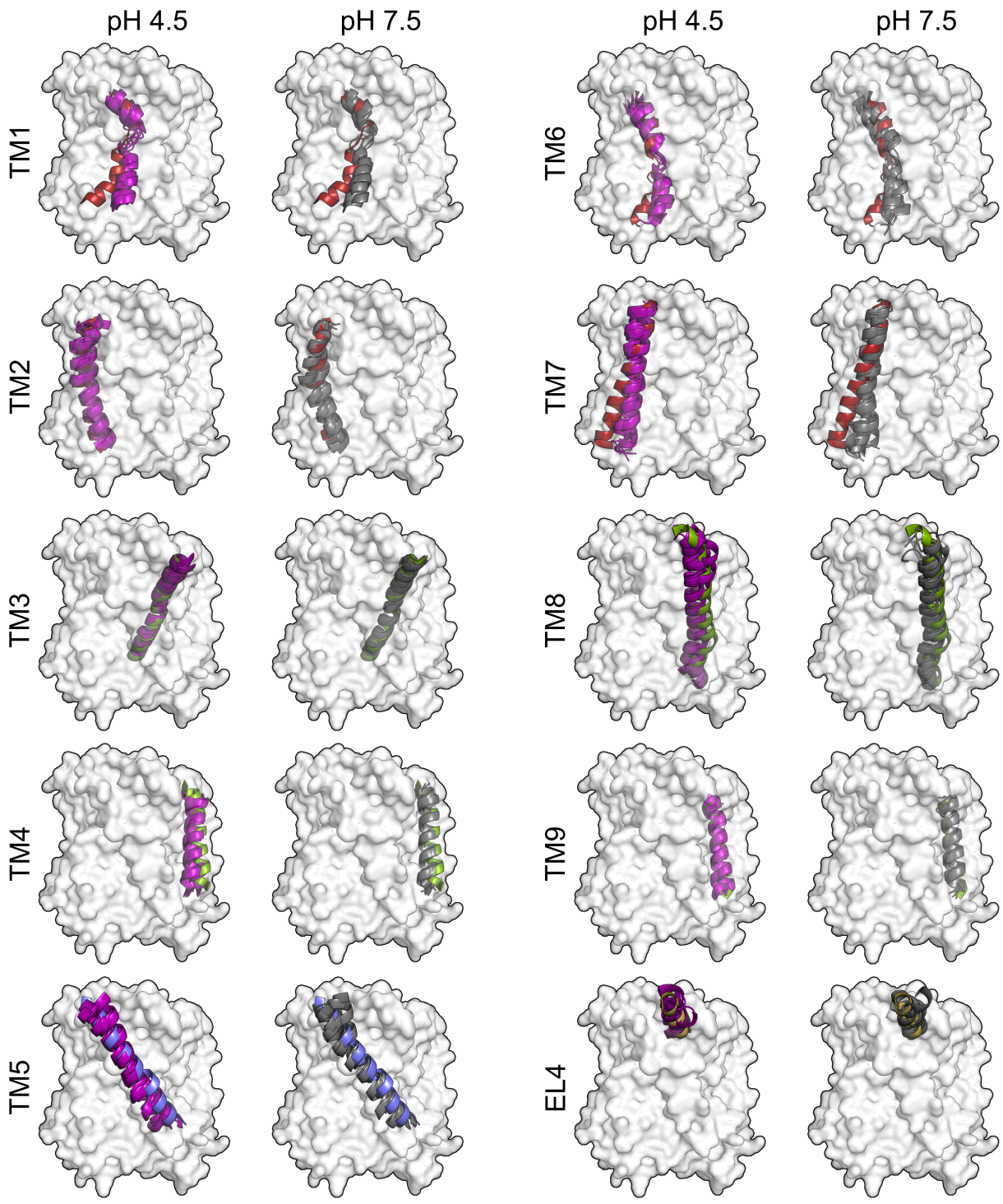

Figure G.12: Positions of the transmembrane helices among the five best low pH (pink) and high pH (gray) models relative to crystal structure (shown in red, green, blue, and yellow for bundle domain, hash domain, gating helices and EL4, respectively).

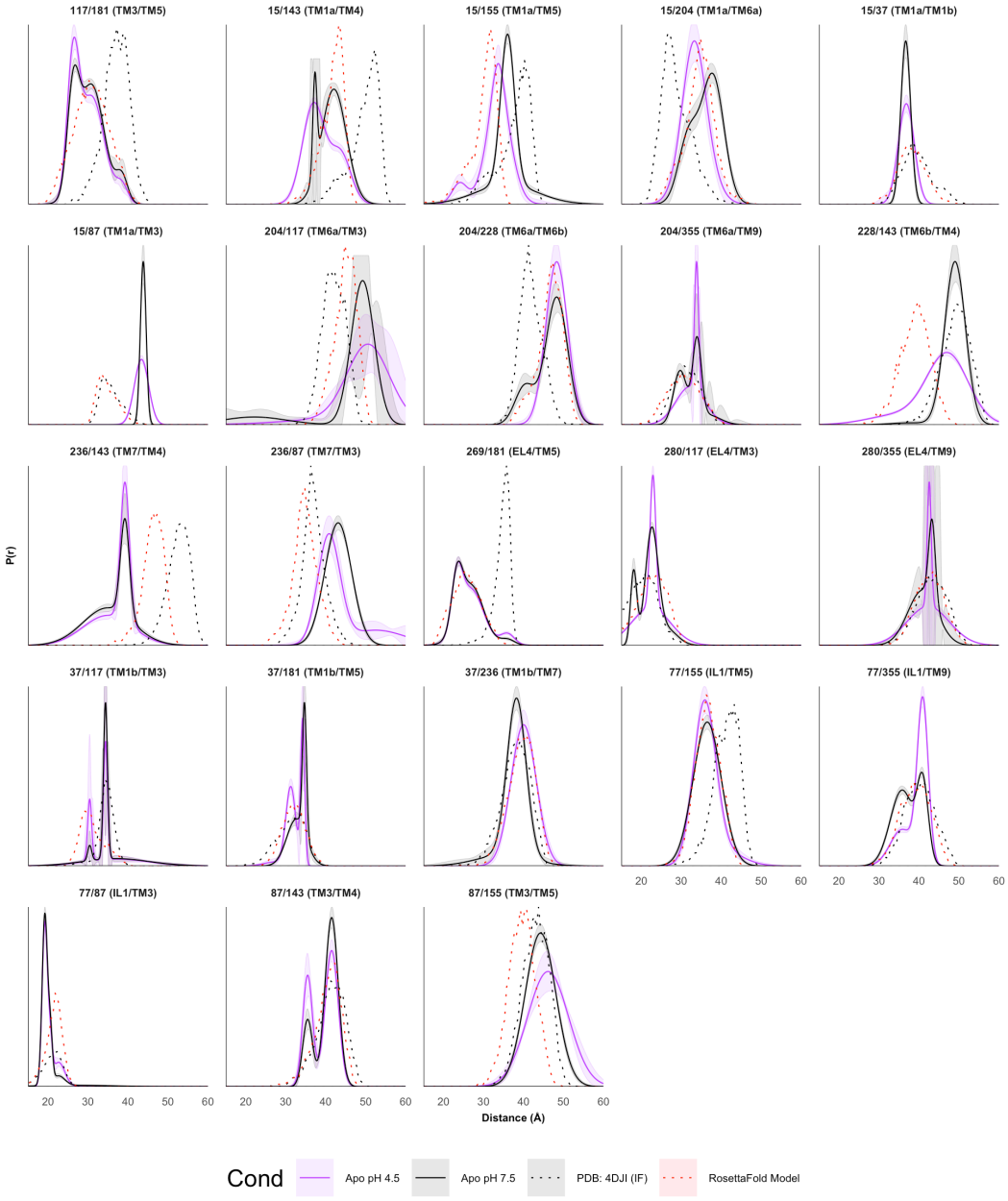

Figure G.13: Experimental DEER distance distributions measured in GadC and compared to a model generated *de novo* using RosettaFold.

Appendix H

Supplement to ”Efficient sampling of loop conformations using conformational hashing and random coordinate descent”

This Appendix contains supplementary information for Appendix D.

Table H.1: Protein structures used in the benchmark of Hash/RCD.

Protein	PDB ID	Length
Soluble proteins		
Ribosomal protein S15	1A32	88
Acanthamoeba castellani profilin IB	1ACF	125
A-spectrin SH3 domain (D48G)	1BK2	57
β-Spectrin (CH domain)	1BKR	109
TRP1/HSC70	1ELW	252
HSD (S46D)	1OPD	85
Streptococcal protein G	1PGX	83
Phage 434 repressor	1R69	69
XcR50	1TTZ	87
Ubiquitin	1UBI	76
Topoisomerase I	1VCC	77
Glia maturation γ-factor	1VKK	154
APE2540	1WDV	304
Acetylcholinesterase	2ACY	98
CheY	2CHF	128
YeeU	2H28	260
SLC9A3R2	2HE4	90
HigA	2ICP	94
Membrane proteins		
Aquaporin	1J4N	271
Mitochondrial ADP/ATP carrier	1OKC	297
Glycogen phosphorylase B	1PY6	498
V-type ATPase	2BL2	1560
GlpG	2IC8	182
DsbB	2K73	183
KdpD	2KSF	107
Rhodopsin II	2KSY	247
GlpG	2NR9	196
ApcT	3GIA	444
Riboflavin uptake protein	3P5N	378