Molecular mechanisms of drug transport through peptide transporters



Simon M. Lichtinger
St John's College

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Supervised by
Professor Philip C. Biggin, Professor Simon Newstead

Department of Biochemistry
University of Oxford

September 17, 2024

May all beings flourish

Declaration

I declare that this thesis is entirely my own work and, except where stated, describes my own research.

Simon M. Lichtinger
St John's College

Acknowledgements

I would like to foremost thank Professors Philip Biggin and Simon Newstead for their support and assistance as supervisors throughout my DPhil project, providing me with an ideal ratio of guidance and independence in driving the project forward in its details. The research presented in this thesis would also not have been possible without substantial training in MD methodology by Dr Zhiyi Wu, binding free energy calculations by Dr Irfan Alibay, python software engineering by Dr Rocco Meli, machine-learning methods by Dr Matthew Warren, wet-lab and cell-based techniques by Dr Joanne Parker and Dr Gabriel Kuteyi, and fluorescence microscopy by Sacha Salphati. I would further like to acknowledge frequent help with technical issues from Daniel Crusius, insightful conversations with all members of the structural bioinformatics and computational biochemistry (SBCB) group, particularly Robert Clark and Dr Jonathan Colburn, as well as advice on cloning and other experimental work from all members of the Newstead group. I would like to acknowledge the Biochemistry Department support staff for maintaining the research infrastructure used in this project, particularly Borut Rozman and John Elder at the IT office. For administrative help in navigating the DPhil I would like to thank the Chemistry in Cells doctoral programme and St John's college, in particular Dr Tracey Marr as programme administrator and Dr Angela Russell doubling as programme director and college advisor. I thank the Wellcome Trust (Grant ID: 218514/Z/19/Z) for funding this DPhil, as well as the High-End Computing Consortium for Biomolecular Simulation (HECBioSim), supported by EPSRC (EP/X035603/1), for compute resources on the EPSRC ARCHER2, Jade 2 and N8 CIR BEDE facilitites. I would further like to thank St John's College for financial support towards conference travel provided via special grants.

The experience in SBCB lives as much from its culture as its academic pursuits. To this end, I would like to express my gratitude towards Dr Michael Horrell for lightening the lockdown winter during which I joined, David Seiferth for his roles in the SBCB Christmas band as well as in the regular SBCB gym crew together with Cyril Schroeder, Mingyu Qin, Sacha Salphati, Nitishwer Anand, Dr Iain Smith, Dr Dheeraj Prakaash, Dr Matthew Warren and Clara Pedersen, and to Robert Clark, Dr Iain Smith and Carl von Hallerstein for essential discussions on the structure of society and the nature of consciousness. Beyond my research group, the DPhil experience was inextricably linked to the connections I have been privileged to make throughout my time in Oxford. Giving an exhaustive list of this network is by all means impossible; to name but a few, I would like to thank the lunch swap community with Matthew Harris and Arundhati Wuppalapati for keeping me well fed, Dr John Colley for numerous wonderful soup-making experiences, Anna Redly for the deeply appreciated support through the first year of my DPhil in lockdown, Gayatri Dangi for the joyful times spent together, the Oxford Triratna Buddhist community, particularly Lottie Oliver and Prajnaketu, for conversations about all matters spiritual, the 'trio' with Andrés Vázquez and Misha Mir for all the music and late nights, as well as Dan Şeremet and Sabina Fîrțală for the continued Romanian presence in my life.

Lastly, I would like to thank my parents Verena and Peter and my siblings Hannah, Jan-Niklas and Malin for their love and support given unconditionally thoughout this life.

Abstract

Proton-coupled oligopeptide transporters (POTs) — members of the major facilitator superfamily (MFS) with homology across the tree of life — couple the intestinal uptake of short peptides to the downhill symport of protons in an alternating access mechanism. They are of great pharmaceutical interest owing to their substrate promiscuity and role in the oral bioavailability of several classes of approved drugs, including β-lactam antibiotics. The molecular mechanism by which they couple proton transfer, substrate binding and large-scale conformational changes have remained elusive, however, and structure-guided or computer-aided drug design for POT-mediated uptake has not yet been accomplished.

Recent cryo-EM structures of mammalian POTs provide a unique opportunity to investigate them using molecular dynamics (MD) simulations, yet MD is beset by sampling problems due to limited accessible time scales. Several classes of enhanced sampling algorithms are available, however they frequently suffer from a bias towards the simulation starting state, a phenomenon known as hysteresis. We therefore first devised a strategy to address this shortcoming which we term MEMENTO (Morphing Endstates by Modelling Ensembles with iNdependent TOpologies), deriving history-free paths for umbrella sampling by using template-based modelling to fix unphysical intermediates from coordinate interpolation. We have developed this into an open-source python package, and validated the approach on four test systems of increasing complexity.

We then applied our method — together with free-energy and constant-pH simulation techniques — to derive a sequence of protonation steps that drive alternating access in POTs, validating our predictions in cell-based transport assays. In particular, we demonstrate the function of an extracellular gate that is only conserved in mammalian POTs, while showing that universally conserved glutamate residues act as intracellular gate-opening triggers. Sequential trans-membrane movement of protons via these residues explains the uni-directional transport exhibited by POTs. We also examined the role played by the presence of substrate in the binding pocket, and suggest that it both exerts a subtle conformational bias towards inwards-facing protein conformations and enables proton translocation by perturbing the pKa value of a key glutamate side chain via a salt-bridge swap mechanism. These results are corroborated by a series of new ligand-bound cryo-EM structures, where we showed using further MD simulations that the ability to engage a conserved arginine residue that triggers the salt-bridge swap distinguishes between good and bad substrates, not substrate affinity to one protein conformation alone.

We lastly benchmarked docking, absolute binding free energy (ABFE) calculations and machine learning models trained using ligand-only features on a retrospective dataset of 184 transport inhibition constants at PepT2. We show how the prediction of accurate protein-ligand binding poses constitutes a crucial limitation for the structure-based methods, and that the ligand-based machine learning models, while giving much better correlations to experimental data, fall short in their ability to capture pharmaceutically important activity cliffs. We have also started work on using these methods prospectively to design prodrugs for transport by POTs, though our efforts thus far have been hampered by solubility problems.

In summary, this thesis covers ground from the development of novel computational tools to their application to the POT transport mechanism, and reaching further towards *in-silico* design of prodrugs for increased bioavailability. It therefore advances the fields of enhanced sampling of conformational changes and membrane transporter biology, with the potential of benefitting their applications to pharmacology in the future.

List of publications

The following academic papers were published as part of the research carried out for this thesis.

- Lichtinger, S. M.; Biggin, P. C. Tackling Hysteresis in Conformational Sampling: How to Be Forgetful with MEMENTO. J. Chem. Theory Comput. **2023**, 19 (12), 3705–3720. https://doi.org/10.1021/acs.jctc.3c00140.
- Lichtinger, S. M.; Parker, J. L.; Newstead, S.; Biggin, P. C. The Mechanism of Mammalian Proton-Coupled Peptide Transporters. eLife, **2024**. https://doi.org/10.7554/elife.96507.2. (reviewed preprint with accepted revisions, version of record pending)
- Parker, J. L.†; Deme, J. C.†; Lichtinger, S. M.†; Kuteyi, G.; Biggin, P. C.; Lea, S. M.; Newstead, S. Structural Basis for Antibiotic Transport and Inhibition in PepT2, the Mammalian Proton-Coupled Peptide Transporter. Research Square, **2024**. https://doi.org/10.21203/rs.3.rs-4435259/v1. (under review)

† equal contribution

A review of enhanced sampling methodology applied to membrane transporters is in preparation in collaboration with Jonathan Colburn and Philip Biggin, and publication is planned as a contribution to a Springer Nature book project in memory of Sergei Noskov. The review may build on portions of the introduction and methods chapters of this thesis. The results presented in chapter 6 are also envisaged to contribute towards a future publication, combined with further data.

Table of contents

Table of contents xiii

List of figures xvi

List of tables xviii

Glossary xix

1 Introduction 1
- 1.1 Peptide transporter biology 1
  - 1.1.1 Alternating-access membrane transport 1
  - 1.1.2 The SLC15 family of proton-coupled peptide transporters 3
  - 1.1.3 Structural and biochemical characterisation of peptide transporters 3
  - 1.1.4 Substrates and the peptide transporter binding site 6
- 1.2 Oral drug bioavailability via peptide transporters 10
- 1.3 Review of peptide transporter simulation studies 12
- 1.4 Project aims and outline of chapters 19

2 Computational Methods 21
- 2.1 Molecular dynamics simulations 21
  - 2.1.1 Theoretical foundations 22
  - 2.1.2 Algorithms 26
  - 2.1.3 Force fields 31
- 2.2 Enhanced sampling molecular dynamics 36
  - 2.2.1 Classification of enhanced sampling methods 37
  - 2.2.2 Collective variable design 39
  - 2.2.3 Driven methods — steered MD 41
  - 2.2.4 Localisation and static bias methods — umbrella sampling 41
  - 2.2.5 Adaptive bias methods — metadynamics 42
  - 2.2.6 Replica exchange 43
- 2.3 Binding free energy calculations 44
  - 2.3.1 Alchemical thermodynamic cycles 44
  - 2.3.2 Sampling the thermodynamic legs 46
  - 2.3.3 Restraints in binding free energy simulations 46
  - 2.3.4 Treatment of charged ligands 47
  - 2.3.5 Further methodological considerations 47
- 2.4 Free energy estimators 49
- 2.5 Constant pH MD 51

3 Tackling hysteresis in conformational sampling — how to be forgetful with MEMENTO 53
- 3.1 The challenge of hysteresis in sampling conformational transitions 53
- 3.2 The MEMENTO method 56
  - 3.2.1 Outline of the basic MEMENTO procedure 56
  - 3.2.2 Adding ligands 58
  - 3.2.3 Adding lipids 59
  - 3.2.4 End-state sampling, replicates and two-dimensional CVs 59
- 3.3 Results on deca-alanine 61
- 3.4 Results on ADK 64
- 3.5 Results on P38α 71
- 3.6 Results on LeuT 76
- 3.7 Discussion 83

4 The alternating-access mechanism of mammalian peptide transporters 84

4.1 Extra- and intra-cellular gate opening triggers in unbiased MD ... 86
4.2 Protonation-state dependent free energy landscapes of PepT2 alternating access ... 94
4.3 Several mechanisms for substrate coupling ... 105
4.3.1 Protonation state-dependent ligand release ... 105
4.3.2 Substrate-modulated alternating-access thermodynamics ... 106
4.3.3 Coupling between substrate binding and ExxER motif protonation ... 106
4.4 Validation of extracellular gating residues in cell-based transport assays ... 112
4.5 Discussion ... 114

5 Substrates and inhibitors — drug dynamics in the transporter binding pocket 118
5.1 Inhibition and transport assays of β-lactams ... 120
5.2 Cryo-EM structures of PepT2 bound to β-lactams ... 120
5.3 MD simulations from β-lactam cryo-EM poses ... 122
5.3.1 Coupling between ExxER motif protonation and drug binding pose ... 122
5.3.2 Binding pose flexibility of the PepT2 inhibitor cloxacillin ... 122
5.4 Discussion ... 127

6 Retrospective and predictive modelling in peptide transporter prodrug design 129
6.1 Two methodological choices ... 130
6.1.1 Force fields for modelling substrates ... 130
6.1.2 Multivalent co-ions for alchemical free energy calculations ... 131
6.2 Retrospective affinity prediction ... 132
6.2.1 Molecular docking ... 132
6.2.2 Absolute binding free energy calculations ... 133
6.2.3 Machine learning affinity prediction ... 134
6.2.4 Limitations of retrospective affinity predictions ... 138
6.3 Prospective application to drug design ... 146
6.4 Conclusion and avenues for future exploration ... 148

7 Conclusions and outlook 150
7.1 Summary of results ... 150
7.2 Outlook ... 151

Appendix A Detailed methods for chapter 3 153
A.1 Deca-alanine ... 153
A.2 ADK ... 154
A.3 P38α ... 155
A.4 LeuT ... 156

Appendix B Detailed methods for chapter 4 157
B.1 Data availability ... 157
B.2 Definition of tip and base bundle CVs ... 157
B.3 MD setup and equilibration ... 157
B.4 Derivation of OF, OCC and IF conformational states ... 158
B.5 Unbiased MD of the OCC state ... 158
B.6 Metadynamics and steered MD ... 159
B.7 MEMENTO path generation ... 159
B.8 1D PMF calculations ... 159
B.9 2D CV derivation ... 160
B.10 2D PMF calculations ... 160
B.11 Absolute binding free energies ... 161
B.12 Constant-pH MD ... 162
B.13 Cell-based transport assays ... 162
B.14 Protein expression controls ... 162
B.15 Protein localisation controls ... 163

Appendix C Detailed methods for chapter 5 164
C.1 Unbiased MD ... 164

C.2 ABFE simulations 164

Appendix D Detailed methods for chapter 6 **165**
D.1 Reference dataset preparation 165
D.2 Ensemble docking 165
D.3 ABFE calculations 165
D.4 Random forest models 166

Bibliography **167**

List of figures

1.1 Schematic overview of alternating access mechanisms 2
1.2 Schematic overview of alternating access in peptide transporters 4
1.3 Structural overview of peptide transporters 5
1.4 Chemical structures of selected peptide transporter substrates 7
1.5 Previous computational studies of POT conformational transitions 15

2.1 Schematic illustration of force field terms 32
2.2 Illustration of selected enhanced sampling methods 38
2.3 Schematic outline of binding free energy calculations 45

3.1 Hysteresis in conformational sampling and illustration of the MEMENTO method 54
3.2 Monte Carlo simulated annealing for MEMENTO path finding 58
3.3 Docking vs morphing for ligand placement in MEMENTO 60
3.4 Comparison of steered MD and MEMENTO on deca-alanine 62
3.5 Convergence analysis of deca-alanine REUS 63
3.6 Comparison of SMD and MEMENTO in 1D REUS on ADK 66
3.7 Convergence analysis on ADK 1D REUS 67
3.8 PMFs from 2D REUS for ADK 68
3.9 Convergence of 2D REUS on ADK for MEMENTO paths 69
3.10 PMF and convergence of 2D REUS on ADK for SMD paths 70
3.11 Comparison of PMFs evaluated for the P38α conformational change using MEMENTO and SMD paths 72
3.12 Unbiased MD of the P38α DFG-out state showing alternative preferred conformation 73
3.13 Convergence analysis of the P38α 2D PMF from MEMENTO paths 74
3.14 Convergence analysis of the P38α 2D PMF from SMD paths 75
3.15 Attempts at evaluating holo PMFs for P38α 76
3.16 MEMENTO applied to the LeuT transporter 78
3.17 Convergence analysis of LeuT 2D PMFs using MEMENTO paths 79
3.18 SMD runs for conformational changes of the LeuT transporter 81
3.19 Convergence analysis of LeuT 2D PMFs using SMD paths 82

4.1 Structural overview of peptide transporters 85
4.2 Membrane embedding of PepT2 structures 87
4.3 Occluded state derivation using metadynamics 88
4.4 Unbiased MD for picking candidate gating salt bridges 89
4.5 Unbiased MD from OCC state — gate opening triggers 91
4.6 Unbiased MD from OCC state — complementary gate opening distances 92
4.7 Unbiased MD from OCC state — all conditions probing the extracellular gate 93
4.8 Hysteresis when sampling PepT2 conformational changes with steered MD or metadynamics 95
4.9 PMFs from MEMENTO with 1D REUS along initial CV guess 96
4.10 2D CVs derived with PCA from initial 1D REUS trajectories 97
4.11 2D PMFs of protonation state-dependent extracellular gate opening 98
4.12 2D PMFs of extracellular gate opening in additional conditions 99
4.13 Collated 1D-reprojections of all OCC$\leftrightarrow$OF 2D PMFs 100
4.14 Analysis of interactions formed by H87 and D342 during extracellular-gate opening 100
4.15 Convergence analysis for extracellular gate 2D PMFs 101
4.16 Representative histograms of 2D REUS windows for extracellular gate opening 102
4.17 2D PMFs of protonation state-dependent extracellular gate opening 103
4.18 2D PMFs of intracellular gate opening in additional conditions 104
4.19 Collated 1D reprojections of all OCC$\leftrightarrow$IF 2D PMFs 104
4.20 Convergence analysis for intracellular gate 2D PMFs 105
4.21 Effects of the presence of Ala-Phe substrate on conformational free energy landscapes 107

4.22 CpHMD and ABFE of substrate binding effects on ExxER protonation 108
4.23 Convergence of CpHMD — E53 109
4.24 Convergence of CpHMD — E56 110
4.25 Alternative analysis of the CpHMD as histograms of short chunk estimates 111
4.26 Validation of extracellular gating residues in cell-based transport assays 113
4.27 Schematic overview of the derived model for PepT2 alternating access 115

5.1 Characterisation of cefadroxil, amoxicillin and cloxacillin in *in-vitro* assays 119
5.2 Cryo-EM structures of PepT2 in complex with cefadroxil, amoxicillin and cloxacillin 121
5.3 Unbiased MD from cefadroxil and amoxicillin poses — violin plots 123
5.4 Unbiased MD from cefadroxil and amoxicillin poses — histograms 124
5.5 Unbiased MD from cefadroxil and amoxicillin poses — plane projections 125
5.6 Unbiased MD from cloxacillin poses — violin plots 126
5.7 Unbiased MD from cloxacillin poses — histograms 127

6.1 Comparison of the AMBER, SAGE and GAFF2 force fields for peptide backbones 131
6.2 Trials of different approaches for multivalent co-ions in ABFE simulations 132
6.3 Overlay of highest-scoring overall and filtered docking poses of Trp-Trp-Trp 134
6.4 Retrospective prediction of substrate affinities at PepT2 using GOLD docking 135
6.5 Retrospective prediction of substrate affinities at PepT2 using ABFE from docking poses ... 136
6.6 Retrospective prediction of substrate affinities at PepT2 using random forest models 137
6.7 Hyperparameter tuning of random forest models for affinity prediction 138
6.8 Interpretation of the Morgan fingerprint random forest models 139
6.9 Interpretation of the 2D pharmacophore random forest models 140
6.10 Overlay comparison of different cefadroxil binding poses 143
6.11 MD equilibrations from cefadroxil docked poses 144
6.12 Transferability of random forest models to untrained chemical space 146
6.13 Valine-penicillin G prodrug trials 148

List of tables

1.1 Literature review of POT transport assays 6

2.1 Summary of common thermodynamic ensembles 25

3.1 Literature review of PMFs published on the ADK conformational change 65

4.1 Ala-Phe peptide affinities to PepT2 in different conformations and protonation states 106

5.1 IC_{50} values for cefadroxil, amoxicillin and cloxacillin in proteo-liposome assays 120
5.2 ABFE results for cefadroxil, amoxicillin and cloxacillin at PepT2 125

6.1 Comparison of ABFE results with Parker et al. 2021 141
6.2 Comparison of ABFE and experimental affinity results for cefadroxil 142
6.3 Comparison of random forest model results for β-lactam antibiotics 145
6.4 Prospective affinity prediction of a penicillin G modification 146

A.1 Overview of all sampling performed on ADK. 154
A.2 Overview of all sampling performed on P38α. 155
A.3 Overview of all sampling performed on LEUT. 156

B.1 Residue numbers used in the defintion of the tip-CV and base-CV. 157
B.2 Overview of all 1D PMF sampling. 160
B.3 Overview of all 2D PMF sampling. 161

Glossary

ABC	ATP binding cassette
ABF	adaptive biasing force
ABFE	absolute binding free energy
ADK	adenylate kinase
aMD	accelerated MD
ANN	artificial neural network
ASBT	apical sodium-dependent bile acid transporter
ATP	adenosine tri-phosphate
AWH	accelerated weight histogram method
BAR	Bennett acceptance ratio
CMAP	correction map
CoMFA	comparative molecular field analysis
CoMSIA	comparative molecular similarity index analysis
CpHMD	constant-pH molecular dynamics
cryo-EM	cryo electron microscopy
CV	collective variable
DEER	double electron-electron resonance
DOF	degree of freedom
DRMSD	different of root mean squared deviations
EAAT	excitatory amino acid transporter
ECD	extracellular domain
EPR	electron paramagnetic resonance
FEL	free energy landscape
FES	free energy surface
GB	generalised Born
IC	independent component
IF	inwards facing

LIE linear interaction energy

MBAR multistate Bennett acceptance ratio
MC Monte Carlo
MD molecular dynamics
metaD metadynamics
MFS major facilitator superfamily
ML machine learning
MRP multidrug resistance-associated protein
MSA multiple sequence alignment
MSE mean-squared error
MSM Markov-state model

NBD nucleotide binding domain
NMR nuclear magnetic resonance

OCC occluded
OCT organic cation transporter
OF outwards facing

PC principle component
PCA principle component analysis
PME particle-mesh Ewald
PMF potential of mean force
POT proton-coupled oligopeptide transporter

QM/MM quantum mechanics / molecular mechanics
QSAR quantitative structure-affinity relationship

RBFE relative binding free energy
RESP restrained electrostatic potential
REST2 replica exchange with solute tempering
REUS replica-exchange umbrella sampling
RF random forest
RMSD root mean squared deviation

SLC solute carrier
SMD steered MD

SPIB state predictive information bottleneck

TI thermodynamic integration

TICA time independent component analysis

TM trans-membrane helix

TMD trans-membrane domain

tMD targeted MD

US umbrella sampling

WHAM weighted histogram analysis method

WT wild-type

WT-metaD well-tempered metadynamics

1 Introduction

Disclaimer

Parts of this introductory chapter on peptide transporter biology (section 1.1) draw on an exposition given for my published work in eLife, Lichtinger et al. (2024), see also in the list of publications. All investigation, writing and figures in that paper were produced by myself with feedback from the co-authors, and will, with adaptations, be used in part here under the CC BY 4.0 license. Images and phrasing may therefore be similar or partially identical to my published work.

The design of new drugs is a long and costly enterprise that nonetheless has the potential of substantial rewards for human health. Prominent among the many challenges researchers commonly face in this process is the question of drug bioavailability. While an oral administration route is generally preferred, this requires a drug to permeate across the intestinal epithelia. In this thesis, we harness state-of-the art molecular simulation techniques together with biochemical experiments as well as — through active collaborations — structural biology and synthetic chemistry to study an important pathway by which drugs cross membranes: active transport through proton-coupled oligopeptide transporters (POTs). The research presented here spans from computational method development to the application of simulations to study the transport of natural peptide substrates and drugs alike, and finally is extended to probe the possibility of computer-aided drug design targeting peptide transporters for improved bioavailability. In this introductory chapter, we present the biological background of membrane transport through POTs and review previous efforts for drug design targeting these transporters as well as computational mechanistic investigations in the literature.

1.1 Peptide transporter biology

1.1.1 Alternating-access membrane transport

Cells require an external lipid membrane to separate their internal cytoplasm from the environment. Since the membrane permeability of common solutes spans ten orders of magnitude, some molecules diffuse readily across the membrane, while the translocation of others requires facilitation by carriers. (Stillwell, 2016; Pizzagalli et al., 2021) Understanding the processes by which small molecules cross membranes is of key pharmaceutical interest owing to their role in drug delivery, which may be mediated by passive diffusion, protein carriers, or a combination of both. (Sugano et al., 2010; Smith et al., 2014)

Membrane transporters are commonly classified as passive and active transporters. The former facilitate the diffusion of substrates down their electrochemical gradients without allowing for concentrative uptake. The latter uses energy either from a direct chemical source such as the hydrolysis of adenosine tri-phosphate (ATP) (primary active) or from the electrochemical gradient of a second, coupled substrate such as Na^+ or H^+ ions (secondary active, symport and antiport mechanisms are possible), as reviewed by Beckstein and Naughton (2022). Acting as non-equilibrium mechanical (and — in the case of primary active transport — also chemical) enzymes, the mechanism of these energy-coupled transport processes have

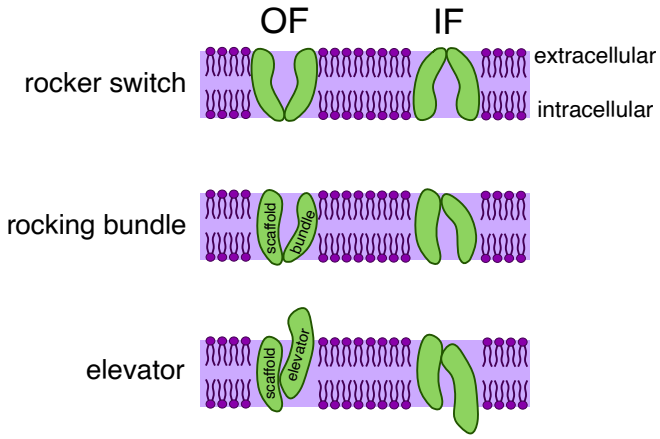


Figure 1.1: Schematic overview of alternating access protein conformations, showing the rocker-switch, rocking-bundle and elevator mechanisms. In a rocker switch mechanism, two bundles undergo approximately equal rigid-body motion to expose the central binding pocket alternatingly to the two membrane faces. A rocking bundle mechanism has the transporter sitting stably in the membrane anchored by the scaffold domain, while a helical bundle rotates to effect alternating access of the substrate binding pocket. The elevator mechanism also uses a scaffold domain to orient the transporter protein, but the elevator domain shuttles up and down parallel to the membrane normal to achieve alternating access.

long been recognised to involve alternating-access conformational changes, where the substrate binding site is exposed to either side of the membrane but not both sides simultaneously. (Mitchell, 1957, 1990; Drew and Boudker, 2016; Drew et al., 2021; Beckstein and Naughton, 2022) Primary active transporters such as P-type ATPases (Dyla et al., 2020) and ATP binding cassette (ABC) transporters (Wilkens, 2015; Furuta, 2021) couple the hydrolysis of ATP in their three cytosolic domains or their two nucleotide binding domains (NBDs), respectively, to alternating access in their trans-membrane domains (TMDs). Since this thesis is largely focussed on a secondary-active transporter, reviewing details of primary-active transport mechanisms is beyond the scope of this introduction.

Although not without exceptions (as, for example, in the ClC family that includes both Cl^- channels and H^+ / Cl^- exchangers, Poroca et al. (2017)), the conformational changes involved in alternating access of secondary-active transporters and some evolutionarily related uniporters (such as a number of sugar porters) are usually seen as falling into one of three canonical alternative conceptualisations (illustrated in figure 1.1). The rocker-switch mechanism involves two bundles of transmembrane helices moving against each other roughly symmetrically to alternatingly expose a central cavity to the opposing membrane sides; this is the case for major facilitator superfamily (MFS) transporters. (Sauve et al., 2023) By contrast, the rocking-bundle mechanism describes a stable scaffold domain against which a flexible helical bundle moves, for example in LeuT-fold transporters. (Forrest and Rudnick, 2009) Lastly, the elevator mechanism — seen for instance in excitatory amino acid transporters (EAATs) — consists of a large-scale translation of the elevator domain along the membrane normal, anchored by a scaffold domain. (Garaeva and Slotboom, 2020)

The capacity for alternating access is encoded in the structural topology of these transporters through inverted repeats. (Forrest et al., 2008; Radestock and Forrest, 2011; Forrest, 2013, 2015; Beckstein and Naughton, 2022) The 12 trans-membrane helices (TMs) of MFS transporters, for example, include 6 TMs in the N- and C-terminal bundles each, giving rise to approximate 2-fold symmetry. Each of the bundles,

in turn, is composed of two 3-TM bundles, which are inverted with respect to each other in that they start at opposing sides of the membrane. Similar arrangements are found in different folds as well, such as the (5+5) inverted repeat arrangement of LeuT-fold transporters. This not only suggests an evolutionary history of duplication-and-fusion but also provides a convenient framework within which to study the inwards facing (IF) and outwards facing (OF) states of alternating-access transporters. The overall asymmetric conformation is swapped (IF $\leftrightarrow$ OF) if the inverted repeat adopts the conformation of the pseudo-symmetric equivalent, and vice versa. This has also been extensively exploited bioinformatically to construct homology models in previously unresolved transporter conformations, in a procedure known as 'repeat-swapping'. (Vergara-Jaque et al., 2015)

Despite the conceptual simplicity of these models for alternating access, complexities and biological constraints arise when considering how transitions proceed between the IF and OF states. (Beckstein and Naughton, 2022) To minimise the occurrence of energetically wasteful states that are wide open at both gates, for example, the transporter preferentially transitions through (a number of) occluded (OCC) states that are at least partially closed at both gates, and often involve bending and disruption of helices within the moving bundles. Plentiful structural and biochemical evidence supports the notion of intermediate OCC states as integral parts of alternating access mechanisms, thus challenging the textbook picture of static helical bundles moving as rigid bodies. (Forrest et al., 2011; Shi, 2013; Henderson et al., 2019; Drew et al., 2021)

1.1.2 The SLC15 family of proton-coupled peptide transporters

The solute carrier (SLC) transporter superfamily encompasses 65 families of more than 450 genes, with substrates ranging in size from simple ions to complex macromolecules used in metabolism and signalling. (Pizzagalli et al., 2021) Within this superfamily, the SLC15 family includes proton-coupled oligopeptide transporters (POTs) that have significant homology through all domains of life and are evolutionarily ancient. (Daniel et al., 2006) Of the four mammalian family members, PepT1 (SLC15A1) and PepT2 (SLC15A2) are the most studied. The former is predominantly expressed in the small intestine and characterised as a low-affinity, high-capacity transporter. (Fei et al., 1994) The latter has a broader expression pattern including the kidneys, lungs and brain and is described as high-affinity, low-capacity. (Daniel and Kottra, 2004) As secondary-active transporters, they couple uphill substrate translocation to the symport of a variable number of protons down their electrochemical gradient. (Fei et al., 1994; Rubio-Aliaga et al., 2000) Figure 1.2 schematically illustrates the alternating access cycle of this coupled symport process.

1.1.3 Structural and biochemical characterisation of peptide transporters

POT family transporters belong to the MFS and share a conserved topology of two six-helix bundles that form the functional transport domain, with their N- and C-termini both facing the cytoplasm. (Newstead et al., 2011) They operate via an alternating access mechanism encoded in four inverted topology repeats, progressively reorienting the N- and C-terminal bundles to cycle through OF, OCC and IF states — see the discussion in section 1.1.1. Since the first structure of a POT family member ($PepT_{So}$) was published (Newstead et al., 2011), many procaryotic (Solcan et al., 2012; Guettou et al., 2013; Doki et al., 2013; Lyons et al., 2014; Guettou et al., 2014; Zhao et al., 2014; Fowler et al., 2015; Boggavarapu et al., 2015; Beale et al., 2015; Parker et al., 2017; Martinez Molledo et al., 2018; Ural-Blimke et al., 2019; Minhas and Newstead, 2019; Stauffer et al., 2022; Kotov et al., 2023) and plant (Parker and Newstead, 2014; Sun et al., 2014) homologues have been structurally and biochemically characterised, all in IF states with varying degrees of occlusion

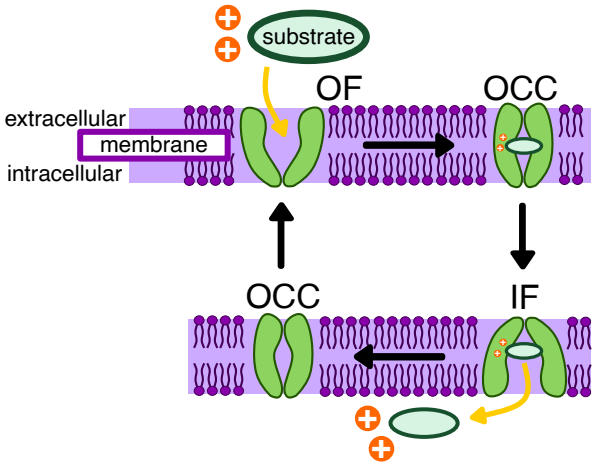


Figure 1.2: Schematic overview of the alternating access cycle of peptide transporters highlighting the symport of substrate and protons through OF, OCC and IF states, here shown for a 2:1 stoichiometry (see discussion in section 1.1.4).

(see figure 1.3a for an overview of available POT structures and their conformational states). Several residues have been suggested to be involved in proton transfer, including a partially conserved histidine on TM 2 (H87; residue numbers refer to PepT2, if not specified otherwise) (Terada et al., 1996; Fei et al., 1997; Chen et al., 2000; Omori et al., 2021; Parker et al., 2021) and two conserved glutamates on TM 1 (E53 and E56) (Jensen et al., 2012; Doki et al., 2013; Aduri et al., 2015). Furthermore, the importance of a number of key residues for POT transport has been established using a range of transport assay protocols, as summarised in Table 1.1. Molecular dynamics (MD) simulations have also helped our understanding of proton-transfer processes and conformational changes, which is discussed in detail in section 1.3.

Cryo electron microscopy (cryo-EM) and Alphafold 2 have recently provided views of mammalian POTs in conformations spanning from OF via inward-facing, partially occluded to fully-open IF. (Parker et al., 2021; Jumper et al., 2021; Killer et al., 2021; Shen et al., 2022) From these structures emerges a picture (see figure 1.3b–c) where the intracellular gate is constituted by broad close-packing of hydrophobic residues on TMs 4, 5, 10 and 11 (F184, Y183, F187, L630 and Y634) that constrict access to the binding site from the intracellular side, with possible stabilisation from the conserved D170–K642 salt bridge. The extracellular gate appears to be spread along the cleft between the N- and C-terminal bundles, with contributions from the H87 (TM 2) – S321 (TM 7) polar interaction network as well as the R206 (TM 5) – D342 (TM 8) and K64 (TM 1) – D317 (TM 7) salt bridges . This is intriguing because the mammalian H87 residue is only conserved in some prokaryotic homologues (such as $PepT_{So}$), and R206–D342 just among mammalian POTs. The extracellular gating mechanism could therefore be less conserved than POT alternating access in general (Parker et al., 2021), which also fits with the fact that these gating residues lie outside the highly conserved motifs of the MFS family.

Aside from the conserved 12 canonical TM helices as shared among all MFS transporters, POTs also contain other structural features that are not universally conserved through the family. For bacterial POT homologues, this includes two additional helices inserted between the N- and C-terminal bundles (Newstead et al., 2011), the function of which is not fully clear. Since they are absent in eukaryotic POTs, they will not concern us further in this thesis. Mammalian POTs, in turn, have an extracellular domain (ECD) between

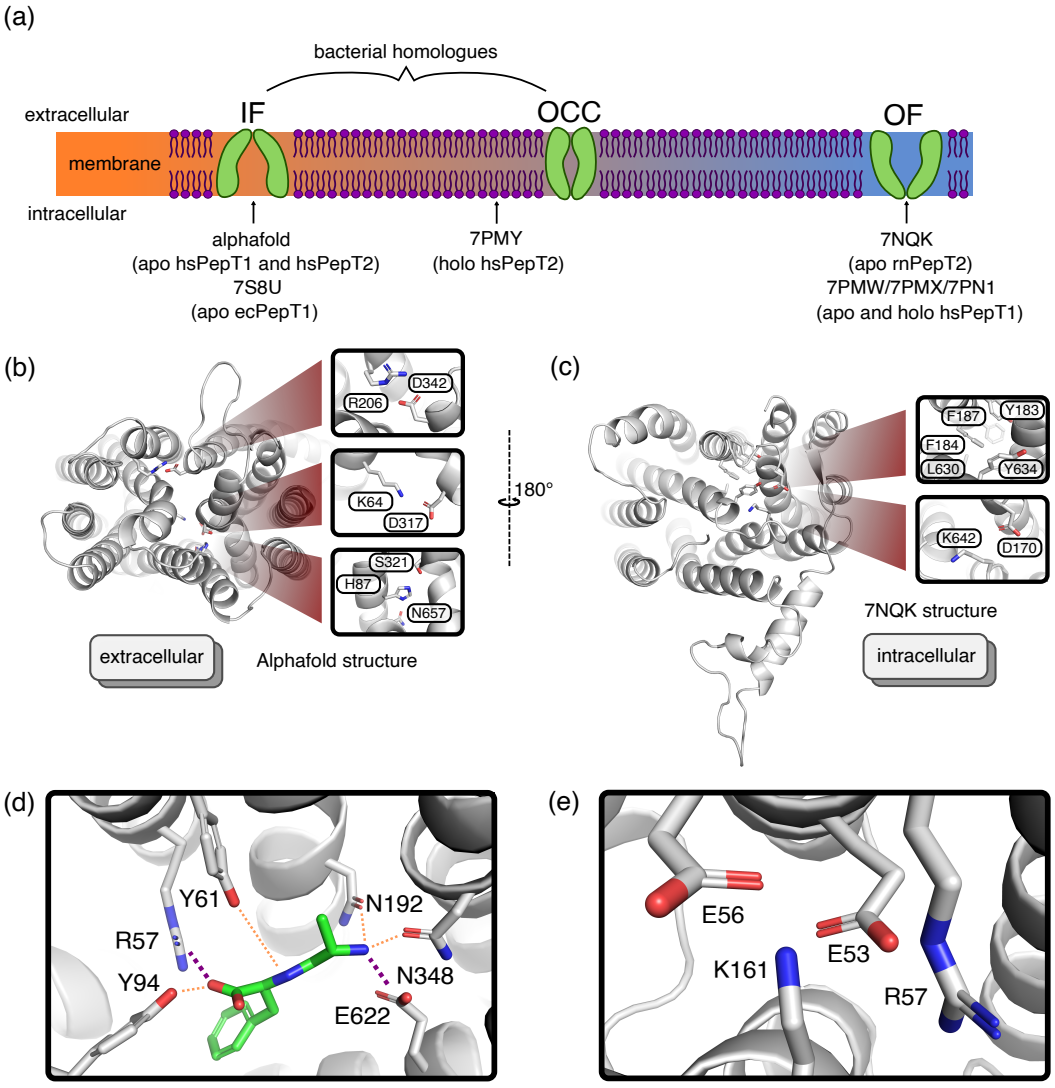


Figure 1.3: (a) Schematic overview of the conformational diversity of available mammalian POT structures. Intermediate positions between states indicate partial gate opening. (b) Alphafold-predicted IF-state HsPepT2 structure (viewed from the extracellular side, with the extracellular domain omitted for clarity), highlighting potential inter-bundle extracellular gate interactions. (c) OF-state Cryo-EM structure of apo RnPepT2 (7NQK, viewed from the intracellular side) (Parker et al., 2021), highlighting potential inter-bundle intracellular gate interactions. (d) Ala-Phe substrate binding pose, representative cluster frame of 1 µs MD simulation from 7NQK structure with added ligand, for setup details see appendix B. Purple dotted lines represent salt-bridge contacts, orange dotted lines other polar contacts. (e) ExxER motif salt-bridge cluster, representative cluster frame of 1 µs MD simulation from 7NQK structure.

Table 1.1: Summary of transport assays in the literature establishing the importance of key residues for transport in POTs by impact of residue mutation on transport activity. Studies are conducted on a range of POT homologues, but for convenience they are grouped by the equivalent residue number in PepT2.

Residue (PepT2)	TM	Studies
E53	1	Solcan et al. (2012); Doki et al. (2013); Sun et al. (2014); Jørgensen et al. (2015); Parker et al. (2017)
E56	1	Solcan et al. (2012); Jørgensen et al. (2015); Sun et al. (2014); Zhao et al. (2014); Parker et al. (2017)
R57	1	Solcan et al. (2012); Guettou et al. (2013); Doki et al. (2013); Jørgensen et al. (2015); Lyons et al. (2014); Parker and Newstead (2014); Sun et al. (2014); Parker et al. (2017); Martinez Molledo et al. (2018)
H87 [1]	2	Fei et al. (1997); Chen et al. (2000); Newstead et al. (2011); Uchiyama et al. (2003); Parker et al. (2017)
Y60	1	Transport affected: Pieri et al. (2009); Solcan et al. (2012); Transport unaffected: Lyons et al. (2014)
Y61	1	Pieri et al. (2009); Solcan et al. (2012); Doki et al. (2013); Lyons et al. (2014); Ural-Blimke et al. (2019); Minhas and Newstead (2019)
Y94	2	Chen et al. (2000); Pieri et al. (2009); Solcan et al. (2012); Lyons et al. (2014); Minhas and Newstead (2019)
D170	4	Fowler et al. (2015); Parker et al. (2021)
Y188	5	Chen et al. (2000); Minhas and Newstead (2019)
N192	5	Transport affected: Doki et al. (2013); Ural-Blimke et al. (2019); Minhas and Newstead (2019); Transport unaffected: Lyons et al. (2014)
D317 [2]	7	Solcan et al. (2012); Doki et al. (2013); Lyons et al. (2014); Parker et al. (2017); Martinez Molledo et al. (2018); Shen et al. (2022)
N348	8	Transport affected: Lyons et al. (2014); Minhas and Newstead (2019); Transport unaffected: Doki et al. (2013)
E622	10	Solcan et al. (2012); Guettou et al. (2013); Doki et al. (2013); Lyons et al. (2014); Zhao et al. (2014); Parker et al. (2017); Minhas and Newstead (2019); Shen et al. (2022)
K642	11	Fowler et al. (2015); Parker et al. (2021)

[1] for those homologues which conserve it
[2] or the equivalent glutamate

TMs 9 and 10 that binds the protease trypsin. (Beale et al., 2015) While in PepT1, the ECD is important for protein expression (Beale et al., 2015) and bridges the N- and C-terminal bundles by interacting with TM 1 in the IF state (thereby impacting transport; Shen et al. (2022)), PepT2 expresses well without the ECD, showing no impact on substrate recognition or transport (Beale et al., 2015), and its movement can largely be described as a rigid body anchored via the connecting loops to the transmembrane domain. (Parker et al., 2021) In the PepT2 simulations in this thesis, we therefore delete the ECD for computational efficiency through smaller box sizes (see appendix B for details).

1.1.4 Substrates and the peptide transporter binding site

The POT substrate chemical space is vast, encompassing most di- and tripeptides (with chemically diverse side chains) (Fei et al., 1994) in animals and bacteria, as well as a range of marketed drugs, including many β-lactam antibiotics, angiotensin-converting enzyme (ACE) inhibitors, and the photosensitising agent aminolevulinic acid. (Luckner and Brandsch, 2005; Biegel et al., 2006; Brandsch, 2013) Figure 1.4 shows the chemical structures of a selection of these substrates. POT homologues in plants have also evolved

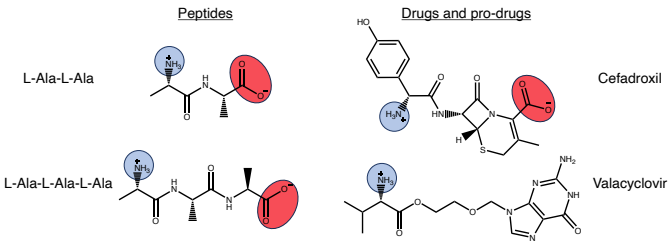


Figure 1.4: Chemical structures of some peptide transporter substrates. The charged N- and C-termini (at neutral pH) are highlighted in blue and red, respectively.

to recognise diverse substrates including nitrate, dicarboxylates, amino acids, glucosinolates and several phytohormones. (Léran et al., 2014) POTs are thus remarkably promiscuous transporters that are thought to accommodate this wealth of substrates by a number of specificity pockets with different chemical properties as well as potentially a diversity of ligand binding modes. (Newstead, 2015)

POTs accommodate their substrates in a highly conserved binding pocket, interfacing between an acidic patch on the C-terminal bundle and a basic patch on the N-terminal bundle (figure 1.3d). For dipeptides, the N-terminus is coordinated by E622 (TM 10) together with N192 (TM 5) and N348 (TM 8), while the C-terminus engages R57 (TM 1) as well as Y94 (TM 2). Another conserved tyrosine, Y61 (TM 1), hydrogen-bonds to features of the peptide backbone. (Lyons et al., 2014; Martinez Molledo et al., 2018; Killer et al., 2021) Tripeptides may adopt a similar orientation as dipeptides (Guettou et al., 2014), or sit vertically in the transporter binding pocket (Lyons et al., 2014), although it has been suggested that this vertical electron density could alternatively be explained by a bound HEPES molecule. (Martinez Molledo et al., 2018) Considering the consensus structural interaction pattern, we decided to investigate in the course of this project primarily the role of E622 and R57 in holding the substrate, and also note that R57 is part of the highly conserved E^{53}xxER motif (figure 1.3e). (Newstead, 2017) The second glutamate in this motif (E56) in particular has been linked to proton coupling experimentally. (Jensen et al., 2012) Since R57 interacts with both the ExxER glutamates and the substrate C-terminus, we hypothesise that it may play an important role in substrate–proton coupling.

The peptide–proton symport stoichiometry is not conserved between different substrates and POT family members. (Parker et al., 2014) For PepT1, stoichiometries of 1:1 and 2:1 have been reported for neutral / basic and acidic di-peptides, respectively. (Fei et al., 1994; Steel et al., 1997) For PepT2, a 2:1 stoichiometry was reported for the neutral di-peptide D-Phe-L-Ala and 3:1 for anionic D-Phe-L-Glu. (Chen et al., 1999) Alternatively, Fei et al. (1999) have found 1:1 stoichiometries for either of D-Phe-L-Gln (neutral), D-Phe-L-Glu (anionic), and D-Phe-L-Lys (cationic). In chapter 4 of this thesis, our results show distinct and additive roles played by two protons in the conformational cycle mechanism, leading us to formulate our conclusions within the framework of a 2:1 stoichiometry for the neutral L-Ala-L-Phe di-peptide we have simulated.

Numerous studies have mapped the substrate structural requirements for transport by POTs in transport assays as well as structural and computational work, and have gained qualitative insight into the peptide transporter pharmacophore. (Tamura et al., 1996; Swaan and Tukker, 1997; Fujita et al., 1997;

Brandsch et al., 1998; Döring et al., 1998; Temple et al., 1998; Brandsch et al., 1999; Terada et al., 2000; Bailey et al., 2000; Meredith et al., 2000; Tateoka et al., 2001; Knütter et al., 2004; Våbenø et al., 2005; Ekins et al., 2005; Vig et al., 2006; Biegel et al., 2006; Guettou et al., 2014; Boggavarapu et al., 2015; Ural-Blimke et al., 2019; Minhas and Newstead, 2019; Parker et al., 2021) From these, a consensus picture emerges, featuring a positively charged amino group and a negatively charged carboxyl — spaced appropriately — that are of central importance to POT binding (although not strictly indispensable). A peptide bond, while not required for binding, tends to enhance affinity. L-amino acids are favoured over their D enantionmers, and particular positions in the substrate may exhibit a preference for hydrophobic side chains.

Several authors have also constructed quantitative models for substrate affinity to POTs — initially as quantitative structure-affinity relationship (QSAR) models when no structural information was available. Gebauer et al. (2003) used comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA) to establish a 3D QSAR model of substrate binding to PepT1. CoMFA calculates steric and electrostatic interaction energies between a given conformation of a molecule and probe atoms placed on a grid. The matrix of these field contributions is then correlated with biological features such as binding affinities using least-squares fitting. In CoMSIA, several more descriptors such as hydrophobicity and hydrogen bond features are included in a workflow based on CoMFA. (Roy et al., 2015) After structurally aligning a range of generated conformers of 79 training-set dipeptides and dipeptide derivatives and locating pharmacophore features, the authors constructed both CoMFA and CoMSIA models and found good predictive power on a test set of 19 dipeptide-type ligands in each case. In the interpretation of their models, they recapitulate the importance of the positively charged N-terminus, the high-electron-density C-terminus, as well the presence of bulky side chains and several hydrophilic and hydrophobic regions. Biegel et al. (2005) employed a similar strategy in constructing a CoMSIA model on a training set of 98 di- and tripeptides as well as β-lactam antibiotics, using data from the same assay (inhibition of radio-labeled Gly-Sar uptake in Caco-2 cells) under comparable conditions. On a test set of 24 molecules of the same types, the authors report a deviation from measured inhibition constants of less than 0.8 log units (that is, less than ≈6-fold deviation in inhibition constant), and concerning a novel dipeptide mimetic they found close agreement between the model's prediction with new experimental data. By including β-lactams in their model, they were able to conclude that the charged amino group or the nitrogen contained in the aminothiazole ring of these drugs have donor properties similar to the N-terminus in peptide substrates. On the basis of data from a dipeptide library comprising 337 K_i datapoints (competitive inhibition of reference substrate transport) by the yeast POT Ptr2p, Ito et al. (2013) trained discrimination analysis models on the peptide amino-acid sequence and physicochemical descriptors to reproduce a low/medium/high experimental affinity classification with good accuracy. When testing their models using physicochemical properties on drug molecules, four out of six predictions were correct. The authors attribute the errors with incorrectly predicted molecules to their being of very different molecular weight to the dipeptide training set. Although the models presented in this paragraph have been superseded in the field with the availability of POT structures, their good performance on ligand-only features as the first available computational approach sets the bar for structure based methods; see also the discussion of our results from ligand-only versus structure-based methods in chapter 6 of this thesis.

Following the more recent availability of POT structures (as reviewed in section 1.1.3), Samsudin et al. (2016) used a set of seven dipeptides and one tripeptide to evaluate the relative performance of several structure-based computational binding free energy approaches on predicting binding to the bacterial POT $PepT_{St}$. Docking scores, several end-state free energy methods and rigorous but expensive thermodynamic

integration are compared, and the results indicate that the more expensive, MD-based methods can more accurately rank the substrates according to experimental IC_{50} values. Using the linear interaction energy (LIE) method (an end-point free energy approach which calculates a binding free energy as the sum of differences in van-der-Waals and Coulomb interactions between the bound and unbound state or between ligands), they also predicted binding of all 400 possible dipeptides with natural amino acids and found that the N-terminal residue plays a particularly important role in determining affinity to the transporter. However, when they attempted to use this method on a homology model of the human PepT1 transporter, the test-set correlation disappeared, thus highlighting the importance of accurate structural models for structure-based affinity prediction. To understand the source of this effect further, the authors also tested homology models of bacterial POTs, for which a range of templates were available. They found that with decreasing sequence identity, the correlation of the calculated free energies with experiment decreases. While from this data alone it cannot be ascertained where in the homology modelling procedure the problem originates, we believe that a number of factors likely operate in parallel, including side chain orientations that can be slow to equilibrate in MD as well as subtle induced fit in the protein following substrate binding that may not be fully conserved between related transporters.

Having solved the first mammalian POT structure, Parker et al. (2021) were subsequently able to accurately correlate results from absolute binding free energy (ABFE) simulations (see section 2.3) of 5 di- and tripeptides to experimental IC_{50} values. Whether correlations to affinity proxies contribute meaningful information regarding transport has been challenged by Kotov et al. (2023), however, working on the bacterial DtpB POT homologue. By comparing affinity from thermal unfolding assays and transport data from proteoliposome assays for 23 di- and tripeptides, they found that medium-affinity substrates are transported best, which they speculate may be a general feature of promiscuous transporters. In the same study, the authors also used docking to make models of all di- and tripeptides in the binding pocket and found no correlation between the docking score and experimental affinity for those compounds with experimental data available, and only very limited correlation with simple chemical descriptors. By training a logistic regression classifier on the Rosetta docking scoring function energy terms, however, they were able to outperform chance in classifying binders vs non-binders. From this, they propose a workflow that includes *in-silico* predictions to narrow down the space for experimental testing.

In summary, while both ligand-only and structure-based quantitative models have shown promise towards predicting binding of substrates to POTs, significant challenges and avenues for further exploration remain. Ligand-only models have so far been trained only on small datasets that have not reflected the full range of data available now (for example with regard to non-β-lactam drugs and non-natural amino acids), and machine learning techniques such as artificial neural networks (ANNs) or random forest (RF) models and cheminformatic descriptors like molecular fingerprints have not been tried on these datasets yet. Furthermore, how well such ligand-based models perform with regard to the kind of activity cliffs common in drug discovery projects (see for instance Stumpfe and Bajorath (2012)) has not been considered yet regarding POTs. Structure- and simulation-based methods, on the other hand, require accurate protein-ligand models, and how well they perform on the recent mammalian POT structures has not been tested with a large dataset yet. It is also crucial to note that even when accurate affinity predictions are available, this does not necessarily translate to accurate predictions of transport. (Kotov et al., 2023) Lastly, as recently demonstrated by Kapoor et al. (2022) on an ABC transporter, for proteins that undergo large-scale conformational changes the entire conformational cycle needs to be considered during *in-silico* modelling of transporter–substrate binding. A thorough understanding of the alternating

access cycle and also its coupling driving forces is required to successfully correlate computational affinity metrics with experimental data — we show this in chapter 5 of this thesis.

1.2 Oral drug bioavailability via peptide transporters

Of the various strategies to deliver drugs into the human body, the oral route is preferred for its convenience, non-invasiveness and good patient compliance. Nevertheless, the oral formulation of a drug may be impeded by a range of factors, including solubility in the intestinal medium, stability in the gastrointestinal tract — especially with regard to the harsh pH environment and various digestive enzymes — as well as permeability through the mucosal and epithelial barriers. (Alqahtani et al., 2021) Of these challenges, we will here be concerned with drug permeation through the intestinal epithelium, which occurs through one of three main mechanisms: passive diffusion across lipid membranes, paracellular diffusion through aqueous intercellular spaces or carrier-mediated transport. (Smith et al., 2014) Which of these routes is favoured depends on the physicochemical properties of a drug in question, and — in the case of carrier-mediated transport — whether any transporters expressed in the intestinal membranes can recognise it as a substrate. Membrane transporters in several families have been found to play roles in intestinal drug absorption through the import of drugs from the intestinal lumen across the enterocyte apical membrane (PepT1, organic cation transporters (OCTs) 1 and 3 and the apical sodium-dependent bile acid transporter (ASBT)) and export from the enterocyte into the portal vein (multidrug resistance-associated protein (MRP) transporters 1 and 3). Flux in the opposite direction, however — back into the intestinal lumen — is also facilitated by a range of transporters, including MRP2 and P-gp. (Müller et al., 2017) Owing to its inward-directed transport direction and high expression in the small intestine in humans, PepT1 therefore plays a significant role in the oral bioavailability of those drugs it can transport. This is particularly relevant for rational drug design, given the widely recognised promiscuity of this transporter. Its ability to accept a large range of chemically diverse substrates (as reviewed in section 1.1.4) gives rise to the possibility of engineering new drugs to be POT substrates, thus increasing oral bioavailability by leveraging an endogenous small-molecule uptake pathway.

Next to many β-lactam antibiotics (Luckner and Brandsch, 2005), this principle of oral drug bioavailability through POT transport is well illustrated by the example of valacyclovir (pictured in figure 1.4). This valine-ester prodrug of acyclovir has a significantly improved uptake profile due to its transport by PepT1. (Beauchamp et al., 1992; Han et al., 1998; Ganapathy et al., 1998; Minhas and Newstead, 2019) POTs can thus be the target of drug and prodrug strategies, with the aim of designing drugs as POT substrates, and constructing prodrugs by attaching amino acids or short peptides to existing drugs. While it should be noted that valacyclovir was developed before the discovery of PepT1 and its role in drug uptake, and it was only realised years after its design that the enhanced bioavailability was through the action of PepT1, similar strategies have since been targeted directly at peptide transporters, evidencing the great pharmaceutical interest in understanding substrate binding to POTs. The use of amino acid conjugation in developing prodrugs has been reviewed by Vale et al. (2018), including for targetting POTs. We will now survey prominent examples of this approach.

Song et al. (2005a) found that valyl and isoleucyl esters of the nucleoside anticancer drug gemcitabine displayed increased transport into HeLa cells through PepT1 (>10 fold for the valine prodrug), and additionally found improved stability against degradation by cytidine deaminase. In a further study, Song et al. (2005b) synthesised various amino-acid ester prodrug candidates based on the antiviral agent BDCRB and observed 2–4 fold higher affinity for PepT1 in a cell-based transport competition assay, where hydrophobic

amino acids in their L-stereoisomers were preferred by the transporter. More recently, Tao et al. (2018) used *in-situ* intestinal perfusion of rats to investigate the pharmacokinetic properties of amino-acid ester prodrugs of the chemotherapy agent decitabine. They found evidence for both passive transport and saturable uptake of the L-valine ester that was partially inhibited by a known PepT1 substrate. Therefore, they suggest a hybrid transport mode of this prodrug across the intestinal enterocyte layer. In another example of the possibility of prodrug design targetting PepT1, Sun et al. (2018) have synthesised seven amino-acid-ester and seven amide prodrugs of the poorly bioavailable neuraminidase inhibitor peramivir and studied permeability across intestinal-derived Caco-2 cell layers. The L-valine ester and L-isoleucine amide were the most permeable prodrugs in the respective groups and uptake of a reference peptide in Caco-2 cells was inhibited by these drugs in a concentration-dependent manner. In rat experiments, the oral bioavailability of the two candidate prodrugs were 65 and 37 % respectively, with differences attributed to the faster bioactivation of the ester prodrug compared to the amide.

Moving beyond the attachment of single amino acids to drugs, Foley et al. (2009) have conjugated thioamide-modified dipeptides (for better resistance to hydrolysis of the peptide bond) to the anti-inflammatory drug nabumetone via a hydroxyimine and PEG linkers. They found their products to be competitive inhibitors of reference-peptide uptake in PepT1-expressing Xenopus oocytes, as well as being active in a trans-stimulation assay (two substrates on both sides of the membrane lead to faster transport) and a Caco-2 cell-based transport assay. In a follow-up study, Foley et al. (2016) attached an equivalent thio-dipeptide to cyclosporin A, demonstrating that it is both an inhibitor (in a Xenopus competition assay) and a substrate (in an electrophysiological transport assay) of PepT1. This is intriguing because this prodrug has a molecular mass of >1000, indicating that very large substrates indeed could be accommodated by PepT1. How such large molecules are bound is not entirely clear in the field at the moment in the absence of structural data. The exceptional ability of the POT binding pocket to binding with tight affinity to ligands ranging in size from small dipeptides like Ala-Ala to very bulky tripeptides such as Trp-Trp-Trp has been noted, however, and a number of distinct binding modes for the non-peptidic parts of a large substrate molecule may co-exist. In a further proof of concept of their thio-dipeptide strategy, Foley et al. (2018) conjugated them to ten approved drugs, of which eight were stable. Five of these compounds (which showed uptake in PepT1-expressing Xenopus oocytes) were assayed for transport across Caco-2 cell layers, and all displayed significant transport. Although, when taking two of their prodrugs to *in-vivo* experiments they found low bioavailabilities and resistance to liver metabolism, the authors found encouragement for the potential of PepT1-targeted prodrug design in their *in-vitro* data and the fact that some prodrug absorption was detected in the *in-vivo* trials.

Taking a different conceptual approach, Colas et al. (2017) conducted virtual screens via docking of FDA-approved drugs and a lead-like library to homology models of PepT1, prioritising those compounds which displayed interactions with those binding-site residues known to coordinate natural substrates. For five compounds, they found inhibition of reference peptide uptake in a yeast transport assay, and one of these molecules was transported by PepT1 in a Xenopus oocyte assay (the others acted as inhibitors only). Notably, these molecules are chemically distinct from canonical peptide-like substrates, suggesting that the chemical space available for POT-targeted drug design may be even wider than anticipated based on known substrates.

In addition to prodrug design targetting peptide transporters for increased bioavailability, it has also been suggested by Arakawa et al. (2020) that a similar strategy could be used for constructing isotope-

labelled MRI probes with desired biodistribution profiles, leveraging for example the high expression of PepT1 and PepT2 in certain cancer cells. The authors synthesised [^{19}F]-containing derivatives of thio-modified L-Phe–L-Ala dipeptides and showed POT transport in PepT1 and PepT2-expressing HeLa cells, as well as metabolic stability in a hepatocyte culture. Although they saw no uptake into tumour cells in a mouse model, they argue that future investigation has potential.

In summary, there is much interest in using peptide transporters as targets for transport-enhanced prodrugs, and promising results have been reported. Nonetheless, rational structure-based design of POT-targetting prodrugs is only starting to be explored, thus warranting the computational investigation of drug binding and transport as well as trials for predictive drug design presented in this thesis to further develop this emerging field.

1.3 Review of peptide transporter simulation studies

Several studies in the literature have used molecular simulations to gain an understanding of the mechanism of POT alternating access and the molecular basis for substrate recognition, including coupling of protonation state changes, substrate binding and large-scale conformational changes. The following paragraphs review papers that have substantially employed MD in the investigation of POTs.

The first MD simulations on a POT were conducted by Doki et al. (2013). The authors begin by revealing IF-state X-ray structures of the bacterial POT homologue GkPOT in the wild-type (WT) apo state and with a sulfate ion bound, as well as the E310Q mutant (PepT2: D317) in both apo / sulfate-bound states and with the bound dipeptide analogue substrate alafosfalin. Since at high sulfate concentrations in the crystallisation buffer a sulfate ion is found close to E310 (in a position not substantially changed following E310Q mutation), E310 is assigned as protonated. In their E310Q-mutated and alafosfalin-bound structure, the substrate C-terminus is located close to E310Q and R43 (PepT2: K64), and the main-chain carbonyl interacts with Y40 (PepT2: Y61). The N-terminus engages N342 (PepT2: N348) and E413 (PepT2: E622). The authors suggest this to be the dipeptide binding mode, although later peptide-bound structures show the substrate C-terminus engaging the R57-equivalent residue instead of K64. They then ran sets of 200 ns-long MD simulations, in the apo E310-protonated and deprotonated states, and with $(Phe)_2$ peptide modelled in the binding pocket while E310 is protonated, finding the IF state to be stable in apo E310 protonated, but sampling OCC-like conformations following large structural changes induced by E310 deprotonation or the inclusion of peptide. In the process of OCC-state formation, a salt bridge between R43 and D317 forms, the importance of which is validated in transport assays. From their data, they propose a substrate-proton coupling mechanism where the transition to the OCC state is only possible when (a) neither proton or substrate, or (b) proton and substrate are bound. This is accomplished by the interplay between the R43–E310 salt-bridge formation and the protonation of E310 and the substrate binding thus enabled, which in turns allows for the occlusion of the intracellular gate.

Aduri et al. (2015) have further investigated the role of the ExxER motif (E^{53}xxER in PepT2) in responding to protonation. They first studied substrate accumulation in the bacterial POT YjdL, which has the unusual QxxEY sequence (instead of E_1xxE_2R). Testing all mutants of this transporter involving combinations of glutamate and glutamine at the '1' and '2' positions, they found that placing glutamates in both '1' and especially '2' positions enhances substrate accumulation with pH-dependent profiles. When the tyrosine of YjdL was replaced with the canonical arginine, substrate accumulation required both glutamates, and any transport is abolished when both glutamates are replaced. For a closer mechanistic investigation

of the ExxER motif coupling, they turned to 35 ns-long MD of the GkPOT apo sulfate-bound structure (see above) with WT sequence and the E_2 to Q mutation, finding a change in conformation of the arginine previously associated with E_2 towards E_1 in the latter case. From this data, the authors suggest a mechanism where peptide binding with its C-terminus near the arginine residue and E_2 enables protonation of E_2, prompting a conformational change in the associated arginine towards E_1 and letting the peptide engage the neighbouring lysine (K64 in PepT2). They also found an opening movement of TM 5 in response to the salt-bridge swap, which they suggest indicates that the salt-bridge status also determines the transporter's preference for IF vs OCC states, thus linking protonation with both local and global conformational changes.

To study the gates that act in POTs to restrict access to the binding pocket from the intra- and extracellular directions, Fowler et al. (2015) — writing when only inward-facing POT structures were available — set out to construct OF-state models for $PepT_{So}$ and $PepT_{St}$ via a repeat-swapping modelling procedure (using a new crystal structure for $PepT_{So}$ as template), and attempted to validate the $PepT_{So}$ OF model with double electron-electron resonance (DEER) spectroscopy, finding partial agreement of their model with the data. As an alternative method for generating new conformational states, they ran triplicate 200 ns-long simulations starting from the $PepT_{So}$ and $PepT_{St}$ IF states. After validating on available MFS transporter structures a metric for determining the opening state of the intra- and extracellular gates based on pore radius analysis, as well as one based on helical tip distances, they projected their MD simulations onto the helical tip distance space, finding that for each protein all of IF, OCC and OF states were sampled. An analysis of formed salt bridges in $PepT_{So}$ predicts that the K84–D79 (PepT2: K110–D105) and D136–K439 (PepT2: D170–K642) salt bridges stabilise the OF state as they are formed in a higher proportion of simulation frames in the OF state compared to OCC and IF, while R52–D328 (PepT2: not conserved) stabilises the IF and OCC states in about half of the simulation frames. In the $PepT_{St}$ simulations, the authors additionally observed, among others, the canonical salt bridge R33–E300 (PepT2: K64–D317). They also highlight the importance of proline-induced helical kinks for the conformational changes, and validate the involved residues in transport assays. Overall, they propose a 'double-scissor' mechanisms for POT alternating access.

The findings concerning proton coupling of global conformational changes from relatively short unbiased MD simulations on GkPOT (Doki et al., 2013; Aduri et al., 2015) have been challenged by Immadisetty et al. (2017). To probe the effect of substrate binding and E310 protonation on GkPOT dynamics, the authors have set up a matrix of 400 ns-long simulations with E310 protonated / deprotonated crossed with the apo as well as Ala-Ala, Phe-Phe and Glu-Glu substrate-bound states, doing two repeats per condition. In these simulations, there are local conformational responses to E310 protonation, including the formation / disengagement of the R43–E310 salt bridge and an effect on substrate orientation within the binding site. They did not, however, find significant global conformational changes in dependence on protonation or substrate-binding states, and the appearance of some correlations in selected simulations disappears when considering the replicate set. While there was a statistically significant correlation between global descriptors of the periplasmic and cytoplasmic gate openings (as expected for an alternating-access transporter), there was no significant correlation between any of the local and global descriptors considered. The authors, while highlighting their findings concerning local conformational changes such as substrate orientation and the formation of salt bridges, conclude that the simulated time scales were too short to capture the alternating-access global conformational transitions and suggest the future possibility of using enhanced sampling simulations to capture them.

Proton movement in $PepT_{So}$ and $PepT_{Xc}$ — both 'mammalian-like' POTs in that they conserve the TM 2 histidine (H87 in PepT2) in the extracellular gate — has been studied by Parker et al. (2017). They first conducted reconstituted transport assays to show that all of H61 (PepT2: H87), the ExxER residues, as well as the putative intracellular-gating K127 (PepT2: K161) and E419 (PepT2: E622) are important for transport, while the mutation of R32 (PepT2: K64) resulted in an active transporter, thus highlighting an alternative extracellular gating mechanism. They also found that the H61D mutation gives a functional transporter but shifts the pH–activity profile, corroborating the role of H61 in proton coupling. A double mutant H61D & D316H (PepT2: D317) also allowed for a partial rescue of transport activity, indicating an important interaction between these two residues. The authors further solved a crystal structure of $PepT_{Xc}$, picked to have the mammalian-style extracellular gate, showing extensive hydration of the binding pocket and also water wires linking D322 (PepT2: D317) and E425 (PepT2: E622) and extending to the ExxER motif. A 100 ns-long conventional MD simulation of $PepT_{Xc}$ (following 350 ns of equilibration) shows waters connecting D322 and E425 almost 90% of the time, thus indicating stable waters in the region. Quantum mechanics / molecular mechanics (QM/MM) simulations combined with umbrella sampling (US) were then used to calculate a potential of mean force (PMF) of proton transfer between D322 and E425. This shows accessible free energy barriers for proton movement, modulated by R37 (PepT2: K64) and K324 (not conserved in PepT2; which do, however, not directly participate in proton transfer). The authors also conducted long, unbiased MD from the IF $PepT_{Xc}$ crystal structure in different protonated states, and found that if H61 is protonated a spontaneous transition to the OF state takes place over 450 ns, while the IF conformation is stable in standard protonation states and with E435 protonated, therefore suggesting coupling between proton binding and conformational changes (figure 1.5a).

Moving towards the explicit investigation of POTs conformational changes with enhanced sampling methods, Selvam et al. (2018) have used an aggregate of $\approx 54\,\mu s$ of adaptive seeding unbiased MD together with an Markov-state model (MSM) to construct a 2D-free energy surface (FES) of apo $PepT_{So}$ (in standard protonation states) along collective variables (CVs) describing the intracellular and extracellular gate opening distances (figure 1.5b). They found distinct free energy basins for all of the IF, OCC and OF states, allowing them to characterise the formed interaction networks, as well as a proportion of states that were wide open at both gates. The OCC state is characterised at the extracellular side by hydrogen bonds linking the N- and C-terminal domains involving N33 (not conserved), S165 (PepT2: T199), S320 (PepT2: S321) and N341 (PepT2: Q345) as well as the R32 (PepT2: K64)–D316 (PepT2: D317) saltbridge. At the intracellular side, the Y431 (PepT2: Y634)–S131 (not conserved) hydrogen bond contributes to stability. The OF state is similar to OCC at the intracellular side, but opens the extracellular gate wide, involving a new interaction between D316, N454 (PepT2: N657) and H61 (PepT2: H87) as well as between K318 (not conserved) and E419 (PepT2: E622). The authors compare their obtained conformational distribution to DEER data, although they do not provide histograms or distribution peaks and only check for overlap of the observed range of values. Likely problems with sampling or the choice of CV space in this study are also indicated by the fact that experimental structures fall far outside the free energy basins in the reported PMF (see the black dots in figure 1.5b).

To investigate how proton and ligand binding impact the relative occupancies of POT conformational states in free energy landscapes, Batista et al. (2019) have used $2\,\mu s$ of adaptive biasing force (ABF) simulations of $PepT_{St}$, each along helical tip distance CVs describing the cytoplasmic and periplasmic gate opening distances, for all combinations of E300 (PepT2: D317) (de)protonation and the presence / absence of an Ala-Phe substrate (figure 1.5c). In the apo-E300-deprotonated state, only IF conformations are accessible,

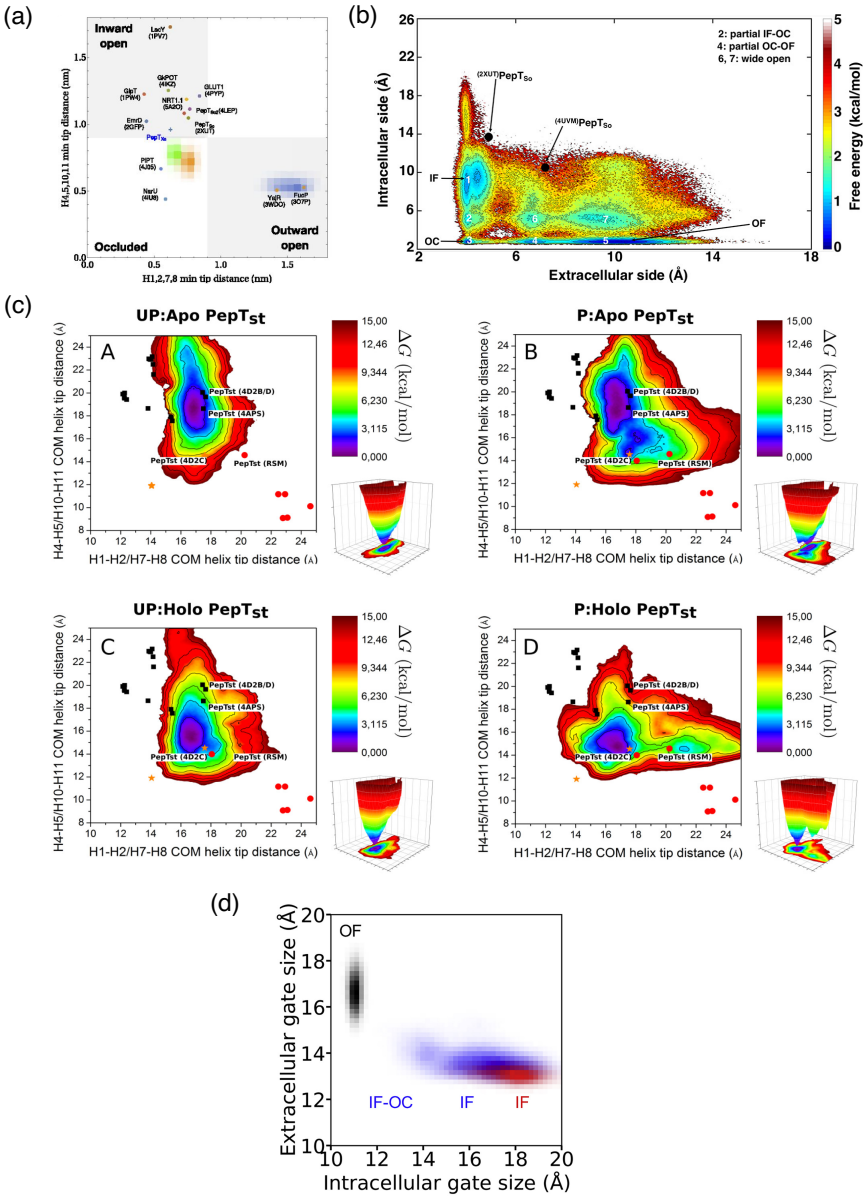


Figure 1.5: (a) Spontaneous conformational transition of $PepT_{Xc}$ in unbiased MD to OF following H61 protonation. Reprinted from Parker et al. (2017) under the CC BY-NC-ND 4.0 license. (b) 2D PMF of $PepT_{So}$ alternating access, showing distinct IF, OCC and OF states. Reprinted from Selvam et al. (2018) under the Standard ACS AuthorChoice/Editors' Choice usage agreement. (c) 2D PMFs of $PepT_{St}$ alternating access, showing the dependence of the conformational preferences on the presence of substrate and the protonation state of E300. Reprinted with permission from Batista et al. (2019). Copyright (2019), American Chemical Society. (d) Spontaneous conformational transition of $PepT_{Sh}$ to OF in unbiased MD (black histogram), starting from the IF-OCC state (left edge of the blue histogram). Reprinted from Li et al. (2022), copyright (2022) with permission from Elsevier.

which broadens towards the OCC state when E300 is protonated. The presence of substrate leads to a shift of the FES towards OCC, thus leading to cytoplasmic-gate closure (while OF is still inaccessible). Protonation of E300 in the holo state lastly makes the OF conformation accessible as a separate free energy minimum. The OF state thus produced is relatively close to the repeat-swapped model, and meta-stable in unbiased MD (partially occludes in one of two 100 ns-long runs). The authors show how the R33 (PepT2: K64)–E300 salt bridge responds to E300 protonation. They also find two salt bridges, D79 (PepT2: D105)–K413 (not conserved) and D145–K347 (not conserved), which are specific for the OF state. The conclusion is that ligand binding predominantly affects the cytoplasmic gate, while E300 protonation disrupts the periplasmic gate. It should be noted that $PepT_{St}$ is not a 'mammalian-type' bacterial POT and so does not have the equivalent of H87 in PepT2.

To probe further the protonation processes at the ExxER motif, Aduri et al. (2019) have investigated the bacterial POT YbgH, which — like YjdL, see the discussion on Aduri et al. (2015) above — has the unusual sequence $QxxE^{21}Y$ at this site. They performed triplicate 100 ns-long unbiased simulations in all combinations of E21 (PepT2: E56) (de)protonation and the presence or absence of an Ala-Ala dipeptide substrate molecule. In the apo-E21-deprotonated condition, a stable salt bridge is formed between E21 and K118 (PepT2: K161) — an interaction which only transiently forms when E21 is protonated. In the presence of substrate with E21 deprotonated, K118 prefers to interact with E21, but switches conformation to engage the substrate C-terminus when E21 is protonated. The authors note that E21 protonation stabilises substrate binding in this fashion. Looking at the possibility of large-scale conformational changes, they report that in the apo state, irrespective of the E21 protonation, the initial IF conformation quickly occludes, forming an R136–E406 salt bridge (not conserved) in the process. Such an occlusion is not observed in the presence of substrate. From their data, the authors propose a two-step peptide binding mechanism, where the peptide initially binds deeper inside the binding pocket near the E21–K118 salt bridge, before E21 protonation produces the stable substrate-bound conformation by releasing K118 to form a salt bridge with the substrate C-terminus.

The first mammalian POT structure was solved by Parker et al. (2021) for PepT2, also representing the first view of an OF state in a POT. After deriving and characterising an inhibitory nanobody, the authors solved an apo cryo-EM structure. A principle component analysis (PCA) of all MD conducted in the paper, as subsequently described, showed that the ECD moves largely as a rigid body with little internal flexibility. The authors then validate in transport assays the importance of the intracellular salt bridge D170–K642. They next used a protocol of docking and MD equilibration (4 replicates of 100 ns, then clustering to obtain stable poses) to place 5 di- and tripeptide substrates in the binding pocket. They note a consensus binding pose involving clamping of the charged termini by E622 and R57, as well as contributions from N192, N348, Y61, Y94 and D317 and W313 to accommodate certain side-chains. From these poses, they ran ABFE simulations, obtaining reasonable correlation (Pearson $r = 0.83$) with experimental IC_{50} values. They also calculated PMFs using US (triplicate 100 ns / window) of Ala-Ala-Ala unbinding, comparing paths generated with steered MD (SMD) to disengage either the N- or C-terminus of the substrate first. The results show that C-terminus unbinding while the N-terminus remains bound is easy, while a free energy penalty is incurred in the reverse ordering. This indicates that the substrate likely enters with its N-terminus first. The authors also show how the prodrugs valacyclovir and valganciclovir inhibit substrate uptake by PepT2, and — using their docking and equilibration protocol — find that they bind with their N-terminus in a similar position to peptides, while the C-terminus reaches further into the binding pocket into the vicinity of the ExxER glutamates.

To study the substrate and proton coupling of protein structural changes in $PepT_{St}$, Immadisetty and Moradi (2021) have run sets of microsecond-long unbiased MD simulations starting in the OCC-apo, IF-apo and inwards-facing OCC Leu-Ala bound states with different protonation-state combinations. For each of the apo-states, they have probed the protonated and deprotonated versions of E300 (PepT2: D317), while for the holo simulation they trialled all possible combinations of E299 (not conserved), E300 and E400 (PepT2: E622) (de)protonation. The authors first investigated the previously proposed R33 (PepT2: K64)–E300 salt bridge, and found that it forms as expected when E300 is deprotonated, but disengages when E300 becomes protonated. They also notice that the presence of substrate has a somewhat destabilising effect on the salt bridge, induced by the substrate being able to interact with E300. Since E299 can become an alternative interaction partner for R33, the authors propose that this residue also needs to become protonated to enable a full transition to OF. They also report that the substrate was not very stable in the inward-facing OCC state, driving full opening towards IF. Further analysing the substrate position within the transporter, the authors found that the substrate sits stably in the binding pocket only when E300 is protonated. When E300 becomes deprotonated, it moves further towards the cytoplasmic side of the transporter, which they attribute to the formation of the R33–E300 salt bridge pushing the substrate down (which also stabilises the protein in the IF state). This position further down into the transporter includes key interactions with E400 and E299. The authors observe that to accomplish full transport of the substrate towards the intracellular side, these residues need to become either protonated or coordinated by Na^+ ions. They also found some movement of the overall transporter conformation in dependence on protonation state changes in a PCA analysis. From the data thus collected, they propose a transport mechanism where protonation of E299 and E300 opens $PepT_{St}$ towards OF. When the substrate engages, this occludes, whereafter the substrate gets released by movement of the protons down from E300 to E400. The authors highlight that the fact that the conformational states sampled in their simulations are distinct from the input crystal structures may suggest the importance of further protonation states they have not considered.

Li et al. (2022) have investigated proton coupling of conformational changes in the non-'mammalian-type' (no H87 equivalent) $PepT_{Sh}$. Starting at the crystal structure bound to the natural S-Cys-Gly-3M3SH substrate, microsecond-long unbiased MD shows how the protonation of E418 (PepT2: E622) is important for maintaining the crystallographic binding pose; when E418 is deprotonated the ligand moves further down towards the cytoplasmic side through coordination of E418 (and remains stably associated over $3.5\,\mu s$). Furthermore, protonation of E418 in this equilibrated state triggers ligand release to the cytoplasm in unbiased MD. To investigate the role of the R44 (PepT2: K64)–E311 (PepT2: D317) salt bridge in $PepT_{Sh}$ conformational changes, the authors conducted $3.5\,\mu s$ of simulation in the E311 deprotonated state and found the salt bridge to be stable, with the transporter remaining in IF most of the time and also transiently sampling an inwards-facing OCC conformation. When E311 is protonated in the IF state, the authors observed no transitions in triplicate $1\,\mu s$-long MD, but starting in the inwards-facing OCC conformation they found transitions to OF (figure 1.5d) in two out of three replicates (the other replicate transitioned back to IF). From this, they conclude that dehydration of the cytoplasmic opening may be the rate-determining step; once this has happened in the inwards-facing OCC state, transition to OF is quick once R44–E311 is disrupted. To test the hypothesis that the proton at E311 (holding the OF state open) would move to E418 to trigger both extracellular gate closure and ligand release, the authors next investigated the E311 $\leftrightarrow$ E418 proton transfer using reactive MD combined with metadynamics (metaD). The obtained PMFs are similar in the IF and OF states, with a slight raising of the E311 well induced by proximity to R44. On the other hand, the process is more strongly modulated by the presence of substrate at E418 and vice versa, raising

the proton free energy at E418 in the holo state, and increasing the ligand distance from E418 when the proton resides at E418. The calculated kinetics on the order of magnitude of microseconds suggest that this proton-transfer would not be rate-limiting to the transport cycle. While providing information on the coupling of protons with the substrate and conformational changes, the authors acknowledge that a further investigation of the ExxER motif protonation states required conformational free energy landscapes (in most simulations, they treated E33 [PepT2: E53] as protonated; when they did deprotonate it they found a higher cytoplasmic gate-flexibility towards exploration of OCC states).

Most recently, published after our study on conformational changes and substrate and proton coupling in PepT2 (chapter 4 of this thesis), Selvam et al. (2024) have extended their investigation of the conformational landscape of apo $PepT_{So}$ using MSMs by including Ala-Ala dipeptide molecules in their simulations. From $85\,\mu s$ of adaptive seeding MD, using the substrate z-coordinate as a CV alongside descriptors of the extra- and intracellular gate opening, the authors constructed a holo conformational landscape of $PepT_{So}$. They found a similar overall profile, but with key differences. The barrier between the OCC and IF states was decreased by about $1\,kcal\,mol^{-1}$, and the profile was generally shifted to wider extracellular gate-opening distances, about which they comment that "These findings together indicate that the presence of the Ala-Ala peptide within the transport tunnel widens the extracellular gate". They then provide a projection of their FES onto the first two time independent component analysis (TICA) components, to show that the states corresponding to OF/OCC and OCC lie close together but are kinetically further from the IF states. The authors then use transition path theory to identify the interactions picked up by the substrate during translocation through the transporter, including D316 (PepT2: D317) and R32 (PepT2: K64) during initial capture, and then R25 (PepT2: R57) engaging the peptide C-terminus and S423 (PepT2: S626) engaging the N-terminus. Lastly, they observed an ordered assembly of three water molecules between D316 and E419 (PepT2: E622), which they argue provides insight into the mechanism of proton transport in $PepT_{So}$. There are, however, substantial problems with this study as it stands as a preprint at the time of writing this thesis. Firstly, the authors do not describe any interactions between the peptide and E419 in their figures or the main text, although this residue has been identified by a number of structural, biochemical and computational studies working on a range of POT homologues to be the key site coordinating the substrate N-terminus (see sections 1.1.3 and 1.1.4 of this introduction). If their simulations indeed show peptide translocation without engaging E419 that either constitutes novel mechanistic insight which the authors have not commented on, or indicates a sampling problem of peptide orientations, possibly arising from only using the z-coordinate of the peptide as a CV in the adaptive seeding procedure. In effect, in the scenario the failure to accelerate a relevant degree of freedom (DOF) in the adaptive seeding procedure that captures binding to E419, would have led to incomplete phase space exploration of the peptide orientations. Moreover, the comparison with the apo PMF published earlier by the same group (Selvam et al., 2018) appears to be fundamentally flawed, because the two studies do not keep the protonation state of the key D316 residue constant. Selvam et al. (2018) write that "The standard protonation states was [sic] used for the titratable groups", while Selvam et al. (2024) explain that "Consistent with established experimental data, Asp316 in helix 7 was maintained in its protonated state". The comparison between the two PMFs is therefore not only an apo/holo comparison, but conflated with the protonation-state change of a key residue that has — in other POTs — been shown to modulate the conformational landscape. (Batista et al., 2019; Immadisetty and Moradi, 2021; Li et al., 2022) This could provide an alternative explanation for the general shift of the PMF to wider extracellular gate openings, and draws into question the conclusions obtained by the study.

In summary, the existing computational literature on POTs (which — with the exception of Parker et al. (2021) — has investigated bacterial homologues) implicates the K64–D317 salt bridge in the extracellular gate of transporters that do not conserve H87, and H87 itself in interaction with D317 for those transporters that do. There is, furthermore, some evidence for coupling between ExxER motif protonation and the bound substrate via R57, as well as for a role of the substrate in regulating the OCC $\leftrightarrow$ IF transition. An investigation of the proton-coupled thermodynamic driving forces for alternating access in mammalian POTs, however, is still missing, as is a thermodynamic description of substrate-proton coupling and the nature of the intracellular gate. Lastly, the molecular mechanisms underpinning substrate promiscuity and selection of some substrates over others have not to date been investigated using molecular simulations.

1.4 Project aims and outline of chapters

Despite substantial previous work spanning biochemistry, structural biology, synthetic chemistry and computational investigations, questions in a number of key areas remain open concerning the alternating access mechanism of POTs, their substrate promiscuity and pharmaceutical applications for oral drug delivery. In this thesis, we use molecular simulations combined with cell-based transport assays, molecular docking and machine learning techniques to provide (partial) answers to the following questions.

1. How does the protonation of key titratable residues couple to conformational changes involved in alternating access? Can a sequence of successive protonation events be derived that captures the thermodynamic driving forces underpinning the full alternating access cycle?
2. What is the nature of the extracellular gate in mammalian POTs? Does the distinction between bacterial homologues that are 'mammalian-type' by containing an equivalent to H87 and those that are not translate to the mammalian proteins? Are there any additional components of the extracellular gate in mammalian POTs that are not conserved in prokaryotic homologues?
3. What constitutes the intracellular gate and its triggers in POTs? Is it conserved across bacterial homologues and the mammalian proteins?
4. How do alternating-access conformational changes couple to the presence of substrate?
5. What is the function of the conserved ExxER motif? How does it couple proton movement to substrate binding?
6. What distinguishes substrates that are transported well from those which are not? What sets apart substrates from inhibitors?
7. Is is possible to predict the affinity of ligands at POTs using computational methods? Which methods perform best, and is explicit structural information regarding ligand binding poses required?
8. Can the attachment of amino acid moieties to poorly bioavailable β-lactam antibiotics turn them into POT substrates? What can computational methods contribute predictively to the design process?

We provide theoretical background and some practical considerations for running MD simulations in chapter 2, before introducing a novel method (which we term MEMENTO) for managing hysteresis problems in enhanced sampling of large-scale conformational changes, as required in this thesis. The method is validated on four example systems, including the membrane transporter LeuT (chapter 3). We then derive, using a range of MD-based methods validated by cell-based transport assays, a model of the alternating-access cycle of PepT2. This includes a full description of the thermodynamic driving forces

that couple large-scale conformational changes to proton movement and substrate binding, a detailed discussion of the extra- and intracellular gates as well information regarding the role of the ExxER motif in coupling substrate binding to protonation-state changes (chapter 4). Next, we provide — through MD simulations in collaboration with structural biologists — a more detailed view of the interplay between ligand binding and ExxER-motif protonation, providing an important point of distinction between high-affinity and low-affinity substrates and inhibitor (chapter 5). We also benchmark a range of computational methods for protein–ligand affinity prediction against available experimental data on ligand binding to PepT2, highlighting the crucial role of diverse experimental datasets as well as the continuing challenge of accurate binding pose prediction in complex binding pockets. In this context, we further present a promising though not yet successful attempt at computational predictive prodrug design for POT transport (chapter 6). Lastly, we summarise the main results obtained in this thesis and provide an outlook of further avenues of investigation (chapter 7).

2

Computational Methods

Disclaimer

Small sections on simulation methodology in this chapter draw on the introduction previously written for my J.C.T.C. publication, Lichtinger and Biggin (2023). All investigation, writing and figures in that paper were produced by myself with feedback from the co-authors, and will, with adaptations, be used in part here under the CC BY 4.0 license. Some phrasing may therefore be similar or partially identical to my published work.

This chapter provides the theoretical background and some general implementation considerations for molecular dynamics (MD) simulations, which form the basis of most research carried out for this thesis. We also introduce enhanced sampling, binding free energy calculations and constant-pH molecular dynamics (CpHMD). Details concerning simulation setup and specific parameter choices, as well as other computational and experimental methods where they were used, are provided in the appendices.

2.1 Molecular dynamics simulations

MD simulations, which integrate the classical-mechanical evolution of complex molecular systems in time using all-atom or coarse-grained representations with differentiable potential energy models (including both classical and quantum-mechanical Hamiltonians), allow both for detailed dynamical insight into the studied systems (functioning as a 'computational microscope') as well as for the calculation of macroscopic thermodynamic properties. They have matured into an essential tool to study complex processes in materials science, nanotechnology and biophysics at a level of dynamical detail often unattainable in experiments. (Steinhauser and Hiermaier, 2009; Hollingsworth and Dror, 2018) Not long after Alder and Wainwright (1957) carried out the first condensed-phase MD simulation of hard spheres, more complex systems like argon atoms (Rahman, 1964; Verlet, 1967) and later proteins (McCammon et al., 1977) were modelled and simulated with empirical potentials. Nowadays, with a range of forcefields available for proteins (Lopes et al., 2015) and other macromolecules (as discussed in section 2.1.3), simulation speeds of hundreds of nanoseconds per day are routinely accessible for small protein systems (on the order of 100,000 atoms) using commercial GPUs. On supercomputer architectures, simulations of boxes containing entire viral envelopes (consisting of hundreds of millions of atoms) have been reported (Durrant et al., 2020; Casalino et al., 2021), and dedicated simulation computers — though not widely available to the academic community — enable speedups on the order of 10^3 compared to commercial hardware. (Shaw et al., 2021) Within this thesis, we investigated the peptide transporter system as introduced in the preceding section using various flavours of MD simulations. Several excellent textbooks exist on the theoretical backdrop of MD, for example by Leach (2001) and Tuckerman (2023), and it cannot be hoped within the scope of this introductory chapter to give a comprehensive review of the statistical mechanics, mathematical derivations and algorithmic complexities that underpin the technique. Nonetheless, we will here give an brief overview of the key principles, simulation algorithms, empirical potentials and enhanced sampling techniques used in the research presented in this thesis. Where not otherwise stated, my discussion in section 2.1 shall be

based on the exposition in Tuckerman (2023).

2.1.1 Theoretical foundations

Quantum mechanics

Microscopic systems evolve in time according to the laws of quantum mechanics, which may be stated as the time-dependent Schrödinger equation (Griffiths, 2005):

$$\hat{H}|\Psi(t)\rangle = i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle, \tag{2.1}$$

or, for the calculation of energies, as the time-independent variant

$$\hat{H}|\Psi(t)\rangle = E|\Psi(t)\rangle, \tag{2.2}$$

where $\hat{H}$ is the Hamiltonian operator, E is the energy and $|\Psi(t)\rangle$ is the wave function, written as a vector in Dirac notation. The details of how one might go about approximating solutions to this equation for a given system are beyond the scope of our very brief discussion; it will suffice to say that for the size of system studied here and the required timescale for simulating biological processes it would be prohibitively computationally expensive. Since atomic nuclei are much heavier than electrons, it is common to separate the motion of nuclei and electrons by using the Born-Oppenheimer approximation. One could then treat the nuclei classically, moving according to Newtonian mechanics in a potential calculated quantum-mechanically as the ground-state electronic energy for clamped nuclei. This approach is called Born-Oppenheimer MD, and may be formulated as

$$\frac{\mathrm{d}\mathbf{P}_I}{\mathrm{d}t} = -\nabla_I V_k^{\mathrm{BO}}(\mathbf{R}_I(t)), \tag{2.3}$$

where $\mathbf{P}_I$ and $\mathbf{R}_I$ are the momenta and positions of the nuclei, respectively, and V_k^{BO} is an effective potential in which the nuclei move, arrived at using an (approximate) quantum-mechanical treatment of the electrons. (Marx and Hutter, 2009) Since this method is still too slow and expensive for most use cases in biophysics, it is common practice to approximate further, by using classical mechanics in an empirical potential (sometimes in combination with a subsystem treated at quantum level of theory in quantum mechanics / molecular mechanics (QM/MM)). The rest of this introductory section will be concerned only with those MD methods where all quantum effects are treated implicitly as part of the force field. While imposing limitations, for example that no bond breaking and formation reactions can be simulated, the classical treatment still produces manifold insights into the molecular mechanisms underpinning biology.

Classical mechanics

The two central tenets of classical mechanics are encapsulated in Newton's second and third laws as published in his 1687 *Principia* (the first law relating to the movement of bodies without external forces can be considered merely a special case of the second, and is thus omitted here).

$$\mathbf{F} = m\mathbf{a} = m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} \tag{2.4}$$

$$\mathbf{F}_{AB} = -\mathbf{F}_{BA}, \tag{2.5}$$

that is, the acceleration **a** or second time derivative of a body's position **r** is proportional to the force **F** acting on it (second law), which can also be formulated in terms of the momentum $\mathbf{p} = m\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}$. Furthermore,

for every force acting on body A by body B there is an equal opposite force acting on body B by body A (third law). With a sufficient set of initial conditions and a specification of how to determine inter-particle forces, these equations determine fully the time-evolution of a system containing N particles. In other words, a trajectory of points $\{\mathbf{r}_1(t), \dots, \mathbf{r}_N(t), \mathbf{p}_1(t), \dots, \mathbf{p}_N(t)\} = \mathbf{x}(t)$ is formed in $6N$-dimensional *phase space*.

Classical mechanics is also often formulated in terms of the *Lagrangian* $\mathcal{L}$ and the *Hamiltonian* $\mathcal{H}$. We will briefly state these here, since they are used in deriving results introduced below. The Lagrangian is the difference between the kinetic energy (calculated from the velocities) and the potential energy (calculated from the positions):

$$\mathcal{L}(\mathbf{r}_1, \dots, \mathbf{r}_N, \dot{\mathbf{r}}_1, \dots, \dot{\mathbf{r}}_N) = K(\dot{\mathbf{r}}_1, \dots, \dot{\mathbf{r}}_N) - U(\mathbf{r}_1, \dots, \mathbf{r}_N). \tag{2.6}$$

The equations of motion of the system can then be derived using the *Euler-Lagrange* equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{r}}_i}\right) - \frac{\partial \mathcal{L}}{\partial \mathbf{r}_i} = 0 \tag{2.7}$$

An equivalent formulation of classical mechanics starts with writing the Hamiltonian

$$\mathcal{H}(\mathbf{r}_1, \dots, \mathbf{r}_N, \mathbf{p}_1, \dots, \mathbf{p}_N) = K(\mathbf{p}_1, \dots, \mathbf{p}_N) + U(\mathbf{r}_1, \dots, \mathbf{r}_N) \tag{2.8}$$

as the total energy, from which it can be shown that

$$\dot{\mathbf{r}}_i = \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} \quad \text{and} \quad \dot{\mathbf{p}}_i = -\frac{\partial \mathcal{H}}{\partial \mathbf{r}_i}, \tag{2.9}$$

which reproduces Newton's second law.

The laws of thermodynamics

Thermodynamics evolved in the 19$^{\text{th}}$ century as a phenomenological theory to describe any processes that involve energy and heat transfer. As such, it is not based on or linked to a particular microscopic theory of matter, but rather describes diverse processes ranging from chemical and biochemical reactions to heat engines in terms of key macroscopic variables. Encapsulating the insight that heat is a form of energy, the *first law of thermodynamics* states that

$$\Delta E = \Delta Q + \Delta W, \tag{2.10}$$

that is, the internal energy of a thermodynamic system will change by ΔE with the heat absorbed by the system (ΔQ) and work done on the system (ΔW). This amounts to a law of conservation of energy that notably applies for both reversible and irreversible paths, as well as spontaneous and non-spontaneous processes.

To make this crucial distinction, the differential $\frac{\mathrm{d}Q}{T}$ (where T is the temperature) plays an important role. Calling the quantity transferred alongside heat in inverse proportionality to temperature *entropy* (S), the *second law of thermodynamics* states that for the thermodynamic universe composed of system and surroundings, the entropy

$$\mathrm{d}S_{\text{tot}} = \mathrm{d}S_{\text{sys}} + \mathrm{d}S_{\text{surr}} \geq 0 \tag{2.11}$$

must always stay the same (for a reversible process) or increase (for an irreversible process) in any thermodynamic transformation. Because the entropy is a state function, only its changes can be measured. The *third law of thermodynamics* is therefore required to set the entropy at absolute zero temperature to a constant, zero for convenience, that can never be physically reached (a property originating from quantum mechanics).

Since the entropy establishes the direction in which physical processes happen and thus puts an 'arrow' on time — while classical and quantum mechanics are time-reversible — this creates a conceptual tension known as *Loschmidt's paradox* that is resolved within the framework of statistical mechanics.

Thermodynamic ensembles

To link the language of thermodynamics with that of classical mechanics, we recognise that a given value of a macroscopic quantity (for example the temperature, related to the average kinetic energy) can be realised by a great number of *microstates*, that is, points in phase space. An *ensemble* of microstates thus refers to all those microstates that correspond to common values of (a set of) macroscopic variables. Using a microscopic function $a(\mathbf{x})$ to calculate A for a given point in phase space $\mathbf{x}$, we can obtain the thermodynamic value of A as an ensemble average over the $\mathcal{Z}$ microstates in the ensemble

$$A = \frac{1}{\mathcal{Z}} \sum_{\lambda=1}^{\mathcal{Z}} a(\mathbf{x}_\lambda), \tag{2.12}$$

which can also be denoted by angle brackets as $\langle a \rangle$. It is possible to show that the *phase-space distribution function* $f(\mathbf{x}, t)$, which gives the fraction of ensemble members in a certain region of phase space, is proportional in general to a function of the Hamiltonian $\mathcal{F}(\mathcal{H}(\mathbf{x}))$:

$$f(x) = \frac{1}{\mathcal{Z}} \mathcal{F}(\mathcal{H}(\mathbf{x})), \tag{2.13}$$

where the normalisation factor $\mathcal{Z}$ given by

$$\mathcal{Z} = \int \mathrm{d}\mathbf{x} \mathcal{F}(\mathcal{H}(\mathbf{x})) \tag{2.14}$$

is known as the *partition function* as a measure of the quantity of microstates in a given ensemble (the integral is taken over all phase space here). Thermodynamic observables appear as partial derivatives of the partition function, and other equilibrium observables can be can be calculated as:

$$A = \langle a \rangle = \frac{1}{\mathcal{Z}} \int \mathrm{d}\mathbf{x}\, a(\mathbf{x})\, \mathcal{F}(\mathcal{H}(\mathbf{x})). \tag{2.15}$$

The nature of $\mathcal{F}$ depends on the details of the ensemble of interest. Ensembles are labelled according to those macroscopic thermodynamic variables that are held fixed in their definition — some common examples with relevance to MD will now be introduced. The NVE ensemble (set particle number N, volume V and energy E) is also referred to as the *microcanonical* ensemble. It is commonly assumed that all microstates thus specified have equal *a priori* probabilities, that is, the system is equally likely to be found in any of them. While this assumption cannot be rigorously proved, it produces results consistent with expectations if used in statistical mechanics. If instead of the energy the temperature is set, the resulting NVT ensemble is called the *canonical* ensemble. The *isothermal-isobaric* or NPT ensemble does not set the volume but instead a value for the pressure. As an alternative to pre-defining the number of particles, it is also possible to set the chemical potential μ in the μVT, or *grand canonical*, ensemble. Table 2.1.1 summarises the state functions and functional forms of the partition functions of these ensembles (without

regard to the values of the normalisation constants).

Ensemble	State function	Partition function
microcanonical (NVE)	$S = k \ln \Omega$	$\Omega \propto \int \mathrm{d}x \delta(\mathcal{H} - E)$
canonical (NVT)	$A = E - TS = -\frac{1}{\beta} \ln Q$	$Q \propto \int \mathrm{d}x e^{-\beta \mathcal{H}}$
isothermal-isobaric (NPT)	$G = H - TS = -\frac{1}{\beta} \ln \Delta$	$\Delta \propto \int_0^\infty \mathrm{d}V e^{-\beta PV} Q$
grand canonical (μVT)	$\Phi = E - TS - \mu N = -\frac{1}{\beta} \ln \mathcal{Z}$	$\mathcal{Z} = \sum_{N=0}^\infty e^{\beta \mu N} Q$

Table 2.1: Summary of common thermodynamic ensembles. Here, β refers to $(kT)^{-1}$, A is the Helmholtz free energy, G is the Gibbs free energy, $H = E + PV$ is the enthalpy and Φ is the grand potential.

Molecular dynamics and ergodicity

Because motion in classical mechanics conserves the total energy of a system as well as the number of particles and — if set up in a box of fixed size — also the volume, there exists a natural link between the propagation through time according to Newton's or Hamilton's equations and the microcanonical ensemble. In an MD simulation, one numerically solves the equations of motion subject to a set of initial conditions, often seeded from an experimental structure and with velocities drawn from the Maxwell-Boltzmann distribution. If the simulation trajectory is able to visit all states in phase space that lie on the ensemble hypersurface, it is called *ergodic* — a property that can be hard to obtain in systems with barriers larger than the thermal energy available (we will return to this issue when discussing enhanced sampling below). From an ergodic trajectory, the ensemble average of a variable A can be obtained as a time-average across the trajectory:

$$A = \langle a \rangle = \frac{1}{M} \sum_{n=1}^{M} a(\mathbf{x}_{n\Delta t}), \tag{2.16}$$

where a simulation of M steps of discrete time step Δt was performed. Next to the calculation of equilibrium properties in this way, MD simulations are also useful for extracting dynamical information from trajectories. To achieve either of these aims in practice, however, we need accurate algorithms to propagate the equations of motion discretely in time (integrator), taking care of constraints where required, as well as to fix thermodynamic quantities such as temperature (thermostat) and pressure (barostat), in cases where we want to not only simulate in the microcanonical ensemble. Furthermore, we require a force field that describes the inter-particle interactions for force calculations with a sufficient accuracy to gain insight into biological processes of interest. [1] In the following sections, all of these will be briefly introduced.

[1] In a practical investigation, a researcher often has to make a trade-off between the level of theory or detail included in the system Hamiltonian and the requirement for obtaining enough sampling of ensembles of interest for statistical accuracy and an approximation to ergodicity within relevant degrees of freedom, for example in the calculation of ensemble averages of observables or free energies. The first such trade-off is usually using classical force fields instead of a quantum-mechanical treatment of the electrons, but further choices are required for example with regard to polarisability (where in application cases such as ions in transporters, the gain in accuracy is sometimes noticeable when comparing with experiment) or further coarse-graining (which is usually necessary to reach sufficient simulation timescales to sample lipid dynamics, if those are of particular interest to a given investigative project).

2.1.2 Algorithms

Energy minimisation

Before presenting the algorithms required to run MD simulations to reproduce the various thermodynamic ensembles, it is necessary to further consider the initial conditions for setting up a simulation. As stated in the previous section, starting coordinates of a protein are usually taken from a resolved structure (or, more recently a predicted structure from tools like alphafold, Jumper et al. (2021)). Usually, such structures do not include hydrogen atoms and can be missing some loops that need to be modelled to have a full representation of the protein as atomic coordinates. Furthermore, solvent and ions (and lipids, for membrane proteins) that are required for making boxes which approximate the biophysically relevant bulk medium are usually not resolved structurally, so that the protein is commonly placed in pre-equilibrated solvent and lipid boxes (for example using tools like CHARMM-GUI, Wu et al. (2014)). The resulting atomic coordinates are therefore most likely not near a local potential energy minimum with regards to the MD force field, and the large forces resulting from such high-energy configurations can lead to instabilities when starting simulations. Therefore, it is common practice to run an energy minimisation before starting MD from an assembled box.

A range of energy minimisation methods exist, as reviewed in chapter 5 of Leach (2001), both with and without using derivative information of the potential. Since the potential in molecular dynamics is differentiable to allow for the necessary force calculations, derivative-based methods are popular to minimize energy in simulation boxes. The simplest, first-order minimization algorithm is steepest descent, where from coordinates at a step $\mathbf{r}_n$, the next set of coordinates is computed as [2]

$$\mathbf{r}_{n+1} = \mathbf{r}_n + \frac{\mathbf{F}_n}{\max(|\mathbf{F}_n|)} h_n, \tag{2.17}$$

where h_n is the maximum displacement and $\max(|\mathbf{F}_n|)$ means the largest scalar force on any atom. If the potential energy decreases with the step it is accepted and the maximum displacement is increased for the next step. Otherwise, the move is rejected and the maximum displacement is decreased.

More sophisticated algorithms like conjugate gradient or second-order methods like the Newton-Raphson algorithm can converge quicker than steepest descent in many cases. However, since a perfect local minimum is not generally required for starting MD simulations and the steepest descent method is robust to remove the highest energy clashes of a molecular geometry, it is used almost exclusively for the purpose of pre-simulation minimisation in the MD field, and thus was our method of choice as this thesis is concerned.

Integrators

A numerical solver, also called integrator, for Newton's equations of motion in a chaotic system needs to be time reversible, symplectic and conserve the volume element of phase space, while being as accurate as possible as measured by the long-term conservation of energy and allowing for a stable simulation with the largest possible time step. In GROMACS (Abraham et al., 2015), the MD engine used for most simulations in this thesis, the default integrator is the leap-frog algorithm, see the discussion in the reference manual

[2] This formulation is taken from the GROMACS reference manual, https://manual.gromacs.org/nightly/reference-manual/algorithms/molecular-dynamics.html.

(as referenced in footnote 2). The positions and velocities of each atom are propagated staggered and separated by $\Delta t/2$ as follows:

$$\mathbf{v}(t + \frac{\Delta t}{2}) = \mathbf{v}(t - \frac{\Delta t}{2}) + \Delta t \frac{\mathbf{F}(t)}{m} \tag{2.18}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \mathbf{v}(t + \frac{\Delta t}{2}). \tag{2.19}$$

This means that the kinetic energy and other properties related to atom positions are calculated half a time step out of sync with the atom positions. While the trajectories generated are identical to alternative formulations like the velocity Verlet algorithm (which works in integer time steps), the latter can be used if, for example, very accurate barostat coupling is required. In the work conducted for this thesis, we chose the default leap-frog algorithm for maximum compatibility with all GROMACS features. Note that for the alchemical free energy simulations, see section 2.3, we used the stochastic dynamics integrator instead of leap-frog. Since it doubles as thermostat, we discuss this integrator in the section on thermostats below.

Constraints

An integrator will update all atom coordinates according to the forces provided. However, it is often convenient to introduce constraints in simulations to eliminate explicit high-frequency motions, for example in bonds to hydrogen atoms that have fast vibrational frequencies because of the small mass of a hydrogen atom (in the simulations conducted for this thesis, we always treated bonds to hydrogen in this way). We can write an example constraint for the bond length b between atoms 1 and 2 as:

$$(\mathbf{r}_1 - \mathbf{r}_2)^2 - b^2 = 0 \text{ or equivalently } |\mathbf{r}_1 - \mathbf{r}_2| - b = 0. \tag{2.20}$$

Here, we use the LINCS algorithm for solving constraints (as implemented in GROMACS). (Hess et al., 1997) In short, LINCS is applied after an unconstrained coordinate update by the integrator. Firstly, atoms in a constrained bond that have moved with respect to each other in the coordinate update are moved in the direction of the original bond direction so that the projection onto the original bond conserves the bond length. Then, a further correction is applied for the amount the rotation of the bond has lengthened it. This step would need to be iterative, since the correction incurs another small rotation, however the accuracy is usually high enough.

Thermostats

To move beyond the microcanonical ensemble, the temperature of the simulation system must be fixed in such a way that the correct (Boltzmann) energy distribution for the canonical ensemble is reproduced, which requires mimicking the effect of a thermal reservoir. The temperature of a simulation box is related to the kinetic energy of its particles via:

$$E_{\text{kin}} = \frac{1}{2} N_{\text{f}} k T, \tag{2.21}$$

where N_{f} is the number of degrees of freedom, calculated as three times the particle number minus number of constraints and degrees of freedom reduced by eliminating centre-of-mass motion.

Several algorithms exist for the purpose of sampling at constant temperature, including the Andersen (Andersen, 1980) and Nosé-Hoover (Nosé, 1984; Hoover, 1985) thermostats. In the simulations conducted for this thesis, we used the velocity-rescaling thermostat with stochastic term by Bussi et al.

(2007), which is based on the Berendsen thermostat (Berendsen et al., 1984) but produces the correct canonical ensemble. The algorithm of this thermostat will now be briefly introduced.

In Berendsen temperature coupling, the kinetic energy of all particles is re-scaled at a certain interval of steps with a factor constructed so that the result is a first-order decay process of the temperature:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau}, \tag{2.22}$$

where τ is a time constant. This suppresses any thermal fluctuations, and so the simulation — while always near the target temperature T_0 does not sample the canonical ensemble. A revision of this formalism (Bussi et al., 2007) that includes a stochastic term can be expressed as the differential of the kinetic energy in terms of a Wiener noise:

$$\mathrm{d}K = (K_0 - K)\frac{\mathrm{d}t}{\tau} + 2\sqrt{\frac{KK_0}{N_\mathrm{f}}}\frac{\mathrm{d}W}{\sqrt{\tau}}. \tag{2.23}$$

Switching off the noise $\mathrm{d}W$ reproduces the Berendsen algorithm, while with the noise the thermostat quickly equilibrates the temperature if it is far from the desired value, but in its proximity the noise-dependent term dominates and a canonical distribution is produced. The coupling parameter τ interpolates between instant thermalisation for small τ and Hamiltonian dynamics at $\tau \to \infty$.

An alternative approach to temperature coupling is to use stochastic Langevin dynamics, where a friction and noise term are included in a modification of the equation of motion of each particle (Goga et al., 2012):

$$m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \mathbf{F}(\mathbf{r}) - m\gamma\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} + \eta, \tag{2.24}$$

where γ is a friction constant and η is a noise process with an autocorrelation function

$$\langle \eta(t)\eta(t+\tau)\rangle = 2\frac{kT}{m}\gamma\delta(\tau). \tag{2.25}$$

This is applied separately to all degrees of freedom without correlation between them. When $1/\gamma$ is large compared to the timescales of the simulated processes, this effectively gives stochastic temperature coupling (while longer time-scale processes are dampened). In alchemical free energy simulations, it is important to generate the correct ensembles for both the coupled and decoupled-ligand states. (Aldeghi et al., 2018) Since in the decoupled state few or no interactions are made with the rest of the system, a separate temperature coupling of the ligand degrees of freedom is required. Stochastic dynamics takes care of this conveniently because it operates on all degrees of freedom individually, hence we use it to replace the thermostat in our binding free energy calculations.

Barostats

The experimental conditions we seek to replicate in MD simulations are often at atmospheric pressure instead of fixed volume, so a pressure coupling algorithm is required. The instantaneous internal pressure in a simulation box can be calculated from the kinetic energy and forces on the particles in the virial form [3]

$$P = \frac{2}{3V}E_\mathrm{kin} + \frac{1}{3V}\sum_i \mathbf{r}_i \cdot \mathbf{F}_i. \tag{2.26}$$

[3] Using the isotropic formulation given as Eq. 1 of De Miguel and Jackson (2006) here for simplicity.

Pressure coupling can then be achieved by scaling the box to match a desired external pressure. The Berendsen barostat (Berendsen et al., 1984) achieves this equivalently to the Berendsen thermostat as a first-order decay. Using a scaling factor

$$\lambda = 1 - \kappa \frac{\Delta t}{\tau_P}(P - P_0), \tag{2.27}$$

where κ is the isothermal compressibility of the system and τ_P is the pressure relaxation time constant, leading to a rate of pressure change

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{P_0 - P}{\tau_P}. \tag{2.28}$$

Similarly to the thermostat equivalent, this barostat produces the right average pressure but not the correct ensemble, and in this thesis we use it only for box equilibrations. Our production simulations use the Parrinello-Rahman barostat (Parrinello and Rahman, 1981), which is a generalisation of the Andersen barostat (Andersen, 1980) to anisotropic pressures. The idea is to modify the system Hamiltonian, making the volume (in the form of the box dimensions) a dynamic variable:

$$\mathcal{H} = E_{\text{pot}} + E_{\text{kin}} + \frac{p_V^2}{2W} + PV, \tag{2.29}$$

here stated for the isotropic case with the momentum associated with the volume p_V and the parameter W as:

$$W = (3N + 1)kT\tau_b^2, \tag{2.30}$$

with τ_b being a time scale for the volume motion. From this, new equations of motion for the particles (the coordinates of which are also scaled) and the box volume arise, giving rise to pressure coupling in the correct NPT ensemble if an adequate thermostat is also employed.

Periodic boundary conditions, non-bonded cutoffs, neighbour lists and long-range electrostatics

Aside from these considerations of the underlying mechanics and thermodynamics, many further algorithmic challenges arise when running an MD simulation. Of these, I will briefly discuss here periodic boundary conditions, the approximation of non-bonded cutoffs, the storing of neighbour lists and the treatment of long-range electrostatics. This exposition is informed by the treatment in Leach (2001), and as far as implementation is concerned by the GROMACS reference manual (see footnote 2).

Since the size (and therefore number of atoms) of the box is a crucial factor limiting simulation speed, one will always want to choose the smallest possible box for a given system of interest. Boundary effects between the explicit solvent and the surrounding vacuum would dictate the use of very large boxes for a centrally solvated protein to achieve bulk solution properties around the protein. The most common way to circumvent this issue is the use of *periodic boundary conditions*: when an atom leaves through one face of the box, it re-enters through an opposing face. There is therefore no solution–vacuum interface and the environment in the box will be much more bulk-like. It should be noted, however, that fluctuations across a length-scale larger than the box-size are suppressed with periodic boundary conditions, which can lead to finite-size effects. The particle interactions are then calculated using the *minimum-image convention*, where all particles only feel the effect of the closest copy of another particle, located in whichever neighbouring periodic image the distance is smallest. For this convention to be appropriate (that is, for an atom to not

see multiple copies of another atom), the box size / image distance needs to be at least twice as large as the non-bonded cutoff (as discussed below). Any box shape is permissible here as long as it is space-filling by translation. GROMACS supports general triclinic boxes, which allows for the use of box geometries including cubic / orthorhombic, hexagonal prism, rhombic dodecahedron and truncated octahedron. While the more spherical box-types are useful for minimizing the number of solvent molecules at constant image distance for soluble proteins, in this thesis we use the orthorhombic box type to include a membrane patch.

As discussed below in section 2.1.3, the force field includes bonded terms (the number of which scales linearly with the number of atoms N) as well as non-bonded terms (that scale with N^2). The latter usually include a short-ranged Lennard-Jones (LJ) potential that falls off quickly with r^{-6} and the much longer-ranged electrostatic potential that falls only with r^{-1}. To avoid calculating the LJ potential for all atom combinations, one can make use of its fast decay with distance (where, for example, only 1 % of interaction energy remains at 2.5σ) by introducing a cutoff. Usually this is combined with a potential shift to stabilise the energy as particles move in and out of the cutoff:

$$\begin{aligned} V(r) &= \mathrm{LJ}(r) - \mathrm{LJ}(r_c) \quad \text{if } r \leq r_c \\ V(r) &= 0 \quad \text{otherwise.} \end{aligned} \tag{2.31}$$

It is also possible to introduce a smooth switching function on the forces to avoid a discontinuity when the force drops to zero at the cutoff boundary, though this is not the default GROMACS behaviour because it is more computationally expensive and not required for energy conservation [4], and we do not use it in this thesis. By itself, this cutoff scheme does not reduce computational load much, since still all inter-atomic distances need to be calculated at every step. This can be avoided by the use of *neighbour lists*. A Verlet neighbour list (Verlet, 1967) stores for each atom those atoms that lie within a given cutoff (plus a buffer). When it comes to doing force calculations, only pairs of atoms in the neighbour lists are considered, and every set number of steps the neighbour list gets reconstructed. In the GROMACS implementation we use here, the buffer size and recalculation frequency of the neighbour list is assigned automatically with a certain energy error tolerance, calibrated on the small jumps in energy caused by particles moving into the cutoff distance while not in the pair list.

Because the electrostatic potential decays much slower with distance, a different treatment is required, especially since the range of these forces can easily be longer than half the box length. The electrostatic potential of a box of N particles that include all its periodic images can be written as a nested sum:

$$V = \frac{f}{2} \sum_{\mathbf{n}} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{|\mathbf{r}_{ij} + \mathbf{n}|}, \tag{2.32}$$

where $\mathbf{n}$ is a vector indicating the position of the periodic image in consideration (the inner sums exclude $i = j$ if $|\mathbf{n}| = 0$), and for convenience we write $f = \frac{1}{4\pi\epsilon_0}$. This sum converges very slowly and it is only conditionally convergent — that is, convergence depends on the order of positive and negative terms being considered. In this thesis we use the particle-mesh Ewald (PME) method for computing this sum, as implemented in GROMACS. (Darden et al., 1993) While a full derivation of PME is beyond the scope of this introductory section, the idea of the Ewald summation method (first developed for the treatment of crystals, Ewald (1921)) is to exploit the identity

[4] See also the discussion at https://manual.gromacs.org/documentation/current/user-guide/mdp-options.html.

$$\frac{1}{r} = \frac{f(x)}{r} + \frac{1-f(x)}{r} \tag{2.33}$$

to decompose the sum into two parts, one of which converges quickly in real space while the other converges quickly in reciprocal space. A choice of the complementary error function $f(x) = \mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} \mathrm{d}t$ achieves this goal, rendering to overall Ewald sum into

$$V = \frac{f}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \begin{cases} \sum_{\mathbf{n}=0}^{\infty\prime} \frac{q_i q_j \mathrm{erfc}(\alpha |\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|} \\ + \sum_{k\neq 0} \frac{1}{\pi L^3} q_i q_j \frac{4\pi^2}{k^2} \exp(-\frac{k^2}{4\alpha^2}) \cos(\mathbf{k}\cdot\mathbf{r}_{ij}) \\ - \frac{\alpha}{\sqrt{\pi}} \sum_{k=1}^{N} q_k^2 + \frac{2f\pi}{3L^3} |\sum_{k=1}^{N} q_k r_k|^2, \end{cases} \tag{2.34}$$

where α is a parameter to pick appropriately so reasonable cutoffs can be used on both real-space and reciprocal sums. The first term in this expression is the real-space sum in terms of box translation vectors $\mathbf{n}$, the second term is the reciprocal-space sum in terms of reciprocal vectors $\mathbf{k}$, and the third and fourth term are corrections to remove the self-interaction of the added Gaussians and to account for the surrounding medium. Unfortunately, this is still quite computationally expensive, scaling with N^2 for a constant α value. Instead of direct summation of the wave vectors in the reciprocal part, in PME the atomic charges are interpolated onto a grid, which allows the fast-Fourier transform algorithm to be used, reducing the scaling to $N \log N$.

2.1.3 Force fields

In the preceding sections, it was assumed that we can state a classical potential V of a certain arrangement of atoms in space, and have required that this potential be differentiable so that the forces on all atoms can be calculated for numerical integration of the positions and momenta as

$$\mathbf{F} = -\nabla V. \tag{2.35}$$

In molecular mechanics, this potential is provided by an empirical force field. We will first give a short overview of the functional forms used in the construction of both bonded and non-bonded terms, before presenting the force fields used in the research for this thesis.

The functional form of force fields

The interactions described with empirical force fields usually fall into bonded terms — including bond stretching, angle bending and several types of torsions — and non-bonded terms, comprising van-der-Waals forces and electrostatics.

$$V_{\text{total}} = V_{\text{bonded}} + V_{\text{non-bonded}} \tag{2.36}$$

The discussion here will draw, where not otherwise stated, on the presentation in Leach (2001), and figure 2.1 illustrates the various terms involved.

Bonded interactions The most elementary term in all force fields is the stretching of the bond between two atoms. An accurate functional form for this is the Morse potential

$$V_{\text{Morse}}(l) = D_e(1 - e^{-a(l-l_0)}), \tag{2.37}$$

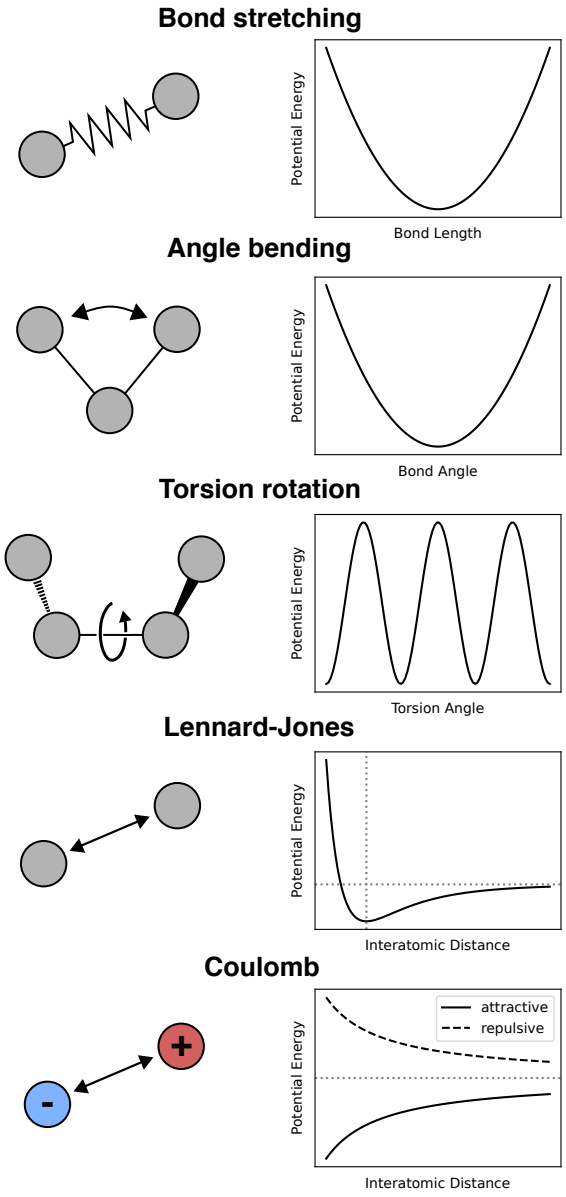


Figure 2.1: Schematic illustration of common functional forms in molecular force fields. Bonded terms are bonds, angles and torsions, non-bonded terms are Lennard-Jones and Coulomb.

where D_e is the depth of the potential well, l_0 is the reference bond length (which need not necessarily be the equilibrium bond length, if other force field terms act on this distance) and $a = \omega\sqrt{\mu/2D_e}$ scales the width of the well. This functional form is not particularly convenient computationally and in MD one does not usually model the extreme lengthening of bonds close to dissociation, so this is usually approximated by a harmonic potential with a force constant k_l:

$$V_{\text{harmonic}}(l) = \frac{k_l}{2}(l - l_0)^2. \tag{2.38}$$

The bending of angles involving three atoms is also usually described with a harmonic potential

$$V_{\text{angle}}(\theta) = \frac{k_\theta}{2}(\theta - \theta_0)^2. \tag{2.39}$$

Bond stretching and angle bending terms hold together the molecular topology, but the relative energies of different molecular geometries that satisfy a given set of bond lengths and angles cannot be described with these functional forms alone. To further capture the preferences associated with rotation around chemical bonds, dihedrals needs to be considered (as terms involving 4 atoms). The case of four atoms forming a linear bonded chain is called a proper dihedral, and the potential surface that describes the rotation around the central bond is usually rather complex. Therefore, a common way to model it is a cosine expansion:

$$V_{\text{prop. torsion}}(\omega) = \sum_{n=0}^{N} \frac{v_n}{2}[1 + \cos(n\omega - \gamma_n)], \tag{2.40}$$

where the relative contributions v_n and phases γ_n are chosen during force-field parameterisation to optimally match a reference, for example from a quantum-mechanical torsion scan. An improper dihedral, on the other hand, is a bending mode in which three atoms that are all connected to a central atom move along the normal vector of the arrangement. This is used to enforce planar geometries, for example in sp^2 hybridised carbon systems, and can be represented using a functional form like

$$V_{\text{improp. torsion}}(\omega) = k_\omega(1 - \cos(2\omega)). \tag{2.41}$$

Some force fields also include further bonded coupling interactions, for example Urey-Bradley terms, which are harmonic-spring 1,3 connections — i.e., direct coupling terms connecting the outer atoms of an angle — that were introduced largely to improve agreement with experimental vibrational spectra. Modern force fields (such as CHARMM since a modification of version 22 (MacKerell et al., 2004) and AMBER since ff19.sb (for all amino acids, Tian et al. (2020)) also include correction map (CMAP) terms, which provide a grid-based correction for backbone ϕ–ψ dihedral angle coupling on a 2D surface for more accurate secondary structure preferences.

Non-bonded interactions Bonded interactions connect atoms within the same molecule, but for a realistic force field one also needs to describe interactions between different molecules and between atoms of the same molecule that lie close in space but far from each other in bonding topology. [5] There are two classes of non-bonded interactions shared by most force fields: van-der-Waals and electrostatic interactions.

Van-der-Waals interactions as commonly treated in molecular force fields include interactions of two distinct quantum-mechanical origins. They incorporate an attractive term to represent dispersion

[5] The overlap with bonded interactions is handled via non-bonded exclusions, where non-bonded interactions are switched off for atoms that are within a certain number of bonds from each other in the molecule topology (on our cases, 3 bonds — that is, within the distance covered by dihedrals).

interactions, which stem from an instantaneous dipole in the electron cloud of one atom inducing a parallel dipole in the electron cloud of another atom. It can be shown either using second-order perturbation theory or from the Drude model of oscillating dipoles that these attractive forces scale as r^{-6}. The second term is repulsive and arises due to Pauli exclusion, which is the principle that two electrons cannot share all quantum numbers, so when they have the same spin they cannot occupy the same regions of space. This has a complex distance dependence involving an exponential term and the functional form used is usually chosen for computational convenience. The most common expression to model these interactions is the Lennard-Jones 12-6 potential:

$$V_{\mathrm{LJ}}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right], \tag{2.42}$$

which depends on two parameters ϵ and σ, and gives a lowest-energy inter-nuclear separation of $r_{eq} = 2^{1/6}\sigma$. Since providing parameters individually for all possible interactions between atom types is not feasible, one often merely sets σ and ϵ parameters for each atom type and combines them using mixing rules such as:

$$\sigma_{AB} = \frac{1}{2}(\sigma_A + \sigma_B) \tag{2.43}$$

$$\epsilon_{AB} = \sqrt{\epsilon_A \epsilon_B} \tag{2.44}$$

Electrostatic interactions in molecular force fields are treated with the Coulomb potential

$$V_{\mathrm{coul}}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.45}$$

for the interactions between charges i and j. Despite this simple functional form, much complexity arises when considering the placement of charges in space to model electrostatic interactions between molecules as arising from non-local electron clouds. The most commonly used model in non-polarisable force fields are fixed partial atomic charges, where fractional charges are located at the center of each atom. More complex approaches exist involving multipoles or partial charges located away from the centre of atoms, but providing details of these approaches is beyond the scope of the present introduction.

Force field parameterisation

The functional forms described in the preceding section are formulated in terms of a range of parameters that need to be empirically determined and assigned to each atom and bonded set of atoms in a simulation box prior to running any MD. Since it is unfeasible to generate a new force field from scratch each time a new system is simulated, force fields are usually formulated with a philosophy of transferability. Benefitting from the fact that the chemical environment of atoms is to a good approximation a local property, determined by surrounding atoms no more than a few bonds away, most force fields contain parameters that are simultaneously applied to a number of atoms in similar chemical environments (for example, side-chain parameters are generally re-used for the same amino acid). Usually this is accomplished in two conceptual stages. A general force field is first developed that includes parameters for all cases that might occur when setting up a simulation of a certain type of biomolecule. A protein force field, for example, will at least contain parameters for all 20 naturally occurring amino acids, while a small molecule general force field will contain transferable parameters for common functional groups and motifs. In the pre-processing stage of setting up MD, one then applies these parameters to a given simulation system. In the case of

proteins, the force field is usually complete in that it contains values for all required parameters, whereas the assignment of force field parameters for small molecules that are much more diverse in their chemical scaffolds usually involves a degree of approximation in transferring parameters from general sets and often requires some *ad-hoc* calculations (for example to determine partial atom charges).

Despite many incremental advances over the last 50 years, the design of force fields remains a highly complex endeavor. (Nerenberg and Head-Gordon, 2018; Van Der Spoel, 2021) The AMBER ff14.sb force field, for instance (which we used for most simulations conducted for this thesis), contains around 1,000 bonded and 100 non-bonded parameters in addition to tabulated partial charges. (Maier et al., 2015) Determining all of these parameters is an extremely non-trivial optimisation problem, and a panoply of design philosophies have been formulated and have contributed over decades to the iterative refinements of several major force field families such as CHARMM, OPLS and AMBER. Surveying these design processes in any amount of detail is far beyond the scope of this introductory chapter. We therefore contend ourselves here with noting that force field design and improvement with regard to bonded and non-bonded parameters can vary in terms of the functional forms employed and parameters chosen for optimisation (torsions, for example, are some of the most common targets for refitting, while other parameters are more often copied over from previous force field iterations), the kind of data to fit to (including structural information and quantum-mechanical calculations at a certain level of theory, often for short peptides in the case of protein force fields) as well as what benchmarks are used to evaluate force field performance (possibly the category with the least consensus approach in the field so far).

It is worth at this stage to highlight the role played by the concept of an *atom type* in assigning force field parameters to a given simulation box. Each atom in a protein, lipid or small molecule has a unique chemical environment (barring symmetry considerations), and to perfectly match this environment a new set of parameters would have to be derived for each bonded and non-bonded interaction. Therefore, one typically approximates each atom's chemical environment to correspond to a certain atom type, providing values for the non-bonded parameters for each atom type and substituting bonded parameters defined between a set of atom types. For a typical protein force field such as AMBER ff14.sb, there are around 50 atom types (representing the comparatively limited space of relevant molecular topologies) connected by highly optimised bonded interactions. In small molecule force fields (such as GAFF, Wang et al. (2004), and the newer version GAFF2), however, the number of atom types quickly increases to take care of the numerous ways in which general organic molecules can be connected. This can be problematic when largely redundant atom types need to be created and the copying of parent atom-type parameters introduces potential for error and can render systematic force fields improvements cumbersome. Mobley et al. (2018) have recently introduced a force field format called SMIRNOFF based on the SMIRKS chemical substructure selection language, using what they term a *direct chemical perception* as an alternative to atom type-encoded bonded parameters. This allows a much more compact statement of the force field and has been the basis for the OpenFF series of ligand force fields. (Boothroyd et al., 2023) It is worth noting here that — although not yet available at the time of writing — the planned OpenFF 3.0 (termed 'Rosemary') will apply this approach to both biopolymers and small molecules in a self-consistent manner.

While bonded and van-der-Waals non-bonded terms are usually fit to collectively match some experimental or higher-level computational data, partial charges are set to reproduce the electrostatic properties of a molecule of interest, either in tabulated form for the amino acids contained in proteins, or when assigning parameters for a new small molecule. In the restrained electrostatic potential (RESP) class of

methods (Bayly et al., 1993; Schauperl et al., 2020), partial atomic charges are fitted to reproduce the electrostatic potential at points outside the van-der-Waals molecular surface calculated from *ab initio* quantum mechanics. Another popular technique are AM1-BCC charges (Jakalian et al., 2000, 2002), calculated from the semi-empirical quantum-mechanical AM1 charges together with a correction to account for charge redistribution due to chemical bonds. Neither of these methods scales well with increasing sizes of molecules, prompting the more recent development of machine learning methods to derive partial atomic charges. (Bleiziffer et al., 2018; Wang et al., 2023)

Force fields used in this thesis

The systems of interest in this thesis mainly are membrane proteins, so that a protein force field and compatible lipid force field are required. Furthermore, peptide and non-peptide ligands are simulated, requiring the use of a small-molecule force field, and lastly water and ion force fields are also required to fill up the simulation boxes with realistic bulk conditions. At the time of starting this project, the most popular protein force fields were several members of the AMBER family (ff99.sb-ILDN, Lindorff-Larsen et al. (2010), ff14.sb, Maier et al. (2015), and the recently released ff19.sb, Tian et al. (2020), though of the latter no GROMACS port was available) and the CHARMM36m force field. (Huang et al., 2017) Compatible with the AMBER force fields were AMBER lipid14 (Dickson et al., 2014) and lipid17 as well as slipids (Jämbeck and Lyubartsev, 2012), whereas CHARMM36 includes lipid parameters. (Klauda et al., 2010). Throughout this thesis we made our choice of force field to maximise consistency with previous work. When validating MEMENTO in chapter 3, we chose force fields to match the studies against the results of which we were comparing, so opted for CHARMM36 for the deca-alanine system, AMBER ff14.sb for ADK and P38α and CHARMM36 with CHARMM lipids for LeuT. In our PepT2 simulations, we adopted the choice of AMBER ff14.sb with slipids as by previous work from our group (Parker et al., 2021). An exception to this are our constant-pH simulations (see the discussion in section 2.5), which require a specific fork of the AMBER 10 force field for the hybrid solvent scheme in the AMBER engine (Swails et al., 2014), and a modified version of CHARMM36 for the all-atom GROMACS implementation. (Buslaev et al., 2022)

Having chosen an AMBER protein force field, we had options to represent ligands using GAFF2 (Wang et al., 2004) or the most recent OpenFF version 2.0 (termed SAGE) (Boothroyd et al., 2023) available ahead of publication at the time of starting these simulations. Furthermore, for the subset of peptide ligands, there was also the option to represent them using the AMBER ff14.sb protein force field. In chapter 3, where we used exclusively peptide ligands to study the ligand-dependence of conformational changes, we opted for the latter variant. In chapters 5 and 6, which include peptide and other small-molecule ligands side-by-side, we chose OpenFF 2.0 since we found a better agreement regarding peptide backbone torsion parameters with AMBER than for GAFF2 (see section 6.1.1). For water and ions, we used the default TIP3P (3 charges, one van-der-Waals sphere) water model (Jorgensen et al., 1983) and ion parameters shipped with the AMBER ff14.sb force field.

2.2 Enhanced sampling molecular dynamics

In section 2.1.1, we have introduced the concept of MD simulations to sample a thermodynamic ensemble by making an assumption of ergodicity — that is, in the infinite time limit, a simulation will fully explore the section of phase space specified by the ensemble in question. However, it is only practically feasible to run limited lenghts of MD simulations, and if the free energy barriers separating two regions of phase space are too high for crossing to be kinetically accessible on the simulation timescale, the ergodicity

assumption often breaks down in practice. Figure 2.2 illustrates the problem with the example of a simple two-well potential, where the thermal energy of a particle moving in one basin is not sufficient to cross the barrier separating it from the second well without the use of enhanced sampling. It is well recognised that many microscopic processes of biological interest, such as protein folding, protein–ligand (un)binding or large-scale conformational changes, to mention but a few, have time scales reaching to the order of seconds and beyond, which is far longer than what is accessible using unbiased MD (where — on current HPC hardware — on the order of ten microseconds are now routinely accessible). [6] Surmounting the barriers involved in sampling such slow processes necessitates the use of *enhanced sampling* methods. (Harpole and Delemotte, 2018; Orellana, 2019; Liao, 2020; Hénin et al., 2022; Ahmad et al., 2022; Pravin and Singh, 2023) Aside from enabling a full exploration of phase space, many enhanced sampling methods also allow for a reconstruction of the free energy landscape of the system.

2.2.1 Classification of enhanced sampling methods

Enhanced sampling methods use a host of different strategies to facilitate the crossing of free energy barriers, connecting regions in phase space that would otherwise be separated during MD simulations. Different classifications of these strategies are possible and may be equally valid, for instance based on whether a given enhanced sampling method uses collective variables (CVs) — a distinction adopted for example by Liao (2020). Here, we will structure the discussion drawing on the exposition by Hénin et al. (2022) throughout, which classifies a large number of methods by the physical and statistical principles underlying the acceleration of sampling. Without being exhaustive, we will now list some of the important categories discussed in that reference.

Driven methods such as targeted and steered MD use a pre-defined schedule to bias a CV or alchemical parameter. Usually this results in an out-of-equilibrium distribution with significant dissipative work being done, though in cases of fast-relaxing orthogonal degrees of freedom (DOFs) free energies may be reconstructed using work relationships like the Jarzynski equation. With static bias methods, a bias potential that does not change over time is applied to the simulation. This may be a potential energy term that does not depend on CVs, such as the dihedral boost potential in the case of accelerated MD (aMD), or a user-defined potential applied along a (set of) CVs, as is done with umbrella sampling using a stratification / localisation strategy. Adaptive bias methods such as metadynamics or adaptive biasing force (ABF), in turn, evolve the bias applied along certain CVs over the course of the simulation to match the structure of the underlying free energy landscape and enhanced the kinetics of a system moving across the landscape. The converged bias potential or biasing forces can then be used to reconstruct free energies. By contrast, generalised ensemble approaches such as replica exchange and expanded ensemble methods enhance configurational sampling of a simulation by allowing it to access alternative ensembles (for example at higher temperatures or modified potentials) where barrier crossing is easier, while keeping the original configurational distribution. It is also possible to accelerate rare events with adaptive seeding, where unbiased trajectories are initiated strategically in regions of phase space that are most likely to lead to transitions in the available sampling time. Lastly, many hybrid methods exist, for example combining bias potentials with replica exchange. After reviewing the notion of a CV and how it may be obtained, we will

[6] Giving figures of timescales accessible to MD is a somewhat ill-defined undertaking, because the running of many simulations is an embarrassingly parallel computational task, so that the sampling accessible to the user becomes a function of available HPC resources. How fast a single trajectory can be propagated in time, however, is limited by parallelisation overheads for CPU architectures or the speed of the GPU used. In benchmarks run by HECBioSim on the UK-based ARCHER2 and JADE2 clusters (https://www.hecbiosim.ac.uk/access-hpc/benchmarks), using the GROMACS 2020.3 software with a system of 61,000 atoms, it is possible to simulate 466 / 132 ns per day on 9 CPU nodes / 1 GPU, respectively. A simulation of 10 µs would thus take ≈ 21 days.

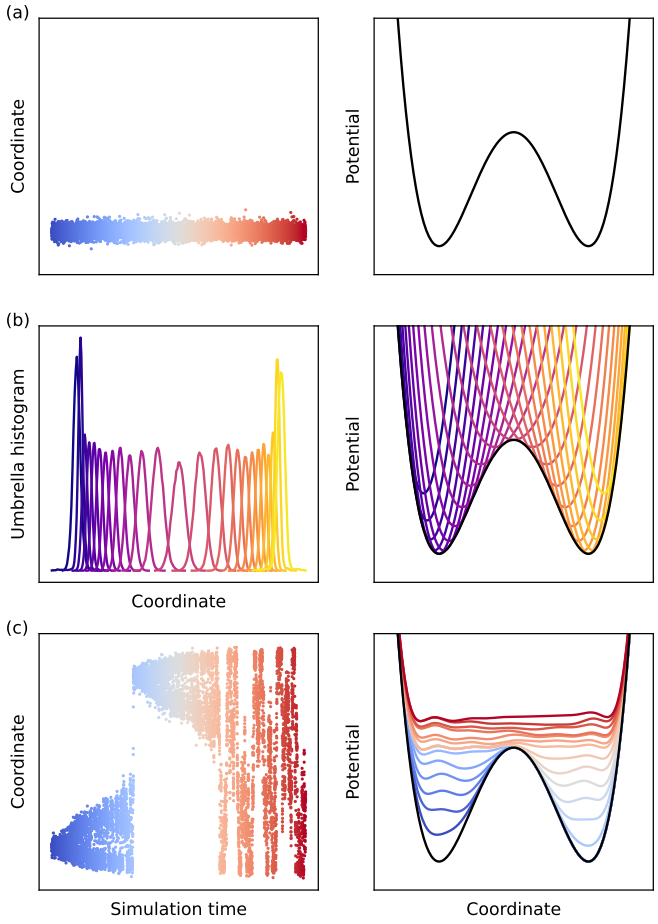


Figure 2.2: Comparison of (a) unbiased MD, (b) umbrella sampling and (c) metadynamics on a simple 1D potential. Unbiased MD never leaves the starting well, while umbrella sampling uses localisation with added harmonic potentials in the coordinate space to ensure sampling across all values of the 1D coordinate. Metadynamics, by comparison, accumulates bias until one well is filled before filling the other well and then diffusing freely. Simulation time is shown in blue to red hues, while umbrella sampling replicate windows are indicated by purple to yellow colours.

discuss in more detail the enhanced sampling methods used in this thesis, employing the classification sketched out in this section.

2.2.2 Collective variable design

A set of CVs ξ maps the configuration $\mathbf{r}$ of a simulation system into a low-dimensional space $\mathbf{z}$:

$$\mathbf{z} = \xi(\mathbf{r}) \tag{2.46}$$

In practice, simulation engine plugins like PLUMED (Tribello et al., 2014) and colvars (Fiorin et al., 2013) allow for functional forms of collective variables ranging from simple inter-atomic distances, angles, quaternions and dihedrals to paths connecting references structures, arbitrary displacements of atomic coordinates and complex order parameters like solvation shells, all of which can be chained together flexibly with user-defined mathematical operations. Despite the ease of naively defining CVs, doing so in a way that will be effective to interpret and bias simulation trajectories is a highly complex task. The notion of optimal CVs can be connected to the *committor* between two states where there are distinct states (that is, the probability of a system in a given configuration relaxing to one state instead of the other), or more generally to the slow eigenfunctions of the MD transfer operator that propagates a molecular configuration forward in time. (Noé and Clementi, 2017) The CVs thus obtained will be optimal in that they capture the slowest movements of a simulation system and can be used to accelerate them with enhanced sampling. The issue is that these eigenfunctions cannot normally be calculated *a priori* for a high-dimensional system of interest, so that the challenge for any researcher wishing to use CV-based enhanced sampling becomes finding workable approximations to the slow modes of a given system. Whether a given CV will be successful generally needs to be evaluated by trial-and-error, often combined with refinement in iterative schemes. We will briefly review some strategies available for CV design, paying particular attention to the methods used in this thesis.

Intuition and biological or physical considerations

Historically (and, to a large extent, still in current practice), CVs were designed intuitively based on physical or biological prior knowledge. This is often related to experimental observables such as different structures resolved for a given protein, for example in the case of the open and closed states of ADK as considered in chapter 3.4. By considering the rigid-body motions of the LID, NMP and CORE domains, one can intuit the simple centre-of-mass distance CVs that form the basis of most enhanced sampling studies on ADK (see table 3.1 for a literature review of enhanced sampling work on ADK). Another example relevant to this thesis is our use of intuitive bundle distances as initial CV guesses, based on the conceptual alternating-access rocker switch mechanism (see section 1.1.1). It is also possible to use structural information to design path-CVs that describe the movement of a system between several known structures derived from experiments or with computational methods. (Branduardi et al., 2007) Intuition-based design can be highly effective for constructing CVs that perform well in enhanced sampling if sufficient prior knowledge is available on the system of interest, given enough trial and error. A recent example of such a procedure applied to a membrane protein is the investigation of the conformational changes associated with thin-gate closure of the Mhp1 transporter. (Meshkin and Zhu, 2021)

Deriving CVs from MD simulations

Without extensive prior knowledge about slow protein motions available, and for improved transferability, many methods have been developed to derive CVs from MD trajectory data, ranging from linear dimen-

sionality reduction to iterative machine learning schemes. (Noé and Clementi, 2017; Sidky et al., 2020; Chen, 2021; Mehdi et al., 2024) Principle component analysis (PCA) and time independent component analysis (TICA) are popular options for design of linear CVs from MD simulations. (Bhakat, 2022) For PCA, starting with a vector of geometric CVs $\mathbf{q}$ (often inter-atomic distances or displacements from an average structure), one computes a variance-covariance matrix of the mean-free vectors $\hat{\mathbf{q}}(t) = \mathbf{q}(t) - \langle \mathbf{q} \rangle$ as

$$\mathbf{C} = \langle (\mathbf{q} - \langle \mathbf{q} \rangle)(\mathbf{q} - \langle \mathbf{q} \rangle)^T \rangle. \tag{2.47}$$

The principle components (PCs) are then the eigenvectors of this matrix, corresponding to orthogonal vectors along the first of which (ordered by eigenvalue) the greatest proportion of the trajectory variance is captured. If the first k PCs explain a satisfactory proportion of the total variance, one may use them as low-dimensional CVs. It should be noted that the PCs are constructed for maximum variance explanation, which does not always yield the slowest modes in a given trajectory. Instead, in performing TICA, one constructs a symmetrised, time-lagged correlation matrix of the mean-free vectors as

$$\mathbf{C}(\tau) = \frac{1}{2}[\langle \hat{\mathbf{q}}(t)\hat{\mathbf{q}}(t+\tau)^T \rangle + \langle \hat{\mathbf{q}}(t+\tau)\hat{\mathbf{q}}(t)^T \rangle]. \tag{2.48}$$

with a lag-time of τ. One then solves the generalised eigenvalue problem

$$\mathbf{C}(\tau)\mathbf{A} = \mathbf{C}(0)\mathbf{A}\Lambda, \tag{2.49}$$

where $\mathbf{A}$ contains the independent components (ICs) that can be used to project the trajectory, and Λ is the diagonal eigenvalue matrix. The ICs can be used as CVs approximating the slowest modes of a system of interest (as far as they were captured in the trajecotry used as input for TICA), and the eigenvalues λ_i relate to the relaxation timescales t_i as

$$t_i = -\frac{\tau}{\ln|\lambda_i|}. \tag{2.50}$$

Both PCA and TICA are linear decompositions, which makes the identified CVs easily interpretable and well-suited for biasing in enhanced sampling. However, the configuration space sampled by an MD trajectory will often be best described by a non-linear representation. Many approaches exist for non-linear dimensionality reduction of MD data that can be used for CV design. These have been extensively reviewed in the recent literature (Sidky et al., 2020; Chen, 2021; Bhakat, 2022; Mehdi et al., 2024) and a survey would be beyond the scope of this introductory chapter, especially since only linear decomposition is used in the research for this thesis. Nonetheless, to give a flavour of what kind of approaches are currently being developed and are starting to be applied to biological systems, we will very briefly introduce the example of the state predictive information bottleneck (SPIB) methodology. (Wang and Tiwary, 2021) SPIB takes as input some MD trajectory data, suitably featurised for each system, and an initial assignment of a small number of discrete meta-stable state labels. An artificial neural network (ANN) encoder transforms this high-dimensional input data into a low-dimensional latent space, and a decoder predicts labels for meta-stable states at a future time. One then optimises the state-label assignment together with the ANN weights, with a user-defined hyperparameter scaling between prediction capacity and model complexity. (Wang and Tiwary, 2021) This latent space representation can then serve as a CV for enhanced sampling. Mehdi et al. (2022) have demonstrated that an iterative protocol of SPIB and metadynamics (starting with unbiased MD data) can sample the chiral transition between a left and right helix in the peptide $(\text{Aib})_9$ and the permeation of benzoic acid through a lipid bilayer. Pomarici et al. (2023) have combined SPIB with bias-exchange metadynamics to accelerate rare transitions in the folding / unfolding of the small proteins chignolin and villin. SPIB-enhanced sampling has also been used together with Alphafold2 (Jumper et al., 2021)

models of increased conformational diversity through reduction of the depth of the supplied multiple sequence alignment (MSA) (del Alamo et al., 2022) to generate thermodynamically weighted ensembles of protein conformational landscapes. (Vani et al., 2023b) This approach has been applied to a DFG-loop kinase, although the authors still found it necessary to manually select knowledge-based CVs as input to SPIB (Vani et al., 2023a), thus highlighting the continuing challenge of identifying good CVs for sampling from sparse data.

2.2.3 Driven methods — steered MD

As discussed by Hénin et al. (2022), driven methods for enhanced sampling use a pre-defined, time-dependent biasing protocol to enhance the exploration of configurational space. In the first formulation of targeted MD (tMD), a moving constraint on the root mean squared deviation (RMSD) to a target structure was used. (Schlitter et al., 1993) Alternatively, in steered MD (SMD) a time-dependent harmonic restraint can be introduced on any CV. It should be noted that while tMD and SMD thus differ in the type of algorithm employed to pull the CV, the name tMD has also been used to denote SMD if the CV is an RMSD to a target state (Apostolakis et al., 1999; Huang et al., 2009), as is mostly the case in this thesis. However, we use the name SMD throughout to highlight our use of a harmonic restraint. tMD and SMD have been shown to give consistent results, so that sampling problems are general and due to the formulation of the CV. (Huang et al., 2009; Ovchinnikov and Karplus, 2012) We can write the general biasing potential of SMD as

$$V_i^{\text{bias}}(\mathbf{r}) = \frac{k_i(t)}{2}|\xi_i(\mathbf{r}) - \mathbf{z}_i(t)|^2, \tag{2.51}$$

where for the i^{th} CV ξ_i, the force constant k_i and the restraint center $\mathbf{z}_i$ can be made to vary in any user-defined way with time. A common protocol, used in this thesis for geometric distance CVs, is to switch on the restraint by using a quick linear increase in k to its final value, then slide the restraint center to the target value, follow by switching off the force and finishing with unbiased relaxation. An alternative, which we use for RMSD-based CVs, is to set the restraint center of the RMSD to zero and linearly increase the force constant over the steer, followed by a fast switch off and unbiased relaxation. Full details can be found in the appendices.

SMD is an out-of-equilibrium technique that can only be used for a direct approximation of the underlying free-energy landscape if it is performed slowly enough for all orthogonal DOFs to relax to an effective equilibrium ensemble around a given CV value. It is not easy to establish that this assumption holds for a given system in practice, although free energies can be estimated from an ensemble of SMD runs, see section 2.4. More commonly, SMD is used to generate initial paths for umbrella sampling or other path-based enhanced sampling methods that allow for a better calculation of free energies.

2.2.4 Localisation and static bias methods — umbrella sampling

A second class of enhanced sampling methods (which we extensively use in this thesis) are called localisation methods in the scheme of Hénin et al. (2022). The principle is to restrict sampling to a subregion of configuration space, usually in several windows so that sampling in regions that would be unfavourable in unbiased MD is enforced in some of the windows. In its most common form, termed umbrella sampling (US), this is achieved with time-independent harmonic restraint potentials. [7] For the n^{th} window in the i^{th}

[7] As noted by Hénin et al. (2022) in their review, the term "umbrella sampling" was first used by Torrie and Valleau (1977) to denote a rather different method. In that paper, a single simulation moved across all CV-space, driven by a manually defined

CV, this potential can be written in a similar way to SMD above:

$$V_{i,n}^{\text{bias}}(\mathbf{r}) = \frac{k_{i,n}}{2}|\xi_i(\mathbf{r}) - \mathbf{z}_{i,n}|^2. \tag{2.52}$$

This is illustrated in figure 2.2b for a two-well potential, where the 1D coordinate doubles as the (only possible) CV. Aside from a definition of CVs, force constants and restraint centers, to start an US simulation one also needs initial frames. As noted in the preceding section, these are often obtained from a driven method such as SMD, although in chapter 3 we show how this can cause severe problems with a sub-optimal (initial) choice of CV and develop an alternative strategy for seeding US simulations. When conducting US simulations, it is crucial to ensure good overlap of the CV distributions of adjacent windows by an appropriate choice of force constant and window spacing, so that the free energy landscape can be reconstructed as explained in section 2.4, often using the weighted histogram analysis method (WHAM) method for US. It is also possible to enhance sampling further by combining US with Hamiltonian replica exchange (see section 2.2.6), known as replica-exchange umbrella sampling (REUS). We use this approach throughout in the work for this thesis, as noted in the appendices.

2.2.5 Adaptive bias methods — metadynamics

While driven methods and US use biases defined ahead of starting a simulation (regardless of whether the bias itself depends on time or not), in the adaptive bias class of methods the biasing potential is learned and refined as the simulation progresses. (Hénin et al., 2022) Convergence is achieved when a certain target sampling distribution is reached, and the underlying equilibrium distribution (that is, the free energy landscape along the CV) can be reconstructed from the final biases. The most commonly used adaptive bias methods are metadynamics and ABF. Of these, only metadynamics was used in this thesis and will be described here.

The idea underpinning metadynamics is to add repulsive potentials (usually chosen as Gaussian kernels) to points in CV-space previously sampled by the simulation (normally amounting to a fixed deposition rate of Gaussian 'hills'). As figure 2.2c illustrates on a two-well potential, this leads to the gradual filling of deep free-energy wells, where — once a basin is filled — the simulation crosses into the next one, finally leading to a flat surface. In the first conceptualisation of metadynamics (Laio and Parrinello, 2002), the Gaussian hills were of constant height H and width σ, so that the bias potential evolves as

$$V_n^{\text{bias}}(\mathbf{z}) = V_{n-1}^{\text{bias}}(\mathbf{z}) + H\exp\left[-\frac{1}{2}\sum_{i=1}^{d}\frac{(z_i - z_{n,i})^2}{\sigma_i^2}\right], \tag{2.53}$$

where $\mathbf{z}_n$ denotes the CV-space vector at metastep n. The average of the bias potential at the end of a converged simulation (disregarding n_0 metasteps as time required to reach the equilibrium distribution) will cancel out the free energy, so that it can be estimated as

$$A(\mathbf{z}) = -\frac{1}{n - n_0}\sum_{k=n_0}^{n} V_k^{bias}(\mathbf{z}). \tag{2.54}$$

Usually — as is the case in this thesis — an asymptotically converging variant of metadynamics is used, called well-tempered metadynamics (WT-metaD). (Barducci et al., 2008; Dama et al., 2014) Instead of having a constant height, the Gaussians are scaled down as the simulation progresses according to

static biasing potential. Terminology has since shifted to refer to the localisation procedure with windows and harmonic restraints described in this section as "umbrella sampling".

$$H_k = H_0 \exp\left[-\frac{1}{\gamma - 1}\beta V_{k-1}^{\text{bias}}(\mathbf{z}_k)\right], \tag{2.55}$$

where γ is the (dimensionless) bias factor. The bias potential does not fully cancel out the free energy wells anymore, but converges to

$$V_t^{\text{bias}}(\mathbf{z}) = -\left(1 - \frac{1}{\gamma}\right) A(\mathbf{z}) + K(t), \tag{2.56}$$

up to a time-dependent constant $K(t)$. The distribution sampled is called the well-tempered distribution, which becomes the flat distribution of classical metadynamics in the limit of $\gamma \to \infty$. As for any other CV-based enhanced sampling method, the speed of convergence of WT-metaD depends on the choice of CV — if chosen poorly it will converge no better than unbiased MD and pathologies such as large fluctuations in the bias potential or simulations that get stuck in one of the minima may arise. (Bussi and Laio, 2020) To make more efficient use of modern, highly parallel compute resources, one can also run multiple-walker metadynamics, where several simulations explore the CV space while depositing a shared bias that is periodically synchronised. (Raiteri et al., 2006) A great number of methods building on modifications of metadynamics exist, for instance for simultaneously using a large number of CVs in bias-exchange metadynamics (Piana and Laio, 2007) or for estimating rare-event kinetics with infrequent metadynamics. (Tiwary and Parrinello, 2013) Reviewing them here would be beyond the scope of this introduction; a non-exhaustive but extensive list of methods and their references can be found with Hénin et al. (2022).

2.2.6 Replica exchange

The final class of enhanced sampling methods used in this thesis is categorised as 'generalised ensemble' by Hénin et al. (2022) — here we focus only on replica exchange. (Hukushima and Nemoto, 1996; Sugita and Okamoto, 1999) While originally formulated for the exchange with replicates simulated at higher temperatures to assist barrier crossing in a low-temperature simulation, the method is general in that simulation systems evolving within any different reduced potentials $u(\mathbf{x}) = \frac{U(\mathbf{x})}{kT}$ can be considered for exchanges, allowing for the combination with enhanced sampling methods that perturb the potential (such as umbrella sampling and metadynamics). Different exchange attempt schedules have been implemented. GROMACS (see footnote 2), which we use through most of this thesis, attempts exchanges with constant frequency, where neighbouring replicates are exchanged in an alternating fashion (that is, on odd exchanges $1 \leftrightarrow 2$, $3 \leftrightarrow 4$ etc, and on even exchanges $2 \leftrightarrow 3$, $4 \leftrightarrow 5$ etc). Exchange attempts are accepted according to the Metropolis criterion (Metropolis et al., 1953), which for a general exchange between two simulations i and j at reduced potentials $u_i(\mathbf{x})$ and $u_j(\mathbf{x})$ can be stated as the probability of accepting an exchange

$$P_{\text{accept}}(i \leftrightarrow j) = \min\{1, \exp\left[u_i(\mathbf{r}_i) + u_j(\mathbf{r}_j) - u_i(\mathbf{r}_j) - u_j(\mathbf{r}_i)\right]\}. \tag{2.57}$$

To be effective, replica exchange simulations thus need to be set up in such a way that significant mixing of replicates takes place. In particular, this means that the spacing of windows in temperature or Hamiltonian needs to be narrow enough for an appreciable average acceptance probability of exchanges. An often-stated target is for an average acceptance rate of 30 %. Another common issue with replica exchange simulations are exchange bottlenecks in extended ensemble space, which curb effectiveness by limiting the expanded configuration space accessible to a given window. (Hénin et al., 2022)

2.3 Binding free energy calculations

The discussion of enhanced sampling methods in the preceding section has been kept entirely agnostic of the biological problem of interest that necessitates their application. As long as it can be expressed in terms of a (set of) CVs or — with methods such as stratification or generalised ensemble — in terms of a parameter λ on which the potential energy function depends, any problem is in principle amenable to study with enhanced sampling. In this thesis, for example, we will mainly be concerned with protein conformational changes, where the above exposition directly applies. It is nonetheless useful to conceptually single out the calculation of protein–ligand binding free energies at this stage, as special techniques have been developed for this purpose.

Small molecule ligands display a wide range of off-rates on protein receptors that often translate to residence times of hours to days for potent competitive inhibitors. (Tummino and Copeland, 2008) Therefore, although in theory it is possible to calculate the binding free energy of a ligand on a protein from the analysis of unbiased MD simulations, this is usually very far from feasible in accessible timescales. Broadly speaking, one may divide MD techniques to sample these slow binding / unbinding processes into geometric and alchemical approaches (see figure 2.3).

In the former family of techniques, one uses enhanced sampling frameworks as presented in the foregoing sections in combination with CVs designed to capture the physical ligand unbinding pathway, often including suitable restraints to deal with the big accessible bulk volume that may hamper sampling of the binding transition. Funnel metadynamics, for example, combines the method of WT-metaD with a flat-bottom funnel-shaped restraint potential that does not hinder ligand dynamics within the binding site but restricts sampling of the bulk solvent to a thin cylinder. This enables reversible sampling of binding and unbinding events, while allowing for an analytical correction to the estimated binding free energy to recover the bulk ensemble. (Limongelli et al., 2013) Tools for setting up the necessary restraint potentials exist (Raniolo and Limongelli, 2020), enabling now-widespread use on protein–ligand binding problems, though as with any CV-based enhanced sampling technique, an appropriate choice of CVs is required to ensure reversible dynamics and convergence. It is also possible to use SMD and umbrella sampling to estimate binding free energies on proteins with a suitable choice of CVs. A notable example of a field where this is common practice is the calculation of lipid affinities to proteins (see the discussion in Corey et al. (2019)), though obtaining converged ligand binding free energy landscapes with umbrella sampling comes with the same challenges as other umbrella sampling applications in terms of CV design and generating starting structures. (You et al., 2019) We will not discuss either of these geometric approaches to calculate binding free energies in more detail here, but focus on alchemical cycles for the rest of this section, as we exclusively use those throughout this thesis.

2.3.1 Alchemical thermodynamic cycles

The conception of alchemy has changed significantly since its exposition in Hellenistic Egypt as ‘the composition of waters, movement, growth, embodying and disembodying, drawing the spirits from bodies and bonding the spirits within bodies’ (Zosimos of Panoplis (300) as quoted in Strathern (2001)), now denoting rather more mundanely the use of unphysical pathways for transforming molecules into empty space or other molecules in MD simulations. Alchemical approaches for estimating protein–ligand binding free energies are used widely on systems of pharmaceutical relevance across academia and industry. (Chodera et al., 2011; Cournia et al., 2017; Mey et al., 2020; Khalak et al., 2021; King et al.,

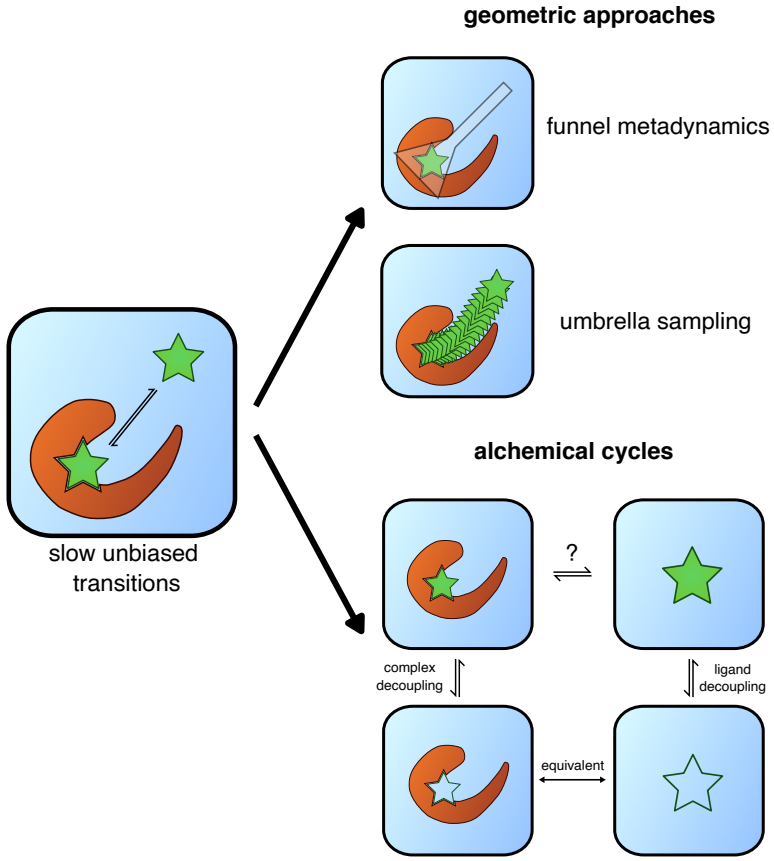


Figure 2.3: Schematic overview of binding free energy simulation methods. Unbiased MD cannot usually sample the binding and unbinding transitions of ligands to protein receptors. Geometric enhanced sampling approaches like funnel metadynamics and umbrella sampling speed up simulations along the physical unbinding pathway, while alchemical thermodynamic cycles decompose ligand binding into the decoupling of the ligand from the protein and the solvent environments.

2021; Cournia and Chipot, 2024) Instead of simulating the actual binding / unbinding pathways a ligand may take in relation to the protein structure, it is possible to instead construct alchemical thermodynamic cycles, where the interaction potentials of the ligand with its surroundings are scaled between a fully interacting state and the decoupled state for absolute binding free energies (ABFEs) (Jorgensen et al., 1988; Aldeghi et al., 2016, 2018; Gapsys et al., 2021), or a fully interacting state with an alternative molecular topology for relative binding free energies (RBFEs). (Tembre and Mc Cammon, 1984; Wang et al., 2015; Cournia et al., 2017; Gapsys et al., 2020) The interpolation in either case is described along an alchemical parameter conventionally denoted as λ. Free energies can be computed along such alchemical pathways by an appropriate simulation protocol (see the following section 2.3.2) together with a free energy estimator (see section 2.4). As illustrated in figure 2.3, by taking the difference between the free energies computed for decoupling or alchemically transforming a ligand in the protein and bulk solvent environments — a scheme known as double decoupling — the binding free energy ΔG (or the difference $\Delta\Delta G$, in the case of a relative calculation) is recovered. ABFEs and RBFEs each face characteristic methodological challenges, which for the former includes the more extensive sampling requirements and for the latter the appropriate construction of ligand-transformation networks and the representation of transformations in molecular topologies. Here, we focus only on ABFEs as these were the method of choice in this thesis.

2.3.2 Sampling the thermodynamic legs

Different enhanced sampling methods can be used to sample the alchemical transformation along the λ parameter (where we can adopt the convention that at $\lambda = 0$ the ligand is fully coupled and the decoupled state is denoted by $\lambda = 1$). In the context of binding free energy calculations, these are usually classed into equilibrium and non-equilibrium approaches. In the equilibrium approach, one uses a stratification strategy where a number of windows are run at fixed successive values of λ, usually defined as a sequence of states that first decouple the Coulomb and then the van-der-Waals interactions. (Aldeghi et al., 2016, 2018) It is also common to combine this with replica exchange to improve phase-space overlap between neighbouring windows. The decoupling free energy calculation can then be accomplished with thermodynamic integration or mBAR (see section 2.4). Alternatively, in a non-equilibrium approach, one runs longer equilibrations at the $\lambda = 0$ and $\lambda = 1$ states and uses them to seed many short driven runs in which the value of λ sweeps from 0 to 1 and vice versa. (Gapsys et al., 2021) From the ensemble of non-equilibrium work values obtained in this way, one can obtain the decoupling free energy via the Jarzynski equality or, taking into account data from both directions, the Crooks fluctuation theorem. In this thesis, we opt for the equilibrium approach as the one somewhat more established in the field, and will not discuss the non-equilibrium method further.

2.3.3 Restraints in binding free energy simulations

The double-decoupling thermodynamic cycle shown in figure 2.3 involves sampling a state where the ligand is fully decoupled from the binding pocket. While in this state it should freely diffuse across the simulation box, the resulting phase-space overlap with the coupled windows would be too small in practice for accurate free energy estimation. Therefore, it is common practice to introduce restraints that are switched on before decoupling takes place, thus holding the ligand in the protein binding pocket. Here, we use the scheme derived by Boresch et al. (2003), which attaches 3 ligand atoms to 3 protein atoms using one bond, two angle and three dihedral restraints. The imposition of this restraint comes with a free energy penalty, since phase space exploration gets restricted. To obtain the correct binding free energy from the full thermodynamic cycle, one therefore needs to calculate correction terms for both the ligand and complex thermodynamic legs. On the complex side, this is usually done by adding λ states to the decoupling protocol in which the restraint gradually gets applied to the coupled ligand positioned in the binding pocket (for a reasonably stable binding pose, this gives sufficient phase space overlap), therefore allowing for a direct inclusion of the restraint in the free energy estimation. For the ligand-only side, one does not usually run simulations with the restraint, but uses an analytic correction term for the restraint imposition that holds for freely moving ligand molecules and can be formulated to include a standard-state correction by using the standard volume V^0 in

$$\Delta A^{\text{restr}} = kT \ln \left[\frac{8\pi^2 V^0}{r_0^2 \sin\theta_{A,0} \sin\theta_{B,0}} \frac{\sqrt{k_r k_{\theta_A} k_{\theta_B} k_{\phi_A} k_{\phi_B} k_{\phi_C}}}{(2\pi kT)^3} \right], \tag{2.58}$$

where r is the restraint bond, θ_A and θ_B are the restraint angles, $\phi_A \dots \phi_C$ are the dihedral restraints and k_x and x_0 denote the force constant and center of a restrained quantity x. Note that while derived in the canonical ensemble by Boresch et al. (2003) as a Helmholtz free energy, the result can equally be used (as it common practice) in the isothermal-isobaric ensemble as a Gibbs free energy term ΔG^{restr}, since in the decoupled state the ligand degrees of freedom do not influence the box volume.

It should be noted that the application of restraints as described in the above paragraph restricts the phase space explored by both the protein and the ligand in all but the fully decoupled window to the vicinity of the binding pose used to derive the restraint. While this has the advantage of facilitating convergence of the binding free energy estimation, it can lead to simulation instabilities if the binding pose is very flexible (such as we encounter in some examples in our large ligand set in chapter 6), and incorrect free energy estimates if multiple binding poses contribute simultaneously as an ensemble. In these cases, alternative restraint definitions and methods based on geometric CVs may allow for sampling the full bound ensemble, albeit at increased computational expense for convergence if the DOFs separating different binding orientations are left to equilibrate, or a ligand-specific CV design task if the required sampling is to be accelerated. In this thesis, after trials with metadynamics (metaD)-based methods that failed to converge owing to the lack of a good CV, the restraints as described above were used throughout.

2.3.4 Treatment of charged ligands

A further complication for double-decoupling alchemical free energy calculations arises when there is a change in charge in the alchemical transformation, for example when calculating an ABFE of a charged molecule to a protein (a case we encounter in chapter 6 of this thesis). When using PME for treating long-range electrostatics, it is a requirement for the algorithm that the simulation box be neutral. While in most cases this can be taken care of using balancing ions in the solvent phase of a simulation, when running alchemical perturbations of charged ligands, the overall charge of the box changes with λ, including fractional changes at intermediate states. The uniform neutralising background charge introduced by default to counteract these changes can lead to artifacts in free energy computations. A semi-analytic correction has been derived by Rocklin et al. (2013), although it is difficult to treat membrane systems in the required implicit-solvent electrostatic calculations. (Wu and Biggin, 2022) It is beyond the scope of this introduction to review the possible correction schemes. In this thesis, we use the method of alchemical transfer of the charge being decoupled to a restrained alchemical co-ion (Na^+ or Cl^-, charge 0 at $\lambda = 0$ and ± 1 at $\lambda = 1$), as recommended by Wu and Biggin (2022). Additionally, we deal here with a number of multiply-charged ligands, which in the charge-transfer scheme could be addressed in one of three ways, in principle: one may simply locate extra charges at the alchemical ion, one could use existing topologies from the literature for physical ions with multiple charges (e.g. Mg^{2+} or SO_4^{2-}), or use multiple single-charge alchemical ions. We show in chapter 6 how the first approach leads to artifacts by rendering the alchemical ions too sticky. The second approach we opted not to take for concerns about extensibility in a field where parameters for ions with all possible higher-valent charges are not necessarily available. By contrast, using multiple alchemical ions is viable if the ions are placed and restrained appropriately far apart from each other (see section 6.1.2).

2.3.5 Further methodological considerations

To conclude the discussion on alchemical free energy simulations, we will briefly touch on three further methodological considerations.

Integrator and thermostat for alchemical simulations

As noted in section 2.1.2, in alchemical simulations it is crucial to ensure that the correct temperature ensembles are sampled in both the protein system and the molecule in its decoupled state, because near $\lambda = 1$ they do not couple to each other anymore. A convenient approach is using a Langevin dynamics

integrator which doubles as a thermostat on all degrees of freedom simultaneously — we do this across all ABFE calculations in this thesis.

Decoupling vs annihilation

In the preceding sections, we discussed decoupling molecules from the surrounding system, which leaves open whether the intramolecular non-bonded interactions are also turned off. Note that both options produce equally valid thermodynamic cycles when applied consistently across the complex and ligand legs. The distinction corresponds to the 'couple-intramol' flag in GROMACS [8]; another way of referring to this choice is to use 'decoupling' for cases where the intramolecular non-bonded interactions are preserved (corresponding to the correct vacuum state of the molecule), and to speak of 'annihilation' when they are switched off with the intermolecular interactions. An issue with decoupling arises for large molecules because, in the GROMACS implementation, intramolecular non-bonded interactions in the decoupled ligand are turned into explicit pair interactions. Molecules that — in any sampled conformation — are larger than the employed pair-list cutoff cannot be supported, and an error is triggered by GROMACS. Since many of our ligands (for example β-lactams) are larger than the pair-list cutoff in our simulations, and to avoid the computationally inefficient increase in pair-list cutoff distance that would be required to accommodate them, we opt in this thesis for an annihilation procedure, as is recommended by the GROMACS developers, and as for example Alibay et al. (2022) have chosen to do.

Softcore potentials

In alchemical transformations, when decoupling the interactions between a ligand and the surrounding system one usually uses a linear scaling of the potential with λ. However, at λ-values where the ligand is very close to fully disappearing (that is, in the notation for ABFE above, λ is very close to 1), this can introduce singularities. Because atoms can get very close when not interacting, the value of $\partial V/\partial\lambda$ can jump a lot in the vicinity of $\lambda = 1$, which can lead to poor convergence and simulation instabilities. To tackle this issue, Beutler et al. (1994) have introduced soft-core interactions for λ-values between 0 and 1 (but not at the end-states, so that the correct distributions and free energies are still sampled). Using a value of 6 for the scaling exponent, a choice made in the GROMACS implementation [9], we can recast the linear interpolation between potentials V^A and V^B as

$$V_{\text{sc}} = (1-\lambda)V^A(r_A) + \lambda V^B(r_B), \tag{2.59}$$

using scaled inter-atomic distances

$$r_A = (\alpha\sigma_A^6\lambda^p + r^6)^{\frac{1}{6}} \tag{2.60}$$

$$r_B = (\alpha\sigma_B^6(1-\lambda)^p + r^6)^{\frac{1}{6}}, \tag{2.61}$$

where the parameter α and the soft-core power p are set by the user, and σ takes the value of the Lennard-Jones parameter at the coupled state and is set by the user at the decoupled state. Since in this thesis we decouple Coulomb and van-der-Waals interactions sequentially, we use soft-core potentials for van-der-Waals interactions only.

[8] See https://manual.gromacs.org/current/user-guide/mdp-options.html for details.

[9] See also https://manual.gromacs.org/nightly/reference-manual/functions/free-energy-interactions.html.

2.4 Free energy estimators

With most of the enhanced sampling methods introduced above and used in this thesis — aside from driven methods for path generation — an important goal of their use in practice is usually the estimation of a free energy landscape along a CV or differences in free energy between two states (as in the case of ABFE). Note here that although the terms free energy landscape (FEL) / free energy surface (FES) and potential of mean force (PMF) are historically distinct notions (see the discussion by Hénin et al. (2022)), they are now used synonymously and will throughout this thesis interchangeably mean

$$A(\mathbf{z}) = -\beta^{-1} \ln \int \delta(\mathbf{z} - \xi(\mathbf{r}))v(\mathbf{r})d\mathbf{x}, \tag{2.62}$$

where the normalised Boltzmann distribution is written as $v(\mathbf{r})$. Further integration over $\mathbf{z}$ can be used to calculate free energy differences between given states. We have already discussed in section 2.2.5 how the bias potential refined over the course of the simulation by adaptive bias methods is related to the FES. In this section, we will give a brief overview of free energy estimators useful in other cases, informing the discussion primarily by the exposition in Hénin et al. (2022).

For an unbiased simulation with sufficient sampling to cover several transitions between two states i and j, the free energy difference ΔA_{ij} can simply be estimated from counting the occurrences N of the two states across the trajectory:

$$\Delta A_{ij} = -\beta^{-1} \ln \frac{N_j}{N_i}. \tag{2.63}$$

Since such a trajectory is not usually available and enhanced sampling approaches that utilise unbiased MD together with adaptive seeding distort the cumulative populations by focussing sampling on rarely explored regions, it is more common to estimate free energies from calculating transition probabilities between states across many shorter trajectories. Notably, this can be achieved by constructing Markov-state models (MSMs). An introduction to MSMs in any detail is beyond the scope of this introduction since we do not use them in this thesis, and the reader is referred to dedicated reviews on the subject. (Pande et al., 2010; Chan and Shukla, 2021) In short, to construct an MSM one first clusters a projection of one's sampling in a suitably-defined CV space (often derived from TICA) into 'microstates' (with a terminology distinct from statistical mechanics). One then computes the transition matrix between these microstates and solves the kinetic master equation to reconstruct equilibrium probabilities of the microstates, thus yielding an FES that can be projected onto any desired CVs.

By contrast, methods that perturb the sampled ensemble by changing the potential require free-energy estimators to recover the unbiased free energy differences between states. Thermodynamic integration (TI) can be used in cases where a coupling parameter λ connects two states (as in the example of alchemical free energy calculations). By recognising that the derivative of the free energy with respect to λ is the ensemble average at a given value of λ

$$\frac{\mathrm{d}A}{\mathrm{d}\lambda} = \langle \frac{\partial U(\lambda)}{\partial \lambda} \rangle_\lambda, \tag{2.64}$$

one can integrate along λ to obtain free energy differences:

$$\Delta A_{ij} = \int_{\lambda_i}^{\lambda_j} \langle \frac{\partial U(\lambda)}{\partial \lambda} \rangle_\lambda \mathrm{d}\lambda. \tag{2.65}$$

This can be formulated for general CVs, where an additional Jacobian term arises for non-linear CVs from the curvature in the CV-isosurface in configurational space. We do not use such an approach in this thesis and will not discuss it further here.

In principle, it is also possible to obtain ΔA_{ij} for from an ensemble average of the difference in potentials ΔU_{ij} sampled at one of the states via exponential averaging, also known as the Zwanzig relationship (Zwanzig, 1954):

$$\Delta A_{ij} = -\beta^{-1} \ln\langle e^{-\beta \Delta U_{ij}} \rangle_i. \tag{2.66}$$

Though this is an exact expression, such an average usually converges badly if there is not much overlap in phase space between the ensembles at states i and j. If sampling at both states is performed, the free energy estimate can be significantly improved over exponential averaging (and in most cases also TI) (Shirts and Pande, 2005) by using the Bennett acceptance ratio (BAR) method (Bennett, 1976), which yields a minimum-variance implicit estimation for the free energy as

$$\sum_{i=1}^{N_i} \frac{1}{1+\exp(\ln(N_i/N_j)+\beta\Delta U_{ij}-\beta\Delta A_{ij})} - \sum_{i=1}^{N_j} \frac{1}{1+\exp(\ln(N_i/N_j)-\beta\Delta U_{ij}+\beta\Delta A_{ij})} = 0, \tag{2.67}$$

where the sums are evaluated over N_i and N_j samples collected in the respective states. An extension to BAR that can use information from any number of sampled intermediate states is multistate Bennett acceptance ratio (MBAR). (Shirts and Chodera, 2008) It is possible for K states with N samples to write a system of coupled equations of the free energies:

$$A_i = -\beta^{-1} \ln \left[\sum_{n=1}^{N} \frac{e^{-\beta U_i(\mathbf{x}_n)}}{\sum_k N_k e^{\beta(A_k - U_k(\mathbf{x}_n))}} \right]. \tag{2.68}$$

Solving for all A_i's simultaneously allows for estimation of any ΔA_{ij} together with its uncertainty. MBAR can further be shown to reduce to BAR in the case of two states.

An alternative but related free energy estimator is WHAM (Kumar et al., 1992), originally derived for the estimation of free energies from US approaches (see section 2.2.4). After running K localised simulations with different restraining potentials, the analysis starts by choosing a number of bins M to bin the sampled data into histograms $n_{i,k}$, counting the number of samples contributed to bin i from simulation k. The unbiased probabilities $P_{i,k}$ can be evaluated with given free energy shifts applied to the windows A_k as

$$P_{i,k} = \frac{\sum_{k=1}^{K} n_{i,k} e^{-\beta U_{k,i}}}{\sum_{j=1}^{K} n_{i,k} e^{\beta(A_j - U_{j,i})}}, \tag{2.69}$$

which in turn yields the free energies

$$e^{-\beta A_k} = \sum_{m=1}^{M} P_{m,k}. \tag{2.70}$$

Solving these equations self-consistently yields an optimal estimate of the PMF, though the formalism is general and can be applied to histograms in any parameter such as temperature. While WHAM can be faster than MBAR when there are many samples, the width of histograms needs to be chosen small enough for energies to be roughly constant in each bin. For zero-width binds, the WHAM equations reduce to

MBAR.

Lastly, it is also possible to estimate free energies from ensembles of non-equilibrium trajectories. (Jarzynski, 1997) It can be shown that the free energy difference $\Delta A_{i,j}$ can be calculated from an exponential average of the work done in non-equilibrium, driven transitions $W_{i,j}$, known as the Jarzynski identity:

$$e^{-\beta \Delta A_{i,j}} = \langle e^{-\beta W_{i,j}} \rangle. \tag{2.71}$$

To improve convergence, it is also possible to calculate free energy differences from an ensemble of transitions in both directions. From the more general Crooks fluctuation theorem (Crooks, 1999), one can derive that the work distributions of forward and backwards transitions cross at the free energy difference between the states (a result commonly used in non-equilibrium alchemical free energy calculations), although providing details would be beyond the scope of this discussion.

As indicated through the preceding paragraphs, these free energy estimators are related to each other, and for a given application case it would be possible to choose one of several methods. In keeping with the majority of studies in the field, we here use MBAR for ABFEs and WHAM for US.

2.5 Constant pH MD

Conventional MD simulations with classical force fields cannot capture chemical reactivity since no bond formation or breaking is possible. Hybrid QM/MM methods can treat subsystems of a simulation box using quantum mechanics and therefore capture reactivity, however the more expensive energy and force evaluation renders them orders of magnitude slower than MD. (Senn and Thiel, 2009) In this thesis, we are concerned with the mechanism of a proton-coupled transporter, thus necessitating the consideration of different protonation states of several key residues, and fundamentally a protonation event is a chemical reaction. We primarily work around the issue by investigating the role of protonation-state changes through running side-by-side simulations in different protonation states, as discussed in chapter 4. Additionally, we make some use of constant-pH molecular dynamics (CpHMD), which will be briefly introduced in this section.

Broadly speaking, CpHMD makes different protonation states of selected residues available during the course of a simulation in one of two ways. (Martins De Oliveira et al., 2022) Discrete CpHMD runs normal MD interrupted by Monte Carlo (MC) moves, where a protonation-state switch is trialled, the associated potential energy change ΔU evaluated and the move is accepted or rejected using the Metropolis criterion

$$P_{\text{accept}} = \min\{1, e^{-\beta \Delta U}\}. \tag{2.72}$$

Continuous CpHMD methods, by contrast, assign a titration coordinate λ to each titratable residue (or several coupled λ's to take into account different tautomeric states) that has a fictitious mass and gets coupled to the system dynamics in an extended ensemble approach. (Henderson et al., 2022) Although this does sample unphysical intermediate protonation states transiently, there is an advantage in propagating the dynamics of coupled residues together.

Since changes in protonation states come with substantial rearrangements of the solvation shell around the affected residues, an important consideration for CpHMD is the water representation used.

With discrete methods, there usually are near-zero acceptance probabilities of MC moves in explicit solvent, necessitating the use of implicit solvent representations. This can take the form of fully implicit-solvent MD or hybrid schemes which only evaluate the protonation-state changes in implicit solvent. Continuous λ-dynamics methods, in turn, have been been used both with implicit-solvent and all-atom potentials.

In this thesis, we make use of two CpHMD implementations, the hybrid-solvent discrete CpHMD method implemented in AMBER (Swails et al., 2014) and the all-atom continuous method recently implemented in GROMACS. (Aho et al., 2022; Buslaev et al., 2022) The former uses an AMBER force field with explicit solvent representation to propagate system dynamics between MC moves, converts to a generalised Born (GB) implicit solvent model to propose and evaluate protonation-state changes in random order and, after restoring the solvent, includes additional solvent relaxation steps if any protonation states have changed. This method supports a combination with replica exchange, and — as we discuss in chapter 4 — is not rigorous for membrane proteins because the membrane is not taken into account when evaluating protonation-state switch probabilities in implicit solvent. (Swails et al., 2014) The latter method uses continuous λ-dynamics (with a multi-site model for histidine) in an efficient implementation with linear scaling of partial charges instead of potentials that notably does not degrade performance with incorporating larger numbers of titratable side chains. (Aho et al., 2022) The authors use a modified version of the CHARMM36 force field, where they have decreased side-chain torsion barriers of aspartate residues to tackle convergence issues in their test systems. (Buslaev et al., 2022) Replica exchange is not supported in this implementation.

When running a CpHMD simulation, one specifies a pH value that affects the energetics of protonation-state changes throughout the simulation. A goal of using CpHMD is often to calculate pKa values of certain residues of interest. This is best done using a computational titration approach, where replicate simulations are run at a range of pH-values (with or without replica exchange, as the implementation permits; see for example Swails et al. (2014)). The deprotonated / protonated fractions f are then collected and fitted to a Hill equation with Hill coefficient n:

$$f = \frac{1}{1 + 10^{n(\mathrm{p}K_a - \mathrm{pH})}}. \tag{2.73}$$

With the statistical-mechanical and computational methodological framework presented in this chapter in place, we can now move to investigating in more detail the problem of hysteresis in enhanced sampling MD before applying the range of simulation methods we have introduced to the transport mechanism of proton-coupled oligopeptide transporters (POTs).

3

Tackling hysteresis in conformational sampling — how to be forgetful with MEMENTO

Disclaimer

The research presented in this chapter has been published in J.C.T.C in 2023 (Lichtinger and Biggin, 2023), see also in the list of publications. All investigation, writing and figures in that publication were produced by myself with feedback from the co-author, and will, with adaptations, be used here under the CC BY 4.0 license. Images and phrasing may therefore be similar or partially identical to my published work.

3.1 The challenge of hysteresis in sampling conformational transitions

Molecular dynamics (MD) simulations promise to place static protein structures into their dynamical context, which frequently involves large-scale conformational changes. (Orellana, 2019) Thanks to rapid advances in structural biology, one may have information about several conformational states of a given protein available, but the mechanism, energetics and kinetics of their interconversion often remain unknown. This establishes a vital, continuing role for MD in the study of protein structure that comes not without challenges. For example, the structural knowledge base about solute-carrier membrane proteins (Pizzagalli et al., 2021) is continuously expanding, but the driving forces of their conformational changes are an active area of research that poses significant challenges for MD. While tens of microseconds can now be simulated on commercial hardware (rising to hundreds of microseconds with parallelisation), this is still orders of magnitude shorter than many conformational changes, which often reach into regimes of milliseconds and beyond. (Jackson, 1993; Meshkin and Zhu, 2021; Tokonami et al., 2022; Chan et al., 2022) Moreover, to accurately describe conformational dynamics at equilibrium, one needs to observe repeated transitions to obtain good statistics. Enhanced sampling methods can help address this issue in various ways: by adaptive spawning of new trajectories, by adjusting potential energy barriers, or by biasing progress along specific reaction coordinates, termed collective variables (CVs). (Hénin et al., 2022) Prominent enhanced sampling methods are reviewed in the introductory section 2.2 of this thesis.

With the goal of obtaining a potential of mean force (PMF) — or free energy surface (FES) — of a given conformational change, it is conceptually useful for the majority of the above methods to separate the task into two parts. One first needs to obtain an initial path connecting two known end states, followed by focusing extensive sampling on the vicinity of this path in phase space to gather a PMF. A popular approach is to use targeted MD (tMD) (Schlitter et al., 1993) or steered MD (SMD) (Grubmüller et al., 1996; Izrailev et al., 1999) (see the discussion in section 2.2.3) — here, we choose SMD with a CV based on the root mean squared deviation (RMSD) to a target state — to generate a path, followed by replica-exchange umbrella sampling (REUS) (Sugita et al., 2000) and processing with the weighted histogram analysis method (WHAM). (Kumar et al., 1992) We note here that although time-dependent biasing schemes like metadynamics (Laio and Parrinello, 2002; Barducci et al., 2008) do not separate path generation and free energy sampling, they still internally construct transition paths by biased sampling and will suffer from

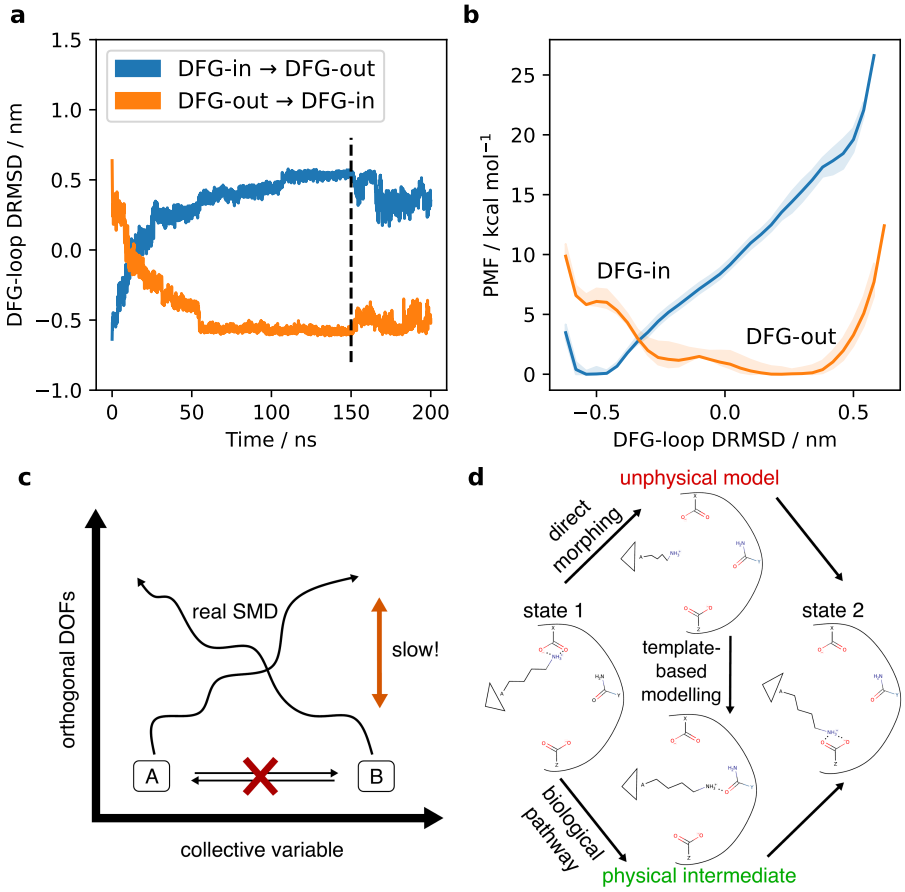


Figure 3.1: (a) Steered MD (SMD) between DFG-in and DFG-out conformations of P38α. Bi-directional steers on the RMSD to the respective target configuration, projected on the difference of RMSDs (DRMSD). (b) Replica-exchange umbrella sampling (REUS) on the obtained transition paths shows strong hysteresis. (c) Schematic representation of the problem of orthogonal degrees of freedom (DOFs) in SMD. (d) Schematic overview of the MEMENTO procedure, fixing unphysical morphing intermediates by template-based modelling.

similar issues to those described below, which we also show for PepT2 in figure 4.8.

In figure 3.1 a-b, we present an example of what the SMD + REUS approach may yield in practice. We ran SMD between the DFG-in and DFG-out conformations of P38α (details of this system are provided in section A.3). In projection onto a simple collective variable (DRMSD $=$ RMSD$_{\text{(DFG-in)}}$ $-$ RMSD$_{\text{(DFG-out)}}$, where the whole proteins are aligned but the RMSD is calculated for the DFG loop only), both biasing directions appear to produce metastable states in their respective target conformations (figure 3.1a). Proceeding to REUS, however, reveals substantial starting-state dependence (or hysteresis), as shown in figure 3.1b. Generally speaking, when a bias is applied to a CV to move between two conformations, the implicit assumption is that orthogonal degrees of freedom (DOFs) will equilibrate within the available sampling. Where this is not the case, as we show schematically in figure 3.1c, the result is a PMF skewed to the starting state of path generation.

It is widely recognised that the quality of paths from tMD and SMD depends strongly on the choice of CV. (Huang et al., 2009; Ovchinnikov and Karplus, 2012; Pan et al., 2014) In particular, global RMSD-based CVs have been shown to be biased in a “large-scale-first” fashion, where large-scale features of a target state

get reconstituted before those at smaller scales as an unphysical artifact of the global best-fit alignment step. (Ovchinnikov and Karplus, 2012) The potential for hysteresis when running SMD between two protein conformations has also been discussed in specific cases. For example, Meshkin and Zhu (2021) note that initially they observed hysteresis of tens of kcal/mol when sampling the transition between the occluded and outwards-facing states of the Mhp1 transporter. Only through "repeated trials and errors" could they design a set of better CVs that reduced hysteresis to an "acceptable level". To our knowledge, however, there has not been a dedicated investigation of hysteresis in conformational sampling of the scope we attempt to provide as the first objective in this chapter. CV choice and path generation are evidently coupled problems, since an ideal CV would generate a history-free equilibrium path between protein conformations with SMD. Because an optimal CV space is not known in most cases, one needs to consider either how to discover or approximate it, or how to build good paths independent of suboptimal CVs to eliminate starting-state bias. We show in this chapter that such paths may perform better than SMD paths in umbrella sampling (which still requires CVs), even without knowing optimal CVs beforehand.

There are therefore, broadly speaking, two ways to address hysteresis in PMFs of conformational changes. A researcher can attempt to design better CVs (see section 2.2.2) that capture all slow movements of the protein by using extensive physical knowledge and restraints (recently demonstrated on a membrane transporter by the aforementioned Meshkin and Zhu (2021)), by combining a larger number of CVs in bias-exchange metadynamics (Piana and Laio, 2007) or — with some limits on intuitive interpretability — using a host of new machine-learning approaches. (Chen, 2021; Mehdi et al., 2024) Alternatively, one may wish to construct a path of intermediate structures that is free of hysteresis. In transition path sampling (Dellago and Bolhuis, 2007; Juraszek et al., 2012), an ensemble of transition paths can be constructed from one rare event trajectory, and the string method (Maragliano et al., 2006; Pan et al., 2008) has been applied with success in transitions as complex as membrane transporter conformational changes. (Moradi et al., 2015; McComas et al., 2023)

Since SMD paths like the P38α runs we presented above can be highly metastable — to the extent that they appear converged in REUS even with hundreds of nanoseconds of simulation time per window — we decided not to attempt refining paths with MD. Instead, we focus on alternative path-generation algorithms. From various takes on morphing (coordinate interpolation) (Weiss and Levitt, 2009; Castellana et al., 2013; Lee et al., 2021) and rigid-body approximations (Wells et al., 2005; Farrell et al., 2010) to elastic network-based models (López-Blanco and Chacón, 2016; Wang et al., 2020a) and elaborate analysis–biasing iterative schemes (Li and Gong, 2022), the field is ripe with ways to connect protein conformations. Unfortunately, the potential for hysteresis in umbrella sampling remains understudied because a full validation of the PMFs is often not undertaken. Furthermore, such validation is hampered by poor code availability and user-friendliness in some of the more complex approaches.

For these reasons, we asked what might be the most conceptually simple, easy-to-implement way to generate paths that can eliminate hysteresis in umbrella sampling. Morphing is by definition history-free, but the intermediates of linear coordinate interpolations are unphysical: bonds and angles are distorted and side-chain interactions are disrupted. Many tools exist to fix morphing intermediates (such as RigiMOL in PyMOL and a panoply of morph servers, Krebs (2000); Seo and Kim (2012); Veevers and Hayward (2018); Castellana et al. (2013)) but they have been developed mainly for visual purposes and lack substantial MD validation. On the other hand, the MODELLER package (Šali and Blundell, 1993) has been used and improved for decades to prepare starting structures for MD simulations. Here, we find that its algorithm

based on satisfying probability distributions for angles and dihedrals together with side-chain interactions is indeed able to fix unphysical morphing intermediates.

This approach for history-free path generation, which we term MEMENTO (Morphing Endstates by Modelling Ensembles with iNdependent TOpologies, illustrated in figure 3.1d), is validated on four example systems: the toy peptide deca-alanine, the enzyme adenylate kinase (ADK), the kinase P38α and the membrane transporter LeuT. At each stage, we rigorously compare our results to what is attainable using SMD alone and find that MEMENTO outperforms SMD on all systems except for deca-alanine, where they behave equally well. We also show how, by combining long end-state sampling with MEMENTO, suitable CVs for umbrella sampling can be developed iteratively. Although we see this research as presenting mostly general findings on how to best sample conformational changes, we also release a documented, tested and user-friendly python implementation of MEMENTO as the path-generation package PyMEMENTO.

3.2 The MEMENTO method

To automatically morph protein end states and fix the intermediates with MODELLER, we designed a python package. The source code, documentation and examples are available on github: https://github.com/simonlichtinger/PyMEMENTO. Our package uses the MDAnalysis (Gowers et al., 2016), numpy (Harris et al., 2020), pandas (McKinney, 2010), matplotlib (Hunter, 2007) and GromacsWrapper (Beckstein, 2021) packages.

3.2.1 Outline of the basic MEMENTO procedure

In the simplest application cases, we perform five key steps.

1. **Coordinate interpolation**
   A linear morph between two protein coordinate files is calculated. If $\mathbf{X}_i$ is the coordinate of the i$^{\text{th}}$ atom of a protein, then:

   $$\mathbf{X}_i(n) = \mathbf{X}_i(\text{start}) + \frac{n}{N_{\text{windows}} - 1} * (\mathbf{X}_i(\text{end}) - \mathbf{X}_i(\text{start})), \tag{3.1}$$

   where n runs from 0 to $N_{\text{windows}} - 1$ and N_{windows} is the number of windows one wishes to use for subsequent umbrella sampling. In this work, we find that 24 windows are usually sufficient for good sampling. This is the number of windows used for all examples unless otherwise stated.

2. **Template-based modelling**
   MODELLER (Šali and Blundell, 1993) is used to generate N_{models} models for each set of morphed atomic coordinates — that is, unphysical geometries from the previous step. This fixes the geometry and provides reasonable side-chain orientations and interactions. We achieve this in the MODELLER package by running model generation with the protein sequence aligned onto itself. The output is an ensemble of models that are close to the morphed intermediate state, but have reasonable geometry and side-chain interactions. In this work, we always generated 50 models per intermediate to provide sufficient side-chain conformational diversity for picking a smooth path in the next step.

3. **Finding a smooth path**
   In our initial trials, we noticed that in the MODELLER ensembles there was spread in some side-chain rotamers, which we feared might prove to be sampling bottle-necks if the path becomes too rough by jumping between these rotamers. There may also be concern that MODELLER could be biased

towards the end-state which will be likely lie lower in energy than transition-state like structures (although we never saw any indication of this being an issue in our work). We therefore wished to pick that model out of each intermediate ensemble (each model being by itself a reasonable geometry fix) which allows for the smoothest overall path and minimises any discontinuities. $N_{\text{models}}^{N_{\text{windows}}}$ paths are possible between the end states, which is usually an astronomically large number. We therefore attempt to find a smooth path through the model space by running Monte Carlo annealing to minimise an energy proxy:

$$E_{\text{proxy}} = \sqrt{\frac{1}{N_{\text{windows}}} \sum_{n} \text{RMSD}(\mathbf{X}(n), \mathbf{X}(n+1))^2}, \tag{3.2}$$

where $\mathbf{X}(n)$ are the heavy-atom coordinates of intermediate n out of N_{windows}. In essence, we minimise the RMSD of RMSDs between neighbouring frames in the path, which gives a penalty to discontinuities with its quadratic term as the equivalent of a strain energy.

At step s of our Monte Carlo chain we have a path $p(s)$, that is a sequence of N_{windows} integers $p_i(s)$ for which $0 \leq p_i(s) < N_{\text{models}}$. We initialise these as random numbers in the appropriate range. A Monte Carlo step is then the random exchange of a model for another one from the ensemble at one of the intermediates, i.e. the random assignment of one element of $p(\text{s})$ to give $p(s+1)$. This is then accepted or rejected according to the Metropolis criterion and the energy proxy stated in equation 3.2. The temperature proxy progression used in the Metropolis algorithm through the annealing procedure was determined by trial and error to work well as

$$T(s) = \frac{T(0)}{(1 + 20 * s/s_{\text{max}})}, \tag{3.3}$$

where the initial temperature proxy is $T(0) = 50$ and the run continues for $s_{\text{max}} = 10^4$ steps. We always run 12 replicates of this procedure and pick the path with minimum energy sampled in the combined runs. Figure 3.2 illustrates the Monte Carlo trajectories we obtained for the adenylate kinase open $\leftrightarrow$ closed transition. While it would be difficult to properly converge to a global minimum path, here we are merely interested in some appropriately smooth path.

4. **Processing of the frames**
   The frames of the determined path are then subjected to common MD simulation preparation steps: capping of termini (if appropriate), solvating MD boxes and adding ions. We have automated these within the package — for this we rely on various 'gmx' tools in the GROMACS simulation engine. (Abraham et al., 2015) Details are provided below for how ligands and lipids are handled. PyMEMENTO supports cubic/orthorhombic and hexagonal boxes, as well as the CHARMM and AMBER forcefields.

5. **Equilibration**
   For each of these boxes, in addition to the equilibration protocol described in appendix A, we run extensive further NPT equilibration with position restraints (1000 kJ mol^{-1} nm^{-2}) on Cα atoms, to ensure that all side chains have found suitable interaction partners. For deca-alanine, ADK and P38α, we use 10 ns per window, for LeuT and holo-ADK we increase this to 90/100 ns to allow for the membrane and the ligand to equilibrate to a new protein conformation.

From these equilibrated path frames we start our umbrella sampling simulations as detailed in the section on MD simulations. PyMEMENTO provides some additional functionality for setting this up with

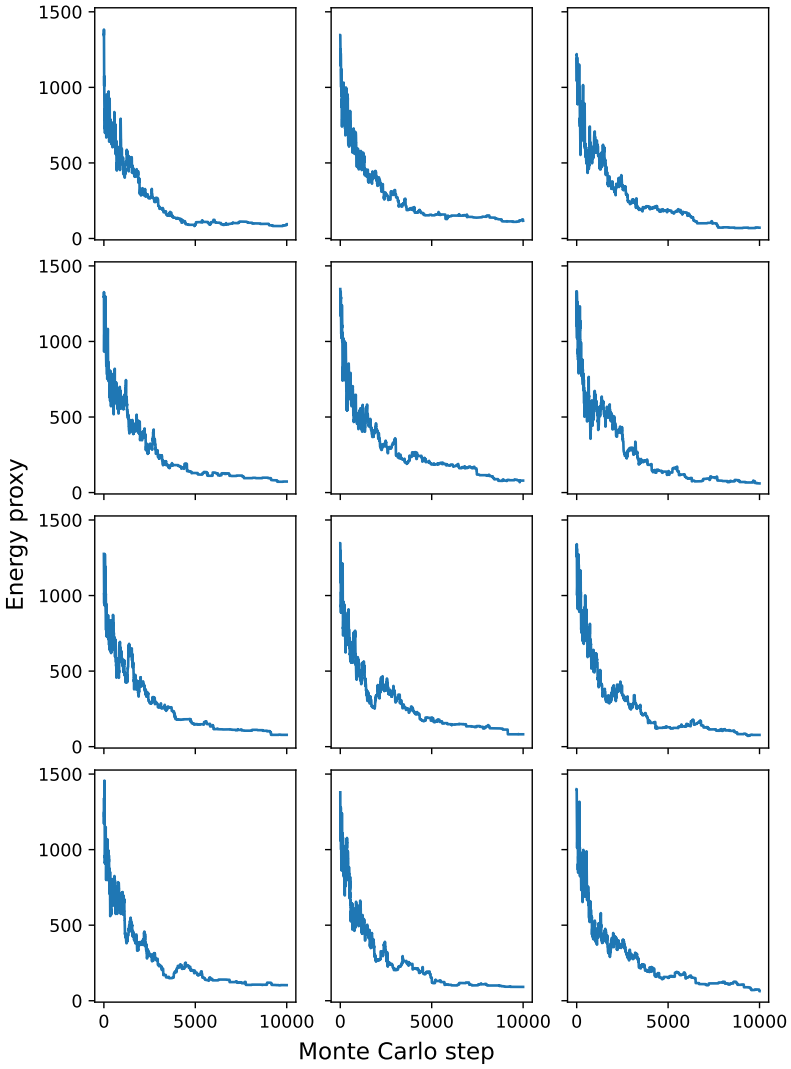


Figure 3.2: Representative Monte Carlo simulated annealing trajectories for the ADK open ↔ closed transition in 12 replicates, as implemented in the PyMEMENTO package. The proxy temperature starts at 50 and is decreased according to equation 3 (main text) through each run.

Gromacs + PLUMED, but the process will fall to the user if a more involved setup (for example 2D umbrella sampling) is required.

3.2.2 Adding ligands

It is often instructive to study how a ligand affects the conformational landscape of a protein, and for our ADK and P38α systems we have investigated this question with crystallographic ligands. However, for both of these proteins the ligand binding pose is only known for one of the conformational states (and may not bind tightly to the other). We therefore aligned the end states and manually inserted the holo-state ligand conformation into the apo state of the end state without known ligand pose, which is not a general approach but in these cases resulted in meta-stable ligand-bound poses after energy minimisation. PyMEMENTO cannot fix internal ligand coordinates, but provided with holo end states thus prepared, it can include the ligand in all intermediate boxes after fractional translation of the centre of mass. The ligand pose can then relax in our long NPT equilibrations with protein Cα restraints. The assumption is therefore

that the ligand does not move much during the conformational change and will relax on the given timescale.

For ADK and P38α, this assumption holds well enough for computing PMFs — we validate this in figure 3.3 against the alternative approach of using the best-scoring docking pose (obtained on the same MEMENTO intermediates using default options and 10 poses per window in the GNINA docking software (McNutt et al., 2021)). We found that the ligand poses produced by MEMENTO after equilibration with protein Cα restraints give docking scores only slightly lower than the best docked poses (figure 3.3a–b), however they are much smoother as measured by an RMSD metric of ligand poses between windows (when aligning the protein, figure 3.3c–d). We also attempted to compute a holo-ADK 2D PMF equivalent to that of figure 3.8d starting from the best docking poses. The shape was qualitatively similar (figure 3.3e), but convergence was much degraded (figure 3.3f), presumably because of slow interconversion between the 'jumpy' docked poses in REUS. For the purpose of this project, we have therefore restricted ourselves to the simple centre-of-mass morphing + equilibration approach. To tackle conformational changes that have a larger component of ligand movement in future applications, a possibility open for exploration could be to implement additional functionality to use an MC-annealing procedure equivalent to the one employed to smoothen modeller ensembles for picking docked poses.

3.2.3 Adding lipids

Our LeuT transporter example case is a membrane protein and therefore needs to be simulated within a membrane box. MODELLER cannot fix lipids and attempting to morph them would likely be futile owing to the flexibility of their rearrangement. Instead, in our approach we can provide one of the end states embedded in a lipid membrane; we recommend using the lipid environment of the protein conformation which occupies the larger area in the membrane. PyMEMENTO uses this pre-equilibrated membrane to embed all other conformational intermediates in lipids in a procedure inspired by the popular inflate-gro script. (Kandt et al., 2007) We first stretch the aligned membrane in its plane to make sure the protein fits (here by 15%), followed by cycles (here 5 cycles) of compression (here by 4% each) and energy minimisation to fit it snugly around the new protein conformation. While this procedure conserves the number of atoms, the box size is not constrained, thus allowing for some conformational states to use a slightly larger membrane patch. The parameters in this procedure are customisable and may need to be adapted for different protein–lipid systems.

3.2.4 End-state sampling, replicates and two-dimensional CVs

As detailed in the sections below, we found great value in running extensive unbiased MD at the end states. Using snapshots from these trajectories to seed MEMENTO, we obtained independent replicates for which we first ran 1D REUS along a naive CV. Where there were significant differences between replicates, we concluded that the end-state sampling had captured a meaningful trend that was propagated through MEMENTO. We therefore concatenated all 1D REUS trajectories (at 5 ns stride for efficiency) and ran principle component analysis (PCA) on the Cα positions using the GROMACS implementation. For P38α, we did this including only the DFG loop (residues 166-176), and also for the entire protein. The two first principle components (PCs) obtained in each of these cases were our 2D CVs for subsequent umbrella sampling.

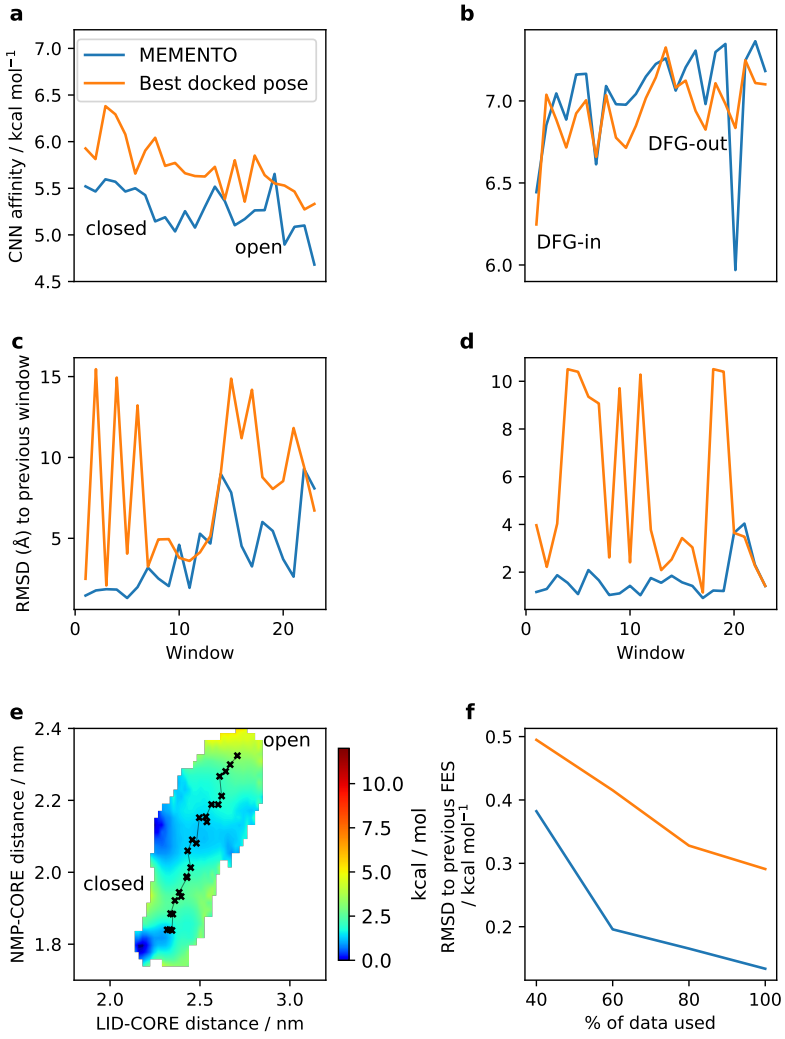


Figure 3.3: Comparison between ligand poses from MEMENTO and the best gnina docking poses (to the MEMENTO protein intermediates) for ADK and P38α. (a) ADK: docking scores for MEMENTO poses (re-score only) and the best of 10 de novo docking poses. Docked poses perform slightly better. (b) P38α: docking scores for MEMENTO poses (re-score only) and the best of 10 de novo docking poses. Docked poses perform similarly except for window 20. (c) ADK: similarity of ligand poses in subsequent windows, as measured by RMSD. Docked ligand conformations jump much more between windows. (d) P38α: similarity of ligand poses in subsequent windows, as measured by RMSD. Docked ligand conformations jump much more between windows. (e) ADK: 2D-holo PMF, starting from MEMENTO protein intermediates and the best gnina docked poses for each window. PMF is qualitatively similar to figure 4d. (f) The 2D PMF from MEMENTO ligand poses converges much better than the one starting from docked poses, as measured by RMSD in the PMF when including successive chunks of data.

For LeuT, using the entire protein for PCA, we found that the first PC unsurprisingly correlated very well with the initial 1D CV; we used it as the first CV for 2D REUS (PC 1). However, the contributions from the remaining PCs trailed off rather slowly, requiring 16 PCs to explain >75% of the variance. We thus decided to make a linear combination of PCs 2–16 that would optimally separate out the replicates (under the assumption that this is where the interesting conformational diversity would sit). We built an in-house script that uses scipy (Virtanen et al., 2020) to maximise by differential evolution (Storn and Price, 1997) an entropy-like metric of distances between MEMENTO path frames:

$$\frac{2}{N_{\text{rep}} * (N_{\text{rep}} - 1)} \sum_{i=1}^{N_{\text{rep}}} \sum_{j=i+1}^{N_{\text{rep}}} \sum_{n}^{N_{\text{windows}}} log(\sqrt{\sum_{\text{PC}} (\mathbf{X}(n,i) - \mathbf{X}(n,j))^2}), \tag{3.4}$$

where N_{rep} is the number of replicates (here 3) and by $\mathbf{X}(n,i) - \mathbf{X}(n,j)$ we denote the distance between two conformational frames in different replicates i and j, evaluated in a projection along a given principal component. The result was termed PC 2 and used as the second CV in 2D REUS.

3.3 Results on deca-alanine

We first sought to verify that MEMENTO can indeed fix unphysical morphs without introducing artifacts. Therefore, we applied the method to a simple toy model: deca-alanine in water (see appendix A for details on simulation setup). This peptide undergoes a transition between α-helical folded and unfolded conformations that has a significant energetic and entropic barrier, but has also been successfully described with umbrella sampling (US) in the literature. (Hazel et al., 2014) Figure 3.4 a-b and the step-through supplementary video 1 of our publication Lichtinger and Biggin (2023) illustrate how reasonable bond lengths, angles and dihedrals are reconstructed by MEMENTO. [1] Note also that the procedure produces evenly spaced intermediates, and thus that there is no bias towards the end-state structures. We ran one-dimensional REUS from these conformations, using the end-to-end distance of the peptide as a CV, and found a sharp free energy minimum at the helical state as well as a broad minimum at an ensemble of extended states (figure 3.4c). We note that the shape of this PMF is compatible with the literature.

To further compare MEMENTO with existing methods, we also ran SMD along the end-to-end distance CV to open up the helical state (figure 3.5a). REUS from snapshots along this trajectory yielded a PMF virtually identical to MEMENTO (figure 3.5b–f). Using three REUS replicates from independent MEMENTO and SMD runs, we also validated that they converge on comparable timescales (figure 3.4d), though convergence is highly variable owing to the inherently stochastic nature of MD. We conclude that MEMENTO performance is on a par with SMD for simple systems, and that it produces high-quality molecular conformers that behave well in US.

[1] The videos referenced in this chapter can be downloaded here.

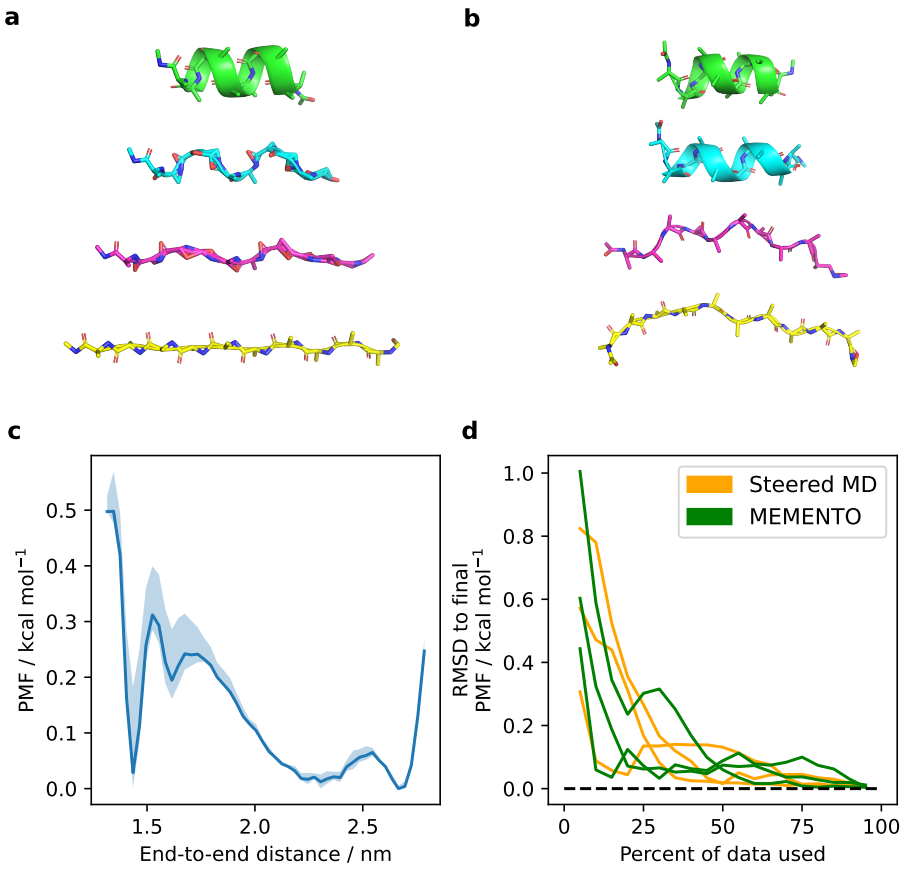


Figure 3.4: (a) MEMENTO windows 0, 7, 15 and 23 after linear morphing, showing unphysical geometry in the intermediates. (b) The same windows after MODELLER processing, re-adding caps and pdb2gmx processing. Unphysical geometry is now fixed. (c) PMF of deca-alanine unfolding in water, sampled by 1D REUS along the end-to-end distance. The shaded area is the range of PMFs observed when taking only the first 60%, the last 60% and the full sampling, which gives an indication of error and convergence. (d) Convergence of triplicate REUS simulations from SMD and MEMENTO paths. Both methods yield paths that converge on comparable timescales for deca-alanine, though convergence is inherently stochastic.

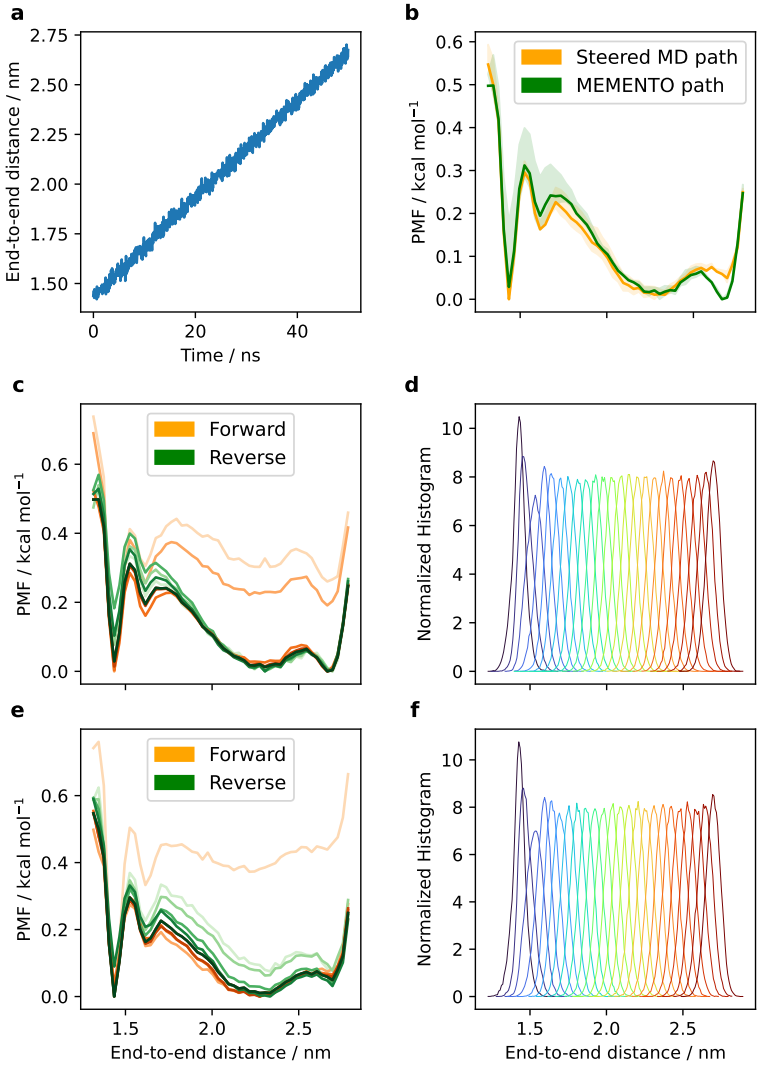


Figure 3.5: (a) SMD progression of deca-alanine along the end-to-end distance CV. (b) PMFs from 1D REUS with SMD and MEMENTO paths overlayed, displaying nearly identical shapes. Shaded area is the range of PMFs observed when taking only the first 60%, the last 60% and the full sampling. (c) Convergence of MEMENTO 1D REUS, shown by the PMF taking (with increasing line saturation) 20, 40, 60, 80 and 100 % of sampling from the front (orange) or the back (green) of trajectories. (d) Normalised MEMENTO 1D REUS histograms, showing excellent overlap. (e)–(f) are equivalent to (c)–(d), but show 1D REUS derived from the SMD path.

3.4 Results on ADK

While the previous example has established that MEMENTO does not degrade sampling of a system where sampling issues have no impact on US within the available timescale, we begin to see its benefits when moving to more complex systems. Adenylate kinase (ADK), an enzyme that catalyses the interconversion of adenosine phosphates, is known to exhibit a large-scale conformational transition between open (Müller et al., 1996) and closed (Müller and Schulz, 1992) states (figure 3.6a). This conformational change may be described either by a varying distance between the LID and NMP domains, or by the distance of LID and NMP domains from the protein CORE. For many years, ADK has served as a benchmark system for computational research on conformational changes and enhanced sampling — we summarise studies which calculated PMFs on ADK in table 3.1. While different studies have obtained substantially different results with various methodologies, the clear literature consensus of atomistic studies is that apo ADK favours the open state while the presence of substrate or inhibitor stabilises the closed conformation. This is also consistent with the crystallographic data and mechanistic intuition. We note that many other papers (Kirillova et al., 2007; Korkut and Hendrickson, 2009; Kubitzki and de Groot, 2008; Ye et al., 2019; Yuan et al., 2020; Kanada et al., 2022) have studied ADK conformational changes without computing a PMF. Systematically comparing these is difficult, however, so within the scope of the current work we focused on broad consensus features of ADK PMFs.

We applied the MEMENTO procedure to ADK (intermediates are shown in figure 3.6b and step-through supplementary videos 2–3 published with our paper, Lichtinger and Biggin (2023)), where we performed independent replicates by equilibration of the open state in unbiased MD (see appendix A for details, closed-state equilibration could not be used to seed replicates due to spontaneous opening in MD). We also ran three replicates of SMD based on a Cα-RMSD CV in the closed $\rightarrow$ open direction, and one replicate in the open $\rightarrow$ closed direction (figure 3.7 a-b; since the closed state opened in unbiased MD, we had no measure for how successful the open $\rightarrow$ closed SMD direction might be — thus we did not pursue it further). Using the LID–NMP domain distance as a simple one-dimensional CV, we computed PMFs for our three MEMENTO and SMD replicates (figure 3.6 c–d, convergence analysis in figure 3.7 c–f).

As expected, the open state was favoured over the closed state by roughly similar free energy differences in all replicates. However, some discrepancies between replicates remained, which prompted us to investigate what a relevant orthogonal DOF may be in this case. We note at this stage, that while it may appear that these inter-replicate differences are more pronounced with MEMENTO paths, in fact SMD appears more consistent here only because it samples the region of phase space responsible for the slow DOF less than MEMENTO — we discuss this in more detail below. By clustering windows 0 (MEMENTO) and 2 (SMD) of the respective first replicates, we found a meta-stable alternative closed state (figure 3.8a), in which the LID is closed on the CORE, but the NMP domain remains open. On inspection, this conformation can be rationalised through a number of positively charged residues that coordinate the highly negatively charged substrate in holo ADK (see figure 3.8c), which can engage in alternative salt-bridges in the absence of ligand: aspartate 33 and arginine 36 (NMP domain) coordinate arginine 131 and aspartates 146 and 147 (LID domain). Several of the studies summarised in table 3.1 have also found that the LID domain can close while NMP remains open in apo ADK simulations. (Shao, 2016; Zheng and Cui, 2018; Peng et al., 2021) This alternative closed state as an orthogonal DOF is sampled to different extents in the MEMENTO replicates (as it was not explicitly included in the simulation setup), leading to the observed differences between our 1D REUS replicates.

Table 3.1: Comparison of our results to published PMFs of the ADK conformational change. We indicate the type of computational model used and which state was found to be favoured in the apo and holo proteins. Our results agree with the consensus of the atomistic studies, while coarse-grained models appear to not capture the energetics of the conformational changes well.

Study	Model	Apo	Holo	Other observations
Arora and Brooks (2007)	atm [1]	open	closed	LID closes before NMP
Lu and Wang (2008)	CG [2]	comp [4]	n/a	LID and NMP closing equivalent
Beckstein et al. (2009)	impl/sol [3]	closed	n/a	open state is plateau
Jana et al. (2011)	atm	int [5]	n/a	LID more flexible
Matsunaga et al. (2012)	atm	open	closed	n/a
Song and Zhu (2013)	atm	open	n/a	n/a
Wang et al. (2013)	CG	comp	comp	LID more flexible
Wang et al. (2014)	atm	open	closed	holo closed has two wells
Li et al. (2015)	atm	open	closed	NMP prefers open in apo
Formoso et al. (2015)	atm	comp	n/a	LID more flexible
Zeller and Zacharias (2015)	atm	open	closed	n/a
Shao (2016)	atm	comp	n/a	Intermediate: LID closed, NMP open
Matsunaga et al. (2016)	CG	comp	n/a	NMP opens before LID
Halder et al. (2017)	atm	int	n/a	NMP more flexible than LID
Zheng and Cui (2018)	atm	open	n/a	LID can close, NMP cannot
Wang et al. (2020a)	atm	open	closed	NMP can close, LID cannot
Wang et al. (2020b)	atm	open	closed	Intermediate at partial LID open
Peng et al. (2021)	atm	open	n/a	LID closure preferred to NMP, but NMP first possible
Lu et al. (2022)	atm	open	closed	native ligand binding pose → closed, non-native → open
This work	atm	open	closed	LID can close in apo, NMP cannot

[1] atomistic
[2] coarse-grained
[3] implicit solvent
[4] comparable, distinct basins for open and closed states of similar depth
[5] intermediate, only one basin that lies between open and closed states

By connecting this alternative closed state to the open conformation with another MEMENTO path, we captured a 2D PMF along the LID–CORE and NMP–CORE distances as CVs (a combination used previously in the literature (Wang et al., 2020a)), shown in figure 3.8b, that converges very well (figure 3.9). In a separate set of simulations including the crystallographic inhibitor AP5A, we likewise achieved a converged PMF — this time favouring the closed protein conformation (figure 3.8d). Consistent with the majority literature conclusions, taken together these PMFs show how the crystallographic closed state is unstable in apo ADK, but also that the LID domain has significant flexibility in the open state — driven by non-native salt-bridge interactions that compensate partially for the lack of substrate. By contrast, holo ADK prefers the closed state, where a number of positively charged residues form salt bridges with the ligand.

We also investigated whether a similar PMF can be recapitulated using only SMD paths. When we projected our SMD runs onto the 2D-CV space, we discovered that the closed → open runs opened the LID before the NMP domain, and the open → closed run closes the LID domain before the NMP domain (figure 3.10). We note that this may have been the reason why inter-replicate differences were

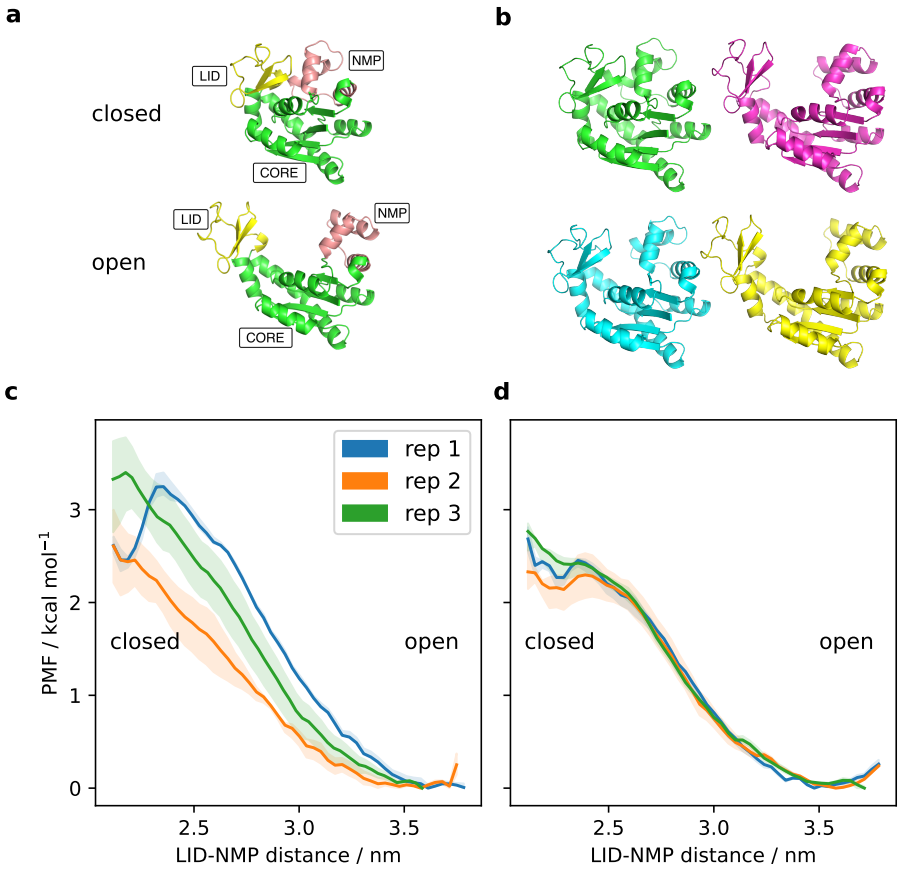


Figure 3.6: (a) An overview of the domain motions in the ADK open ↔ closed conformational change. (b) Representations of MEMENTO ADK intermediates 0, 7, 15 and 23, displaying the required domain motion. (c)–(d) PMFs from 1D REUS of the ADK conformational change along the LID–NMP CV, with paths from (c) MEMENTO and (d) SMD. Shaded areas are the ranges of PMFs observed when taking only the first 60%, the last 60% and the full sampling.

less pronounced in the 1D REUS from SMD paths (figure 3.6d) compared to MEMENTO. Since we only used the opening-direction paths for those calculations, the alternative closed state was sampled less compared with MEMENTO paths. In light of our MEMENTO results regarding LID flexibility, the SMD paths look encouraging, however an attempt to compute a 2D PMF by REUS (figure 3.10b) converged poorly (figure 3.10 c–d), even if the rough shape of the PMF was consistent with our previous results. Notably, our MEMENTO paths are much smoother in CV space (figure 3.8b, as indicated by black crosses and connecting lines on the 2D PMFs) than the SMD paths, which move less evenly (figure 3.10b) and with more noise. Furthermore, we did not succeed in incorporating the alternative closed state explicitly into the PMF, because REUS from bi-directional SMD between the open and alternative closed states showed significant hysteresis (figure 3.10 e–f). We therefore conclude that MEMENTO not only produces results in line with expectations for ADK, but also that there is a tangible advantage over SMD for path generation in this case. The example further highlights how MEMENTO can help identify orthogonal DOFs and thus design better CVs through the possibility of incorporating extensive end-state sampling to propagate conformational diversity that would not otherwise be feasible to achieve across many stratified windows. There is also full flexibility to add extra paths to an existing PMF, which is useful if one does discover relevant alternative conformations.

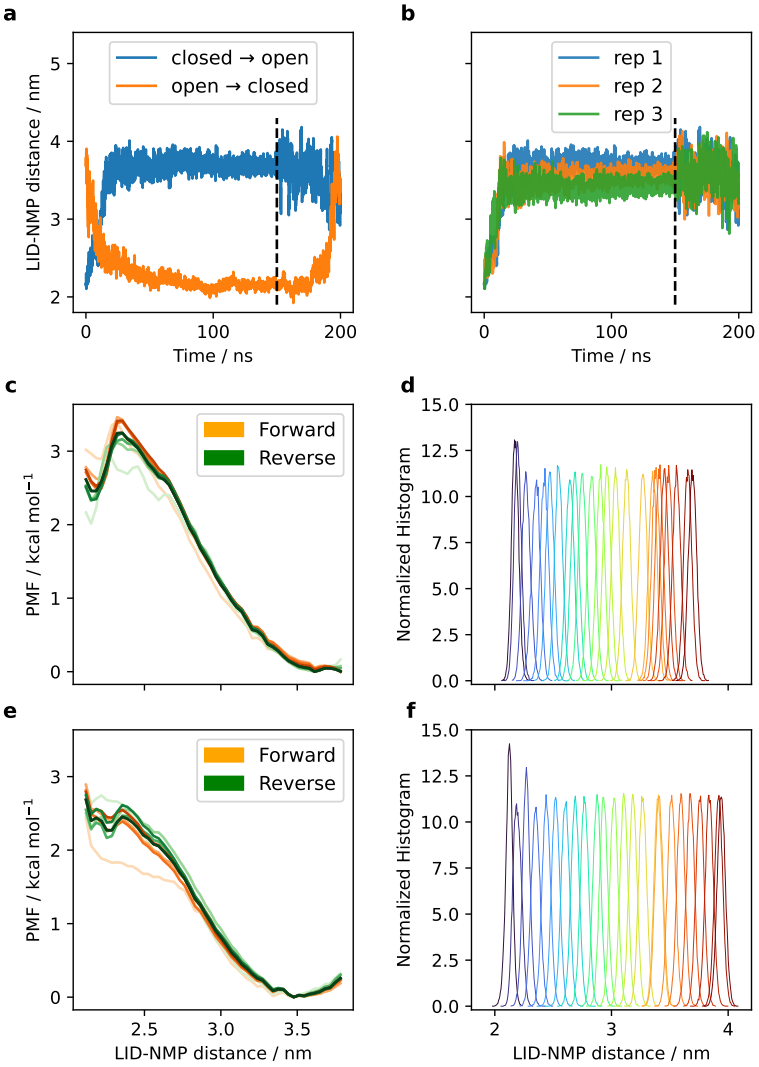


Figure 3.7: (a) ADK SMD biased along the Cα-RMSD, projected onto the LID–NMP distance. The obtained closed conformation is not stable. (b) Three replicates of closed → open SMD all give stable open states. (c) Convergence of MEMENTO 1D REUS rep 1, shown by the PMF taking (with increasing line saturation) 20, 40, 60, 80 and 100 % of sampling from the front (orange) or the back (green) of trajectories. (d) Normalised MEMENTO 1D REUS rep 1 histograms, showing sufficient overlap. (e)–(f) are equivalent to (c)–(d), but show 1D REUS rep 1 derived from a SMD path.

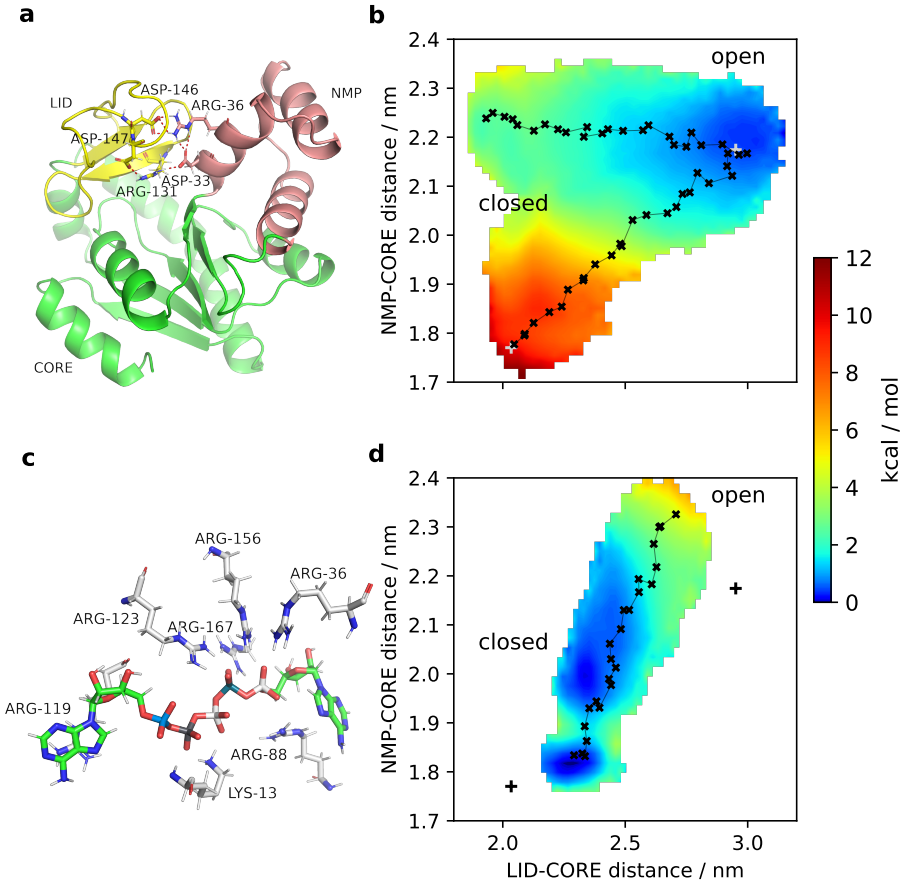


Figure 3.8: (a) An alternative closed state identified by clustering trajectories from 1D REUS with MEMENTO paths. The LID domain is closed while the NMP domain remains open. (b) PMF from 2D REUS along the LID–CORE and NMP–CORE CVs, with MEMENTO paths connecting the closed, open and alternative closed states. Crosses indicate the REUS window starting frames, connected in sequence. Plus-symbols (+) indicate the location of the crystal structures in the CV space. (c) AP5A inhibitor binding pose in the 1AKE crystal structure, illustrating how the negative ligand charge is accommodated by multiple positively charged protein residues. (d) PMF from 2D REUS in the presence of AP5A, displaying a switch of the conformational preference to the closed state.

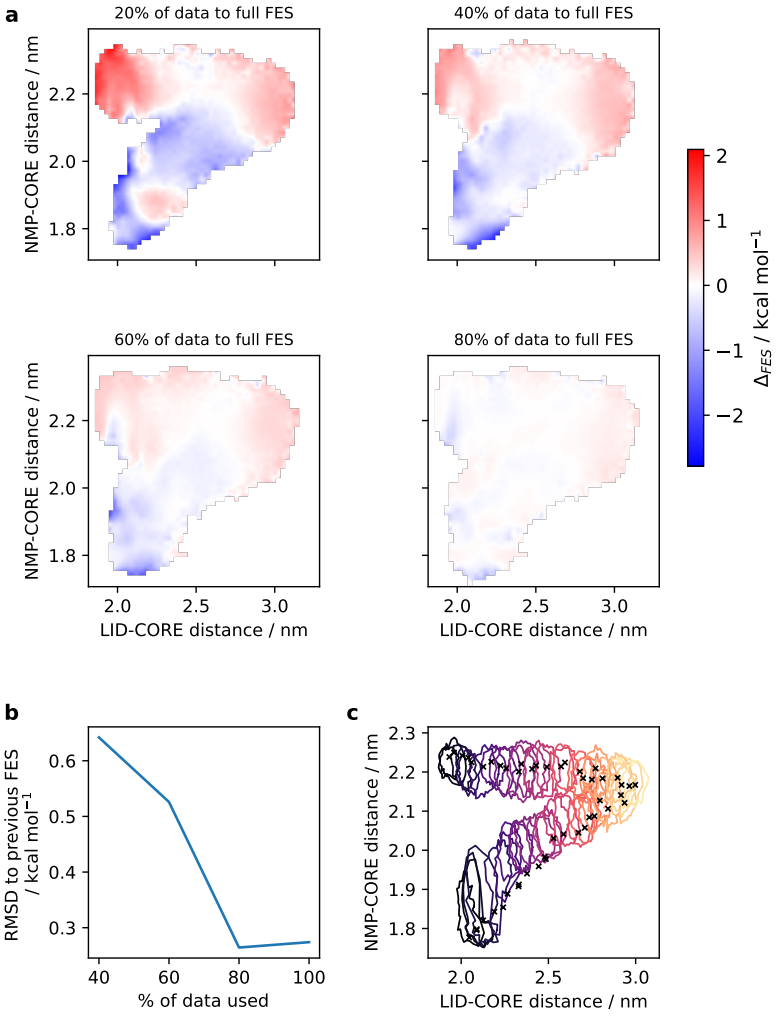


Figure 3.9: (a) Convergence of ADK 2D REUS from MEMENTO paths, shown as the difference between the PMFs including 20, 40, 60 and 80% of the data and the full PMF, thus highlighting regions of faster and slower convergence. (b) Alternative representation of convergence as the RMSD from adding successive 20% chunks of data into the PMF computation. (c) 2D-histograms shown as paths encircling the normalised histograms at 30% of their peak height, displaying satisfactory overlap. Crosses indicate the REUS window starting frames.

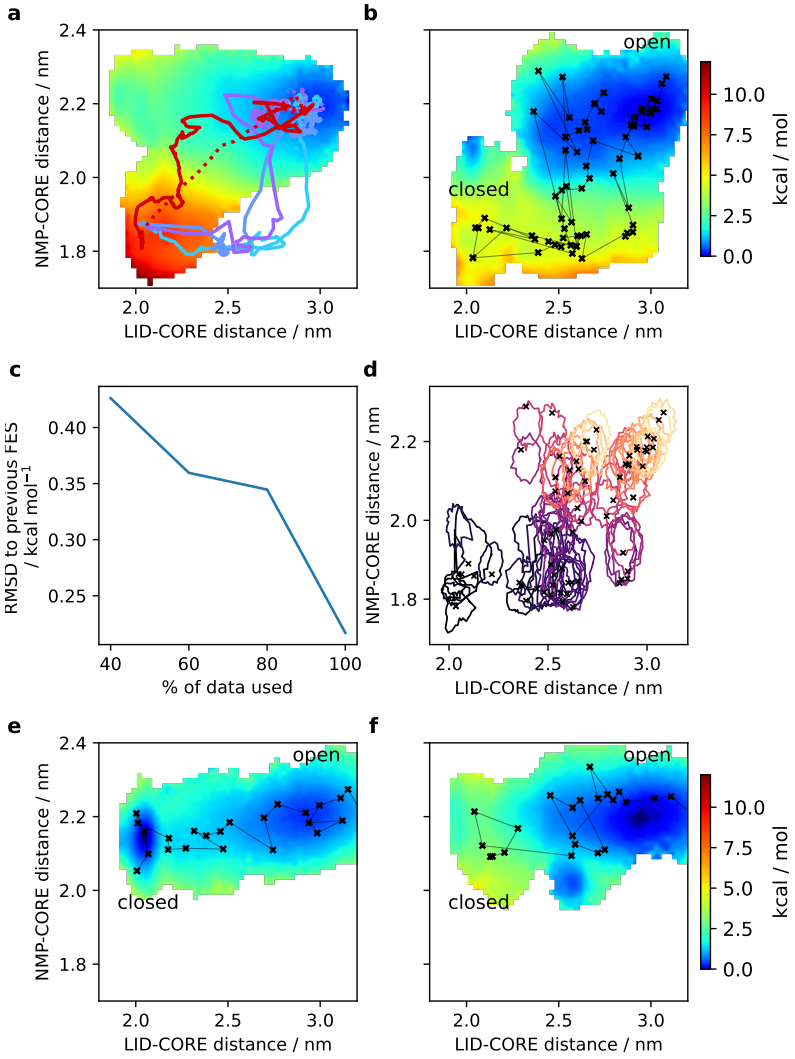


Figure 3.10: (a) Projection of the ADK SMD runs onto the LID–CORE and NMP–CORE distance 2D-CV space. Blue hues represent closed → open runs, the red line is the open → closed run. Solid lines are the biased portion, dashed lines the relaxing portion of the runs. The 2D PMF from MEMENTO is shown as reference in the background. (b) PMF from 2D REUS based on closed → open SMD runs. Crosses indicate the REUS window starting frames, connected in sequence. (c) Convergence of the 2D PMF as the RMSD from adding successive 20% chunks of data into the PMF computation. (d) 2D histograms shown as paths encircling the normalised histograms at 30% of their peak height, overall satisfactory overlap with issues in some regions. (e)–(f) PMFs from 2D REUS from the (e) alternative closed → open and (f) open → alternative closed state SMD runs, showing significant bias towards the SMD starting state (hysteresis).

3.5 Results on P38α

To provide a second example of a conformational change in a globular protein (that is not as much of a canonical example in the enhanced sampling field as ADK, but for which previous work still exists) we next focused our efforts on P38α, a mitogen-activated protein kinase with roles in heart physiology and implications in heart disease. (Marber et al., 2011) P38α undergoes a conformational change in its DFG loop, where the DFG-in conformation (Wang et al., 1997) is active, while the DFG-out state (Gill et al., 2005) is inactive (illustrated in figure 3.11a). It is known from nuclear magnetic resonance (NMR) (Vogtherr et al., 2006) and electron paramagnetic resonance (EPR) (Roser et al., 2018) spectroscopy studies that apo P38α displays a conformational equilibrium between DFG-in and DFG-out geometries, and that type 2 inhibitors (which bind to the DFG-out conformation) reduce the accessibility of the active DFG-in conformation. Several computational studies have broadly come to the same conclusions using various enhanced sampling methods (Wang et al., 2014, 2020a; Peng et al., 2021), although inhibitors — where they were simulated — are not found to abolish completely the (meta-)stability of a DFG-in-like state, but merely raise it in energy relative to DFG-out.

As for ADK, we first ran 500 ns of unbiased equilibrations at all combinations of DFG-in/out and apo/holo states. In the introductory figure 3.1 a-b, we have already demonstrated how SMD suffers from large hysteresis in 1D REUS along a DRMSD calculated for the DFG loop only. To see how a history-independent path would perform in comparison, we ran three MEMENTO replicates (step-through supplementary videos 4–5 of our publication, Lichtinger and Biggin (2023)), followed by 1D REUS along the DRMSD (figure 3.11b). While the results already look more reasonable than SMD, the substantial differences between replicates we found led us to hypothesise that our end-state sampling included significant conformational relaxation which was propagated through MEMENTO, but was of too large a timescale to be equilibrated in each REUS window. To find a set of CVs which includes this slow DOF, we opted to extract relevant information from the 1D REUS sampling we had already accumulated, in an iterative fashion via PCA, see section 3.2.4 for details. By including the DFG loop only or the whole protein in the PCA, we hoped to obtain PCs that separate the DFG-loop motion from a superimposed protein conformational bias (figure 3.11c and supplementary videos 6–7 of our publication illustrate the highest contributing PCs in each case, which we used as CVs in subsequent simulations). The loop PC expectedly covers a motion that would interconvert DFG-in and DFG-out loop conformations. The protein PC also carries an element of DFG-loop motion, but in conjunction with a twist in the protein along the thin hinge that lies under the DFG loop.

Subsequent 2D REUS we ran along the same MEMENTO paths with the new CVs produced a PMF that was highly instructive about the processes at work (figure 3.11d). Projecting our initial apo-P38α unbiased MD (that started from the crystallographic DFG-out conformation) onto this PMF (figure 3.12), one can clearly see how the protein assumes an alternative, less extreme DFG-out geometry on removal of the crystallographic ligand. Replicates 2 and 3 therefore start from this free energy basin, which explains why these replicates showed an increasing bias towards DFG-out in 1D REUS, and also why replicate 1 converged poorly in the DFG-out region. Because the relaxation movement is now included in the PMF, convergence stands much improved and is no longer worse in some regions of the PMF than others (figure 3.13).

With a working set of CVs at hand, we further wished to investigate whether the 1D REUS hysteresis with SMD paths we observed was only caused by our poor initial 1D CV guess for umbrella sampling, or whether the path exhibited starting-state dependence also in our more detailed 2D-CV space. We therefore projected our SMD replicates onto the 2D PMF (figure 3.14a) and found that replicates 1 and 3 appeared

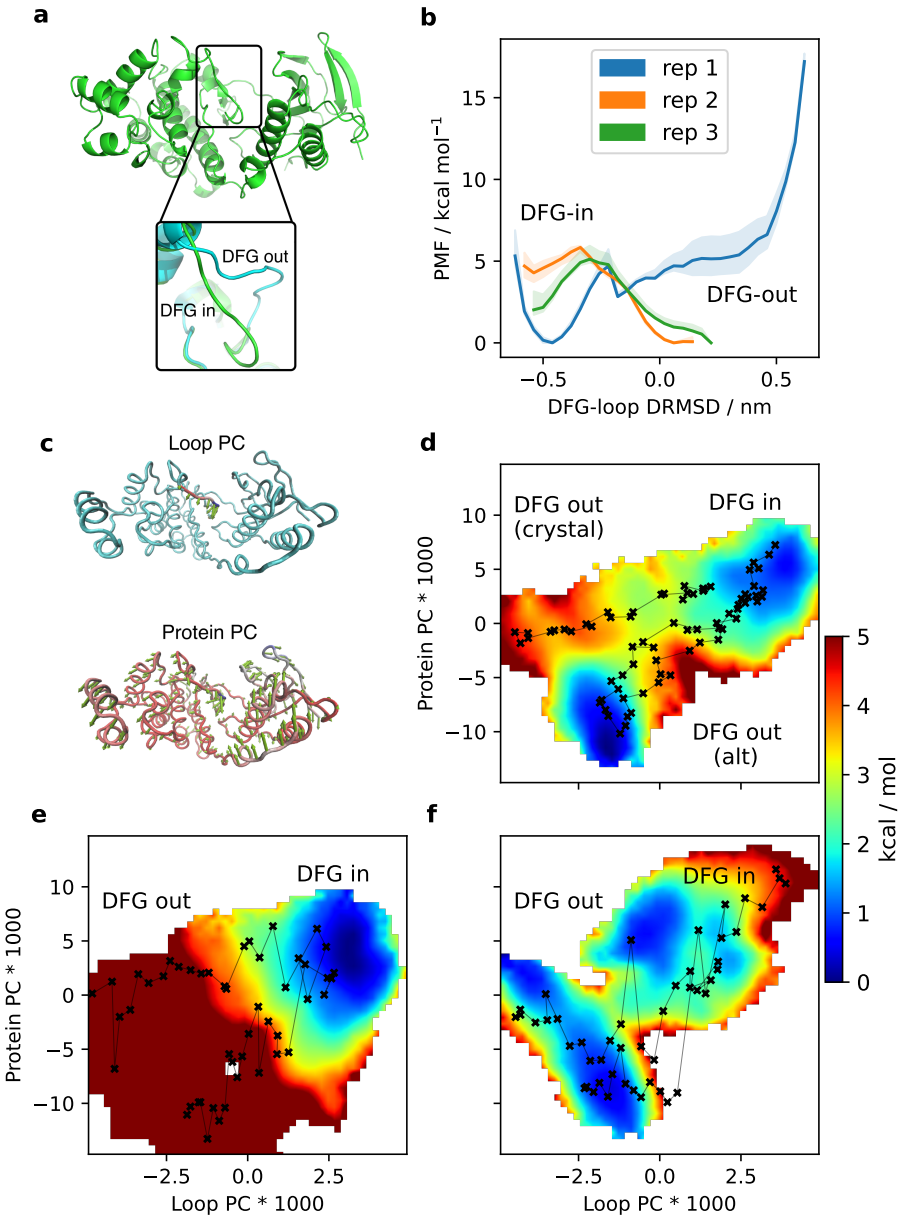


Figure 3.11: (a) Overview of the P38α system, highlighting DFG-in and DFG-out states. (b) 1D REUS along the DRMSD CV from MEMENTO paths, showing big differences between replicates. Shaded area is the range of PMFs observed when taking only the first 60%, the last 60% and the full sampling. (c) PCA results, showing the two CVs obtained for 2D REUS sampling. (d) 2D REUS along PCA-CVs with MEMENTO paths. Crosses indicate the REUS window starting frames, connected in sequence. (e)–(f) 2D REUS from SMD paths, in the (e) DFG-in $\rightarrow$ DFG-out and (f) DFG-out $\rightarrow$ DFG-in directions.

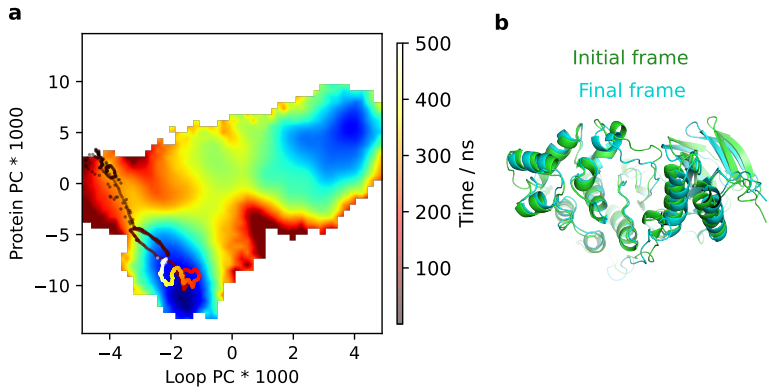


Figure 3.12: (a) Projection of the apo-P38α unbiased MD run starting from DFG-out onto the 2D PMF with PCA-CVs, showing a relaxation towards an alternative DFG-out state. (b) Rendering of the alternative DFG-out state, showing a lesser extent of DFG-loop opening and a general twist in the protein.

to lead to meta-stable states in the vicinity of their respective targets, even though the paths were not as smooth as MEMENTO in this CV space. However, in 2D REUS, the paths still showed significant bias towards the respective SMD starting states (figure 3.11 e–f). This is not a simple convergence issue within sampling time-scales similar to those we used for MEMENTO REUS (figure 3.14 c–f), although the DFG-out $\rightarrow$ DFG-in SMD direction has some issues with REUS histogram overlap leading to noticeably slower convergence. We also investigated whether the hysteresis might be predominantly due to slow solvent DOFs that could be addressed by re-solvating the SMD intermediates, but found that this does little to improve the situation (figure 3.14b). In conclusion, while better CVs may well theoretically capture those DOFs that produce hysteresis in SMD for P38α, they are not as simple to obtain as with MEMENTO. We have not managed to develop any here, and we also note that in a prospective study these would have to be known before running SMD. It would not be possible to re-use paths in iterative CV refinement as easily as we do with MEMENTO, because hysteresis is intrinsic to the paths and does not disappear when switching to better CV spaces for REUS.

Finally, we wished to explore the role of the crystallographic type 2 inhibitor that was bound in the DFG-out structure (Gill et al., 2005) (figure 3.15a). Using the same 2D CVs as before, we obtained a converged PMF (figure 3.15 b–d) that showed a free-energy basin near the DFG-out crystal structure, in contrast to the apo state. However, the PMF has low free energy barriers to conformations that register as DFG-in to our CVs, which contradicts the experimental evidence. Noting that other MD studies (using 1D US) (Wang et al., 2014, 2020a) also found DFG-in conformations to be accessible in holo P38α, we judge that this may either be an issue with MD models of this system, or that what we observe as DFG-in-like in our simulations may not be experimentally relevant DFG-in conformations, thus hinting at a separate CV problem specific to the holo state. To our knowledge, this issue has not been addressed before and it would be worthwhile to do so in the future. However, we see this as beyond the scope of the present study, given a general investigation of hysteresis in enhanced sampling, not the specific FES of the P38α system is of interest here.

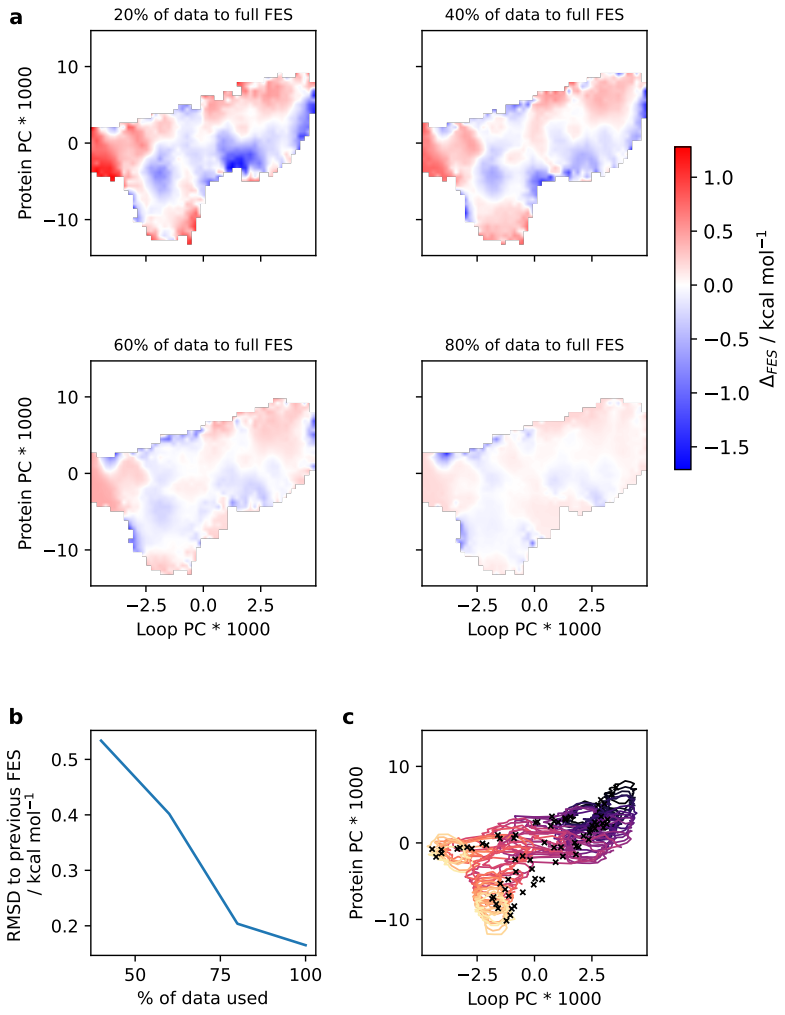


Figure 3.13: (a) Convergence of P38α 2D REUS from MEMENTO paths, shown as the difference between the PMFs including 20, 40, 60 and 80% of the data and the full PMF, thus highlighting regions of faster and slower convergence. (b) Alternative representation of convergence as the RMSD from adding successive 20% chunks of data into the PMF computation. (c) 2D histograms shown as paths encircling the normalised histograms at 30% of their peak height, displaying satisfactory overlap. Crosses indicate the REUS window starting frames.

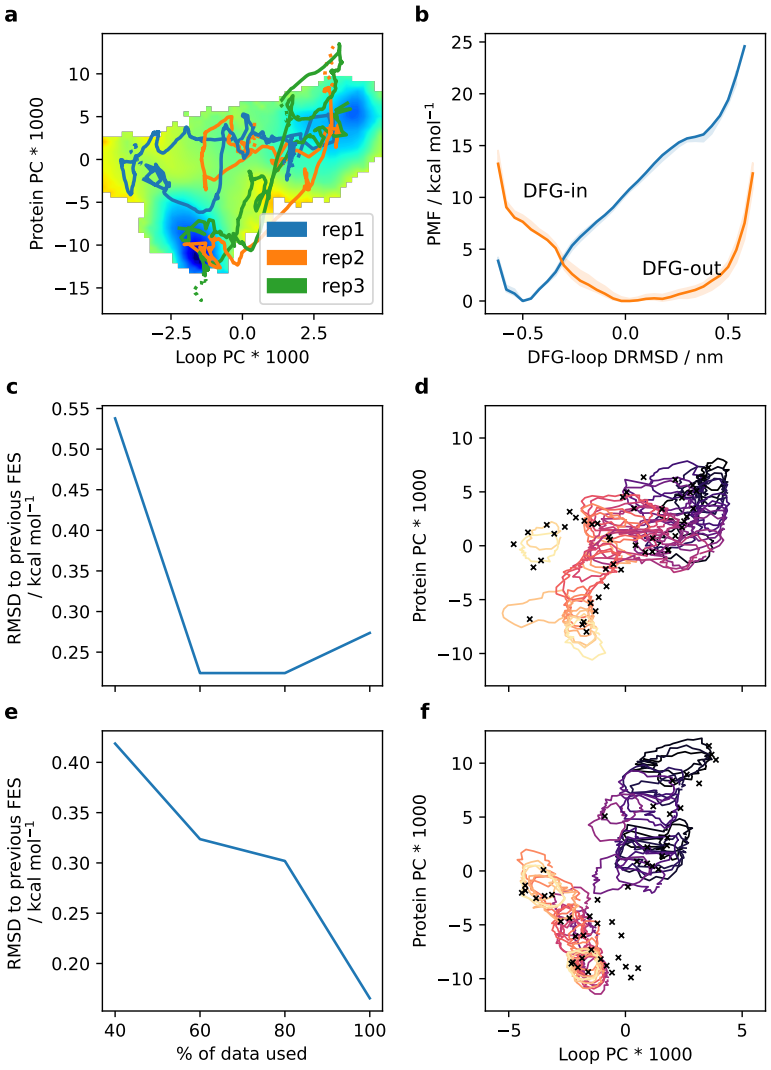


Figure 3.14: (a) Projection of the P38α SMD runs onto the PCA CV-space. Solid lines are the biased portion, dashed lines the relaxing portion of the runs. The 2D PMF from MEMENTO is shown as reference in the background. Rep 1 and 3 approach their respective target states in both directions, while the rep 2 SMD fails in the DFG-in $\rightarrow$ DFG-out direction. (b) 1D REUS from SMD rep 1 paths after re-solvation of all intermediates. (c) Convergence of the 2D PMF for DFG-in $\rightarrow$ DFG-out SMD, as the RMSD from adding successive 20% chunks of data into the PMF computation. (d) 2D-histograms for the same PMF, shown as paths encircling the normalised histograms at 30% of their peak height, with some overlap issues. (e)–(f) Same as (c)–(d), but for the DFG-out $\rightarrow$ DFG-in SMD direction.

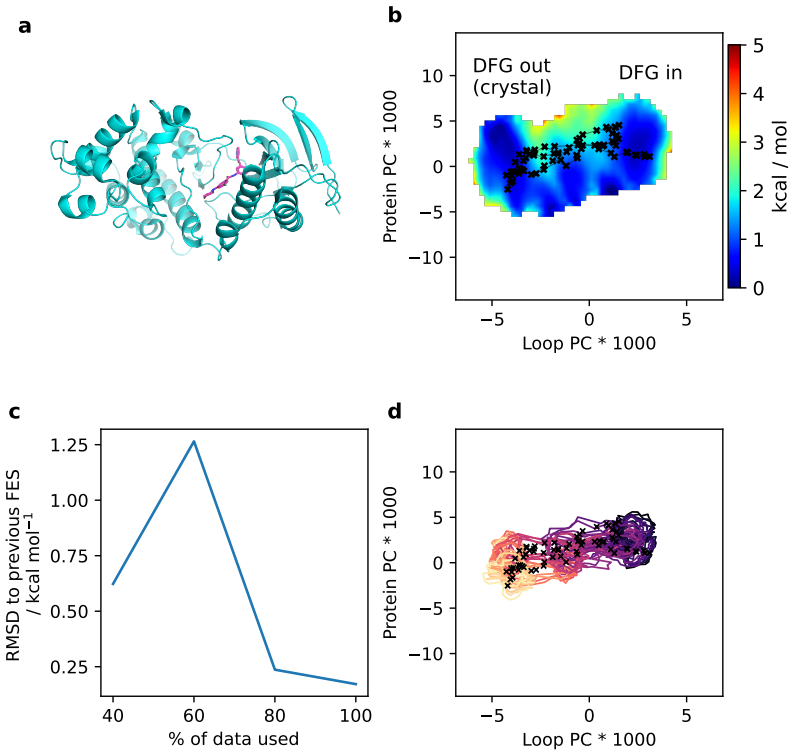


Figure 3.15: (a) P38α inhibitor binding pose in the DFG-out conformation. (b) 2D REUS along PCA CVs with MEMENTO paths for holo P38α. Crosses indicate the REUS window starting frames. (c) Convergence of the 2D PMF as the RMSD from adding successive 20% chunks of data into the PMF computation. (d) 2D histograms, shown as paths encircling the normalised histograms at 30% of their peak height, with satisfactory overlap

3.6 Results on LeuT

Thus far, our validation examples were all soluble proteins. However, conformational dynamics of other types of systems such as membrane proteins also play a crucial role in biology. For example, the solute carrier (SLC) superfamily encompasses 65 families of more than 450 transport proteins, moving substrates of diverse sizes and chemical properties across biological membranes. (Pizzagalli et al., 2021) These transporters function by exposing a substrate binding site in turn to the intra- and extracellular medium, in what is termed an alternating access mechanism (see section 1.1.1 for a more detailed exposition). Among the SLC transporters, many structurally diverse variants have been described, which operate by the rocker-switch, rocking bundle and elevator mechanisms. (Drew and Boudker, 2016; Drew et al., 2021; Beckstein and Naughton, 2022) A dynamical representation of these conformational changes is of great pharmacological interest since many drugs are carried by SLC transporters, and may help design better drug delivery approaches in the future. For more details, see the discussion in chapter 1.

Leucine Transporter (LeuT) — a bacterial sodium-coupled leucine transporter — is an archetypal SLC transporter with a rocking bundle mechanism (the folds observed in the SLC 6, 7, 11, 12 and 38 families are its namesakes). Therefore, understanding the LeuT conformational changes is likely to facilitate insight into the transport of various amino acids, neurotransmitters and inorganic ions. Here, we apply MEMENTO and SMD to the conformational change between the inwards facing (IF) (pdb: 3TT3) (Krishnamurthy and Gouaux, 2012) and outwards-facing occluded (OCC) (pdb: 3F3E) (Singh et al., 2008) conformations (figure

3.16a). Note that while the IF state is captured in the apo condition, the OCC-state crystal structure is bound to a leucine molecule. We decided to sample the apo conformational transition here to keep the test case simpler (without the need for modelling a ligand binding pose in the IF state, that is), although this necessitates the removal of the ligand in the MD preparation step. Doing this has the potential to lead to artefacts, for example if the protein conformation is substantially stabilised by the ligand (see for example the discussion on P38α in the preceding section). While we found no evidence of such a problem here — since the stable OCC-state basin in the PMF indicates the stability of the generated apo state, this is an adequate test case for the MEMENTO method — our PMF needs to still be interpreted bearing this detail of the setup process in mind.

As in the previous examples, we first ran long (here $1\,\mu s$) unbiased MD on lipid-embedded boxes of these two conformations (both as apo protein). While the OCC state is stable in MD, we noticed that in the IF state simulations, TM 1 collapsed at its intracellular end to an orientation that takes it closer to the opposing bundle — that is, we observed partial spontaneous closure (figure 3.16b). Propagated through MEMENTO paths (step-through supplementary videos 8–11 of our publication, Lichtinger and Biggin (2023)) and 1D REUS along a DRMSD CV (based on the original IF and OCC coordinates), this leads to big differences between replicates. Notably, in replicates 2 and 3 (the starting states of which have TM 1 closed), the PMF is not sampled near the original IF structure (figure 3.16c). Since we wished to determine the energetics of the full connecting path between the IF and OCC structures, we once again set out to construct 2D CVs using PCA based on the 1D REUS trajectories. The details of this procedure are described in section 3.2.4; in short, we found that too many PCs were required to explain a substantial proportion of the variance to be directly useful in US. Therefore, we took PC 1, the highest contributing direction, as one CV, while constructing the second CV as a linear combination between the next 15 PCs that separates the replicate MEMENTO paths best (we termed this CV 'PC 2' for simplicity). These CVs are global and complex, but roughly speaking PC 1 covers the conformational changes occurring both at the intra- and extracellular sides between IF and OCC, while PC 2 focuses more on the TM 1 movement with less contributions at the extracellular face (figure 3.16d and supplementary videos 12–15 of our publication).

Equipped with these CVs, we proceeded to computing the 2D PMF of the overall conformational change. In our first attempts, we experienced issues with undersampling the IF $\leftrightarrow$ 'collapsed TM 1' region (unsurprisingly, since it was not originally included as an explicit path) as well as parts of the OCC $\leftrightarrow$ 'collapsed TM 1' transition region (it appears that 24 windows were not quite enough for this conformational change). We therefore used MEMENTO again to supplement our PMF with extra paths between the crystallographic and 'collapsed TM 1' IF states, as well as additional 16 windows along the replicate 2–3 paths within the transition region. The final PMF is shown in figure 3.16e and converges reasonably well (figure 3.17), though somewhat worse than previous examples even with $500\,ns$ / window of sampling. The PMF reproduces the observation from unbiased MD that the IF state is not stable in the simulated conditions but collapses into a broad basin of an inward-facing, partially occluded state. There is then a second free energy barrier between this basin and the crystallographic outwards-facing OCC state, which corresponds to the rocking bundle motion. Since this barrier is significantly smaller, however, than for the direct IF $\leftrightarrow$ OCC interconversion, our simulations suggest a preferred sequential mechanism for the overall conformational change proceeding via first closing TM 1 before engaging in the rocking bundle movement.

This PMF agrees well with previous data published on LeuT. Gur et al. (2015) simulated a set of

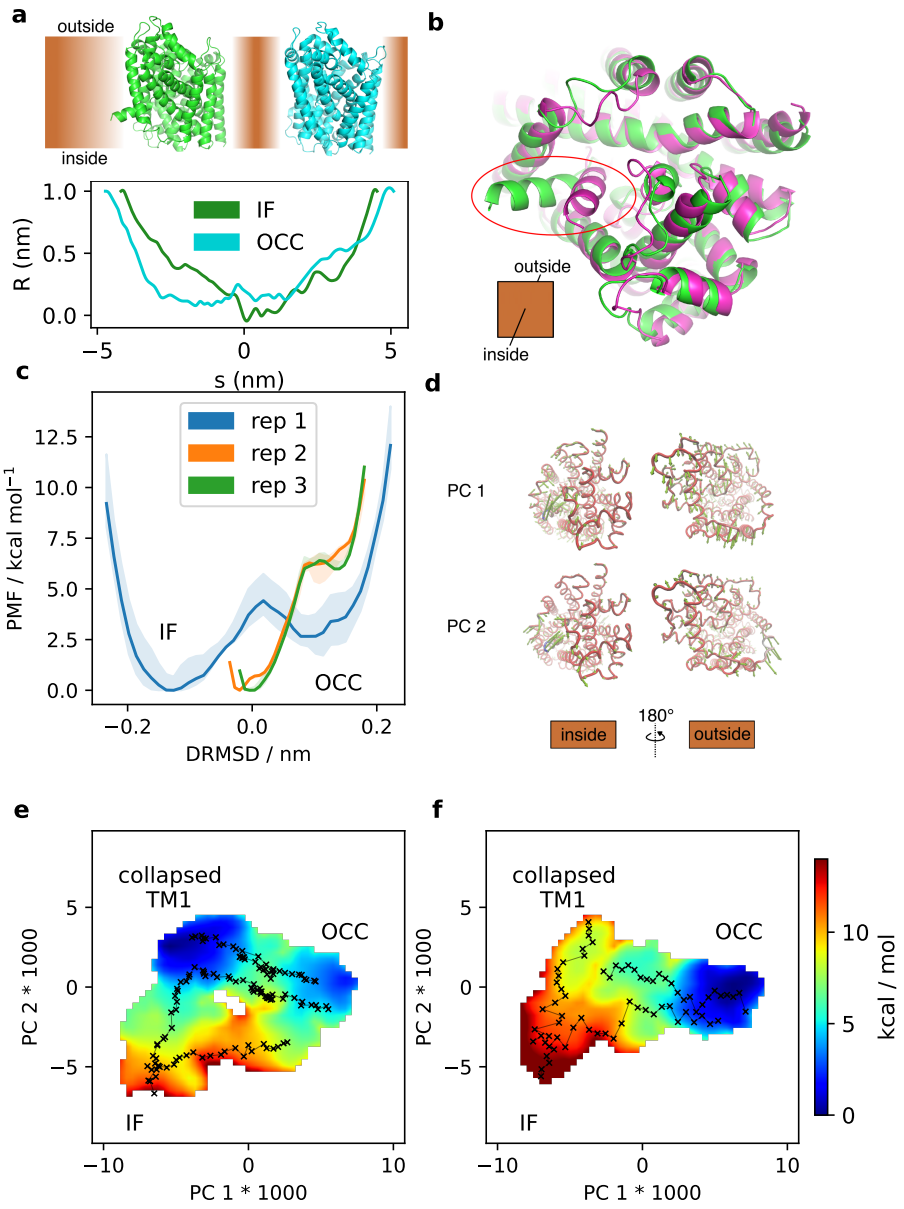


Figure 3.16: (a) Overview of the IF and OCC structures of LeuT, and pore radius profiles calculated with CHAP. (Klesse et al., 2019) (b) Illustration of the closing motion of TM 1, before (green) and after (pink) 1 µs unbiased MD. (c) 1D REUS along the DRMSD CV from MEMENTO paths, showing big differences between replicates. Shaded area is the range of PMFs observed when taking only the first 60%, the last 60% and the full sampling. (d) PCA results, showing the two CVs obtained for further sampling. (e)–(f) 2D REUS from (e) MEMENTO and (f) SMD paths. Crosses indicate the REUS window starting frames, connected in sequence.

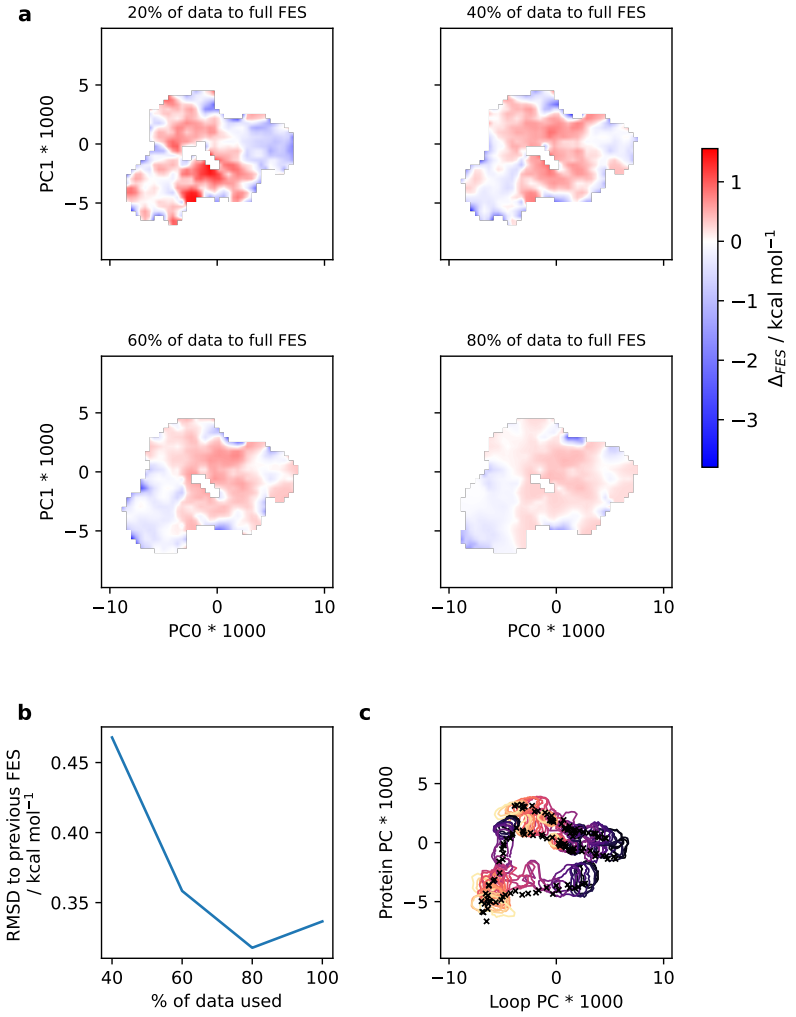


Figure 3.17: (a) Convergence of LeuT 2D REUS from MEMENTO paths, shown as the difference between the PMFs including 20, 40, 60 and 80% of the data and the full PMF, thus highlighting regions of faster and slower convergence. (b) Alternative representation of convergence as the RMSD from adding successive 20% chunks of data into the PMF computation. (c) 2D histograms shown as paths encircling the normalised histograms at 30% of their peak height, displaying satisfactory overlap. Crosses indicate the REUS window starting frames.

unbiased MD trajectories, starting from crystal structures and paths obtained using an anisotropic network model-based biasing scheme termed coMD. They found two free energy basins of inward-facing and outward-facing conformations that interconvert most favourably via occluded states (although obtaining sufficient sampling in the transition region was challenging). They also found that the intracellular side of TM 1 tends to partially close up during their situations. This observation had also been made in previous experimental (Kazmier et al., 2014) and computational (Grouleff et al., 2015) investigations, and is reproduced in this work with the 'collapsed TM 1' state. The literature therefore supports the shape of the PMF we have obtained for apo LeuT. Moreover, to our knowledge this study contributes the most extensive sampling of LeuT IF $\leftrightarrow$ OCC conformational dynamics expended to date, therefore corroborating earlier results. However, investigating the role of substrate and coupled sodium ions in the transport cycle — as other studies have attempted — was beyond the scope of the present work, since we are chiefly concerned with sampling methodology here.

In order to compare our MEMENTO 2D PMF to what would be achievable with SMD alone for path generation, we ran three replicates of long ($500\,\mathrm{ns}$) SMD, with initial and target states set to the same frames of unbiased MD as we used for MEMENTO above (figure 3.18). We found that only SMD OCC $\rightarrow$ IF replicates 1–2 and no IF $\rightarrow$ OCC runs reached meta-stable conformations near their target states. When we calculated a 2D PMF along the successful paths, supplemented with extra frames from the apo IF unbiased run (in place of the extra MEMENTO path used above to close the gap between IF and 'collapsed TM 1'), we found a strong bias towards the OCC state (figure 3.16f) and convergence worse than with MEMENTO paths (figure 3.19). The fact that the windows from unbiased MD and SMD OCC $\rightarrow$ 'collapsed IF' did not overlap well with each other (see figure 3.19 b–c), made us suspect that hysteresis is at play here. A comparison to the MEMENTO 2D PMF is also suggestive of strong hysteresis, if we assume the MEMENTO results as ground truth. We cannot explicitly show the starting-state bias by contrasting PMFs from SMD in opposite directions, since we were unable to perform SMD in the IF $\rightarrow$ OCC direction at all. This in itself can be taken as strong indication for hysteresis, nonetheless.

We have demonstrated in conclusion that MEMENTO — together with an iterative approach for deriving CVs to match DOFs sampled in long unbiased MD — can provide valuable insight into complex and global protein conformational changes in LeuT. SMD, in turn, fails again to provide paths of matching quality for the purposes of US, due to starting-state bias.

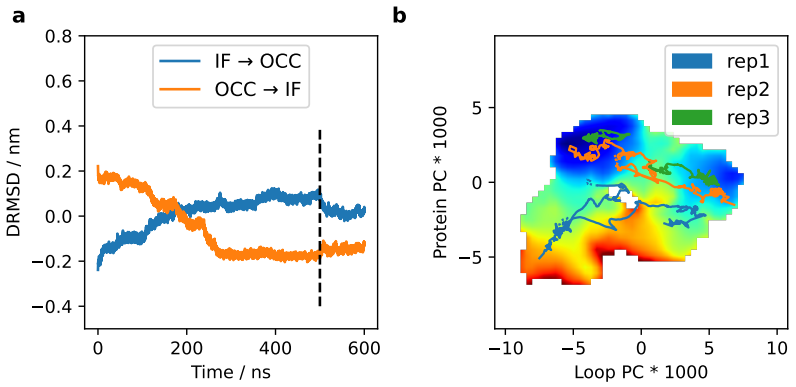


Figure 3.18: (a) LeuT SMD between IF and OCC states, replicate 1, projected onto the DRMSD CV. The dashed line represents the end of the restraint, followed by unbiased relaxation. (b) Three replicates of LeuT SMD, projected onto the MEMENTO 2D PMF, showing how only replicates 1–2 in the OCC → IF direction reached their respective target states. Solid lines are the biased portion, dashed lines the relaxing portion of the runs.

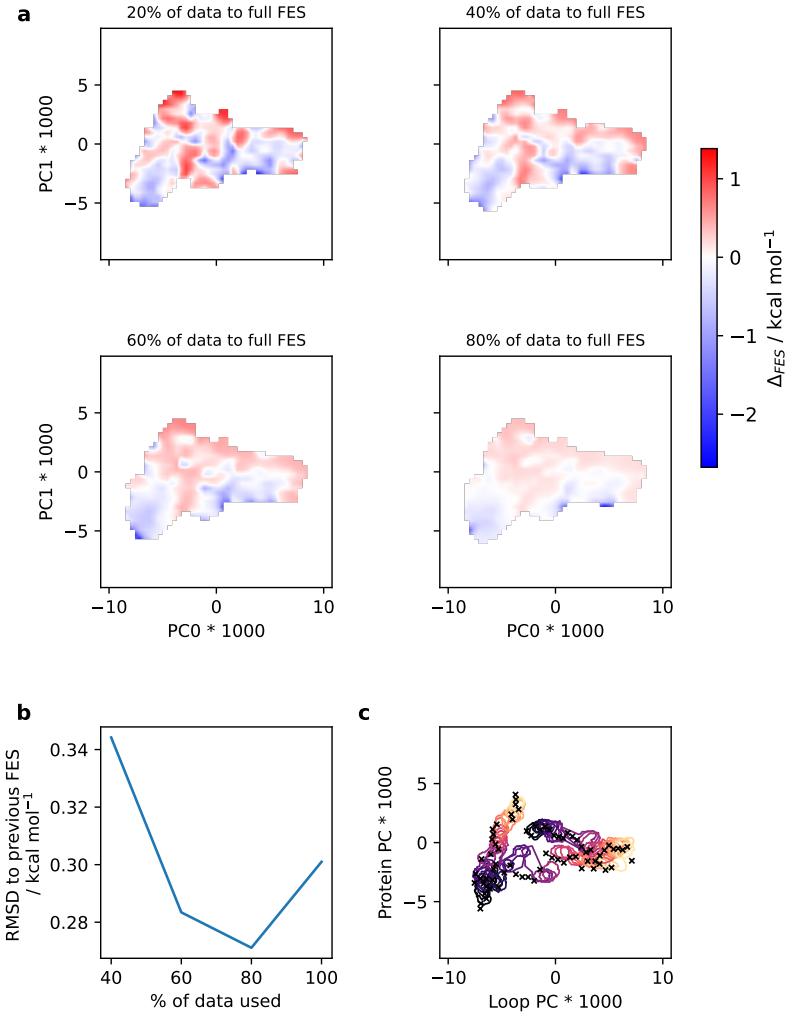


Figure 3.19: (a) Convergence of LeuT 2D REUS from SMD paths, shown as the difference between the PMFs including 20, 40, 60 and 80% of the data and the full PMF, thus highlighting regions of faster and slower convergence. (b) Alternative representation of convergence as the RMSD from adding successive 20% chunks of data into the PMF computation. (c) 2D histograms shown as paths encircling the normalised histograms at 30% of their peak height, displaying insufficient overlap. Crosses indicate the REUS window starting frames, connected in sequence.

3.7 Discussion

The work presented in this chapter investigates four protein systems that exhibit conformational changes for which structural data is available at both end states: deca-alanine, ADK, P38α and LeuT. Furthermore, for each of these systems, previous studies that calculated PMFs for the conformational changes are available in the literature. The first aim of this part of my DPhil project was to determine the extent to which SMD suffers from hysteresis when used to seed umbrella sampling simulations for computing PMFs. We also wished to study whether these adverse effects can be mitigated by the use of paths generated with MEMENTO (implemented and freely available as PyMEMENTO) — a simple procedure we devised for crafting history-independent paths using a combination of coordinate morphing and template-based structure modelling.

On deca-alanine, we found no difference in the quality of SMD and MEMENTO paths: the simple nature of the system evidently does not produce significant hysteresis. The example therefore serves as validation of the PyMEMENTO code, showing that it does not introduce artifacts in an easy test case. Moving on to ADK, SMD and MEMENTO paths can both yield converged PMFs along physically intuitive CVs. When attempting to investigate the role of an alternative closed conformation, however, hysteresis appeared in bi-directional SMD, thus conferring an advantage to MEMENTO for a thorough exploration of the system. For P38α, SMD suffered from large-scale hysteresis. MEMENTO, in turn, was able to generate converged and consistent PMFs, albeit only once we had derived custom CVs using an iterative scheme that utilises MEMENTO, extensive end-state sampling and PCA. Lastly, concerning the membrane transporter LeuT, this approach for generating custom CVs was again successful with MEMENTO, while SMD suffered from hysteresis even when using the consistent CV space obtained from MEMENTO results. We note that for all examples, our results are in good agreement with previously published data in the computational and experimental literature.

In conclusion, SMD is prone to hysteresis when combined with umbrella sampling for all but the simplest conformational changes, at least without substantial prior knowledge regarding CV choice. We therefore suggest as best practice to all investigators who use SMD for generating paths for refinement and PMF computations to carefully use controls in both targeting directions, where possible. Hysteresis — if it is an issue within the relevant CV space — will become apparent as substantial differences between the obtained PMFs. It is unclear still whether this is also an issue when using SMD as a part of more elaborate biasing schemes, however our results warrant a degree of caution. It is left for future work to establish this for each of the many methods using SMD as part of their pipelines. We also show that MEMENTO, despite its conceptual simplicity and easy implementation, is a powerful and flexible tool for conformational sampling. This is especially the case due to the ease of its combination with long end-state sampling and dimensionality reduction for CV derivation, and because newly discovered conformational states can effortlessly be integrated into existing ensembles of paths. Nonetheless, evaluating the free energy profiles of protein conformational changes remains a formidable challenge that requires substantial expertise and system-specific knowledge or deep *ad-hoc* investigation on the part of the researcher.

4

The alternating-access mechanism of mammalian peptide transporters

Disclaimer

The research presented in this chapter has been published in eLife in 2024 (Lichtinger et al., 2024), see also in the list of publications. All investigation, writing and figures in that publication were produced by myself with feedback from the co-authors, and will, with adaptations, be used here under the CC BY 4.0 license. Images and phrasing may therefore be similar or partially identical to my published work.

As presented in detail in chapter 1 of this thesis, proton-coupled oligopeptide transporters (POTs) are of great pharmaceutical interest owing to their promiscuous substrate binding site that has been linked to improved oral bioavailability of several classes of drugs (sections 1.1.4 and 1.2). Members of the POT family are conserved across all phylogenetic kingdoms and function by coupling peptide uptake to the proton electrochemical gradient (section 1.1.2). Cryo-EM structures and alphafold models have recently provided new insights into different conformational states of two mammalian POTs, PepT1 (SLC15A1) and PepT2 (SLC15A2), see section 1.1.3. The structural information available on POTs and the location of key residues are illustrated in figure 1.3, which is reprinted here as figure 4.1 for convenient reference. Previous studies nevertheless have left open important questions regarding the mechanism of proton and substrate coupling, while simultaneously providing a unique opportunity to investigate these processes using molecular dynamics (MD) simulations (section 1.4). In this chapter, we employ extensive unbiased and enhanced-sampling MD, totalling close to 1 ms of sampling, to map out the full PepT2 conformational cycle and its thermodynamic driving forces. By computing conformational free energy profiles in different protonation states and in the absence or presence of peptide substrate, we identify a likely sequence of intermediate protonation steps that drive inward-directed alternating access. In particular, protonation of H87 and D342 together triggers extracellular-gate opening, whence the protons move down into the transporter to the ExxER motif and E622, which also triggers substrate release. Our simulations highlight key differences in the extracellular gate between mammalian and bacterial POTs, which we validate experimentally in cell-based transport assays. We also clarify the thermodynamic role that the ExxER motif plays in the intracellular gate. Our results from constant-pH molecular dynamics (CpHMD) and absolute binding free energy (ABFE) calculations further establish a mechanistic link between proton binding and peptide recognition at the ExxER motif, revealing key details underpinning secondary active transport in POTs.

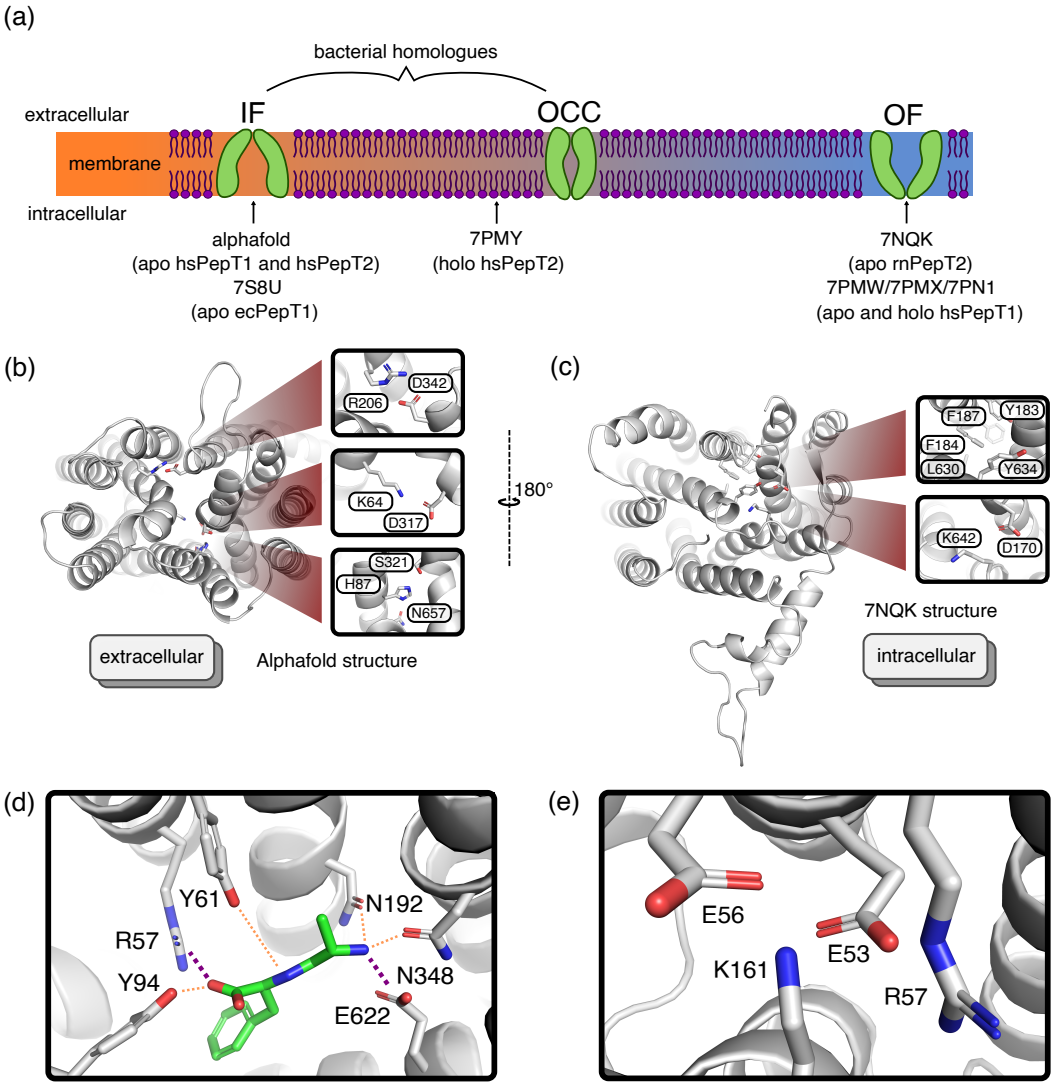


Figure 4.1: (a) Schematic overview of the conformational diversity of available mammalian POT structures. Intermediate positions between states indicate partial gate opening. (b) Alphafold-predicted IF HsPepT2 structure (viewed from the extracellular side, with the extracellular domain omitted for clarity), highlighting potential inter-bundle extracellular gate interactions. (c) OF Cryo-EM structure of apo RnPepT2 (7NQK, viewed from the intracellular side) (Parker et al., 2021), highlighting potential inter-bundle intracellular gate interactions. (d) Ala-Phe substrate binding pose, representative cluster frame of 1 µs MD simulation from 7NQK structure with added ligand, for setup details see appendix B. Purple dotted lines represent salt-bridge contacts, orange dotted lines other polar contacts. (e) ExxER motif salt-bridge cluster, representative cluster frame of 1 µs MD simulation from 7NQK structure.

4.1 Extra- and intra-cellular gate opening triggers in unbiased MD

We began our computational investigation by embedding PepT2 structures (using the sequence of the rat homologue) in the OF (Parker et al., 2021), IF (alphafold prediction, Jumper et al. (2021)) and inwards-facing partially occluded (Killer et al., 2021) conformations in 3:1 POPE:POPG membranes (figure 4.2, details in appendix B). While we were able to obtain stable wide-open OF and IF simulation boxes, the occluded (OCC) state required further attention as the MD simulations from the inwards-facing, partially-occluded structure showed embedding artifacts, including extracellular-gate instability and intracellular gate hydrophobic collapse (see figure 4.2a). We therefore opted to derive an OCC state using metadynamics simulations in 5 replicates (figure 4.3, further details in appendix B), using stability in unbiased MD to select the best among the obtained candidates. We note that steered MD (SMD) could have produced similar results here; our choice was determined by having tried metadynamics first in the course of this project. The OCC state thus developed is validated by the further work in this paper, showing it to be a stable conformational basin that is functionally occluded in that it can open both towards OF and IF in different protonation-state conditions (see figures 4.5, 4.11 and 4.17).

Equipped with models of the protein conformations required for PepT2 alternating access (OF, OCC and IF), we ran triplicate sets of 1 µs-long MD simulations in a range of conditions. To decide which conditions to probe apart from the apo, standard protonation states as obtained above, we investigated the extent to which the H87–S321, D342–R206, K64–D317 and D170–K642 interactions (see figure 1.3b–c) are formed in the OF (closed intracellular gate) and IF (closed extracellular gate) conformational states (figure 4.4). In the IF state, we found that H87–S321 and D342–R206 are consistently interacting, whereas the K64–D317 interaction, while formed in $\approx$ 72% of MD frames, is unstable and of a transient nature, and therefore unlikely to contribute much to extracellular gate stability. The D170–K642 saltbridge, in turn, is only formed in $\approx$ 22% of OF-state frames, thus likely not substantially adding to the stability of the intracellular gate. We therefore decided to mainly focus on probing the H87–S321 and D342–R206 interactions with respect to control of the extracellular gate. Since no salt bridges or other specific interactions involving protonatable residues seem to demarcate the intracellular gate, we decided to focus on the ExxER glutamates and E622 for their effects on intracellular gate opening.

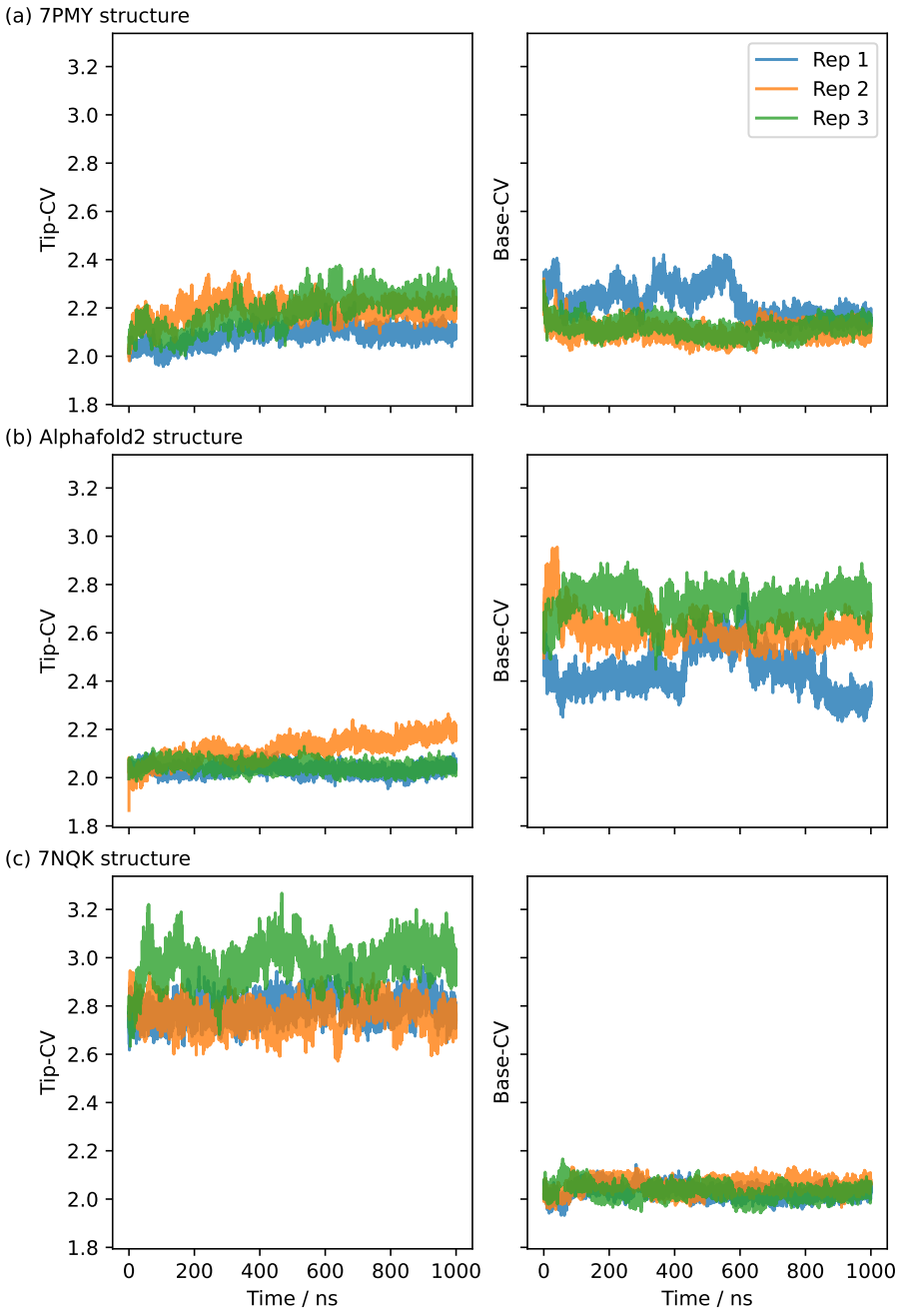


Figure 4.2: 1 µs-long MD simulations starting from CHARMM-GUI-embedded and equilibrated PepT2 structures, projected along CVs describing the opening of the extra- and intracellular gates, see also appendix B for CV definitions and figure 4.5a for an illustration. (a) The inward-facing, partially occluded cryo-EM structure (7PMY) moves towards an occluded state via closure of the intracellular gate. However the extracellular gate partially opens in the process. Rep 1 also displays a partial helical unfolding near the intracellular gate (see provided coordinate files in the supplementary data of our publication, Lichtinger et al. (2024)). (b) Alphafold-based IF embeddings 1 and 3 explore a range of IF conformations while maintaining a stable extracellular gate, whereas rep 2 partially opens the extracellular gate. (c) The OF cryo-EM structure (7NQK) remains stable with a tight intracellular gate.

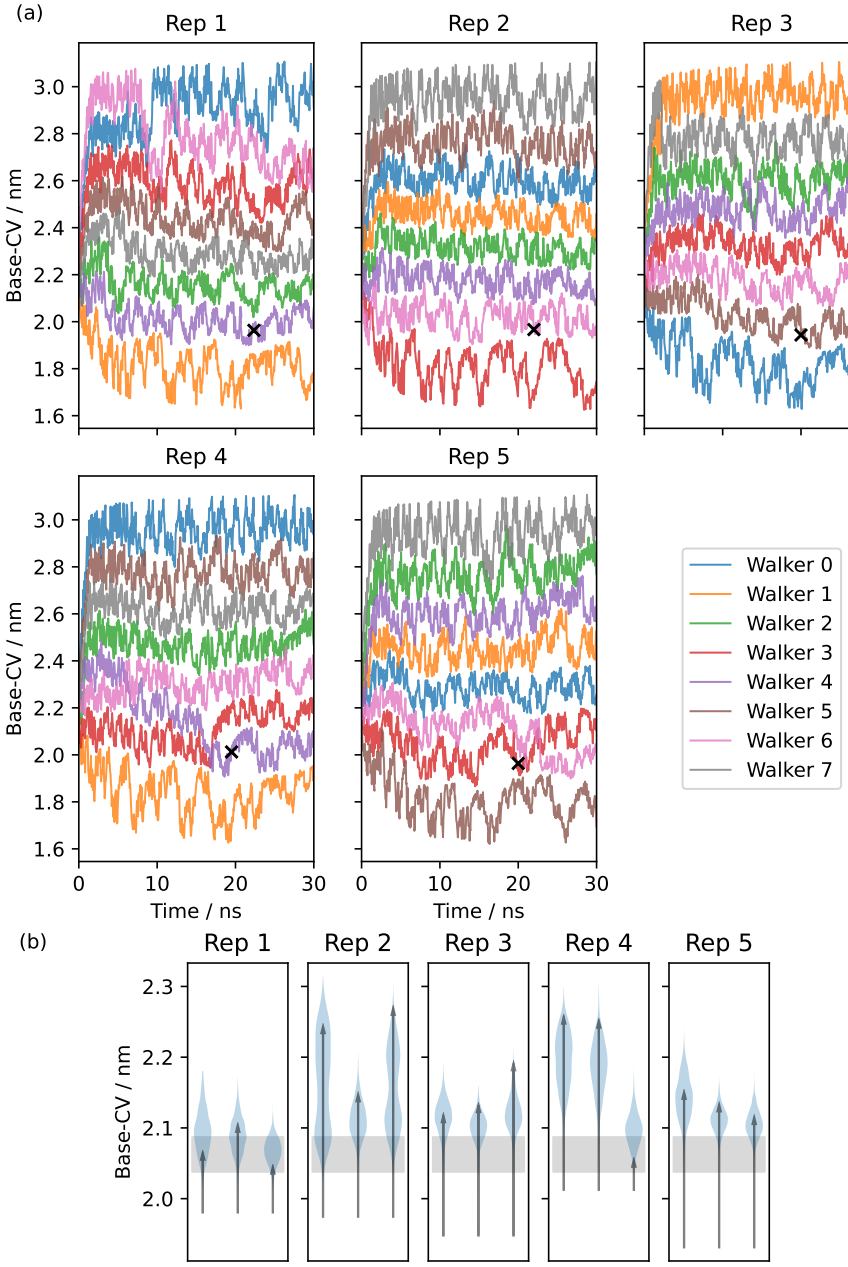


Figure 4.3: (a) Five replicates of multiple-walker metadynamics along the base-CV, starting from the Alphafold IF equilibration rep 1. Black crosses indicate potential OCC states picked around a base-CV value of 2 nm and at a simulation time of around 20 ns. (b) Equilibrations (3 per metadynamics replicate) of the candidate OCC states in 100 ns unbiased MD. Histograms of trajectory projections onto the base-CV are shown as violin plots, arrows link the first and final frames. The grey shaded area corresponds to the range of base-CV values sampled in our OF simulation. Rep 1 is a stable OCC state, whereas the other replicates display partial intracellular-gate opening.

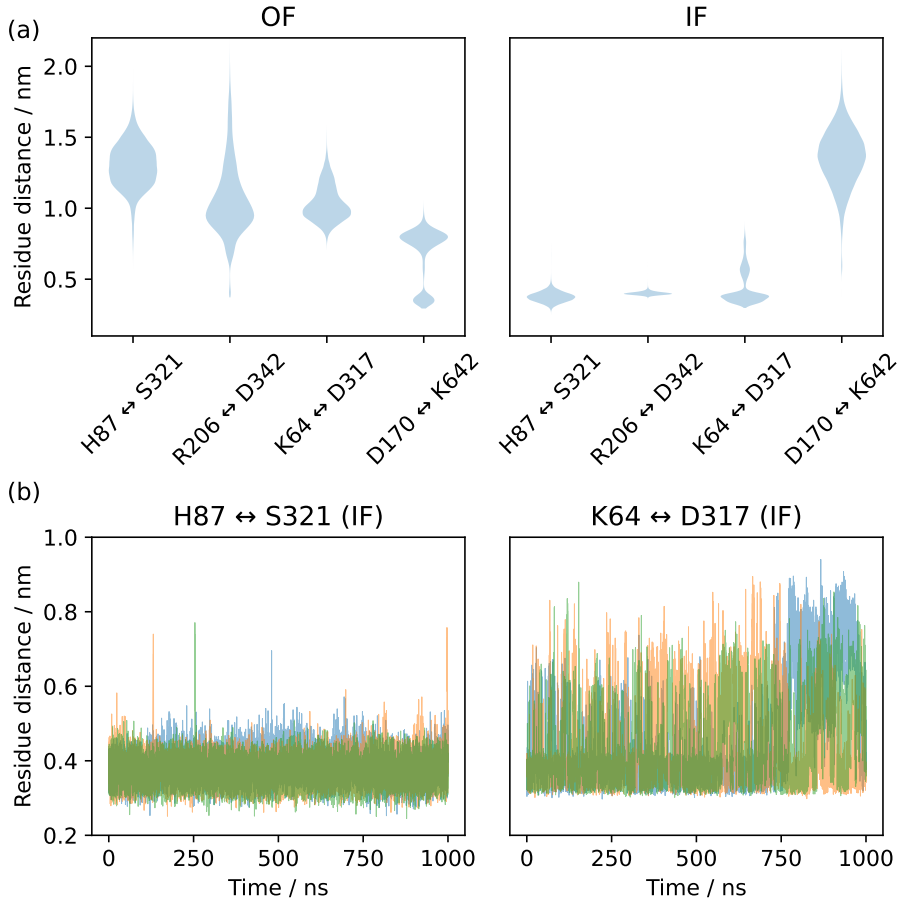


Figure 4.4: Inter-residue heavy atom (H87: NE2, S321: OG, R206: CZ, D342: CG, K64: NZ, D317: CG, D170: CG, K642: NZ) distances for several possible gating interactions. (a) Histograms from pooled triplicates of 1 µs MD as violin plots in the top row, starting from alphafold-derived IF and cryo-EM derived OF conformations. H87↔S321 and R206↔D342 are always formed in the IF state but not in the OF state, while K64↔D317 and D170↔K462 show a preference for the IF and OF states respectively but are not formed in all trajectory frames. (b) Triplicate time series of the H87↔S321 and K64↔D317 interactions in the IF state, showing how the former is a tight interaction, while the latter is unstable and only transiently formed.

Guided thus in our choice of which residues to investigate, we probed whether the OCC state opens spontaneously to OF or IF states in a range of different protonation-state and mutation conditions (as assessed by projection of unbiased MD runs onto intuitive collective variables (CVs), defined as the centre-of-mass distance between the tips and bases of the N- and C-terminal bundles, see figure 4.5a). We found that the extracellular gate remains stably closed in triplicates of 1 µs-long MD when H87 or D342 are protonated individually, but the OCC state can open spontaneously on the simulated time scale to an OF conformation when both are protonated simultaneously (figure 4.5b; figure 4.6a for plots of the opposite gate in the same trajectories, showing how flexibility of the intra- and extracellular gates is anti-correlated). A comparable effect is found in the presence of the physiological peptide substrate L-Ala–L-Phe (figure 4.5b, panels 5–6). We have also tested further combinations of mutations and protonation state changes relating to the putative extracellular gating interactions (D317 protonation and mutations of R206, S321, D342 to alanine, with and without H87 protonation) as well as some control mutants which we did not expected to have an effect (I135L, T202A, Q340A and the saltbridge-swapped mutant R206D & D342R, combined with H87 protonation). Across our $48 * 1$ µs unbiased MD runs collated in figure 4.7, we observed 3 full extracellular-gate opening events, in conditions where H87 was protonated and the D342–R206 saltbridge

was also disrupted either by D342 protonation or mutation to alanine (D342A). We also saw one partial opening event when in addition to H87 protonation we also mutated S321 to alanine (S321A). The data thus suggests that for spontaneous extracellular-gate opening to occur on this time scale in unbiased MD, disruption of the OCC-state H87 interaction network is essential, and the D342 salt bridge appears to make an additive contribution towards extracellular gate stability (though this is not a strict correlation, as illustrated by the S321A partial opening event). The intracellular gate, by contrast, is more flexible than the extracellular gate even in the apo, standard protonation state; however, following either protonation of the conserved E53 and E622 residues or the insertion of Ala-Phe substrate, the intracellular gate becomes more flexible and can spontaneously open (figure 4.5c; see figure 4.6b for the corresponding plots of the extracellular gate opening). It should be noted that no direct comparison with experiment can be drawn from the time scales observed in the unbiased MD here. While our data does show that once certain protonation state changes have taken place the conformational transition occurs on the order of 10s to 100s of nanoseconds, this does not include time taken for these residues might become protonated — in kinetic experiments, these would be measured in a coupled fashion.

Although these unbiased simulations show a large amount of stochasticity and drawing clean conclusions from the data is difficult, we can already appreciate a possible mechanism where protons move down the transporter pore, first engaging H87 and D342 to favour the OF state and then moving to ExxER and E622 to favour the IF orientation, driving successive conformational changes along the way. The initial unbiased approach taken in this section therefore serves to generate hypotheses for testing by a more rigorous investigation of the protonation state-dependent conformational changes. To this end, we set out to employ enhanced sampling simulations for obtaining conformational free energy landscapes in dependence on a range of protonation state and substrate binding conditions.

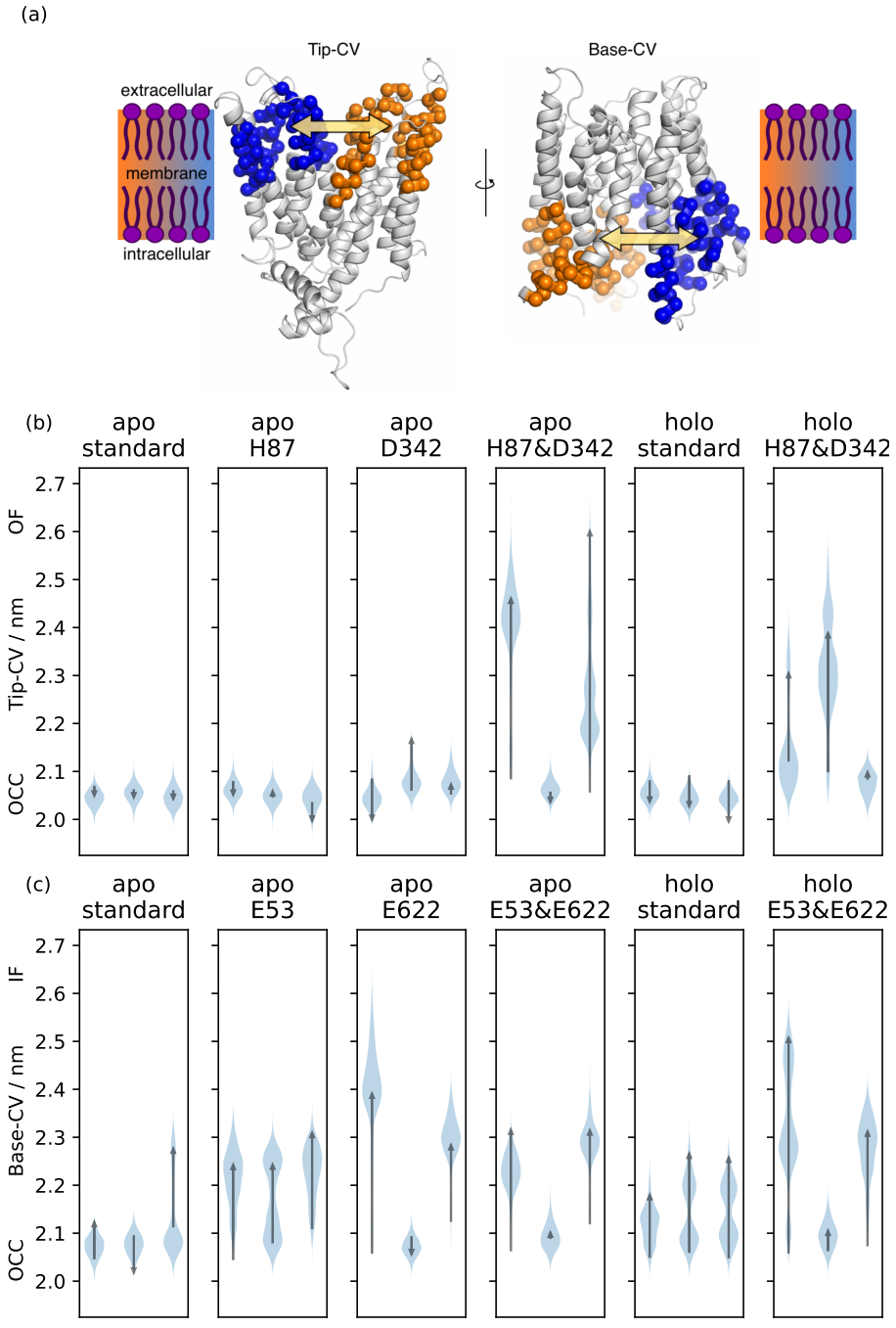


Figure 4.5: (a) Illustration of the CVs used to quantify extra- and intracellular gate opening, consisting of inter-bundle centre-of-mass distances between the helical tips (top 11 residues) and bases (bottom 11 residues). (b)+(c) Triplicate 1 µs-MD simulations starting from OCC, showing the effects of different protonation and substrate binding states, projected onto the (b) tip-CV and (c) base-CV respectively. Violin plots are trajectory histograms, arrows link the CV values of the first and last frames. The complementary gates, i.e. the base-CV projection for (b) and tip-CV projection for (c) are shown as controls below in figure 4.6.

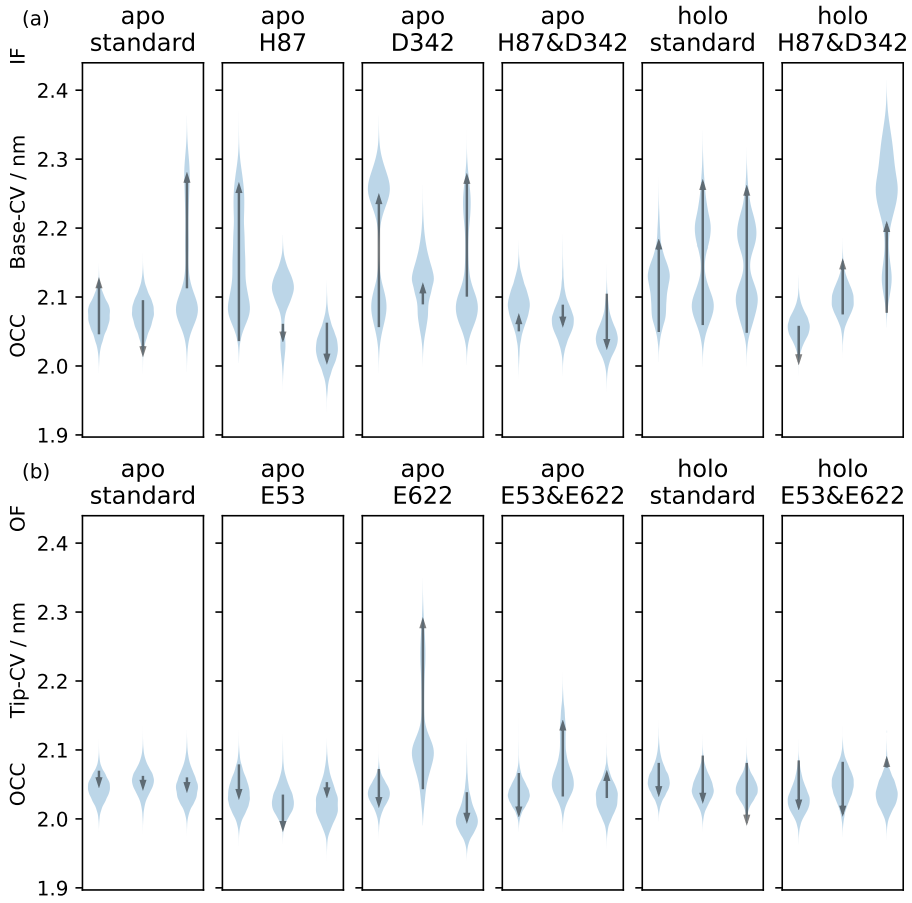


Figure 4.6: Triplicate 1 µs-MD simulations starting from OCC, showing the effects of different protonation and substrate binding states, projected onto the (a) base-CV and (c) tip-CV respectively. Violin plots are trajectory histograms, arrows link the CV values of the first and last frames. Intracellular gate flexibility is suppressed by conditions that favour extracellular gate opening and vice versa.

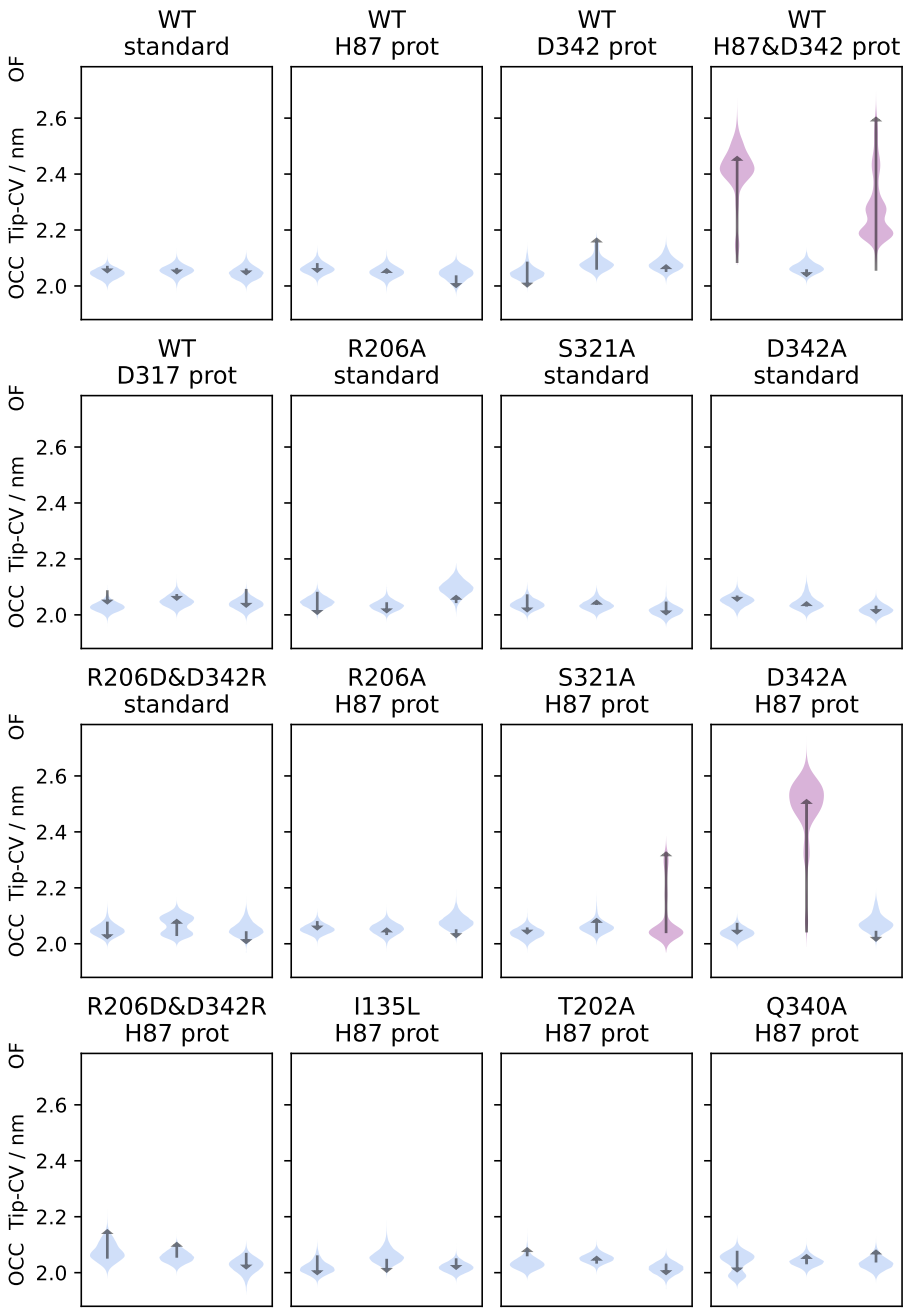


Figure 4.7: Triplicate 1 µs-MD simulations starting from the OCC state, showing the effects of different protonation states and mutations projected onto the tip-CV. Violin plots are trajectory histograms, arrows link the CV values of the first and last frames. Trajectories which displayed significant extracellular gate opening are highlighted in purple. Spontaneous extracellular gate opening requires H87 protonation, and the disruption of the D342–R206 salt bridge also makes a significant contribution, either by mutation or protonation.

4.2 Protonation-state dependent free energy landscapes of PepT2 alternating access

To overcome the time-scale limitations of MD simulations and sample important slow degrees of freedom, many enhanced sampling approaches have been developed and employed in the computational study of membrane proteins. (Harpole and Delemotte, 2018) We provide a review of enhanced sampling methodology in section 2.2 of this thesis. An important class of methods that includes (among others) the popular techniques of SMD, umbrella sampling (US), metadynamics, adaptive biasing force (ABF) and the accelerated weight histogram method (AWH) uses a small number of CVs along which to bias the simulation, thus improving exploration of important regions of phase space if the CV includes the relevant slow degrees of freedom (DOFs). If the CVs are not optimal, problems can manifest in the form of hysteresis (starting-state dependence) when moving between known end-states (see the discussion in chapter 3 of this thesis). This is the case for the PepT2 OCC↔OF transition with the simple tip-CV illustrated in figure 4.5a. Using either metadynamics or SMD with replica-exchange umbrella sampling (REUS), the end-state from which the simulations were started is always favoured in the resulting potential of mean force (PMF), with the hysteresis effect totalling $\approx 40\,\mathrm{kcal\,mol^{-1}}$ for each method (figure 4.8).

We have developed a strategy to overcome such hysteresis issues in conformational sampling which we call MEMENTO (Morphing Endstates by Modelling Ensembles with iNdependent TOpologies), see chapter 3 and the corresponding publication. (Lichtinger and Biggin, 2023) The method is implemented as the freely available and documented PyMEMENTO package. In brief, MEMENTO generates paths between known end states by coordinate morphing followed by fixing the geometries of unphysical morphed intermediates. Since these paths by definition connect the correct end states (unlike biased MD methods like SMD, where not reaching the target state in slow orthogonal DOFs is a common source of hysteresis), they can drastically reduce hysteresis in enhanced sampling compared to SMD as a path generation method. We have shown this for several validation cases, including a large-scale conformational change in the bacterial membrane transporter LeuT. After running initial 1D REUS from MEMENTO replicates along a simple CV guess, we can use the generated MD data to iteratively improve CVs using principle component analysis (PCA), thereby capturing slow motions from long end-state sampling that propagate through MEMENTO as differences between replicates.

Here, we ran triplicates of MEMENTO for the OCC↔OF (standard protonation states and H87&D342 protonated) and OCC↔IF (standard protonation states and E53 protonated) conformational changes, followed initially by 1D REUS along the tip-CV (figure 4.9). The results are much more consistent than SMD or metadynamics along the same biasing CV, and the shapes of the PMFs fit well with the trends we previously observed in unbiased MD from the OCC state. [1] Since, however, significant differences between replicates remained, we used PCA (see appendix B for details) on the sampling collected of the standard protonation state transitions to derive sets of 2-dimensional CVs (figure 4.10 and supplementary videos 1–4 of our publication, Lichtinger et al. (2024) [2]) that capture the main gate-opening motions in the first principle component (PC), and the direction along which the differing replicates can be best separated out as the second PC (these correspond to cleft sliding and twisting motions).

[1] Note that no advantage of US itself over metadynamics is to be inferred from this comparison. While the MEMENTO + US approach uses the same CV in the PMF evaluation as metadynamics, it additionally slices through orthogonal DOFs independently of the CV choice. The point therefore is that using MEMENTO to generate the paths has an advantage over not supplying a path and using metadynamics just from the end states. For this, no prior knowledge of a better CV is required.

[2] The videos referenced in this chapter can be found online at https://elifesciences.org/articles/96507/figures.

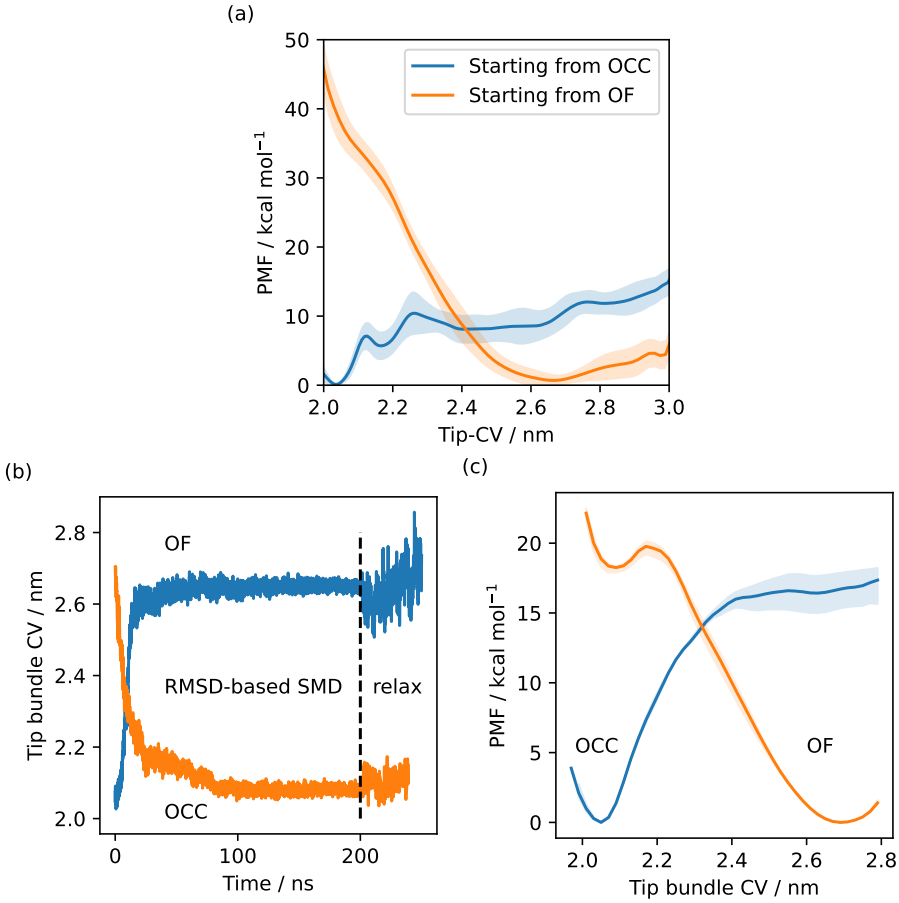


Figure 4.8: (a) Free energy profiles from metadynamics simulations (8 walkers) along the tip-CV, starting simulations from the OCC (total sampling 1.7 µs) or OF state (total sampling 860 ns). Solid lines are the free energy estimates using the second half of the data only, shaded area is the standard deviation of free energy estimates with respect to sequential data chunks. The disparity between the curves indicates a significant hysteresis problem, favouring the initial state of the respective metadynamics run. (b) SMD runs between OF and OCC states, biasing the heavy-atom RMSD to the respective target state, shown as projections along the tip-CV. Metastable OCC and OF states are formed that remain stable in 50 ns unbiased MD. (c) REUS along the tip-CV with starting conformations picked to be equidistant in the CV from two SMD runs. Sampling was using 48 windows for a total of 4.4 µs (OCC→OF path) and 6.1 µs (OF→OCC path). Solid lines are PMFs calculated using all sampling, the shaded areas are error ranges obtained by omitting either the first 40% or the last 40% of sampling. The disparity between the curves indicates a significant hysteresis problem, favouring the initial state of the respective SMD path-generation run.

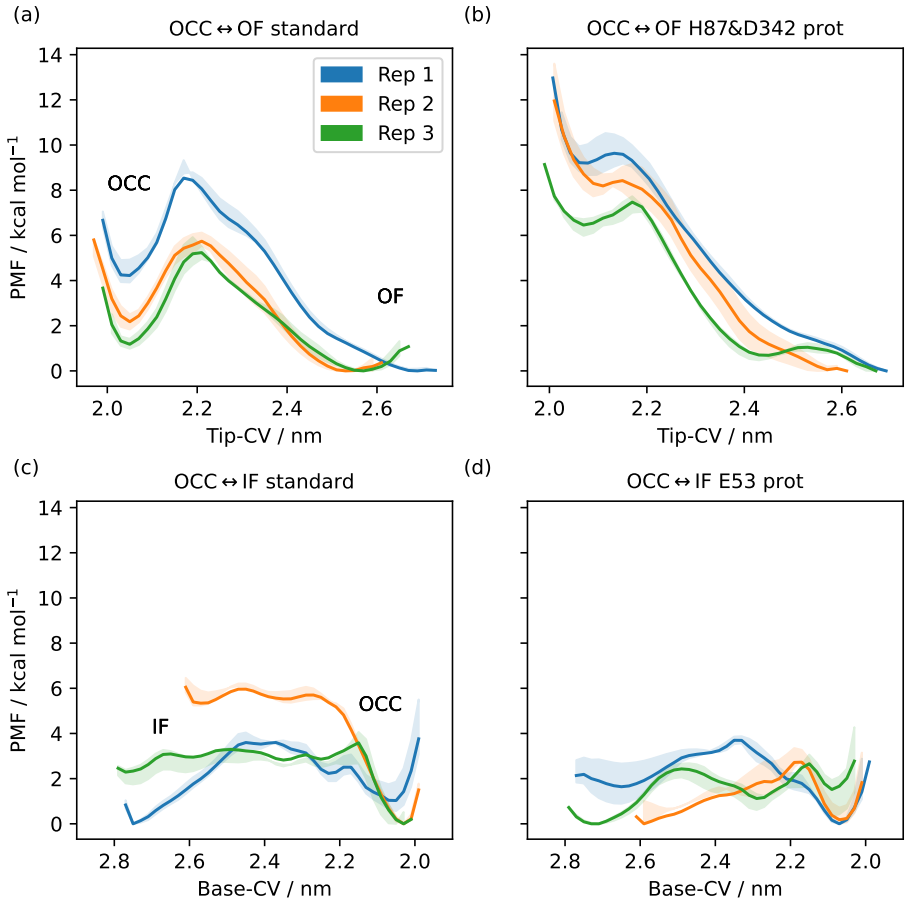


Figure 4.9: 1D PMFs along the tip-CV or base-CV (as indicated), from REUS starting at MEMENTO intermediates. The central line is calculated using all sampling, whereas the shaded areas are enclosed by curves derived from omitting the first 40% or the last 40% of sampling, giving a sense of apparent convergence (while substantial inter-replicate differences remain). (a) OCC↔OF transition, standard protonation states. Distinct OCC and OF basins with separating barrier. (b) OCC↔OF transition, H87 and D342 protonated. The OCC-state basin and the separating barrier largely disappear. (c) OCC↔IF transition, standard protonation states. Distinct OCC basin and raised IF plateau. (d) OCC↔IF transition, E53 protonated. The IF plateau and the barrier are lowered with respect to OCC.

Equipped with these CVs, we first studied the protonation-state dependence of the OCC↔OF transition. As figure 4.11 shows (2D PMFs in panel a, projections onto PC 1 in panel b), the OCC state in standard protonation states forms a basin that is metastable with respect to OF (lying $\approx 3\,\mathrm{kcal\,mol^{-1}}$ higher than OF, separated by a barrier of $\approx 3\,\mathrm{kcal\,mol^{-1}}$). Protonation of H87 still leads to a metastable OCC-state basin, although it is raised by $\approx 4\,\mathrm{kcal\,mol^{-1}}$ and the barrier is decreased to $\approx 1.5\,\mathrm{kcal\,mol^{-1}}$. Protonating D342, in turn, does not affect the relative free energies of the OCC and OF states, but does lower the transition barrier by $\approx 1\,\mathrm{kcal\,mol^{-1}}$. These effects are approximately additive, so that protonation of both H87 and D342 abolishes the metastable OCC state — an observation which agrees with the ability of the OCC state thus protonated to spontaneously transition to OF in unbiased MD (see section 4.5 for a comparison with the results obtained by Parker et al. (2017) on $\mathrm{PepT_{So}}$ on this point).

We have also computed 2D PMFs in further protonation-state and mutation conditions to gain a better understanding of how the H87 and D342–R206 interaction networks control the extracellular gate (figure 4.12 for the 2D PMFs; all 1D reprojections are shown for reference in figure 4.13). From the data presented

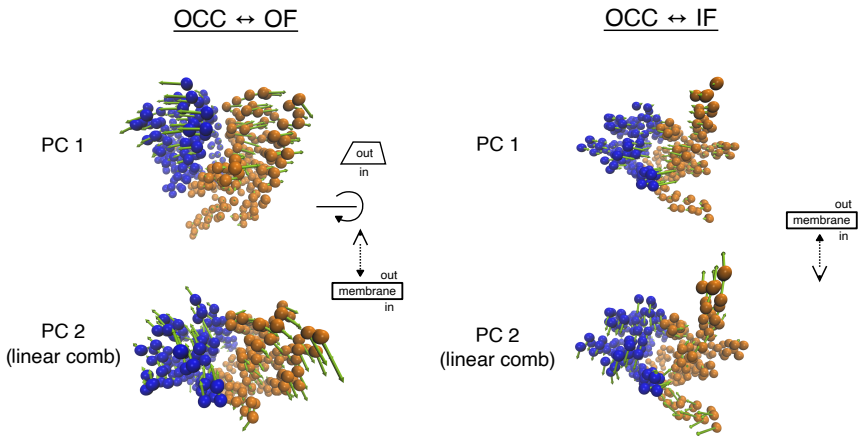


Figure 4.10: Illustration of the PCA-derived CVs for 2D REUS. TM-helix CA-atoms are shown as spheres, coloured blue for the N-terminal bundle and orange for the C-terminal bundle, while arrows show magnitude and direction. For the OCC↔OF transition, PC 1 corresponds to the gating motion, while PC 2 is a cleft-sliding movement. For OCC↔IF, PC 1 corresponds to the gating motion, while PC 2 is a twisting movement. See supplementary videos 1–4 of our publication (Lichtinger et al., 2024) for animated versions of the same representations.

thus far, it is not clear whether the effect of H87 protonation on the OCC → OF transition is due merely to the loss of hydrogen-bond interactions with S321, or whether the introduction of a positive charge in this location makes a significant mechanistic contribution. To address this question, instead of protonating H87 we mutated it to alanine (H87A). In the resulting PMF, the OCC state is raised less with respect to OF compared to the protonated version, and the transition barrier increases to $\approx 5\,\mathrm{kcal\,mol^{-1}}$, suggesting that there exists an interaction made by positively charged H87 that becomes particularly relevant in the transition region. Further analysis of the H87 interaction networks in our 2D REUS trajectories (figure 4.14a) reveals that when H87 is protonated, the interaction with S321 is substituted by an interaction with D317 that is strongest in the transition region. Interestingly, this suggests an alternative mechanistic role for the essential D317, which — as discussed above — we have not found forming the structurally observed saltbridge with K64 consistently in our simulations.

An equivalent investigation of the D342A mutation results in a PMF that shows both a decrease in the transition barrier, and — as opposed to D342 protonation — also raises the OCC state in energy with respect to OF. This may be explained by the fact that protonated D342 can still hydrogen-bond with R206, so although the interaction is less prominent it presumably still contributes somewhat to OCC-state stability (figure 4.14b). As a control, we also show that although the protonation of the ExxER glutamates (E53 and E56) can affect the relative OCC vs OF free energies, there is no lowering of the transition barrier (while they do have an effect on the transition barrier separating OCC from IF, as discussed below). Convergence analysis and representative 2D REUS histograms for our OCC ↔ OF PMFs can be found in figures 4.15 and 4.16.

We next employed an equivalent approach to investigate the OCC↔IF transition. As can be seen from figure 4.17, in standard protonation states, the OCC state forms a well-defined basin, connecting to a broader and shallower IF-state basin raised $\approx 2\,\mathrm{kcal\,mol^{-1}}$ over OCC via a barrier of just over $\approx 3\,\mathrm{kcal\,mol^{-1}}$. When either E53 or E622 are protonated, the IF state drops to a similar free energy level as OCC and the transition barrier lowers by $\approx 1\,\mathrm{kcal\,mol^{-1}}$. These effects are additive, so that when both E53 and E622 are

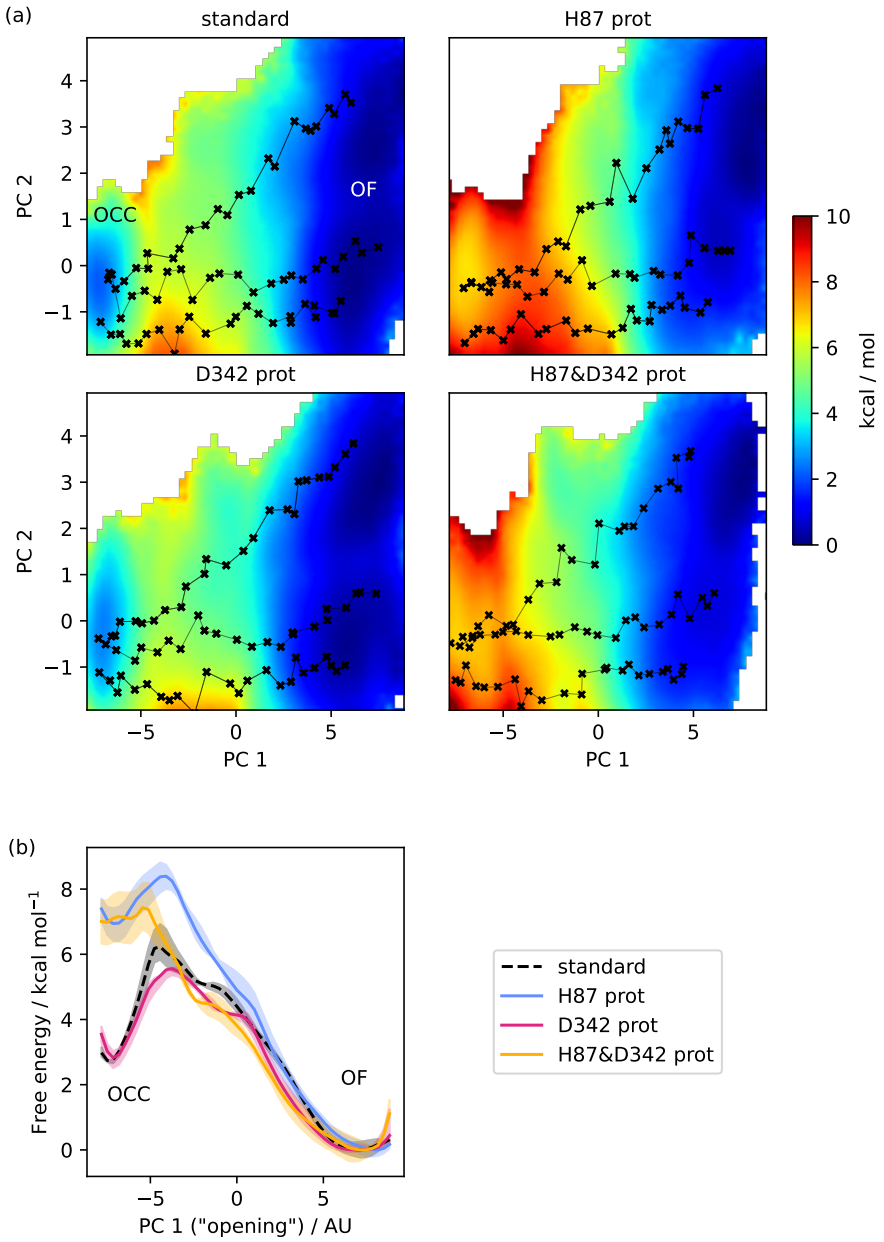


Figure 4.11: (a) 2D PMFs from REUS starting with MEMENTO paths, in different protonation states of candidate extracellular gating residues. (b) Projection of the 2D PMFs in part a onto PC 1 using Boltzmann reweighting. Shaded areas indicate convergence errors as the range of PMF values for a given CV value obtained with the first 40 %, the last 40 % and 100 % of sampling included (after alignment to the 100 % curve). H87 and D342 form an additive extracellular gate, where H87 protonation changes the relative OCC–OF state energies as well as the transition barrier, while D342 protonation only contributes in the transition region. Note that the individual PMFs are only determined by our REUS approach up to additive constants, and are shown aligned here at the OF state for convenience of comparison.

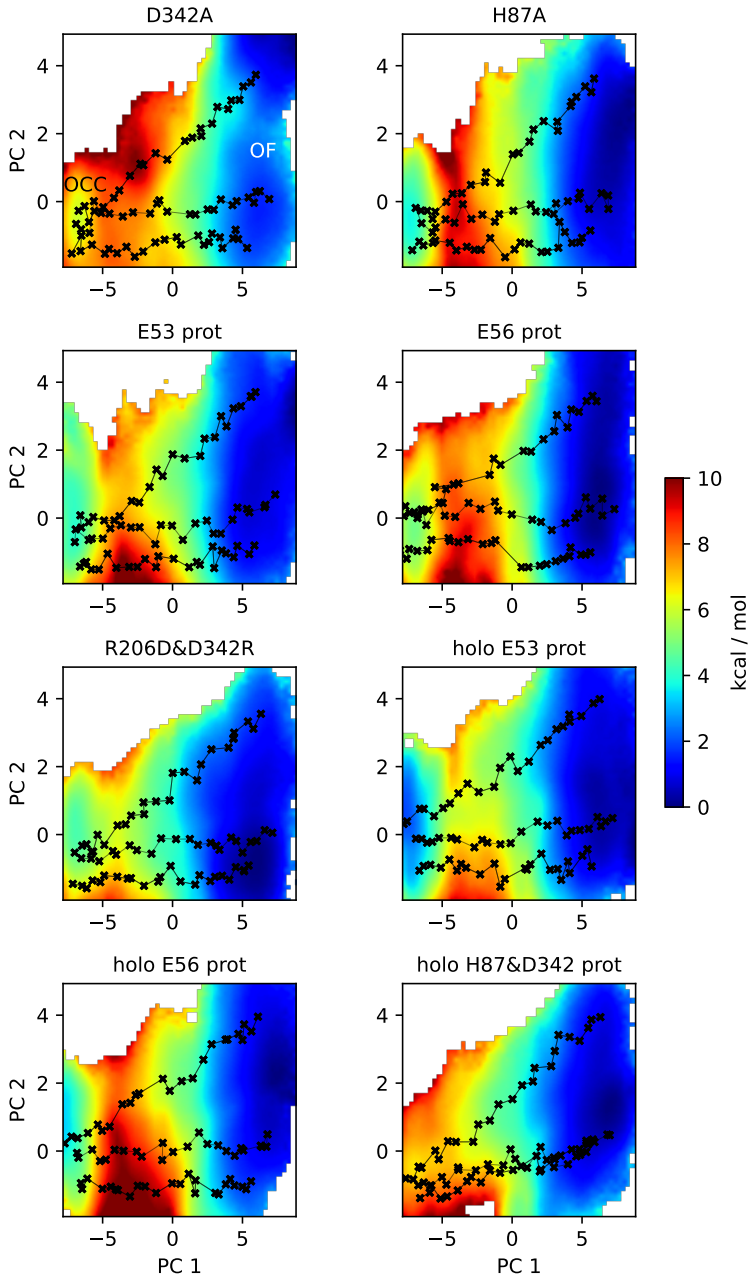


Figure 4.12: 2D PMFs of the OCC↔OF transition from REUS with MEMENTO paths in additional protonation state and mutation conditions.

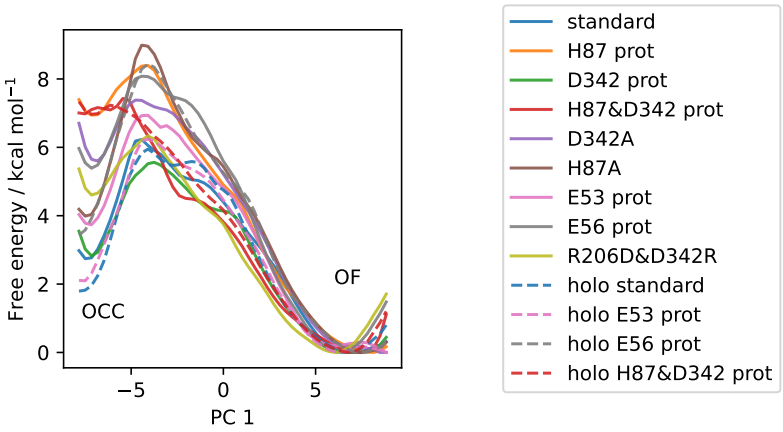


Figure 4.13: All available OCC↔OF 2D PMFs, projected onto the first CV (PC 1). Solid lines are apo PMFs, dashed lines are Ala-Phe substrate-bound and color-matched to the respective apo PMF. Note that the individual PMFs are only determined by our REUS approach up to additive constants, and are shown aligned here at the OF state for convenience of comparison.

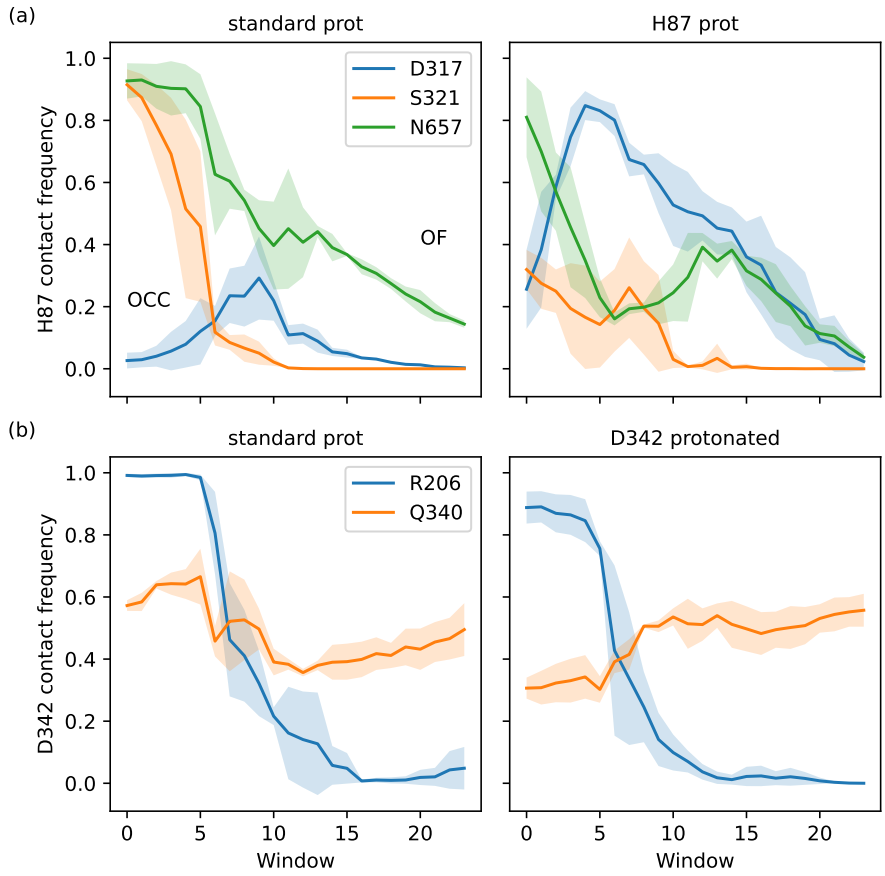


Figure 4.14: Interaction plots of the 2D PMF trajectory data (high force constant windows only), calculated as frequencies of finding inter-residue heavy-atom distances smaller than 0.35 nm, shown as a line for the average across 3 replicates with shaded standard deviations. (a) H87 interactions with D317, S321 and N657. Protonation of H87 replaces the interaction with S321 by one with D317. (b) D342 interactions with R206 and Q340. The tight salt bridge D342–R206 is disrupted by D342 protonation, but the residues still interact in the OCC state via hydrogen bonds.

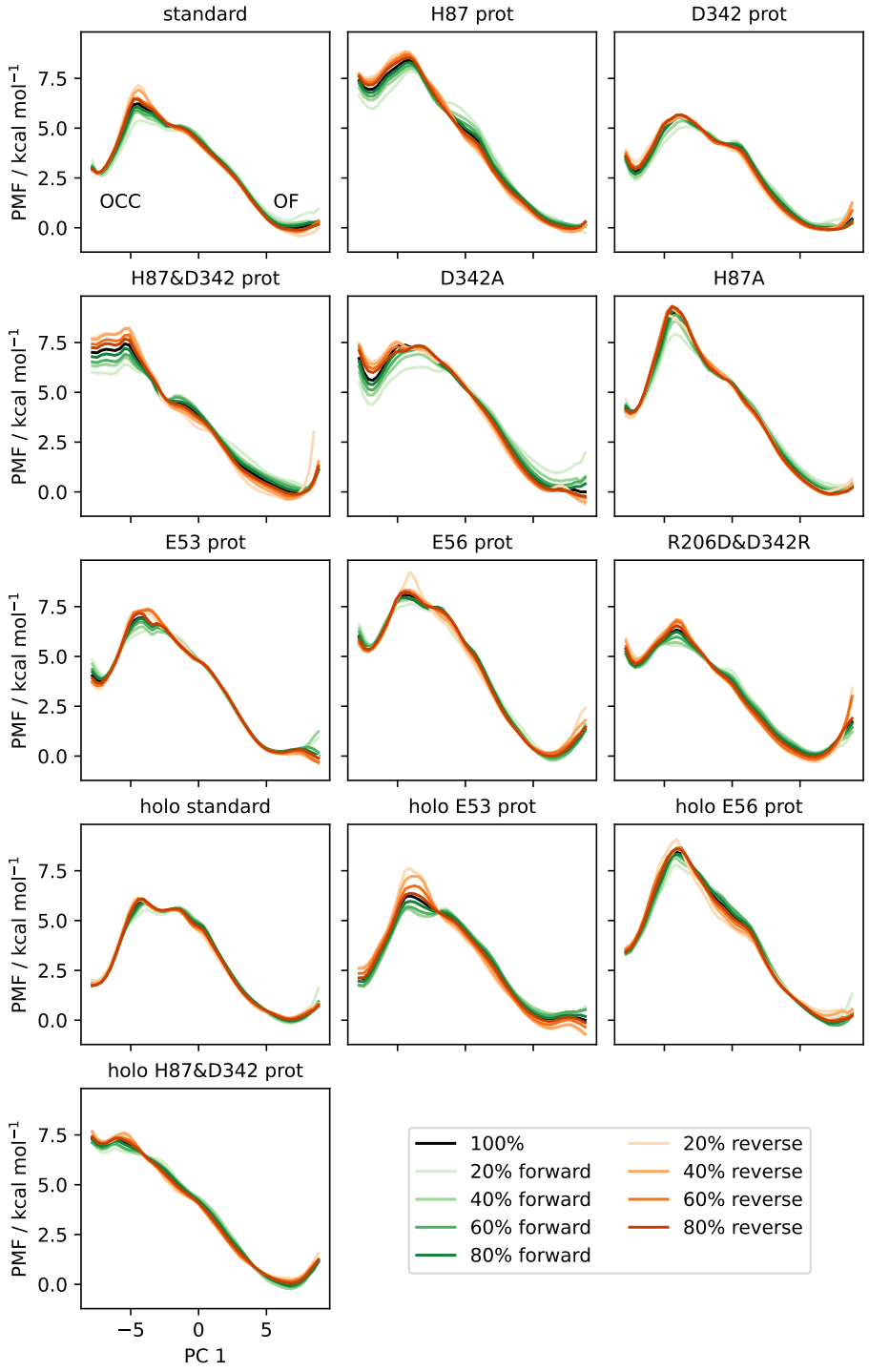


Figure 4.15: Convergence plots of all OCC↔OF 2D PMFs, shown as projections onto PC 1 including successively (increasing saturation) more data points, starting from the first frame (green) or from the last frame in reverse (orange). The PMF using all data is shown in black.

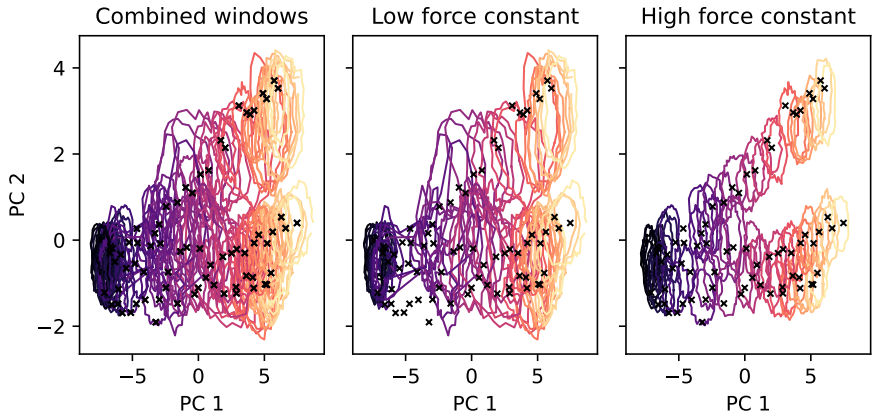


Figure 4.16: 2D REUS histograms for the OCC↔OF standard protonation state 2D PMF, drawn as contour lines at 30% of the maximal histogram height for each window (coloured by window, from purple at OCC to yellow at OF). Black crosses indicate the MEMENTO-derived REUS starting frames. Using the lower force constant (see Materials and Methods) gives good overlap in the basin regions while the transition region is undersampled. Using the higher force constant gives good overlap in the transition region while the basins are not sampled widely enough. Combining all windows results in good overlap across the 2D-CV space.

protonated, the IF state is favoured by $\approx 3.5\,\mathrm{kcal\,mol^{-1}}$, and is accessible from OCC via a barrier of only $\approx 1\,\mathrm{kcal\,mol^{-1}}$. As for the OCC↔OF transitions, these results explain the behaviour we had previously observed in the unbiased MD of figure 4.5c. The stochastic partial intracellular gate opening seen with those runs can be rationalised through the lower transition barrier from OCC to IF compared to the transition to OF in our PMFs, together with the broad and flat shape of the IF-state basin. Additionally, we have also computed the equivalents of these PMFs in the presence of Ala-Phe substrate (figure 4.18), which we will discuss in the section below. All projections onto PC 1 are shown in figure 4.19, and convergence analysis is provided in figure 4.20. Taken together, the PepT2 apo 2D PMFs provide a detailed view of the way the alternating access cycle is driven by proton movement from the extracellular to the intracellular side of the transporter that fits well with its biological function of using two protons from the extracellular medium to energise cycling from OF to IF, and back spontaneously once the protons have left (see the discussion in section 4.5).

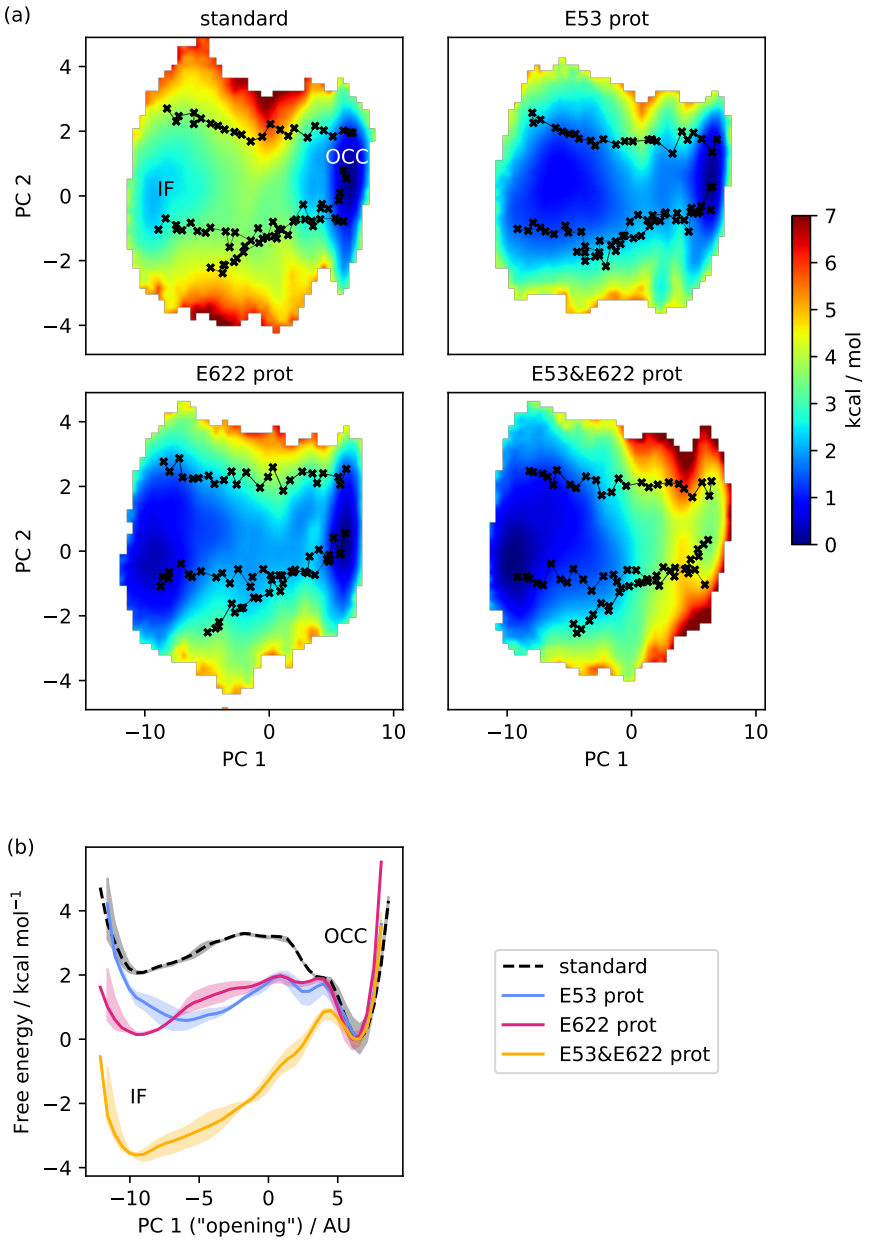


Figure 4.17: (a) 2D PMFs from REUS starting with MEMENTO paths, in different protonation states of candidate intracellular gate-controlling residues. (b) Projection of the 2D PMFs in part a onto PC 1 using Boltzmann reweighting. Shaded areas indicate convergence errors as the range of PMF values for a given CV value obtained with the first 40 %, the last 40 % and 100 % of sampling included (after alignment to the 100 % curve). E53 and E622 protonation have additive and approximately equal effects on driving the OCC→IF transition. Note that the individual PMFs are only determined by our REUS approach up to additive constants, and are shown aligned here at the OCC state for convenience of comparison.

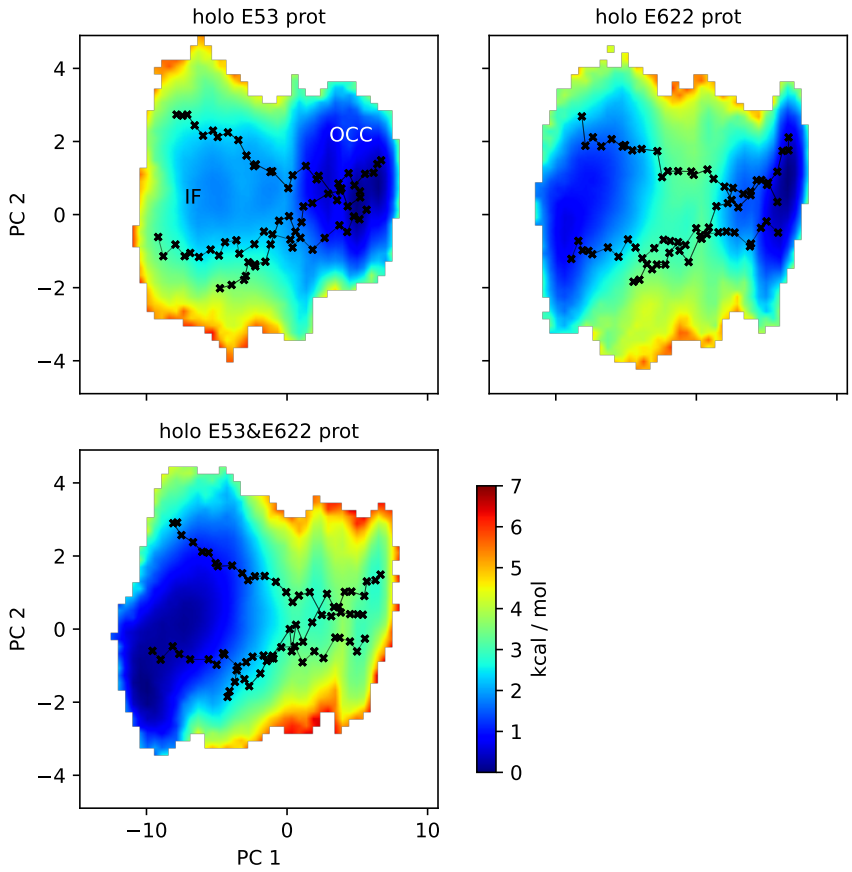


Figure 4.18: 2D PMFs of the OCC↔IF transition from REUS with MEMENTO paths in additional substrate-bound protonation state conditions.

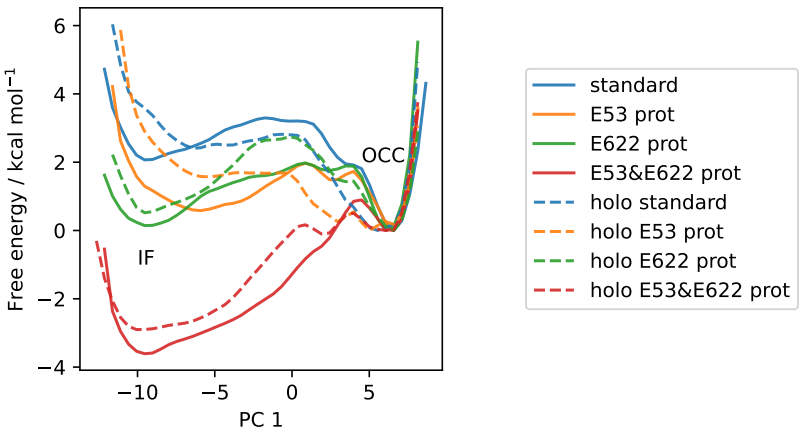


Figure 4.19: All available OCC↔IF 2D PMFs, projected onto the first CV (PC 1). Solid lines are apo PMFs, dashed lines are Ala-Phe substrate bound and color-matched to the respective apo PMF. Note that the individual PMFs are only determined by our REUS approach up to additive constants, and are shown aligned here at the OCC state for convenience of comparison.

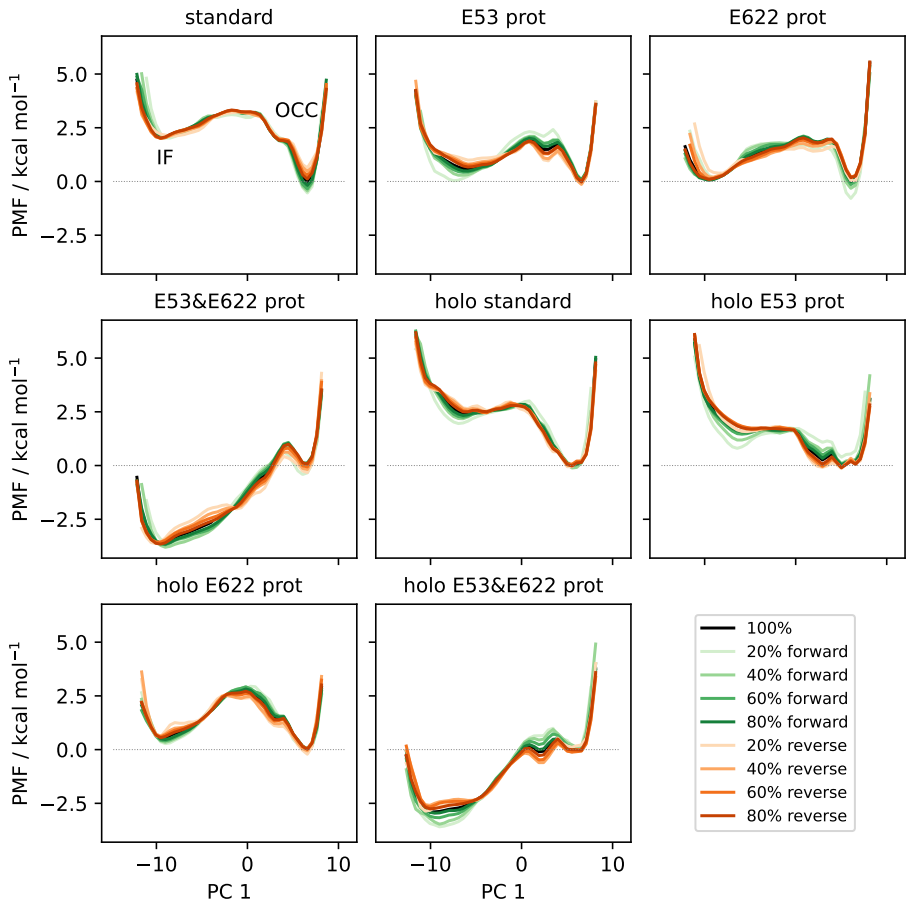


Figure 4.20: Convergence plots of all OCC↔IF 2D PMFs, shown as projections onto PC 1 including successively (increasing saturation) more data points starting from the first frame (green) or from the last frame in reverse (orange). The PMF using all data is shown in black.

4.3 Several mechanisms for substrate coupling

Given the evidence presented so far, which provides a plausible model for how protons drive alternating access based just on an investigation of the apo states, it remains unclear how coupling of proton transport to the substrate is achieved — that is, why the transporter cannot be driven by protons without the presence of substrate (and would thus just leak protons). To investigate the mechanism underpinning peptide–proton coupling, we constructed simulation boxes that included a bound Ala-Phe substrate molecule (see appendix B for details).

4.3.1 Protonation state-dependent ligand release

With these boxes at hand, we then calculated the Ala-Phe affinity using ABFE simulations in different protein states (table 4.1). Firstly, we observed that the affinity is similar in the OF and IF states, indicating that the binding of substrate alone does not thermodynamically drive the transporter from OF to IF. We did find, however, that protonation of E622 (i.e., the saltbridge partner of the substrate N-terminus) significantly decreases substrate affinity. Given that protonation of E622 also favours the OCC→IF transition, we suggest a dual function of E622 protonation that includes both stabilising the IF state with respect to OCC and facilitating substrate release from holo IF.

Protein state	$-\Delta G_{\mathbf{bind}}$ / kcal mol^{-1}
OF	8.0 ± 0.3
OF & E53 prot	8.9 ± 0.5
IF	7.0 ± 0.4
IF & E622 prot	2.9 ± 0.2

Table 4.1: Results of ABFE calculations, showing that the affinity of Ala-Phe substrate does not depend much on the conformational state (OF vs IF), but is significantly decreased on E622 protonation.

4.3.2 Substrate-modulated alternating-access thermodynamics

To explore what effect the substrate has on the PepT2 conformational landscape, we repeated a set of our 2D PMFs in the presence of substrate (figure 4.21). For both the OCC↔OF and OCC↔IF transitions, the 2D PMFs have similar shapes in the apo and holo states. For OCC↔OF, however, we found an increased width of the OCC state basin in the direction of PC 2, which — as is evident after projection onto PC 1 — stabilises OCC with respect to OF by $\approx$1 kcal mol^{-1} as an effect of increased OCC flexibility in the orthogonal DOF. We find a similar stabilisation when E53 or E56 are protonated, but not when both H87 and D342 are protonated (figure 4.12 and 4.13). This indicates that the presence of substrate in conformations approaching the OCC state from OF may trigger proton movement further down into the transporter — driven by entropic gains from increased flexibility in orthogonal DOFs.

The OCC↔IF PMF also presents a broader OCC-state basin in the presence of substrate, this time in the direction of PC 1 (consistent with the higher OCC-state flexibility directed towards IF observed in unbiased MD, see figure 4.5c). In the projection onto PC 1, this manifests as a broader free energy well and a lower barrier towards IF, even if the relative energies of the basins are not significantly affected. This effect is similarly apparent when E53 is protonated, but not with E622 protonation, which instead leads to a raised transition barrier (these opposing trends continue to be additive when both residues are protonated; see figure 4.18 and 4.19). We reason that this is due to a more flexible substrate orientation (disengaging the N-terminus) when E622 is protonated. While, taking these observations together, substrate binding does lead to an OCC→IF bias, it also seems unlikely that E622 is protonated at the moment when the OCC→IF transition happens in the holo transporter. This may suggest a further intermediate protonation step that our simulations have not captured (see the discussion in section 4.5).

4.3.3 Coupling between substrate binding and ExxER motif protonation

As noted in the introductory section 1.1.3 and as suggested by some previous computational work on bacterial homologues (see the discussion in section 1.3), another possibility for the substrate to engage with the transport cycle is found in R57 interacting both with the substrate C-terminus and with the ExxER glutamates, protonation of which drives the transporter towards the IF state, as our PMFs demonstrate. A natural hypothesis then is that substrate binding — which engages R57 — loosens the R57 interaction with E53 and/or E56, thus enabling the protonation of those residues and progress along the alternating access cycle. To test this hypothesis, we conducted CpHMD simulations (Swails et al., 2014) to probe the E53 and E56 pKa values in all combinations of OCC / OF and apo / holo conditions (figure 4.22a). Concerning the E53 pKa value, we see a potential response to substrate binding in the OF state (though the error bars calculated from the triplicate standard deviations overlap), but not in the OCC state. On the other hand, we do see a raising of the E56 pKa beyond error in holo OF or OCC states compared

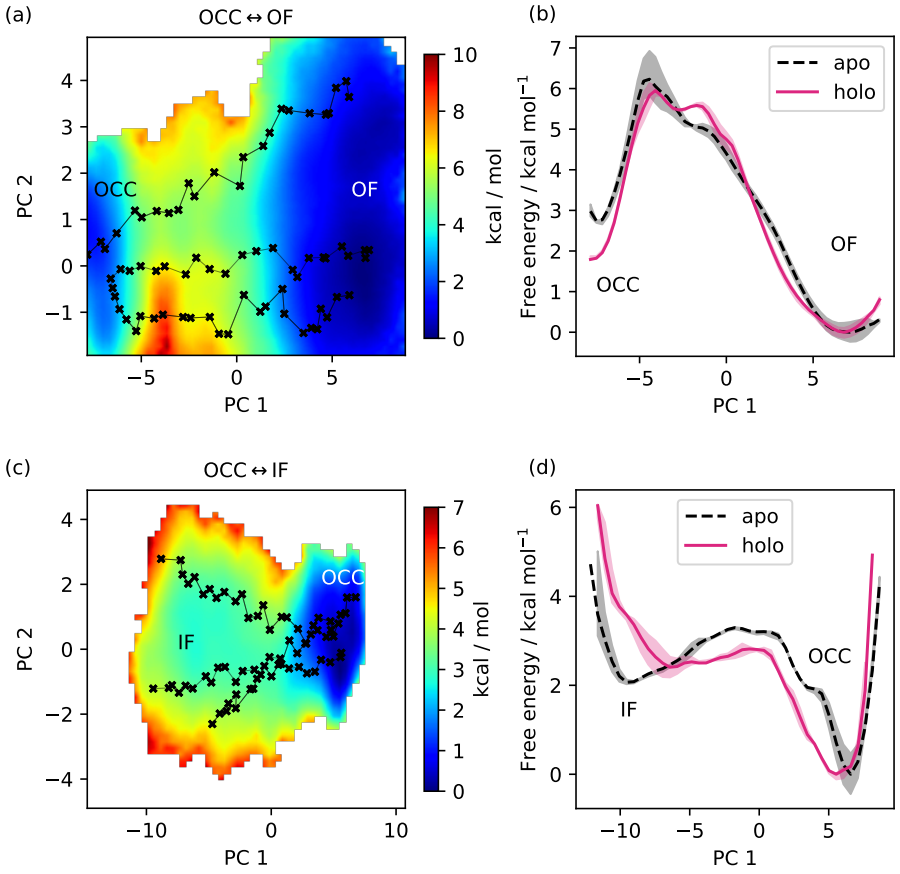


Figure 4.21: (a) 2D PMF for the OCC↔OF transition from REUS starting with Ala-Phe-bound PepT2 MEMENTO paths. The OCC state has an increased basin width in PC 2 (compare for figure 4.11a), and a transition path shifted in PC 2. (b) Projection of the PMF from panel a onto PC 1, showing how in holo PepT2, the OCC state is stabilised by $\approx 1\,\mathrm{kcal\,mol^{-1}}$. Shaded areas indicate convergence errors as the range of PMF values for a given CV value obtained with the first 40 %, the last 40 % and 100 % of sampling included (after alignment to the 100 % curve). Note that the individual PMFs are only determined by our REUS approach up to additive constants, and are shown aligned here at the OF state for convenience of comparison. (c) 2D PMF for the OCC↔IF transition from REUS starting with Ala-Phe-bound PepT2 MEMENTO paths. The structure of the IF plateau is not significantly affected, but OCC is more flexible in PC 1. (d) Projection of the PMF from panel c onto PC 1, showing how in holo PepT2, the OCC state has a broader basin, corresponding to intracellular gate flexibility. Convergence error and alignment of PMFs is shown as in panel b.

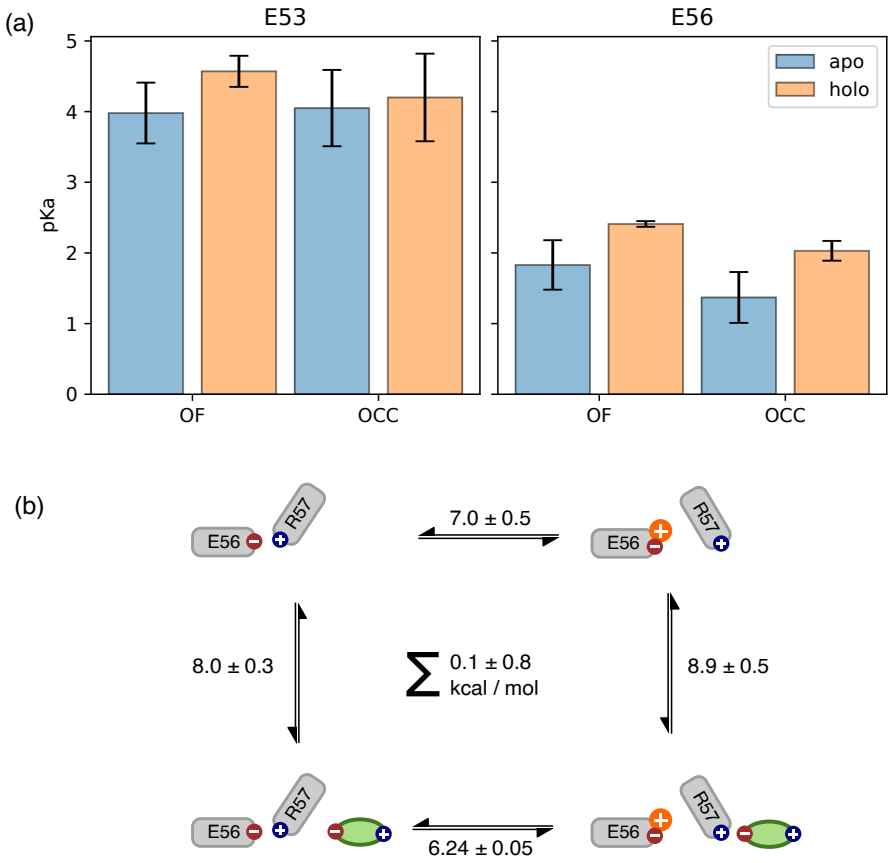


Figure 4.22: (a) E53 and E56 pKa values from CpHMD simulations, in the apo and holo as well as the OF and OCC states, estimated as mean ± standard deviation from triplicate runs (using the full simulation data for fitting the titration curves). The presence of substrate raises the E56 pKa in either conformational state, while some effect on the E53 pKa may also exist in the OF state. (b) Illustration of a thermodynamic cycle of E56 protonation and Ala-Phe binding, with edges filled in via CpHMD (converted into kcal mol^{-1} at pH 7) for the top and bottom transitions, and ABFE for the left and right edges. Notably, ABFE displays a response of Ala-Phe affinity to E56 that is consistent with the CpHMD results, and the cycle closes very well. The error in the cycle closure residual is estimated as a square root of the sum of squared standard deviations of the individual edges.

to apo, amounting to ≈0.6 log units in both cases. As shown in figures 4.23 and 4.24, the pKa values estimated for successive data chunks across the CpHMD trajectories vary significantly with simulation time in a complex superposition of timescales, and with a dynamic range larger than the replicate error bars. As an alternative to the replicate-based representation in figure 4.22a, we have therefore also analysed the pooled data for each condition as histograms of pKa values estimated from short chunks of our simulations (figure 4.25). From this, we recover the same effect of substrate binding on the E56 pKa in the OF and OCC states, as well as a potential effect on the E53 pKa in the OF state only. Since the shift in E56 pKa was more robust across conformational states, we focus on this residue in the following validation and discussion, although we note that if there was in fact a significant raising of the E53 pKa as well, this would further strengthen our conclusions about substrate coupling to ExxER motif protonation.

It is important to recognise here that the hybrid-solvent CpHMD method as implemented in AMBER is not rigorous for membrane proteins, since the membrane is not taken into account while evaluating proposed protonation state changes in implicit solvent. On the other hand, we were not able in initial

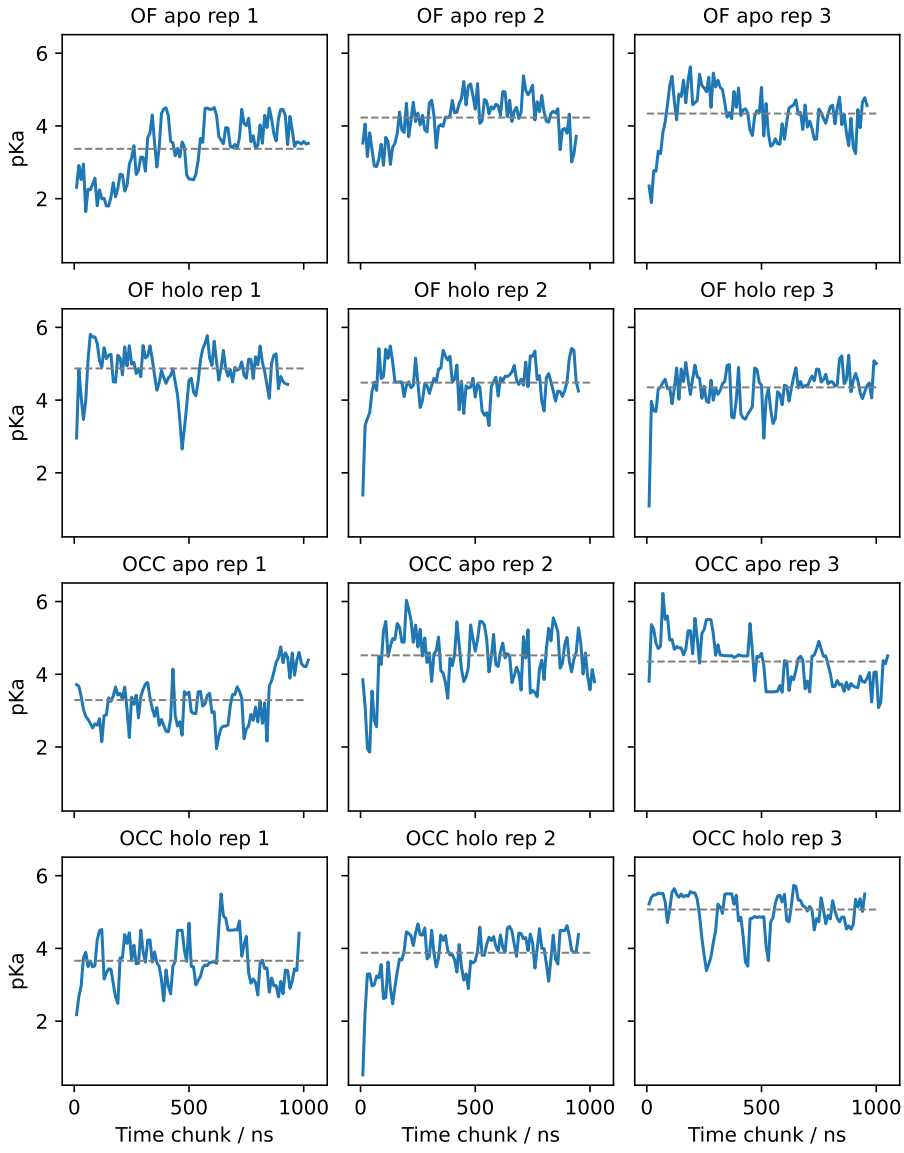


Figure 4.23: E53 pKa values estimated for (separate) successive chunks of 80 ns (10 ns per pH window) of CpHMD via fitting the Hill equation. Grey lines indicate the pKa resulting from fitting the total simulation data. The pKa values appear highly dynamic and superimpose several relaxation timescales. Convergence seems to be achieved in most replicates but is not straightforward to assess.

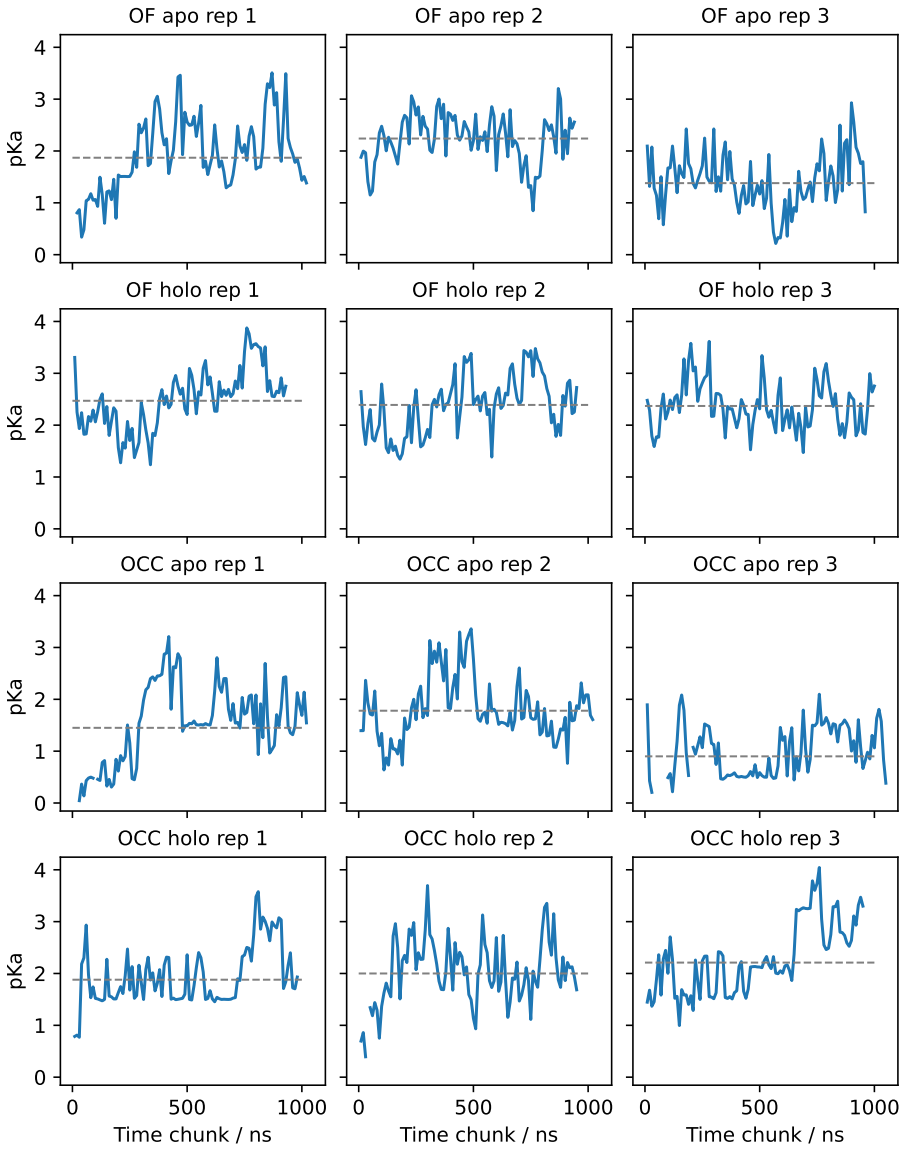


Figure 4.24: E56 pKa values estimated for (separate) successive chunks of 80 ns (10 ns per pH window) of CpHMD via fitting the Hill equation. Grey lines indicate the pKa resulting from fitting the total simulation data. The pKa values appear highly dynamic and superimpose several relaxation timescales. Convergence seems to be achieved in most replicates but is not straightforward to assess.

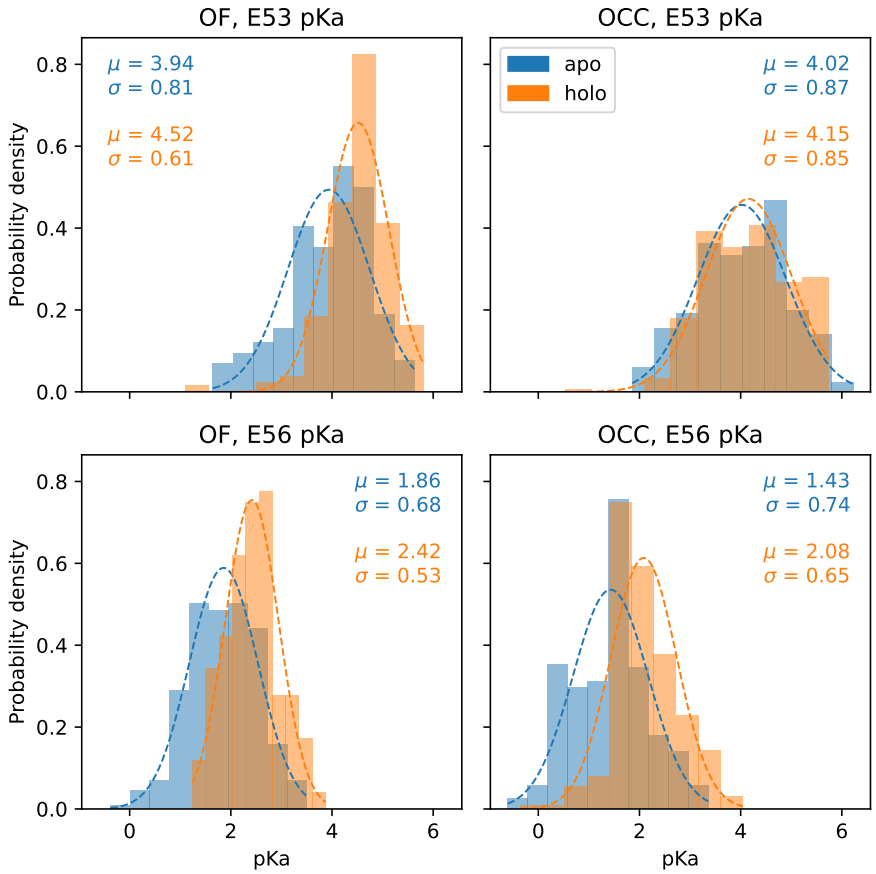


Figure 4.25: Histograms of the pKa values estimated from chunks of 80 ns (10 ns per pH window) of CpHMD, pooled for all trajectories of a given condition. Apo conditions are shown in blue, holo (Ala-Phe bound) are shown in orange. Gaussian fits are indicated by dashed lines and the resulting fit parameters are given in the respective color for each histogram.

trials to obtain sufficient transition counts to converge an alternative explicit-solvent CpHMD method as implemented in GROMACS. (Aho et al., 2022) [3] To validate our results, we therefore also constructed a thermodynamic cycle of substrate binding and E56 protonation in the OF state by including data from separate ABFE calculations as connecting thermodynamic legs (figure 4.22b). The ABFE results show a complementary effect of the E56 protonation state on substrate affinity, closing the thermodynamic cycle and validating our CpHMD simulations using an orthogonal MD technique. We conclude that substrate binding does indeed facilitate protonation of E56 (whence, we stipulate, the proton moves to E53, which is situated in close proximity). If the neglected membrane environment in the hybrid-solvent CpHMD did produce significant artifacts in the pKa values, then it would appear that there is error cancellation when assessing the impact of substrate binding as a difference of pKa values in the apo and holo conditions.

It should be noted that — as throughout this thesis, see the discussion below — in studying the coupling between substrate binding and protonation-state changes at the ExxER motif we have not made the substrate C-terminus protonatable. Since, in order to induce E56 protonation, the substrate C-terminus

[3] We have discussed this in more detail in the author responses to reviewer comments on our publication. (Lichtinger et al., 2024) In short, the all-atom CpHMD simulations become stuck in local minima of the coupled phase space spanned by the E56 protonation state and the side-chain conformational changes associated with R57 rotation away from E56 in response to protonation. This leads to very slow protonation-state transitions and the impossibility of estimating converged pKa values on the accessible time scales.

needs to engage R57 in a salt bridge, its pKa is likely to be low, rendering the assumption reasonable for those substrate conformations. However, it is possible that the system could also adopt states in which the ExxER motif salt-bridge network is stable in a way similar to the apo condition while the substrate gets protonated when oriented away from R57. If such conformations contribute significantly to this semi-grand canonical ensemble, the ExxER motif pKa values estimated without taking them into account may exhibit some bias. By undersampling more apo-like conformations in the holo state in this way, it is possible that the calculations presented here overestimate the substrate-induced pKa shift of E56, although the direction of change would be expected to be the same, because the substrate can still engage R57 when it is deprotonated (we speculate that the histograms in figure 4.25 for the holo state may become bi-modal in this case). While the possibility would need to be taken into account for a more rigorous quantitative estimate of the ExxER pKa value shifts, the pKa calculations would become much harder to converge since the slow DOF of substrate re-orientation would need to equilibrate to the protonation-state changes (which happen fast since they are treated as Monte-Carlo moves in an implicit solvent at regular intervals). Here, we contend ourselves with the more qualitative observation that an appropriately positioned substrate in the canonical, structurally observed binding pose facilitates ExxER-motif protonation.

4.4 Validation of extracellular gating residues in cell-based transport assays

To experimentally validate the results of an MD investigation, a first step is to probe the importance of the key implicated residues for protein function. We note that the literature already contains ample data to show that E53, E56, R57, H87 (for those homologues which conserve it), D317 (or the equivalent glutamate) and E622 are important for transport through POTs (see the literature review in table 1.1). Of the residues implicated by our simulations, therefore, only S321 and D342 have not been studied before, and thus serve as predictive validation test cases here.

Using cell-based transport assays (see appendix B for details), we tested the transport activity of rat PepT2 wild-type (WT) and several mutants: H87A (as a positive control known from the literature), I135L (as a negative control, without any expected effect), as well as the mutants of interest D342A and S321A (figure 4.26). We note that all our mutants expressed slightly less compared to the WT at the same amount of transfected DNA ($0.8\,\mu g$), but more than WT at a reduced transfection DNA level ($0.5\,\mu g$) (figure 4.26b). To control for this difference in expression levels, we took WT ($0.5\,\mu g$), which transports $\approx$20% less than WT ($0.8\,\mu g$), as a lower bound for the WT transport activity and as the point of comparison for statistical tests. We found that all mutants of residues predicted to be involved in transport displayed significantly reduced transport activity (p-values: 2.2×10^{-5} for H87A, 1.6×10^{-4} for D342A, 6.9×10^{-5} for S321A, while I135L is indistinguishable from WT at p = 0.79). We also note that D342A, although its activity is significantly reduced, still transports more than H87A (p = 3.9×10^{-7}). This fits well with our 2D PMF results, where H87 protonation does more than D342 protonation to stabilise OF with respect to OCC.

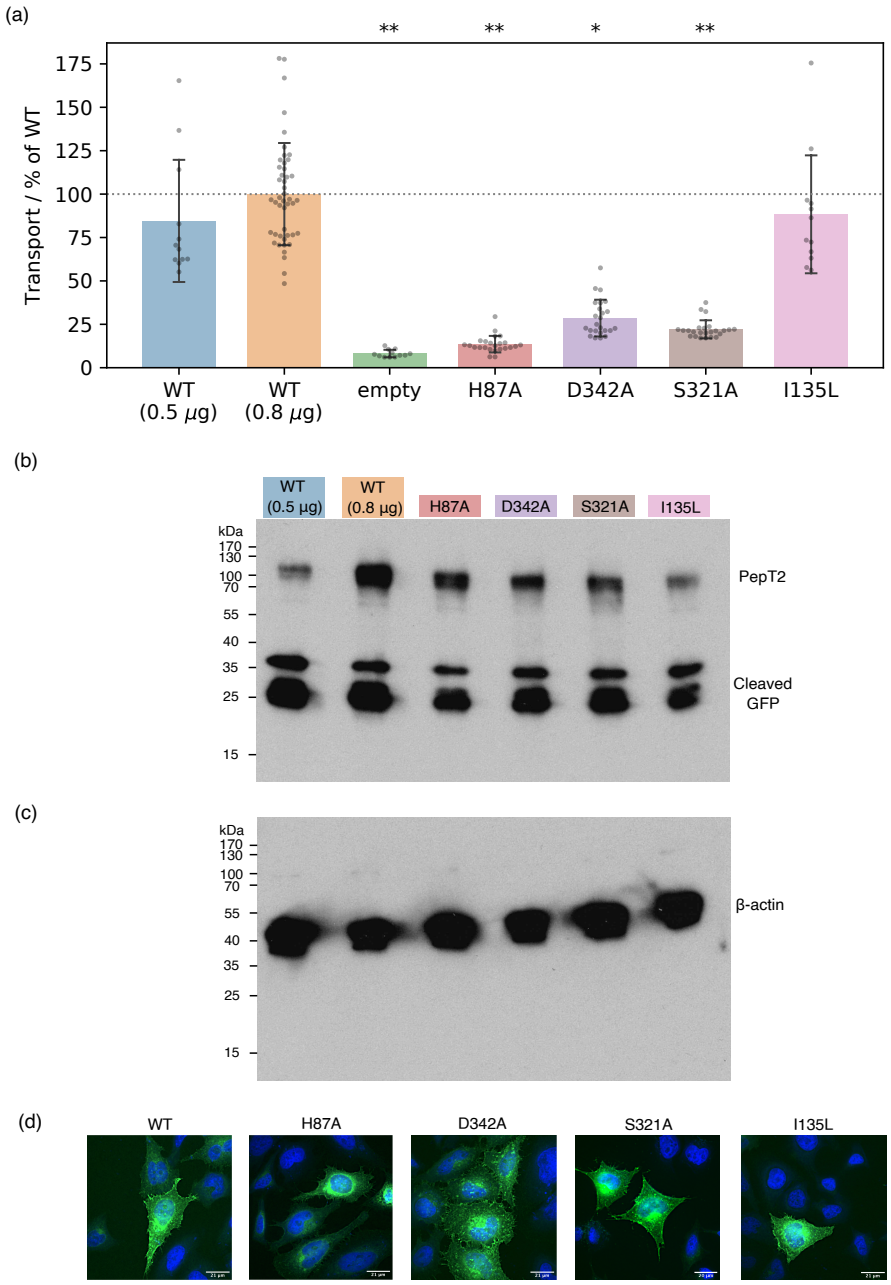


Figure 4.26: (a) Cell-based transport assays for PepT2 WT (transfected with 0.5 µg, n=12, and 0.8 µg, n=46, of DNA per well), empty plasmid vector (n=12) and PepT2 H87A, D342A, S321A (n=24 each) and I135L (n=12) mutants, all transfected with 0.8 µg of DNA. Diagram shows transport as fluorescence in post-assay lysate divided by total protein concentration, normalised to the WT (0.8 µg) mean. Bars are mean values ± standard deviation, and swarm plots show samples corresponding to individual wells. Single asterisk indicates $p < 10^{-3}$, double asterisks $p < 10^{-4}$ significance levels for difference compared to (weaker-transporting, 0.5 µg-transfected) WT, as evaluated using a two-tailed t-test. (b) Western-blot showing expression levels of WT and mutant GFP-labelled PepT2, with an anti-GFP primary antibody. All mutants express at levels between the WT transfected with 0.5 µg and 0.8 µg plasmid DNA. Cleaved GFP is also visible at low molecular weight, at levels comparable for WT and mutants. (d) Loading control of the Western blot shown in (b), using an antibody against β-actin, showing even loading of the gel. (e) Fluorescence microscopy images, overlaying GFP-labelled PepT2 (green) with DAPI-labelled DNA (blue). Membrane expression is qualitatively shown for WT and all mutants by the thin cell outline of GFP.

4.5 Discussion

Integrating the results from extensive sampling across several MD methods, covering all stages of the PepT2 alternating access cycle, we are now in a position to propose a detailed molecular mechanism of the complete transport cycle, including accounts both of proton coupling to conformational changes and of substrate–proton coupling (figure 4.27). Starting at the apo OCC state without any protons bound, we find in our 2D PMFs that proton binding to H87 and D342 stabilises OF with respect to OCC (with H87 being the major contributor). Given that H87 and D342 are accessible from the acidic extracellular bulk, and in light of the transition region stabilisation from the H87 (protonated)–D317 interaction we have identified, our simulations suggest an interpretation where protonation happens in the OCC state, driving the OCC→OF transition by stabilising the OF state over OCC. However, if the OCC→OF transition is kinetically accessible on experimental timescales without prior protonation events (beyond what our MD was able to sample), it would also be consistent with our data that OCC→OF is spontaneous in standard protonation states, and H87 and (possibly) D342 are merely the initial sites of protonation once OF is reached, providing further stabilisation. [4] We note in this context the limitation of using non-reactive MD methods for sampling the PepT2 conformational changes. In principle, a multi-dimensional PMF calculated with a reactive MD method where one CV is an explicit proton movement coordinate could disambiguate between the possible scenarios here. However, we believe that such a fully coupled treatment of proton movement and large-scale conformational changes is not yet computationally feasible, and focussed here on achieving convergence of the conformational sampling in discrete protonation-state combinations. The question of whether proton binding happens in the OCC or OF state therefore warrants further investigation, and indeed the co-existence of several mechanisms may be plausible. Nonetheless, this study contributes important details to a mechanistic understanding of the thermodynamics of proton-coupled alternating access.

Our results with regard to the driving forces behind the OCC→OF motion agree with the work of Parker et al. (2017), who found spontaneous opening of $PepT_{Xc}$ towards OF after protonating the equivalent histidine to H87 (PepT2). In transporters which do not conserve the mammalian histidine at the TM 2 position such as $PepT_{St}$ (Batista et al., 2019) and $PepT_{Sh}$ (Li et al., 2022), on the other hand, previous simulation studies have implicated protonation of the glutamates equivalent to D317 (PepT2) in the opening of the extracellular gate (see section 1.3 for a more detailed discussed of these studies). Our simulations suggest that the involvement of D317 in the extracellular gating mechanism of PepT2 is by interacting with the protonated H87 to stabilise the transition region of extracellular gate opening. The mechanism of extracellular gating is therefore conserved less widely than the overall alternating access mechanism. This point is further highlighted by our computational and experimental results indicating a role of the D342–R206 saltbridge, which is conserved only among mammalian POTs, but not in $PepT_{So}$ and $PepT_{Xc}$ which do have the TM 2 histidine. This may explain why spontaneous opening in unbiased MD was observed by Parker et al. (2017) for $PepT_{Xc}$ following just H87 protonation, while for PepT2 in this work protonation of D342 is also required.

Once the transporter is in the OF conformation, the substrate enters the binding site. We have treated

[4] It may seem like the latter model is favoured by the fact that in our PMFs, OCC lies higher than OF, even when neither H87 nor D342 are protonated. We believe, however, that there is a danger of overinterpreting this feature of the PMF. Any combination of effects from forcefields, lipid composition and the population shifts afforded by transmembrane electrochemical gradients could perturb the conformational equilibria. It would therefore not be meaningful to interpret the shape of a single PMF in this way; only responses of the PMF to protonation-state or substrate-binding changes should be used to inform our view of the conformational cycle, since these are likely to benefit from error cancellation with respect to factors that act on the overall protein conformations (which are conserved between the different conditions in which the PMFs were sampled).

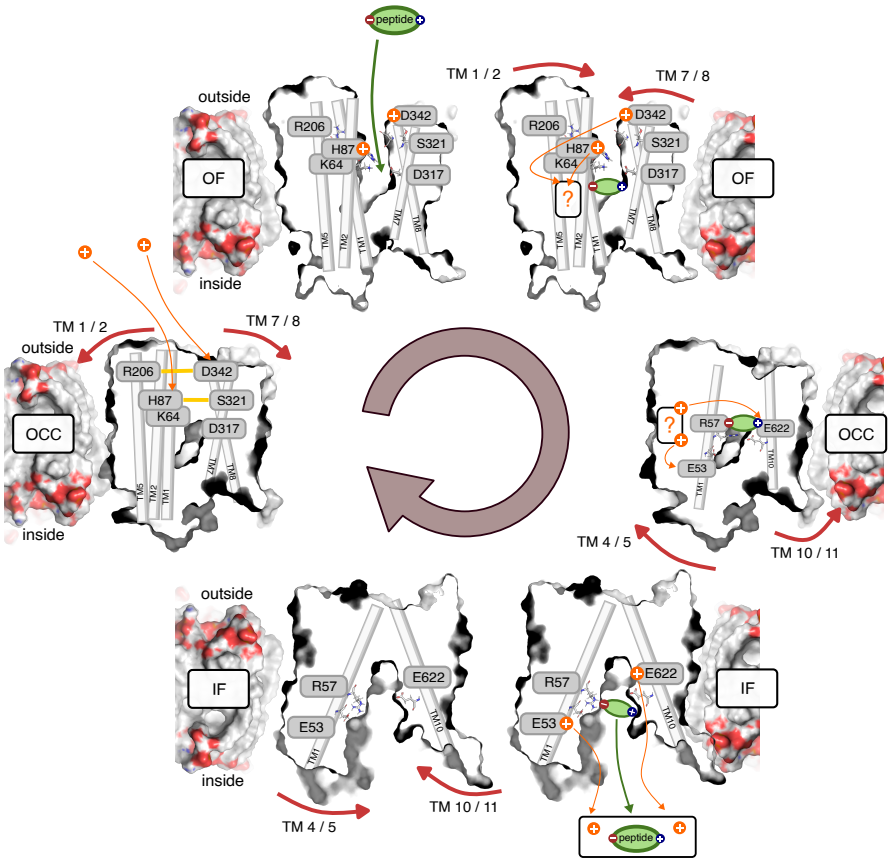


Figure 4.27: Schematic overview of the PepT2 alternating access transport cycle proposed in this work. Protons located at a question mark indicate a proton-transfer step with an as-of-yet unknown mechanism regarding intermediate residues.

this step alchemically in our ABFE simulations, so that the data presented here is agnostic of the orientation of the entering substrate and the sequence of engagement of the binding-site residues; previous MD simulations have suggested, however, that positioning of the peptide N-terminus precedes the C-terminus moving into place. (Parker et al., 2021) Substrate binding then has two distinct effects: firstly, it exerts a small thermodynamic bias towards the OCC state via increased flexibility in degrees of freedom orthogonal to the overall OF→OCC transition. Secondly, through engaging R57, substrate entry increases the pKa value of E56 in the ExxER motif, thus thermodynamically facilitating the movement of protons further down the transporter cleft. We can also speculate that — in addition to this thermodynamic favouring of E56 protonation — there might be a kinetic effect on proton transfer from moving the positively charged R57 out of the way of the incoming proton. This could be investigated using reactive MD or quantum mechanics / molecular mechanics (QM/MM) simulations (both approaches have been employed for other protonation steps of procaryotic peptide transporters, see Parker et al. (2017) and Li et al. (2022)), however the putative path is very long (≈ 1.7 nm between H87 and E56) and may or may not involve a large number of intermediate protonatable residues in addition to binding site water. While such an investigation is possible in principle, it is beyond the scope of the present study. Likewise, a coupled enhanced-sampling treatment involving both proton movement and large-scale conformational changes — as discussed above for the ordering of steps in the OCC→OF transition — would make for interesting future work, once it becomes computationally tractable.

Our data is not fully determinate with respect to whether protons move to E56 before or after the OF→OCC conformational transition (our CpHMD, for example, remains agnostic on this matter since the shift in the E56 pKa value induced by the substrate is evident in both the OCC and OF states). We may interpret the fact that OCC is raised in energy while H87 is protonated and that substrate-induced OCC stabilisation is not found when H87 and D342 are protonated (but does occur when E56 is protonated) as an indication that proton movement is favoured before the transition into OCC is complete. On the other hand, the transition-region interaction between protonated H87 and D317 could also be interpreted as a potential facilitator of the OF→OCC transition. We thus speculate that the proton movement processes may happen as an ensemble of different mechanisms, and potentially occur contemporaneously with the conformational change. This, in addition to a flexible binding pocket, may also contribute to the substrate promiscuity mechanism.

We note at this stage that — throughout our study — we have not investigated the possibility of the substrate C-terminus itself becoming (transiently) protonated. This would need to be taken into account when treating proton movement through the transporter explicitly in the future (see the discussion of such approaches above). There is evidence in our simulations that an additional protonation site — aside from H87, D342, E53, E56 and E622 — may be involved in the mechanism, since E622 protonation, while biasing the transporter towards IF, also increases the OCC→IF transition barrier if Ala-Phe substrate is bound (we therefore indicate the proton movements at these stages with a question mark in figure 4.27). There is thus the intriguing possibility that the substrate itself may temporarily hold the proton, although given the nature of the data presented here only speculation is currently possible on this point. What is clear from our 2D PMFs, regardless, is that protonation of ExxER does act as an intracellular gate trigger (and may also pull the transporter through the chemical equilibria all the way from OF). Taking together our 2D PMFs and ABFE simulations, it is also clear that E622, in addition to being essential for peptide recognition, plays two further roles: its protonation both facilitates substrate release and makes an additive contribution to the IF-directed bias exerted by the intracellular gate triggers (whether E622 forms 'part' of the intracellular gate remains then as a merely linguistic question). At this stage, we do not yet have an understanding of how exactly intracellular gate opening (which involves breaking an assembly of several hydrophobic residues) is effected by the protonation of these glutamate residues. This question should prove interesting to study in future work. Once the substrate (driven by E622 protonation) and the protons (driven by their electrochemical potential gradient) have left through the open intracellular gate into the intracellular bulk, the resulting apo, standard-protonation-state IF conformation has a thermodynamic preference to return to the OCC state as evidenced in our 2D PMFs. We thus arrive back at the starting state and have completed the proton- and substrate-coupled alternating access cycle.

In support of our MD data, we present mutational analysis in cell-based transport assays. Mutations of H87, S321 and D342 to alanine all significantly decrease transport activity, with H87A having the strongest effect. Taken together with similar results in the literature on E53, E56, R57, D317 and E622 (as referenced above), all residues implicated by our study have therefore been confirmed in their importance for transport via mutagenesis. While the cell-based assay used here cannot differentiate for example between proton-coupling and non-proton-coupling residues, our results still provide a useful first step towards the validation of the gating mechanisms we propose with our PMFs, and should prove informative for the future design of more in-depth experiments. Part of a future study may also be investigating potential rescue mutants that were beyond the scope of this chapter, for instance swapping the D342–R206 salt bridge. A further piece of validation concerns the role of R57 in responding to the incoming substrate

by freeing up E56 for protonation (as indicated by our CpHMD and ABFE simulations). The cryo electron microscopy (cryo-EM) structures of drug-bound PepT2 and corresponding MD simulations presented in chapter 5 of this thesis reinforce our understanding of this mechanism.

In conclusion, the research presented in this chapter utilises the recent wealth of bacterial and mammalian peptide transporter structures to construct a model of their alternating access mechanism. We explain how the movement of two protons through the transporter drives the accompanying conformational changes, as well as how conformational changes and proton-movement events are coupled to the presence of substrate. Questions regarding some of the finer details, notably the precise sequence of proton transfers and conformational transitions (if a single such sequence exists) and whether further protonation sites contribute to the mechanism remain open for future investigation. Nonetheless, the evidence supplied here addresses the alternating access proton-symport mechanism in unprecedented detail, particularly through the extensive use of free-energy simulation techniques. This information will prove useful for the project of employing peptide transporters as vehicles for drug delivery — especially since what determines the efficacy of a transporter substrate is not only related to affinity but crucially also to an ability of the substrate to move through the steps of the alternating access cycle once bound to the transporter.

5

Substrates and inhibitors — drug dynamics in the transporter binding pocket

Disclaimer

The research presented in this chapter has been published (Parker et al., 2024), see also in the list of publications. All MD investigation and related figures in that publication were produced by myself with feedback from the co-authors, and will, with adaptations, be used in part here under the CC BY 4.0 license. All experimental work discussed in this chapter was conducted and analysed by the collaborators. Most of the writing in the publication was drafted by the co-authors, or collaboratively for the sections concerned with simulation results, and will therefore not be verbatim reproduced in this thesis. However, parts — including the method section — were written by myself and will be used directly.

In section 1.1.4 of this thesis, we have introduced the broad substrate specificity of proton-coupled oligopeptide transporters (POTs), ranging from natural di- and tripeptides to a range of marketed drugs, particularly β-lactam antibiotics. The affinities with which POTs bind such diverse substrates (often reported as inhibition constants of the transport of a labelled reference substrate) range from 10's of nM to mM values. (Biegel et al., 2006) This substrate promiscuity, together with their expression pattern across the small intestine and kidneys, renders POTs targets of interest for the development of orally bioavailable pharmaceuticals, as we discussed in section 1.2 and illustrate in figure 5.1a. Nevertheless, the mechanisms by which the peptide binding pocket (see section 1.1.3) accommodates drugs such as β-lactams, how substrates are distinguished from inhibitors and what determines the affinity and ability of a substrate to be transported remain poorly understood.

In this chapter, we present a collaborative study to address these questions by an investigation of the differences in how three β-lactam drugs (figure 5.1b) engage PepT2. We used *in-vitro* assays to confirm that the cephalosporin cefadroxil is a high-affinity, transport-effective substrate of PepT2, while the penicillin amoxicillin shows lower affinity and transport. Cloxacillin, a penicillin that — in contrast to amoxicillin — lacks a charged amino terminus, acts an inhibitor of transport only. To explain these differences, we solved cryo electron microscopy (cryo-EM) structures of the three drugs bound to PepT2. We then started extensive unbiased molecular dynamics (MD) simulations from these binding poses, showing how cefadroxil and amoxicillin are largely stable in engaging the acidic-patch E622 residue, but — as is the case in our structures — the drug carboxyl groups do not bind to R57 (which is known to clamp peptide substrates together with E622, see section 1.1.3). If, however, E56 in the conserved E^{53}xxER motif is protonated, R57 is freed up to engage the drug carboxyl groups. For cefadroxil, this is sterically possible while keeping the interaction between the drug and E622 intact, while in the case of amoxicillin this interaction weakens when coordinating R57 at the other end of the drug molecule, thereby suggesting a mechanism of substrate discrimination explaining the relative transport activities of these two molecules. We corroborate these results from unbiased MD using absolute binding free energy (ABFE) calculations. Cloxacillin, on the other hand, is very flexible in the binding pocket without picking up interactions with E622, thus establishing an

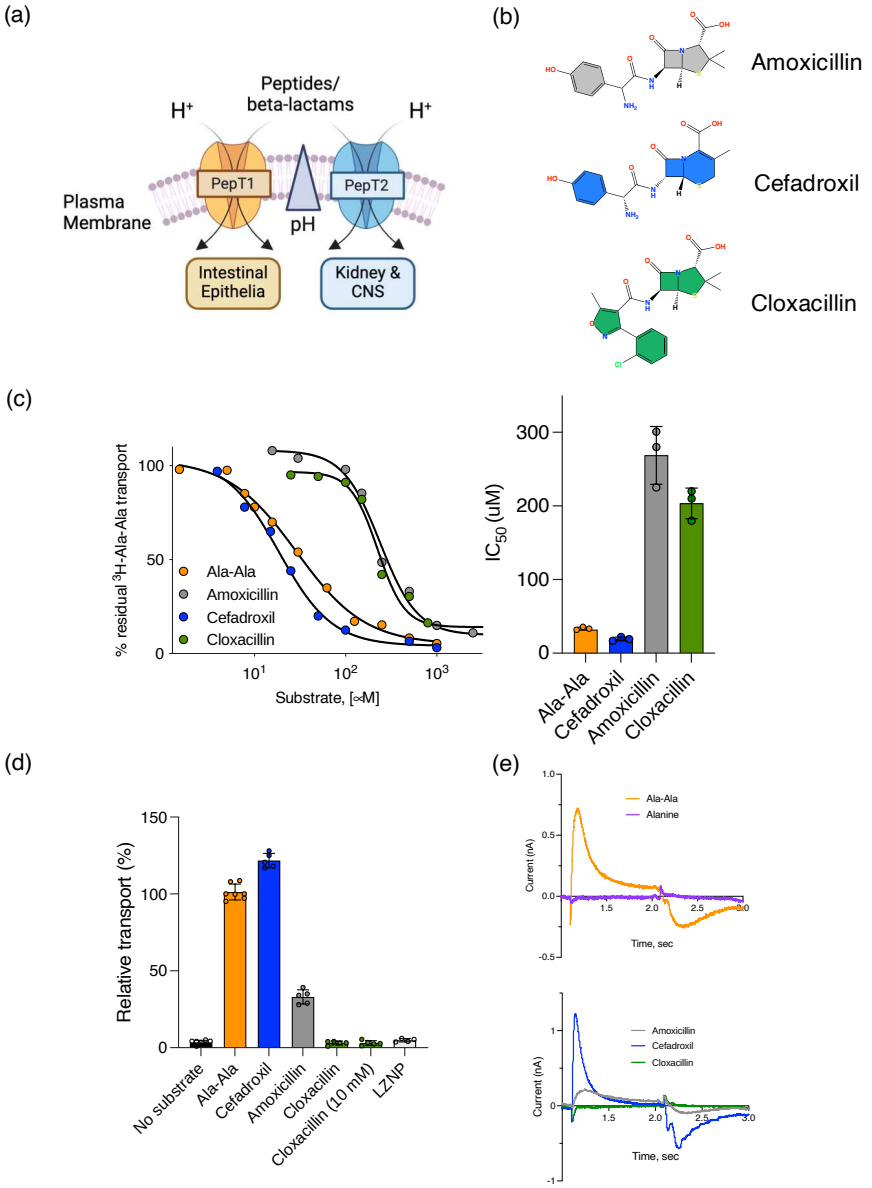


Figure 5.1: (a) Schematic overview of the physiological roles played by PepT1 and PepT2. (b) Chemical structures of the β-lactams cefadroxil, amoxicillin and cloxacillin. (c) Proteoliposome transport inhibition assays and the calculated IC_{50} values for the β-lactams and the Ala-Ala dipeptide. (d) Counterflow assay results, showing transport of Ala-Ala, cefadroxil and amoxicillin, but not cloxacillin or the known inhibitor LZNP. (e) Solid support membrane recordings showing capacitance changes with Ala-Ala, cefadroxil and amoxicillin, but not cloxacillin. Figure adapted from Parker et al. (2024) under the CC BY 4.0 license.

inhibitory mechanism by steric occlusion.

In the following sections, I will first briefly introduce the transport assay results and cryo-EM structures that were produced by the experimental collaborators. I will then, for the purpose of this thesis, predominantly focus on the MD investigation starting at these structures, which is my own work, and discuss the results in light of the available transport data.

5.1 Inhibition and transport assays of β-lactams

Affinities measured at PepT2 for a range of β-lactams have been reported in the literature, primarily as inhibition constants of reference peptide uptake (usually radiolabelled Gly-Sar) in cell-based assays. These range from high-affinity ligands (such as cefadroxil, $K_i = 3\,\mu M$) via those with intermediate affinity (for example amoxicillin, K_i =0.2–0.43 mM) to low-affinity binders (such as cloxacillin at $K_i = 1$ mM). (Luckner and Brandsch, 2005) We first set out to validate these K_i values and the relative classification of affinities by determining IC_{50} values for cefadroxil, amoxicillin and cloxacillin on uptake of radiolabelled Ala-Ala dipeptide via rat PepT2 reconstituted into liposomes. The results (see figure figure 5.1c and IC_{50} values summarised in table 5.1) confirm that cefadroxil binds to PepT2 as strongly as dipeptides, while amoxicillin and cloxacillin have weaker affinities. This assay can however not distinguish between substrates and mere inhibitors of transport, so we tested the compounds for transport activity in a counterflow assay, where uptake of ^{3}H-Ala-Ala is driven by a high concentration of the molecule to be tested on the inside of the liposomes (figure 5.1d). We found that while cefadroxil transports slightly better than Ala-Ala dipeptide (in line with the IC_{50} values) and amoxicillin transports at about 35 % of Ala-Ala activity, cloxacillin is inactive in this assay, even when raising its concentration 20-fold over the other molecules. The same observation is made for the known POT inhibitor LZNP. Therefore, cloxacillin, while having a comparable affinity at PepT2 to amoxicillin acts only as an inhibitor of transport, without itself getting transported. We also validated these results in a solid support membrane assay, where all of Ala-Ala, cefadroxil and amoxicillin (the latter at a lower level) produced noticeable capacitance changes, while cloxacillin did not (figure 5.1e).

Ligand	IC_{50} / μM
Cefadroxil	19 ± 3
Amoxicillin	270 ± 39
Cloxacillin	203 ± 21
Ala-Ala	32 ± 3

Table 5.1: IC_{50} values for cefadroxil, amoxicillin, cloxacillin and unlabelled Ala-Ala for comparison calculated for uptake inhibition of radiolabelled Ala-Ala substrate by rat PepT2 reconstituted into liposomes. Data is shown as mean $\pm$ standard deviation of triplicate experiments.

5.2 Cryo-EM structures of PepT2 bound to β-lactams

To investigate the molecular basis for these distinct recognition properties revealed in the *in-vitro* assays, we solved cryo-EM structures of rat PepT2 in complex with cefadroxil, amoxicillin and cloxacillin, using the same inhibitory nanobody as in a previous study. (Parker et al., 2021) PepT2 appeared in an outwards facing (OF) conformation very similar to the previously published structure (pdb: 7NQK), with Cα root mean squared deviation (RMSD) values of 0.618, 0.451 and 0.351 Å, respectively, although the extracellular domain (ECD) could not be modelled due to a lower resolution of that region. The observed electron density

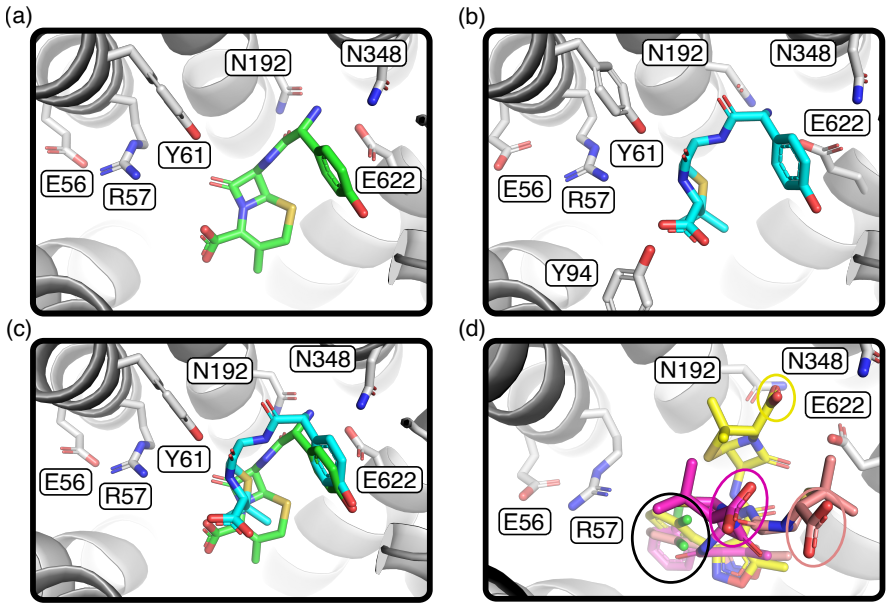


Figure 5.2: Cryo-EM binding poses of β-lactam binding to PepT2, with (a) cefadroxil, (b) amoxicillin, (c) overlay of the cefadroxil and amoxicillin poses and (d) cloxacillin in 3 distinct models fit to the electron density map, with fixed PheCl ring, black circle, but flexible β-lactam ring (including the C-terminus), circles coloured pink, salmon and yellow.

in the binding pocket allowed us to clearly model cefadroxil and amoxicillin in similar orientations, as shown in figure 5.2a–b, although the β-lactam ring of amoxicillin is rotated compared to cefadroxil (figure 5.2c). The drug amino groups in each case sit in the triad of E622, N192 and N348 that also coordinates the N-terminus of peptides. (Parker et al., 2021; Killer et al., 2021) Y61 sits near the peptide bond carbonyl group (for cefadroxil) and the β-lactam carbonyl group (for amoxicillin), respectively, similar to how it coordinates the backbone amide carbonyl in peptides, and the hydroxy-phenyl rings of cefadroxil and amoxicillin engage D317 further up in the binding pocket. Noticeably, however, the drug carboxyl group, equivalent in molecular topology to the C-terminus of peptides (and termed the drug C-terminus in this chapter for convenience), does not bind closely to R57 in our structures, merely sitting in the rough vicinity of the basic path of R57, K161 and K64 but located more centrally in the binding pocket. R57, which in previous structures of Ala-Phe binding was coordinating the peptide C-terminus, instead forms a salt-bridge with E56 in our drug-bound structures. We discuss this feature in detail below.

The cryo-EM dataset for cloxacillin, by contrast, revealed three distinct binding poses in two separate maps, see the models in figure 5.2d. The chlorophenyl group is roughly anchored to a position reaching towards the intracellular side of the binding pocket below the ExxER motif and close to a number of hydrophobic residues (including I191, which we use below to quantify the positioning of this ring across MD trajectories). Aside from this anchor, there is great conformational flexibility, and few specific interactions are made. Importantly, neither E622 nor R57 appear to be engaged by cloxacillin. The distinct and more flexible binding mode likely is a result of the absence of a free amino group that anchors peptides as well as cefadroxil and amoxicillin to E622.

5.3 MD simulations from β-lactam cryo-EM poses

Starting at the cryo-EM models of cefadroxil and amoxicillin as well as the three cloxacillin poses, we ran unbiased MD in 6 replicates of 1 µs each. Motivated by the results shown in chapter 4 with regard to the coupling between ExxER glutamate protonation and peptide binding, we also performed unbiased MD in the same setup but with E53 or E56 protonated (only E56 was probed for cloxacillin). More details on simulation setup can be found in appendix C. The following sections present our results regarding proton coupling of substrate binding and substrate discrimination for cefadroxil and amoxicillin, as well as the inhibitory mechanism of cloxacillin.

5.3.1 Coupling between ExxER motif protonation and drug binding pose

As figure 5.3a–b shows for individual replicate trajectories and figure 5.4 highlights as histograms of pooled replicates, both cefadroxil and amoxicillin, when the ExxER motif glutamates are deprotonated, have their N-termini stably bound to E622 while the C-termini are more flexible and do not approach R57 closely. This is in line with the interactions observed in the cryo-EM structures and previous results that emphasised the role of E622 in capturing the N-terminus of peptides as a first step to binding. (Parker et al., 2021) If, however, E56 or — to a lesser extent — E53 are protonated, R57 moves towards the center of the binding pocket and coordinates the drug C-terminus. In the case of cefadroxil, the drug remains stably attached at the N-terminus following E56 protonation (with some flexibility in the precise positioning of the N-terminus with respect to N192, see figures 5.3–5.4c). For amoxicillin, however, the amino group moves away from E622 following E56 protonation. Thus, unlike cefadroxil, amoxicillin cannot stably interact with both E622 and R57 through its N- and C-termini, respectively. We show this trend graphically as 2D-plane-projected histograms of important chemical group positions in figure 5.5.

We next sought to verify these observations from unbiased MD thermodynamically by performing ABFE calculations to determine the affinities of cefadroxil and amoxicillin in both E56 protonation states (see appendix C for details). The results are given in table 5.2 and show that when E56 is not protonated, cefadroxil and amoxicillin have similar binding affinities (in orientations that resemble the cryo-EM poses), with a slightly higher affinity for cefadroxil by $\approx 1.4\,\mathrm{kcal\,mol^{-1}}$. When E56 is protonated, however, cefadroxil displays a gain in affinity of $\approx 6\,\mathrm{kcal\,mol^{-1}}$, while the amoxicillin affinity decreases by $\approx 1.7\,\mathrm{kcal\,mol^{-1}}$. The tighter binding of cefadroxil following E56 protonation is explained by its ability to interact with both E622 and R57 simultaneously in this condition, while the decrease in affinity of amoxicillin is in line with the reduced stability of the N-terminus–E622 interaction following E56 protonation. [1]

5.3.2 Binding pose flexibility of the PepT2 inhibitor cloxacillin

Compared to cefadroxil and amoxicillin, the cryo-EM pose of cloxacillin displayed none of the peptide-specific interactions picked up by the other β-lactams, and showed large orientational flexibility, leading to three distinct binding poses observed in the electron density maps. Our unbiased MD simulations confirm this point (figure 5.6 for individual trajectories and figure 5.7 for a histogram analysis of pooled replicates). The cloxacillin binding mode in our unbiased MD runs was very dynamic, as illustrated by only occasional interactions of the drug C-terminus with R57, without a discernible increase following

[1] Note that in this chapter, we are using the ΔG values obtained from ABFE calculations only qualitatively by comparing different substrates and showing changes in ΔG triggered by protonation-state changes. For a discussion of how these values can be compared to experiments, see section 6.2.4.

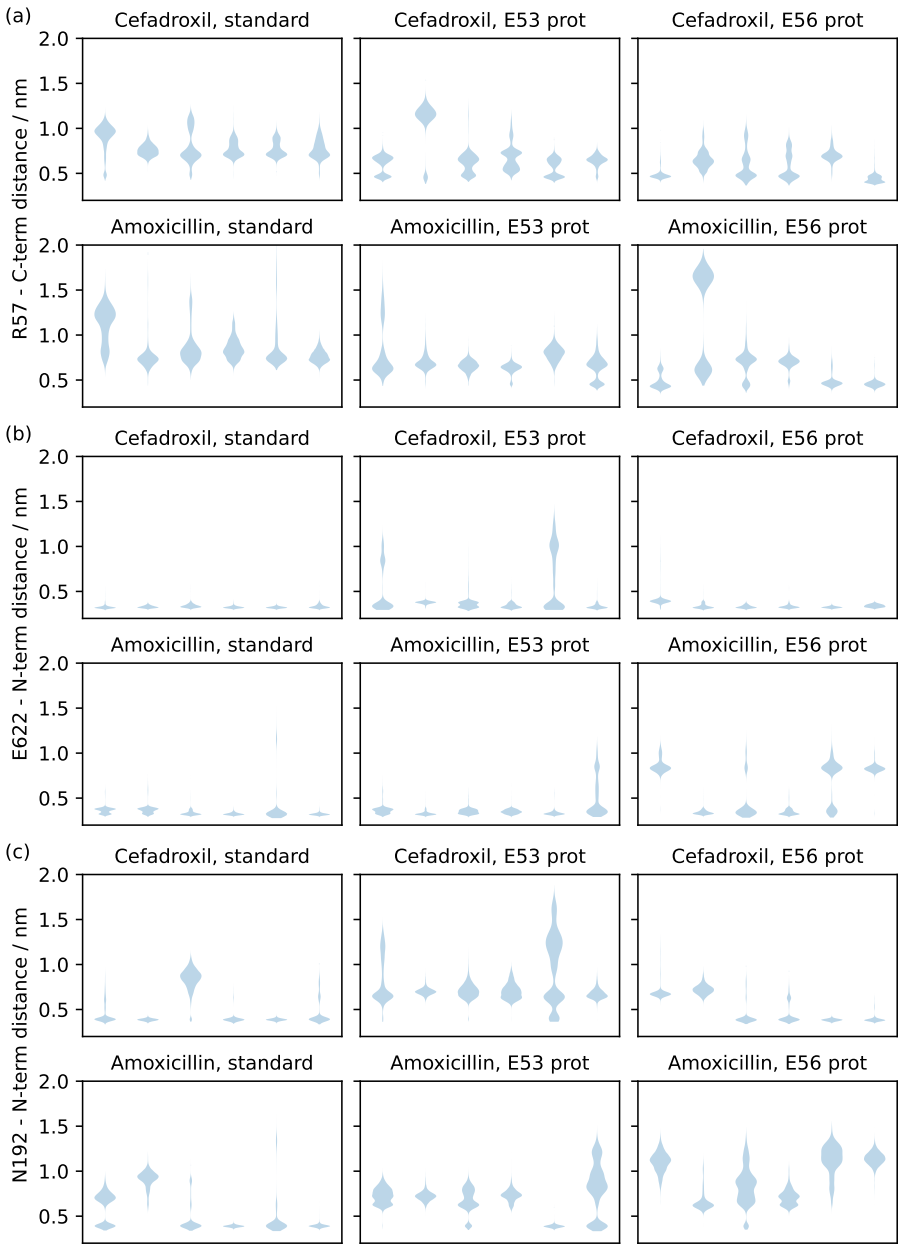


Figure 5.3: Violin plots showing replicate histograms of key ligand–protein distances calculated across unbiased 1 µs MD trajectories, for standard protonation states, E53 protonated and E56 protonated. (a) Drug carboxyl (central carbon) – R57 (Cζ) distance, (b) drug N-terminus (amino N) – E622 (Cδ) distance, and (c) drug N-terminus (amino N) – N192 (Cγ) distance.

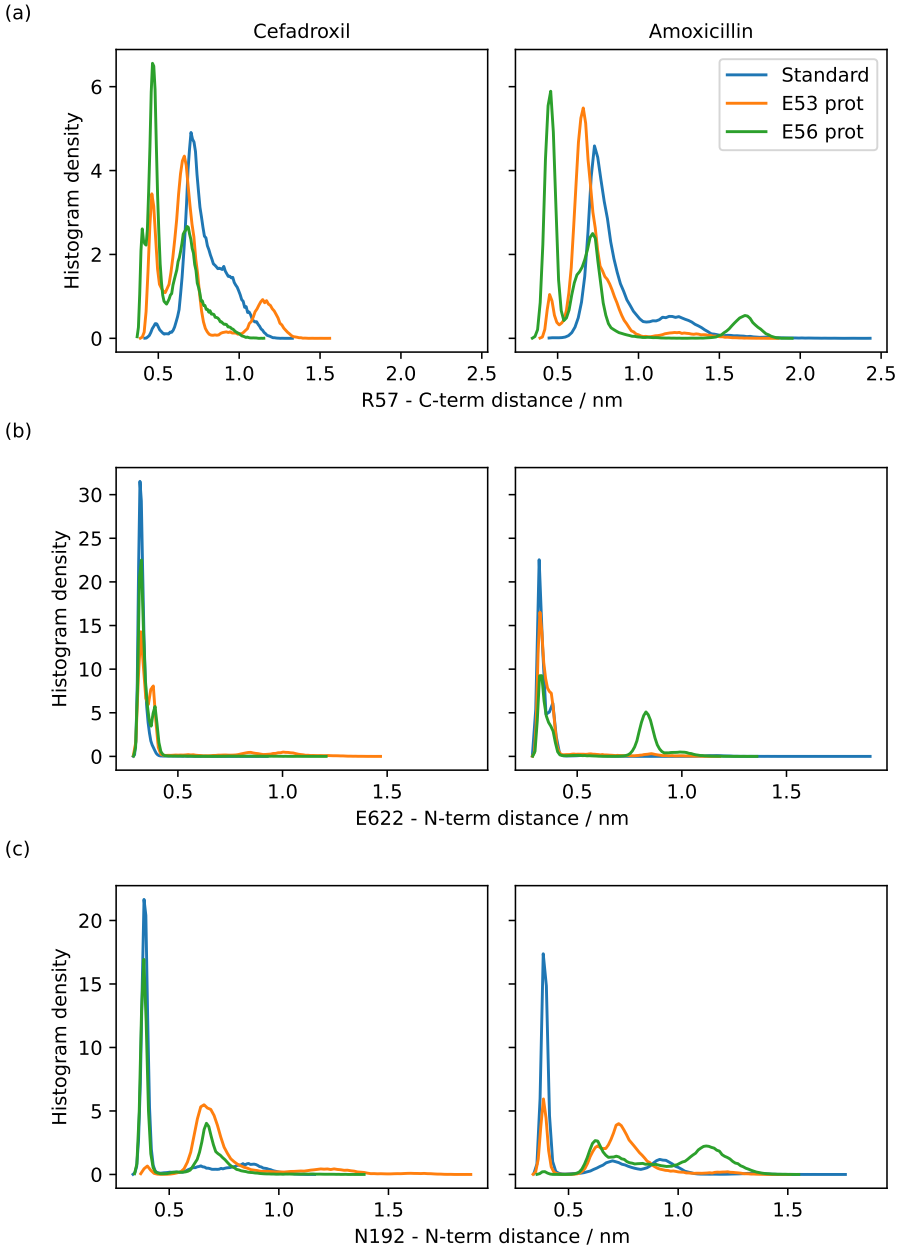


Figure 5.4: Histograms of key ligand–protein distances pooled across unbiased 1 µs MD trajectories in 6 replicates, as shown individually in figure 5.3, for standard protonation states, E53 protonated and E56 protonated. (a) Drug carboxyl (central carbon) – R57 (Cζ) distance, (b) drug N-terminus (amino N) – E622 (Cδ) distance, and (c) drug N-terminus (amino N) – N192 (Cγ) distance.

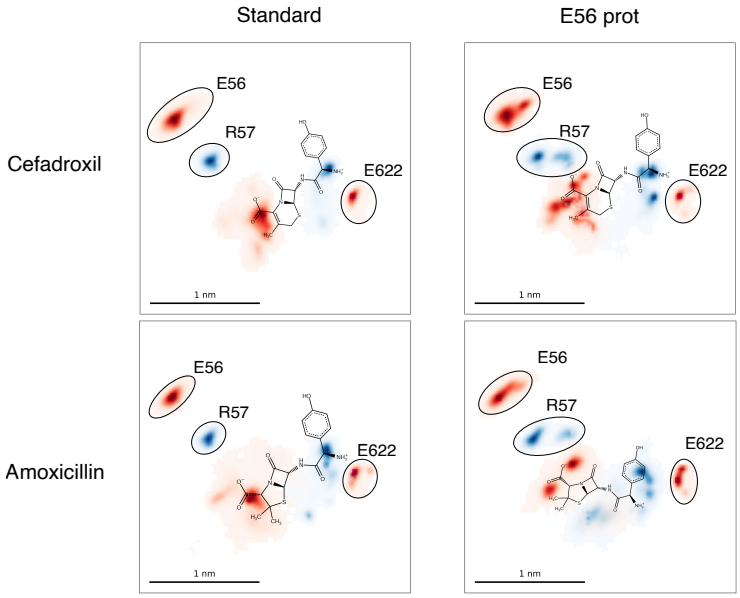


Figure 5.5: Pooled replicate trajectories (same trajectories as figures 5.3 and 5.4) projected onto the plane spanned by the Cα atoms of E622, R57 and W313. Positions of E622, R57, the drug N- and C-terminus (all as defined in the captions to the figures above) and E56 (Cδ) are shown as 2D histograms in the plane. Protein residue densities were surrounded with black ovals, with the chemical structures of cefadroxil and amoxicillin overlaid for illustration.

Ligand	E56	-ΔG / kcal mol^{-1}
Cefadroxil	not protonated	12.0 ± 1.5
	protonated	18.0 ± 0.5
Amoxicillin	not protonated	10.6 ± 1.7
	protonated	8.9 ± 1.1
Cloxacillin	not protonated	≈ 6.8
	protonated	≈ 5.5

Table 5.2: Affinities of cefadroxil, amoxicillin and cloxacillin at PepT2, evaluated using ABFE calculations in different protonation states of E56. Tabulated values are mean ± standard deviation for cefadroxil and amoxicillin from 5 replicates started from the unbiased MD replicate with the highest ABFE affinity, while for cloxacillin we report just the highest ABFE affinity among the unbiased MD-equilibrated replicate poses. See appendix C for details.

E56 protonation (figure panels 5.6–5.7a), and the unstable chlorophenyl ring positioning (panels 5.6–5.7b). Our ABFE data underlines the looser nature of the cloxacillin binding mode (table 5.2) by showing an affinity $\approx 3.8\,\text{kcal mol}^{-1}$ lower than for either cefadroxil or amoxicillin, and without stabilisation from E56 protonation.

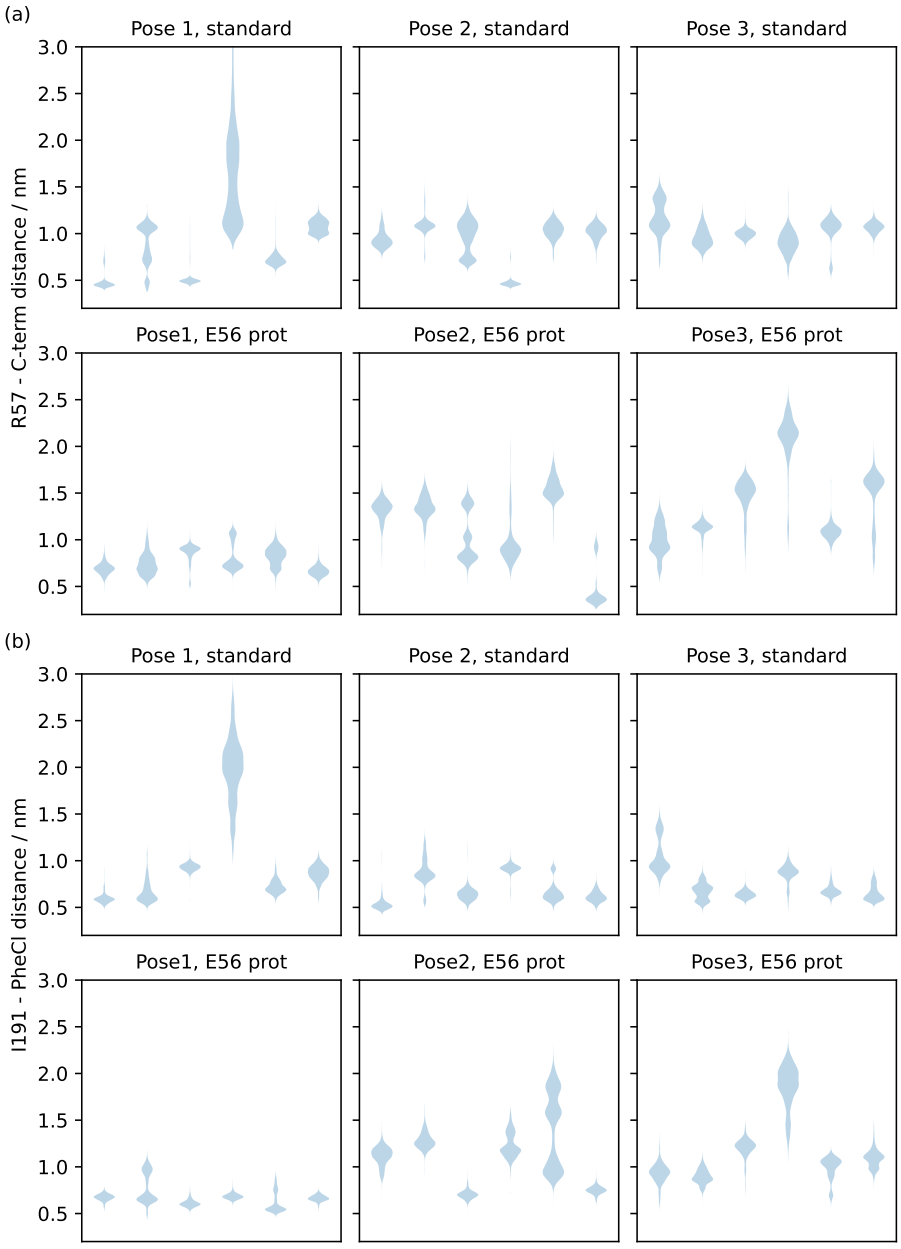


Figure 5.6: Microsecond-long unbiased molecular dynamics (MD) simulations starting at the 3 cryo-EM models of cloxacillin, using 6 replicates for each model and condition (standard protonation states, E56 protonated), shown as violin-plot histograms of individual replicate trajectories. (a) Drug carboxyl carbon – R57 (Cζ) distance, (b) drug chlorophenyl - I191 distance (calculated as the centre-of-mass distance between the I191 side-chain atoms and the heavy atoms of the chlorophenyl ring).

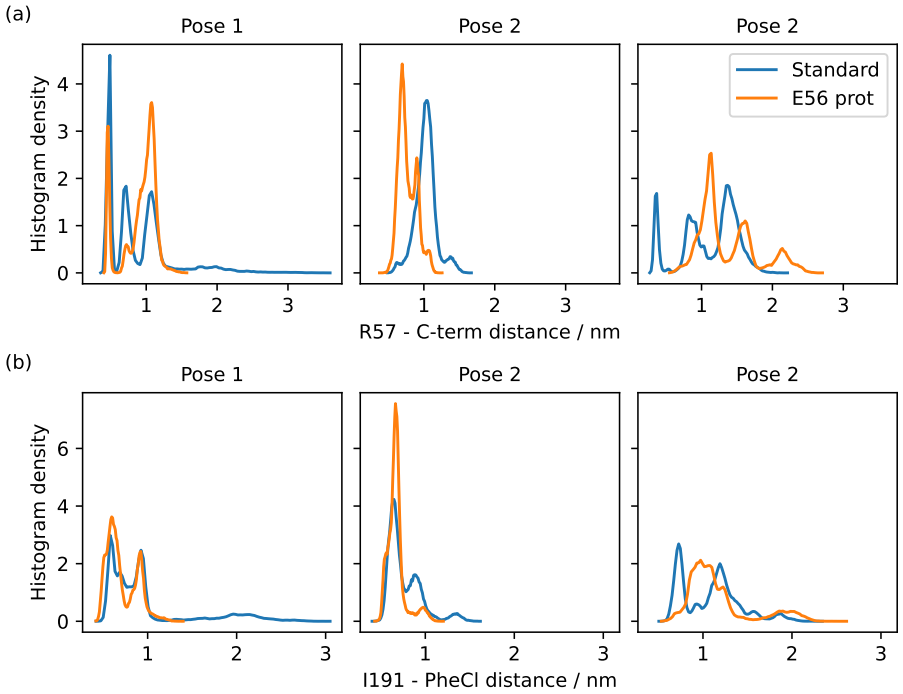


Figure 5.7: Microsecond-long unbiased molecular dynamics (MD) simulations starting at the 3 cryo-EM models of cloxacillin, using 6 replicates for each model and condition (standard protonation states, E56 protonated). Data is shown pooled for each pose / condition. (a) Drug carboxyl carbon – R57 (Cζ) distance, (b) drug chlorophenyl - I191 distance (calculated as the centre-of-mass distance between the I191 side-chain atoms and the heavy atoms of the chlorophenyl ring).

5.4 Discussion

The results of our study provide a structural basis for understanding β-lactam recognition by POTs. Cefadroxil, as the best substrate tested, adopts a binding pose similar to dipeptides, while the lower affinity of amoxicillin can be explained by a rotated position in the binding pocket. MD simulations confirm the overall stability of these binding modes, with a slightly higher affinity estimated in ABFE calculations for cefadroxil in line with the experiments. Notably, unlike previous peptide-bound structures (Killer et al., 2021), our cryo-EM structures do not show close contacts between the drug C-termini and R57 in the ExxER motif — an observation that is reproduced in MD simulations with standard protonation states for the ExxER glutamates. When we assigned E56 as protonated, however, we observed a salt-bridge swap in unbiased MD, where R57 moved away from E56 and towards the center of the binding pocket to engage the drug C-terminus. The cryo-EM structures therefore appear to represent a trapped intermediate state of the transport cycle, at an intermediate position of substrate binding prior to conformational rearrangements (which require proton movement to the ExxER motif, as shown in chapter 4). This provides a further layer of evidence for the model of substrate-coupled proton movement presented in section 4.5. We suggest that proton transfer from H87 to the ExxER glutamates (which drives inwards-directed conformational changes) is regulated by the presence of the substrate via a swap in coordination of the E56–R57–substrate salt bridge triad (E56–R57 …substrate → E56P …R57-substrate). Aside from the structural evidence the cryo-EM poses constitute for an intermediate substrate binding pose that does not interact with R57, the ABFE results presented in table 5.2 for cefadroxil correspond directly to and corroborate those in shown in table 4.1 for Ala-Phe.

Moreover, we found that cefadroxil can engage in this coupling mechanism more effectively than amoxicillin, as indicated by the stability of a cefadroxil pose simultaneously engaging E622 and R57 in unbiased MD (when E56 is protonated), whereas the amoxicillin N-terminus is unstable following E56 protonation. The gain in ABFE affinity of cefadroxil upon E56 protonation reiterates this point in contrast with the loss of affinity for amoxicillin through the same protonation-state change. In terms of molecular topology, this difference appears to be due to the changed β-lactam core between the two drugs, where the C-terminus of cefadroxil is attached to an sp^2-hybridised carbon in a 6-membered ring as opposed to an sp^3 carbon in a 5-membered ring for amoxicillin. The consequently altered steric preferences of the drugs determine their ability to engage in the substrate–proton coupling mechanism, thus rationalising the experimental results showing ≈ 3-fold higher transport for cefadroxil than amoxicillin. It is, in short, not the binding free energy to a particular conformational state that is predictive of drug transport, but the ability of the drug to participate in the dynamic transport mechanism. Our results may thereby provide a mechanistic rationale for the observations of Kotov et al. (2023), who found that medium-affinity peptides are transported best by a bacterial POT homologue (see section 1.1.4 for details). We also provide a step towards a qualitative β-lactam pharmacophore, by elucidating a geometric requirement for the orientation of the N- and C-termini in the drug molecule. While in theory this could make for a testable hypothesis regarding β-lactams we have not studied here, ascertaining which ligand conformations are available to substrates in the POT binding pocket is a complex task with many pitfalls in the absence of structural data, as the work in chapter 6 of this thesis demonstrates.

Lastly, we have shown that cloxacillin, a β-lactam that does not have a free amino group analogous to a peptide N-terminus, is an inhibitor but not a substrate of PepT2. Considering the binding pose flexibility exhibited in cryo-EM and MD simulations, we conclude that cloxacillin inhibits transport by PepT2 via steric occlusion. While it can sit in the binding pocket with some stabilisation with respect to leaving the pocket, its binding mode is highly dynamic without pick up the necessary interactions (with R57 and E622) connecting the N- and C-terminal bundles to become a transported substrate (which are picked up by cefadroxil, and less stably by amoxicillin). Since it nonetheless sterically takes up space required for the binding of substrates, it acts as an inhibitor.

6

Retrospective and predictive modelling in peptide transporter prodrug design

The design of new drugs and prodrugs as substrates of proton-coupled oligopeptide transporters (POTs) is of great pharmaceutical interest for enhancing oral drug delivery, as we have reviewed in section 1.2 of this thesis. As discussed therein, previous efforts in this direction have yielded promising results, although attempts at computer-aided predictive drug and prodrug design targeting peptide transporters are still sparse. Notably, the need for accurate structural models has been highlighted by Samsudin et al. (2016) in their study employing molecular dynamics (MD)-based free energy calculation methods to predict POT substrate affinities, and the work of Kotov et al. (2023) has shown the difficulties with obtaining accurate affinity predictions from molecular docking.

In chapters 4 and 5 of this thesis, we have demonstrated how the use of absolute binding free energy (ABFE) calculations can provide insight into the mechanisms of substrate selectivity for PepT2. In particular, we have shown how computing the binding affinity of substrate and inhibitor molecules in different protein conformational states and protonation-state conditions of the key E56 residue allows conclusions to be drawn concerning the role of the E56–R57 saltbridge motif in responding to the incoming substrate, and how this mechanism discriminates between good and bad substrates. Earlier work from our group (Parker et al., 2021) has also shown that good correlations between experimentally determined K_i values (for inhibition of transport of a reference peptide) and ABFE predictions can be obtained without available structurally resolved ligand binding poses, although that study included only a few peptide substrates. The intriguing upshot of such successful applications of binding free energy methods to research questions concerning POT substrates is that predictive modelling of substrate binding to POTs may be within reach of current computational methodologies, with potentially significant applications in the drug discovery pipeline.

In this chapter, we sought to explore this possibility in more depth. We benchmarked retrospective affinity prediction using molecular docking, ABFE calculations and machine learning with random forest models on the largest available dataset of PepT2 affinities (184 compounds, Biegel et al. (2006), see appendix D for details). Our results indicate only low correlations (Pearson r = 0.431, Spearman ρ = 0.433) between docking scores and experimental affinities, and any correlation disappears for the subset of β-lactam antibiotics. The situation is not improved by calculating affinities using the much more expensive ABFE method (r = 0.30, ρ = 0.26). By considering a direct comparison with the ABFE data presented in chapter 5 concerning the antibiotic binding poses determined by cryo electron microscopy (cryo-EM), we conclude that one of the factors currently limiting accuracy is the prediction of accurate binding poses for highly flexible ligands in a dynamic binding site that may receive the substrate with substantial induced fit. Furthermore, the nature of the experimental data may reduce what correlation can be obtained, as inhibition constants capture binding to the full ensemble of protein states while the ABFE is evaluated only against one protein conformation. Random forest models trained on the ligand structures alone were able to give much better correlations with experimental data (r = 0.81, ρ = 0.82), although activity cliffs in

structurally similar molecules are systematically underpredicted and the method does not transfer well to chemical space not included in the training dataset.

Lastly, moving towards prospective prediction, we attempted to design a penicillin G prodrug derivative to target POTs for transport. After having partnered with collaborators for a synthesis of the new compound, we unfortunately encountered problems with solubility that precluded testing in the available PepT2 transport assays. Further attempts with modified chemical structures were beyond the scope of this doctoral research project, although we are optimistic that the method may bear fruit in the future.

The computational work and solubility characterisation presented in this chapter were performed by myself, the synthesis of the modified penicillin G was devised and carried out by Dorian Munțean in the group of Chris Schofield (Department of Chemistry, University of Oxford).

6.1 Two methodological choices

Before we could move to affinity predictions using MD techniques, we had to clarify two methodological choices: which force fields to use for modelling the substrates, and how to treat charge transformations in ABFE simulations.

6.1.1 Force fields for modelling substrates

As we have detailed in section 2.1.3 of this thesis, several possibilities were available for representing substrates in our simulations. Peptides could be parameterised using the AMBER 14.sb force field (Maier et al., 2015) as well as the GAFF2 (Wang et al., 2004) or OpenFF 2.0 (SAGE) (Boothroyd et al., 2023) small-molecule force fields. In chapter 4, we have chosen to model peptides using the AMBER force field for consistency with previously published work from our group. (Parker et al., 2021) The dataset studied in this chapter, however, contains both peptides comprised of canonical amino acids and chemically modified peptides, as well as other drug molecules, thus making the use of a small-molecule force field necessary throughout (including for peptides, to obtain consistent results).

During initial trials with the GAFF2 force field, we noticed inconsistent conformational preferences of peptide substrates compared to AMBER. We traced the issue to the peptide backbone dihedral conformational landscape, determined using well-tempered metadynamics (WT-metaD) as free energy surfaces for Ala-Ala dipeptide in figure 6.1. [1] While AMBER and SAGE produce similar profiles, the locations of minima and barriers as well as the barrier heights are significantly changed with GAFF2, compared to the consensus of the other two force fields. We therefore chose to use SAGE for all ligands in this chapter as the most compatible with our previous simulations, and because we suspected based on these results (although we have not determined the quantum-mechanical ground truth) that GAFF2 may be assigning inaccurate dihedral parameters for peptides.

[1] Similar results could have been obtained, traditionally, with torsion scans, although they would need to be performed in two dimensions in this case, given the coupling between the two dihedrals. Since the molecules were small enough and no collective variable (CV) problems prevented convergence on the accessible time scale here, we chose to compute the full 2D free energy profiles for the comparison.

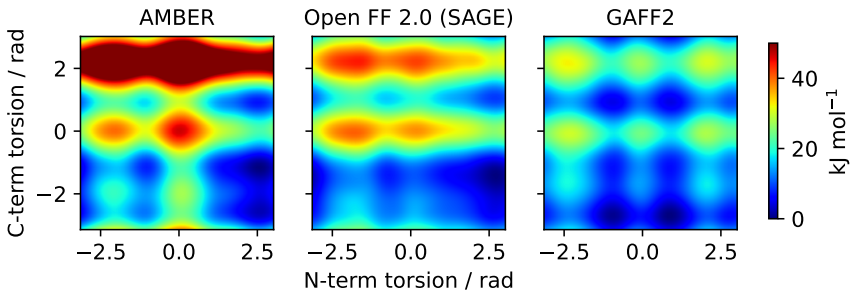


Figure 6.1: Comparison of the backbone free energy landscapes of Ala-Ala dipeptide in solution, evaluated via 50 ns of WT-metaD, for the AMBER 14.sb, Open FF 2.0 (SAGE) and GAFF2 force fields. N-term torsion is the dihedral of atoms N(1)–Cα(1)–C(1)–N(2), C-term torsion of atoms C(1)–N(2)–Cα(2)–C(2). Numbers in brackets indicate dipeptide residues.

6.1.2 Multivalent co-ions for alchemical free energy calculations

A second question that arose in the work for this chapter concerns the handling of charge transformations in the ABFE simulations. As explained in section 2.3, the long-range electrostatic treatment using the particle-mesh Ewald (PME) method requires a neutral box, and the correction scheme proposed by Rocklin et al. (2013) for the artifacts introduced by changing the charge of a ligand during an alchemical free energy calculation cannot be readily applied to membrane proteins. (Wu and Biggin, 2022) We therefore adopt the suggestion of Wu and Biggin (2022) to use charge transfer to a restrained alchemical co-ion to keep a neutral box. The exact nature of this ion does not matter for the thermodynamic cycle, and so for the treatment of multiply charged ligands several options arise. One may (unphysically) scale up the charge of the alchemical ion to match the ligand charge, use physical ions of matching charges, or use multiple alchemical ions places appropriately far apart. We trialled all approaches for the case of a double-negatively charged ligand (as it occurs repeatedly in our dataset). Our results (figure 6.2) show that the hypothetical Cl^{2-} with van-der-Waals parameters taken from regular Cl^- cannot be used because it stably and tightly associates with 4 Na^+ ions from the solution over a timescale of ≈ 10 ns, leading to artifacts in the free energy estimation. A sulfate (SO_4^{2-}) ion with available literature parameters (Kashefolgheta and Vila Verde, 2017), on the other hand, does not show this pathology, and neither does the combination of two Cl^- ions when spaced more than ≈ 8 Å apart. Although alchemical co-ion parameters would have been available for all ligand charges studied in this chapter and this treatment would thus have been possible here, we opted for the more generalisable approach of using multiple alchemical ions, to allow for consistency with potential future work.

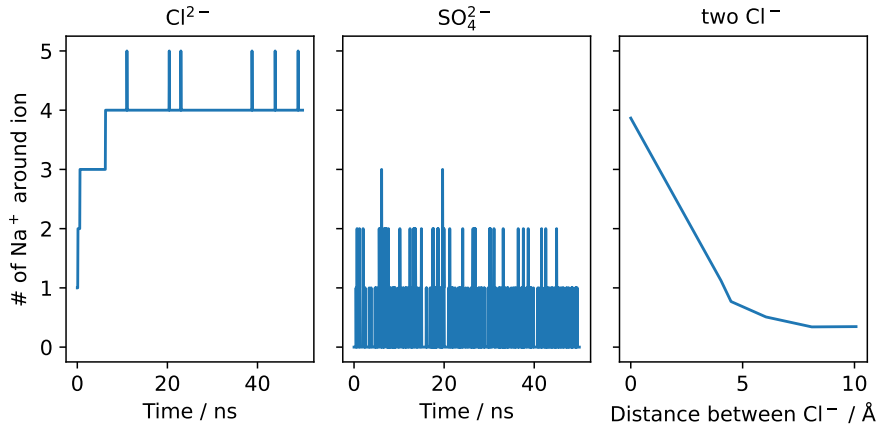


Figure 6.2: Aggregation of Na^+ counterions around double-negatively charged ions (expressed as number of ions found within 5 Å of the negative ion) trialled as co-ions for alchemical calculations. In 50 ns-long MD simulations, 4 Na^+ ions stuck to the unphysical Cl^{2-}, while no such aggregation was observed around SO_4^{2-}. Using two Cl^- ions solves the aggregation problem if they are more than ≈ 8 Å apart (the data point at zero distance corresponds to using a single Cl^{2-}).

6.2 Retrospective affinity prediction

A first step in the development of a predictive model for a biophysical property like small-molecule affinity to a target is the retrospective prediction of a set of known affinities. In the cases of molecular docking and ABFE calculations below, this serves as a benchmark set to evaluate the quality of the predictions, whereas in the case of machine-learning methods, random subsets of the experimental dataset form the training data. For substrate binding to PepT1 and PepT2, the most extensive collection of substrate affinity data can be found in Biegel et al. (2006), where the dataset of K_i values for [^{14}C]Gly-Sar uptake inhibition at SKPT cells was most comprehensive at 184 molecules, encompassing di- and tripeptides with natural and modified amino acids and backbones, as well as a range of drug molecules (see appendix D). Importantly, we note that while the assay protocol is the same, the measurements in the dataset were carried out across a number of studies that we pool here for a maximally sized dataset. We touch again on limitations of the dataset in section 6.2.4; for now we adopt it as the best available option at the time of writing.

6.2.1 Molecular docking

Of the computational approaches commonly employed to predict small-molecule binding poses and affinities to proteins, molecular docking stands out as inexpensive, fast and easy to use. With wide applications in the drug discovery field, docking is particularly suited (but not limited) to 'virtual screens' of large libraries of compounds to a given protein target. (Pinzi and Rastelli, 2019; Bender et al., 2021) While traditionally docking is performed as fitting a flexible ligand into a rigid protein (sometimes incorporating local flexibility in selected protein side-chain orientations) and then evaluating affinity using an empirical scoring function, approaches have been devised to incorporate enhanced protein flexibility. Ensemble docking, for instance, where an ensemble of receptor conformations is generated using MD simulations, allows for conformational selection of ligand binding and has been widely used where protein flexibility is key to ligand binding. (Amaro et al., 2018) A number of packages to perform docking have been developed over the last three decades, but a review of the available methods and implementations is beyond the scope of thesis chapter. Here, we use the GOLD docking software, which employs a genetic algorithm for

searching the conformational space accessible to the combined flexible ligand and selected protein side chains. (Jones et al., 1997) For GOLD docking scores, a higher numerical value corresponds to a tighter predicted affinity.

To maximise conformational exploration within our docking protocol (following initial unsuccessful trials without protein and / or ligand flexibility), we opted for treating key side chains known to coordinate POT substrates (R57, Y61, Y94, D317 and E622) as fully flexible, and used an ensemble approach using a total of 45 protein conformations from the ensemble of triplicate 1 µs runs performed in chapters 4–5 in the apo, Ala-Phe and cefadroxil bound states (see appendix D for details). Figure 6.4a shows the maximum obtained docking scores plotted against the experimental pK_i values, displaying only a very modest correlation with a Pearson correlation coefficient of r = 0.35 and a Spearman rank correlation coefficient of ρ = 0.34. A visual inspection of several high affinity docking poses revealed that occasionally the key canonical interactions with E622 (and N192 / N348) in the acidic patch and R57 in the basic patch (see the discussion in the introductory section 1.1.4) are not formed in the highest scoring docked pose, despite the molecule in question having an amino terminus and / or a carboxyl group (an example for this was the Trp-Trp-Trp tripeptide, see figure 6.3). Given our results in chapters 4–5, which demonstrate the importance of these interactions in engaging with the POT transport mechanism, we decided to explore whether such high-affinity poses may be artifacts that — while plausible in docking — are unlikely to contribute to substrate recognition by POTs. To this end, we filtered the docked poses by these canonical contacts, where they are possible in the presence of appropriate charged chemical groups in the ligands (see appendix D for details). This procedure significantly reduced the predicted affinities of several outliers (including the Trp-Trp-Trp example from 136.1 to 96.3), leading to a somewhat improved correlation with experimental data at r = 0.431 and ρ = 0.433, prompting us to use the filtered poses and scores in the rest of this work. We also sought to explore which types of ligands are predicted the best, and whether docking scores provide any information on drug binding to POTs, as would be desirable for applications in the design of orally bioavailable medications. We found (see figure 6.4c–d) that the group of dipeptides with natural amino acids has the best correlation with experimental affinities (r = 0.45, ρ = 0.46), while tripeptides correlate less well (r = 0.28, ρ = 0.25). The set of β-lactam drugs, on the other hand, displayed no correlation whatsoever between docking scores and the experimental data (r = -0.14, ρ = -0.22), indicating that docking by itself is not likely to be useful in predicting the affinities of molecules with diverse chemical topologies to POTs.

6.2.2 Absolute binding free energy calculations

As introduced in section 2.3 and applied in chapters 4 and 5 of this thesis, alchemical binding free energy calculations provide a thermodynamically rigorous way of estimating binding affinities between proteins and small molecules, limited by the accuracy of the classical force field representation and the accessible sampling. Compared to molecular docking, they are orders of magnitude slower, but incorporate flexibility of the protein, explicit solvent molecules and entropic contributions to binding from capturing dynamic information. Nonetheless, they require an initial pose of the ligand within the protein, because sampling the full binding process is usually unfeasible on timescales accessible to unbiased MD. Where structural information is available, such as the cryo-EM poses in chapters 4 and 5, this can be used, but otherwise one usually has to resort to molecular docking for identifying a binding pose that is then equilibrated in MD and scored with ABFE.

Since for the vast majority of substrates in this chapter no binding poses have been structurally

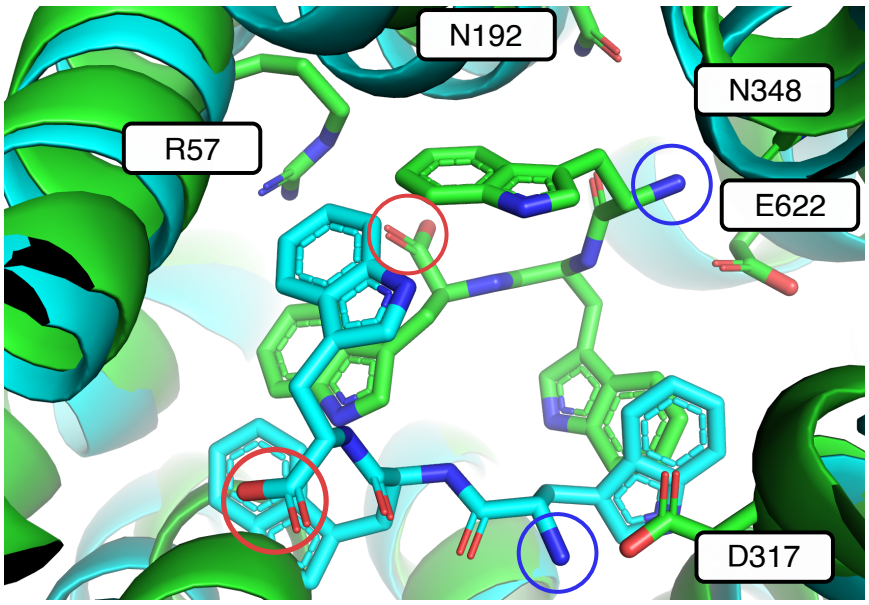


Figure 6.3: Overlay of highest-scoring overall and filtered docking poses of the Trp-Trp-Trp tripeptide. The overall highest-scoring docked pose (docking score 137.1) is shown in blue, the highest scoring docked pose among the poses with the highest filtering score (docking score 96.3) in green. The N- and C-termini of the tripeptide are highlighted with blue and red circles, respectively. While the highest-scoring filtered pose has its termini positioned close to the canonical E622/N192/N348 and R57 protein residues, the overall highest-scoring pose does not adopt this orientation. Instead, it locates the N-terminus at D317 and is stabilised by a range of hydrophobic contacts with the walls of the binding pocket.

resolved, we started 50 ns-long unbiased MD equilibrations from the filtered highest-scoring docked poses, as presented above. From these boxes, we set up ABFE calculations in single replicates (for computational efficiency, see appendix D). In figure 6.5a, we plot the ABFE results as ΔG_{bind} against the corresponding experimental value calculated as RT $\ln(K_i)$ (see section 6.2.4 for a discussion of this conversion). The computed binding affinity values present themselves as very noisy, and correlation with experimental data is poor and not improved compared to docking (r = 0.30, ρ = 0.26). There is also only limited correlation with the docking scores of the initial poses the ABFE runs were based on (figure 6.5b, r = 0.46, ρ = 0.44), showing that the ABFE results are neither predictive of the experimental data, nor do they simply reproduce the docking data. When considering, as above, the ligand classes of di- and tripeptides and β-lactams, correlation is best for the tripeptides (r = 0.43, ρ = 0.45), degraded for dipeptides (r = 0.20, ρ = 0.15) and absent for the drug molecules (r = 0.06, ρ = 0.06; figure 6.5 c–d). While this ordering is somewhat different from the one obtained with docking alone (where dipeptides slightly outperformed tripeptides), the low correlations in all classes of molecules led us to not investigate this difference between the scoring methods further but instead focus on the sources of the overall poor correlation. We conclude therefore that a combination of molecular docking (as described above) with ABFE for scoring the poses is not more accurate for retrospective affinity predictions at PepT2 than docking alone. This hints at the identification of good protein–ligand binding poses as a limiting factor, as we discuss in more detail in section 6.2.4.

6.2.3 Machine learning affinity prediction

Both docking and ABFE calculations operate on the protein–ligand complex, using structural information of the ligand binding site and the (likely) orientation of a ligand therein to predict a binding free energy. Machine learning (ML) approaches have also been applied to the structure-based affinity prediction problem

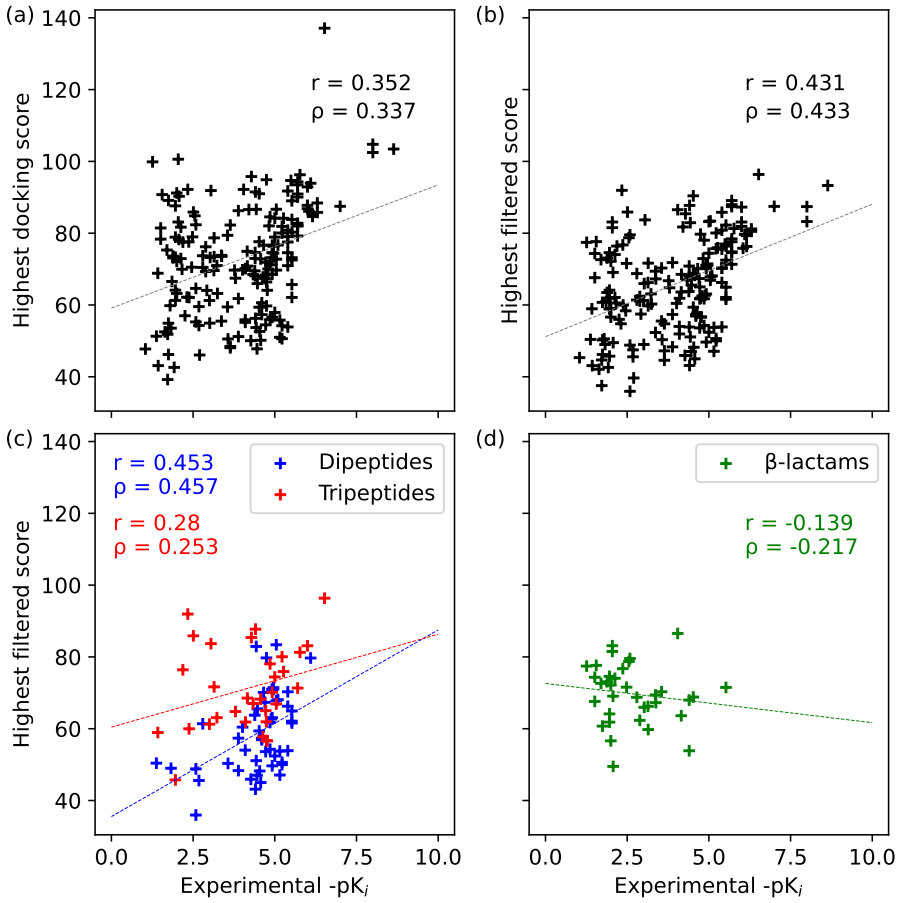


Figure 6.4: GOLD ChemPLP docking scores (flexible docking to an ensemble of apo and holo protein conformations, see appendix D) plotted against experimental pK_i values of reference peptide transport by PepT2 measured at SKPT cells (Biegel et al., 2006), showing (a) the highest docking score across all poses, or (b) the docking score of poses filtered by contacts to the canonical residues R57, N192, N348 and E622 (used in the following panels). (c) and (d) highlight the how the best correlation is achieved for dipeptides, which degrades for tripeptides, and is lost completely for β-lactams.

thus formulated, for example in the development of ML-based scoring functions for docking. (Meli et al., 2022) Since we reasoned above (and discuss in more detail below) that finding appropriate ligand binding poses may be the factor limiting prediction accuracy in the case of PepT2 substrates, we did not pursue such directions. Instead, we decided to explore whether ML models trained within a cheminformatics paradigm on ligand-only features could yield accurate affinity predictions for PepT2 with the available dataset.

A host of ML methods have been employed in the cheminformatics field. (Lo et al., 2018; Rodríguez-Pérez et al., 2022) Here, we focus on a simple yet powerful model architecture: random forest regressors. (Breiman, 2001; Svetnik et al., 2003) In short, a random forest consists of an ensemble of decision trees, where the predictions from all trees get averaged — a procedure that helps to reduce overfitting which is otherwise prevalent for single decision trees. Here, we trained random forest regressors on three sets of features: (1) 64 low-dimensional molecular descriptors as computed by the rdkit package, (2) 1024-bit extended-connectivity fingerprints (commonly known as Morgan fingerprints), and (3) 2D pharmacophore fingerprints. The molecular descriptors include all single-valued descriptors that rdkit can generate, ranging, among others, from mass and number of hydrogen-bond donor and acceptor atoms, as

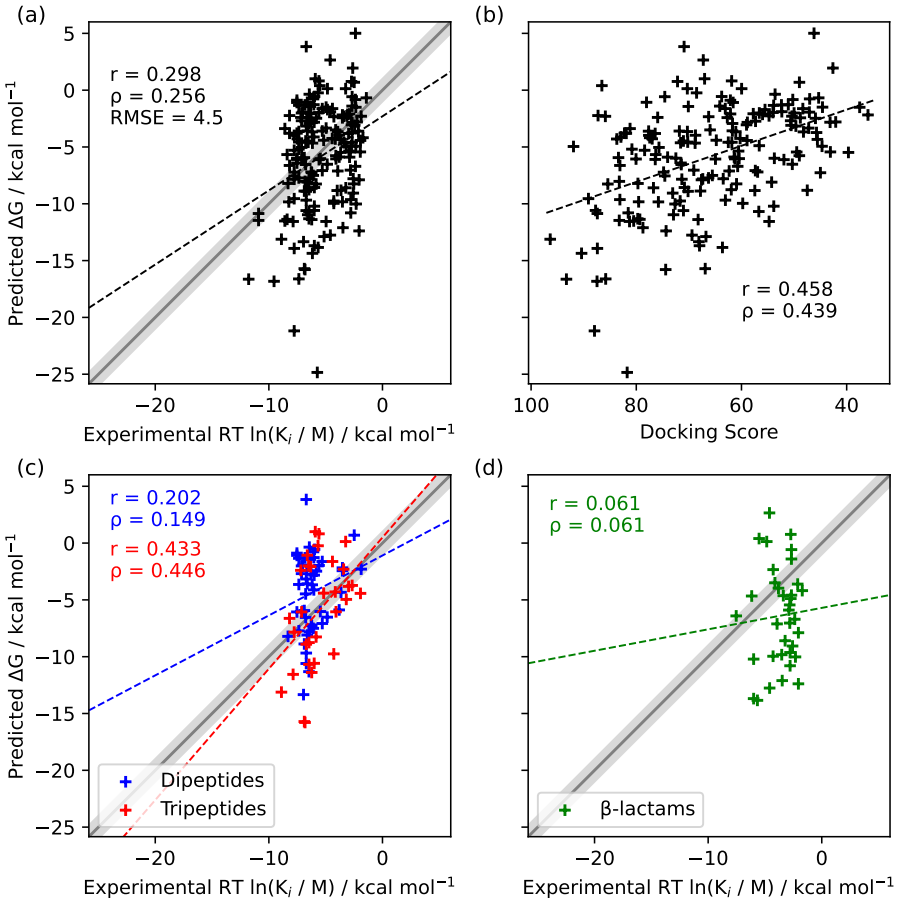


Figure 6.5: Results of ABFE calculations in the PepT2 binding site, starting from MD-equilibrated docked poses (see appendix D), compared with SKPT reference peptide inhibition RT ln(K_i) values from Biegel et al. (2006). (a) Full dataset of 180 ligands, after 4 our of 184 simulations failed to give stable Boresch restraints and have been excluded, (b) comparison of ABFE values to docking scores of the initial poses, (c) subsets of di- and tripeptides, and (d) subset of β-lactams.

well as other categories of functional groups, to predictors of physicochemical properties such as logP and molecular topology order parameters. We have used the list of descriptors as curated by Warren et al. (2023). Morgan fingerprints are calculated by deriving identifiers for atoms present in a molecule that are then hashed onto a bit vector of given length. (Rogers and Hahn, 2010) Following initial assignment of integer identifiers based just on properties of an atom within the molecular topology that do not depend on numbering (atomic number, valence, number of neighbours, ...), they are iteratively refined by appending neighbouring atom identifiers, bond order and stereochemistry, followed by re-hashing onto an integer. 2D-pharmacophore fingerprints, as implemented in rdkit, use a set of functional group definitions combined into 2-point and 3-point molecular features (using distance bins in molecular bonding topology). These features get ordered into a bit-vector that describes whether a given feature is present or not. [2]

Since in this approach to affinity prediction our experimental dataset needs to be used as training data for the ML models, and a model cannot be evaluated based on performance on molecules it has seen as part of the training set, we used k-fold cross-validation (see, for example, Refaeilzadeh et al. (2009)).

[2] See https://github.com/rdkit/rdkit/blob/master/Docs/Code/Chem/Pharm2D.signatures.jpg for an illustration of the bit-vector structure used.

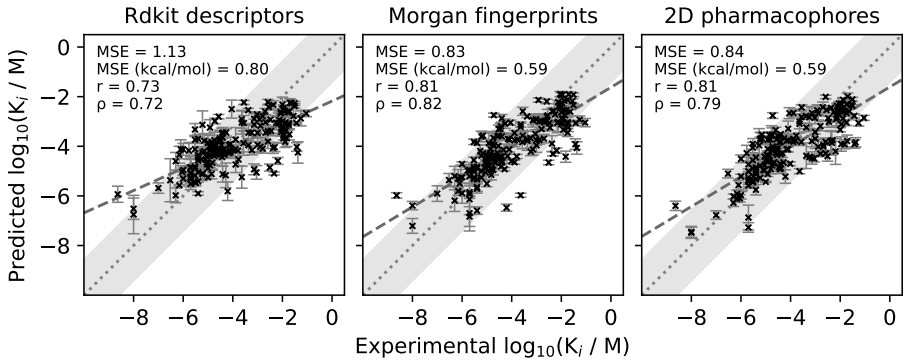


Figure 6.6: Prediction of SKPT reference peptide inhibition $\log(K_i)$ values using random forest models. Features are 64 low-dimensional molecular descriptors calculated by rdkit, 1024-bit Morgan fingerprints and 2024-bit 2D pharmacophores. Correlations are reported as MSE in $\log(K_i)$ values and a conversion to kcal mol^{-1} for reference, as well as the Pearson correlation coefficient. All 184 datapoints with SKPT data from Biegel et al. (2006) were used in 3-fold cross-validation, and each compound was in the test set 5 times, giving error bars reported as mean $\pm$ standard deviation.

Picking $k = 3$ for our case, we randomly split the dataset into 3 parts of equal size, where each part can act as the test set of a model trained on the other two parts. By repeating the random 3-fold split 5 times, we calculate mean and standard deviation for the times a given molecule was in the test set.

Figure 6.6 shows the obtained correlations of the predictions with experimental data, at the optimal hyperparameters determined in figure 6.7 (see also appendix D for details). At Pearson r = 0.73, Spearman ρ = 0.72 and mean-squared error (MSE) = 1.13 in the pK_i, the rdkit descriptors model significantly outperforms docking and ABFE at a fraction of the computational cost. The correlation is further improved by using the more detailed feature space of Morgan fingerprint and 2D pharmacophores at r = 0.81, ρ = 0.82 / 0.79 and MSE = 0.83 / 0.84 respectively. While we note that the small computational effort compared to the structure-based methods is somewhat misleading, given the reliance on extensive prior experimental data, we were impressed by the high baseline performance of the random forest models using the dataset and set out to explore in more detail how it is achieved.

As shown for our Morgan-fingerprint and 2D-pharmacophore-based models respectively in figures 6.8a and 6.9a, random forest regressors come with an importance weight for each feature in the trained model. Morgan fingerprints bits are hashed from 32-bit integer atom identifiers, and as such cannot be unambiguously interpreted without context. One can, however, find instances of molecules in the dataset triggering positive values of certain fingerprint bits and plot the relevant molecular substructures. We do this for the top 25 fingerprint bits in figure 6.8b, showing how the model picks up on the amino terminus, peptide bond and C-terminus of ligand molecules, as well as parts of the β-lactam core. There are also features corresponding to common chemical modifications present in the dataset, for example bits 32 and 190, indicating a possibly limited scope of the model to the chemical space explored in the experimental studies contributing to the dataset. The 2D-pharmacophore features, in turn, are unique in the bit vector so that a direct interpretation is possible, although not very intuitive. For illustration purposes, we show the 20 most important pharmacophore features instantiated in the Asp-Phe dipeptide in figure 6.9. Bit 556, for example (the bit with the second-highest weight), is a 3-point pharmacophore

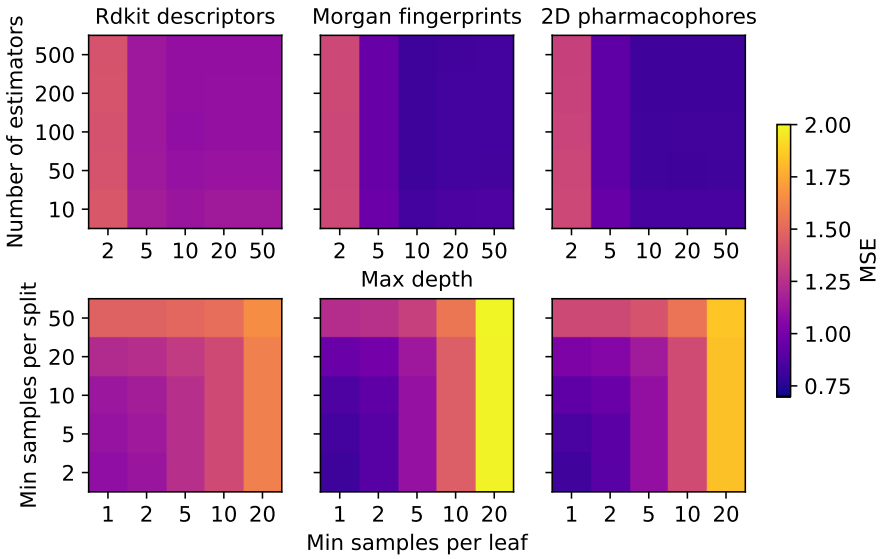


Figure 6.7: Hyperparameter tuning for random forest models concerning the number of estimators, maximum depth, minimum number of samples per split and minimum number of samples per leaf. Benchmarking is shown through the MSE to experimental data as a target function. The mean of triplicate grid searches is shown. Performance does not improve with number of estimators greater than 100 or maximum depth greater than 10, and the minimal values for the minimum number of samples per split and per leaf are optimal.

involving a hydrogen-bond acceptor, a donor and a positive ionisable group, thus corresponding to the combination of the amino terminus and the adjacent peptide bond in POT substrates. Some pharmacophore bits can also be instantiated multiple times in a given molecule. Bit 1610, representing a 3-point donor–hydrophobe–negative ionisable feature, for example, corresponds to both the Phe hydrophobic side-chain together with the peptide-bond C-terminus and the peptide bond amide, as well as the Asp side-chain carrying its own negative ionisable group. Taking these example together, in either the case of Morgan fingerprints or 2D-pharmacophores, some interpretation of the ML model is possible that points to known features of peptide transporter substrates such as the charged termini and peptide bonds. An intuitive rationalisation of the finer details relating to amino-acid side chain selectivity and the recognition of non-canonical substrates, however, is not easily possible.

6.2.4 Limitations of retrospective affinity predictions

In the preceding sections we have shown how docking and ABFE calculations starting at docked poses fail to provide good correlations with experimental K_i data for the PepT2 system, while random forest regressors trained on ligand-only features perform significantly better in terms of a correlation to experimental data, as determined through 3-fold cross-validation. We will now explore, using case studies of molecules with structurally resolved binding modes, possible reasons for the failure of the structure-based methods, and also evaluate whether the good global correlation observed with the ML methods can translate to practical use cases.

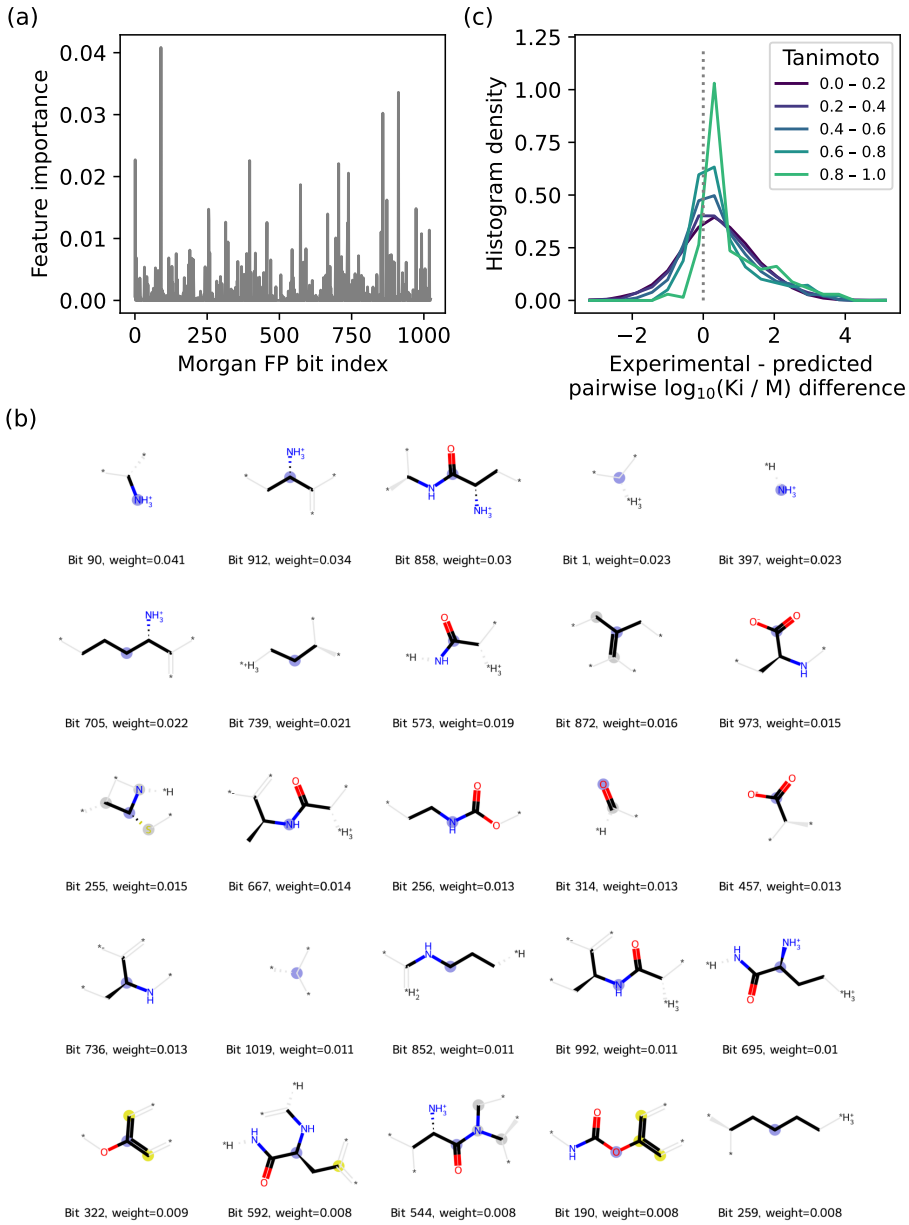


Figure 6.8: Characteristics of the Morgan fingerprint random forest regressor. (a) Random forest feature importance scores of all 1024 fingerprint bits. A number of features are picked out as significantly more important than the background. (b) Molecular interpretations of the 25 most important fingerprint bits as example molecular substructures in the dataset contributing to triggering positive values for the bits in question. Blue circles indicate the central atom in each fragment, yellow atoms in aromatic rings and gray aliphatic ring atoms. (c) Histograms of the experimental minus the predicted log(K_i) values differences for all pairs of molecules, grouped by Tanimoto molecular similarity. For similar molecules, the model predictions become noticeably flat, indicating a general inability to capture activity cliffs.

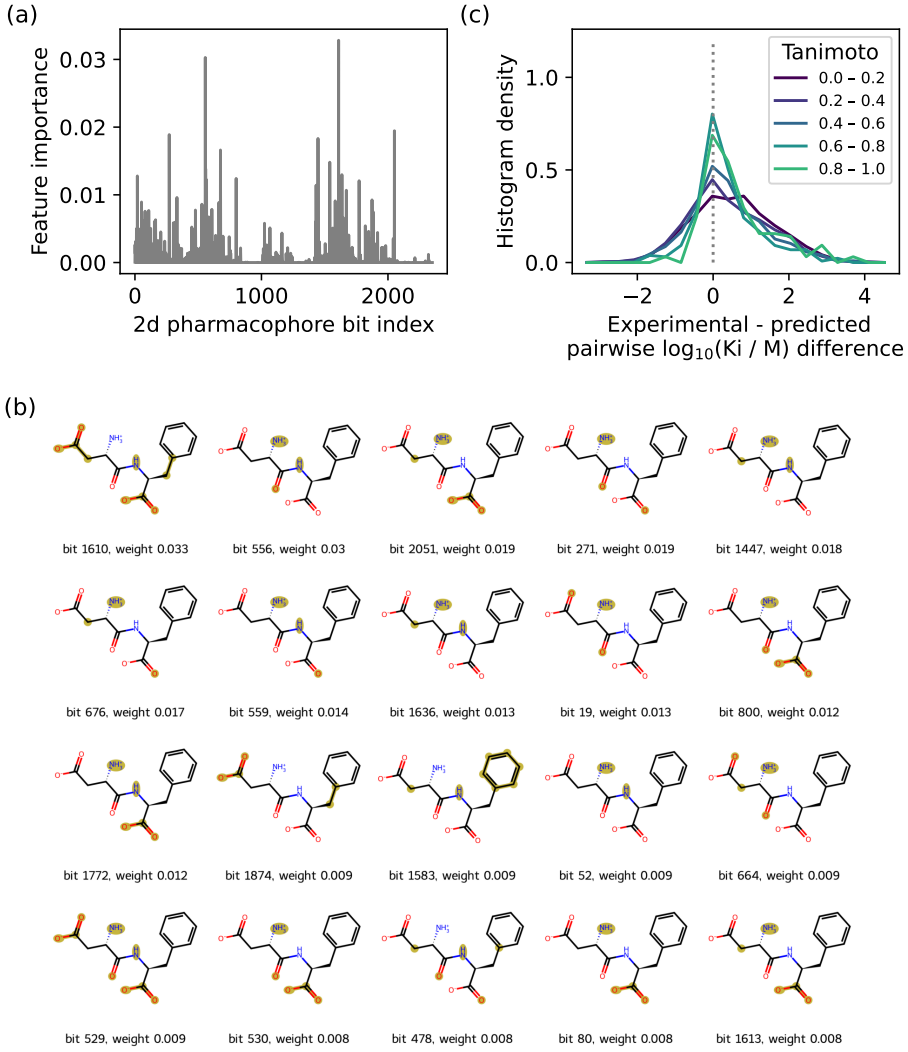


Figure 6.9: Characteristics of the 2D pharmacophore random forest regressor. (a) Random forest feature importance scores of all 2048 fingerprint bits. A number of features are picked out as significantly more important than the background. (b) Asp-Phe dipeptide with atoms contributing to the most important pharmacophore features highlighted in yellow. In cases where there were several instances of a pharmacophore feature across a molecular, all atoms involved are highlighted. (c) Histograms of the experimental minus the predicted $\log(K_i)$ values differences for all pairs of molecules, grouped by Tanimoto molecular similarity. For similar molecules, the model predictions become noticeably flat, indicating a general inability to capture activity cliffs.

Limitations of structure-based affinity predictions

As we have hypothesised above, the fact that the physically motivated and computationally expensive ABFE method cannot improve correlation with experiment compared to docking scores using the same poses may be an indication that the problem lies with identifying stable protein–ligand binding configurations. We will now explore this possibility further, using di- and tripeptides for which previous ABFE data exists in the literature and the β-lactam antibiotics cefadroxil, amoxicillin and cloxacillin as example cases.

Parker et al. (2021) have previously performed ABFE calculations on 5 di- and tri-peptides at PepT2, with binding poses obtained from docking followed by several replicates of MD equilibration and clustering. While the authors report only correlations with (unit-less) experimental log(IC_{50}) values, for 3 of their 5 studied ligands (Ala-Ala, Ala-Ala-Ala and Phe-Ala), K_i values are available in the dataset by Biegel et al. (2006), which we have used here. In table 6.1 we compare these previous ABFE results with experimental RT ln(K_i / M) data and the results of our own pipeline (which uses less MD sampling as necessitated by the much larger number of ligands). For the subset of these 3 ligands, there does not appear to be correlation between the ABFE of Parker et al. (2021) and the K_i-derived experimental data with r = -0.46 and RMSE = $5.2\,\mathrm{kcal\,mol^{-1}}$, and neither is there a correlation with the log(IC_{50}) data — as used by the authors — for this subset (r = 0.09, RMSE not applicable). For comparison, on the full set of 5 ligands and experimental IC_{50} values, the authors report a correlation of r = 0.83, underlining the challenges associated with establishing correlations from small datasets. Not only is the dynamic range of ABFE values reported by Parker et al. (2021) ($8\,\mathrm{kcal\,mol^{-1}}$) much larger than the dynamic range of the experimental data ($0.6\,\mathrm{kcal\,mol^{-1}}$), but also for all but one ligand the ΔG reported from ABFE simulations has a larger numerical value than the experimental estimate calculated from K_i values. The ABFE results on these 3 peptides from our own study (highest filtered docked poses) correlate equally badly with experiment (r = -0.78, RMSE = $4.6\,\mathrm{kcal\,mol^{-1}}$), although in this case we underestimate instead of overestimate affinities. This hints further at a possible problem with pose prediction with our pipeline. Regarding the overestimation of affinity by Parker et al. (2021), we cannot attempt to troubleshoot this here. We note, however, that while in principle K_i values are thermodynamic quantities that can be converted to a binding free energy, this is only the case when the inhibition mechanism is well defined, and binding to the outwards facing (OF) state of PepT2 in the unprotonated condition may not be a good proxy for binding to the complete protein conformational ensemble — in several protonation states — that receives ligands which inhibit reference peptide transport (we have shown an example of this in chapter 5, where cefadroxil and amoxicillin exhibited opposing trends in the change of ABFE following protonation of E56). A general treatment of such cases is beyond the scope of this work, and we will in the analysis below contend ourselves with discussing correlation coefficients between the computed and experimental ΔG values, and comparing ABFE results from different starting states to each other.

Ligand	$-\Delta G_{bind}$ / kcal mol^{-1}		
	RT ln(K_i / M)	Parker et al. 2021	This work (highest dock)
Ala-Ala	7.1	10	2.0
Ala-Ala-Ala	6.5	15	10.7
Phe-Ala	6.7	7	2.4

Table 6.1: Results of ABFE calculations in this study, compared with results reported by Parker et al. (2021), and experimental data from Biegel et al. (2006) on the three shared peptides across the studied sets.

Moving to a comparison between our ABFE data from docked and cryo-EM poses, in table 6.2, we list

as the first three columns the experimental RT ln(K_i) affinity values, ABFE-derived ΔG values starting from cryo-EM poses and also from the filtered highest-scoring docked poses of the β-lactams. We note firstly that the ΔG values obtained by ABFE starting from cryo-EM poses are significantly larger than those calculated from experimental K_i values. There is precedent for such an observation in the data of Parker et al. (2021), as discussed in the preceding paragraph. We next note that while amoxicillin has comparable ABFE values from the cryo-EM and docked poses, the cefadroxil filtered highest-scoring docked pose produces an ABFE estimate that implies $\approx$6 kcal mol^{-1} less tight binding than the cryo-EM-derived pose. To a lesser extent ($\approx$3 kcal mol^{-1}), we also observe an underestimated affinity for cloxacillin compared to the cryo-EM pose. We decided to investigate the case of cefadroxil in more detail as an example we hoped may help troubleshoot the poor ABFE performance in our retrospective affinity prediction pipeline.

Ligand	$-\Delta G_{bind}$ / kcal mol^{-1}				
	RT ln(K_i / M)	cryo-EM	Docked pose	Second-highest dock	Third-highest dock
Cefadroxil	7.5	12.0 $\pm$ 1.5	6.4	18.1	14.4
Amoxicillin	4.6	10.6 $\pm$ 1.7	12.7	10.9	4.9
Cloxacillin	4.1	6.8	3.5	-4.5	-2.0

Table 6.2: Results of ABFE calculations, showing how the use of the highest-scoring docked poses leads to a significant underestimation of binding affinity compared to available cryo-EM poses for cefadroxil. Alternative cefadroxil binding poses from lower-scoring docked poses have a much increased affinity in ABFE.

By comparing the cefadroxil docked pose to the frame picked by MDRestraintsGenerator from the MD equilibration as a starting point for ABFE (see figure 6.10a), we found that the docked pose is not fully stable at the drug N-terminus. While in the docking pose it sits in the expected triangle between E622, N192 and N348, this arrangement dissociates in the MD equilibration, where most noticeably N348 moves away from the drug N-terminus. By contrast, figure 6.10b shows how in the MD-equilibrated cryo-EM pose the N-terminal interaction network is stable, which we reason contributes to the higher ABFE estimate. This comparison shows how the PepT2–cefadroxil phase space contains at least one protein–ligand configuration region that has a significantly higher affinity than the docked pose, which implies that the docked pose gives underestimation in ABFE by not identifying a more relevant conformation. The comparison may not be entirely fair at this stage, however, because the cryo-EM pose does not engage R57. We have discussed in chapter 5 how this may be an intermediate step of the transport mechanism, where the substrate next pulls R57 with its C-terminus, freeing up E56 for receiving a proton and thus driving the alternating-access protein conformational changes. But since we filtered by contacts including between the C-terminus and R57, the docking procedure could not possibly have identified this particular conformation.

We therefore probed whether the poor stability of the filtered highest-scoring docked pose (docking score 71.5) could be improved by running MD equilibrations and ABFE from the next-highest-scoring poses (docking scores 71.3 and 70.0). As shown in the 4th and 5th columns of table 6.2, this makes a stark difference to the calculated affinities, with results of -18.1 and -14.4 kcal mol^{-1}, compared to -6.4 kcal mol^{-1} from the highest-scoring pose. Figure 6.10c shows a comparison between the highest and second-highest-scoring docked poses, revealing only small structural differences, most notably in the positioning of E622 and the rotamers of flexible dihedrals surrounding the peptide bond adjacent to the β-lactam core. As shown in figure 6.10d, this pose, while changing significantly in MD equilibration, remains stably associated at its N-

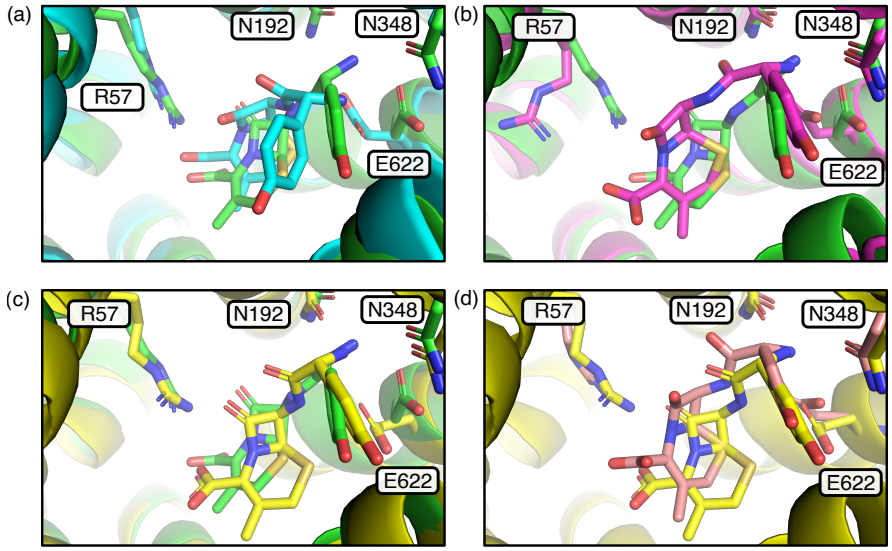


Figure 6.10: Comparison of different cefadroxil poses in the PepT2 binding pocket, overlayed after best-fit alignment of the protein. (a) Filtered highest-scoring docked pose (green) and MD-equilibrated frame picked by MDRestraintsGenerator for seeding ABFE (blue). (b) Filtered highest-scoring docked pose (green) and MD-equilibrated frame starting at the cryo-EM pose (highest-affinity starting frame for ABFE among the replicates, see chapter 5). (c) Filtered highest-scoring (green) and second-highest-scoring (yellow) docked poses. (d) Filtered highest-scoring docked pose (yellow) and MD-equilibrated frame picked by MDRestraintsGenerator for seeding ABFE (salmon).

and C-termini to E622/N192/N348 and R57. The ABFE affinity of $-18.1\,\mathrm{kcal\,mol^{-1}}$ from this pose is in excellent agreement with the $-18.0\pm0.5\,\mathrm{kcal\,mol^{-1}}$ found in chapter 5 for the cefadroxil cryo-EM pose after E56 protonation and the subsequent release of R57 to engage the ligand C-terminus, thus suggesting that this docked pose may represent a reasonable approximation to the important region of protein-ligand phase space. Considering the small differences in docking score between these three cefadroxil poses that lead to vastly different ABFE results, we recognise a high degree of stochasticity in the docking pipeline. For amoxicillin and cloxacillin, there is also high stochasticity in the calculated ABFE values, as shown in table 6.2, although in those cases the highest-scoring docked poses did yield the highest affinity in ABFE.

A promising avenue of exploration in light of these findings on peptides and β-lactam antibiotics — with the aim of improving correlation with experimental data — may therefore be to run ABFE calculations starting at several docked poses to sample more protein–ligand phase space (a reasonable minimum could be 3 poses). To see whether it may be possible to devise a protocol to save a large chunk of the increased requirements for compute resources, we checked whether the ABFE affinity from the docked poses (pose 2 > pose 3 > pose 1) was something that could have been predicted clearly from the 50 ns-long MD equilibrations. To this end, we computed root mean squared deviation (RMSD) values from the energy-minimised docked poses across the 50 ns equilibrations, fitting the trajectories using the whole protein for alignment, and computing the RMSD of the cefadroxil and protein R57, N192, N348 and E622 atom coordinates. As shown in figure 6.11, poses 1 and 2 have similar deviations from the initial docked pose by RMSD at ≈0.4 nm, while pose 3 is more stable at RMSD values resting mostly below 0.25 nm. This trend does not align with the ABFE values, where a discrepancy of $\approx 11\,\mathrm{kcal\,mol^{-1}}$ is observed between poses 1 and 2, and pose 2 (not pose 3, as might have been expected from the RMSD plot) has the highest affinity. While it may be possible in principle to find a metric that performs better than RMSD at picking

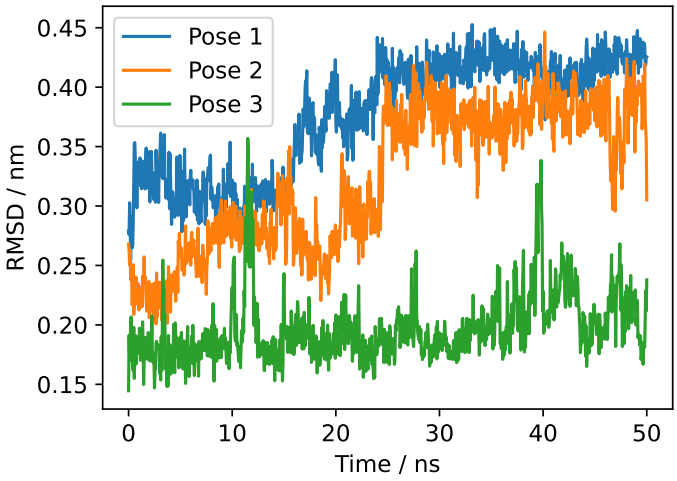


Figure 6.11: MD equilibrations from the top three filtered docked poses of cefadroxil at PepT2, shown as RMSD of the ligand and R57, N192, N348 and E622 atom coordinates with alignment of the whole protein.

out the right docked poses, we were not able to do so within the scope of this thesis, leaving only the option of running the full ABFE protocol on the second and third-best docked poses for all ligands. Within the scope of this thesis, such an endeavor was not possible due to its high computational cost and time constraints on the doctoral project. Whether it would improve correlation with experimental data is therefore an open question, albeit one we are actively investigating still. In principle, CV-based enhanced sampling methods could also help sample ligand binding poses, but — as the example of cefadroxil with the slow degree of freedom (DOF) around the peptide bond shows — the required CVs would be specific to the individual ligands. The CV design method in the case of a large dataset would therefore need to be automatic and pick out the relevant CVs from a feature space that is agnostic of the ligand or slow DOF in question. While several methods based on machine learning have been proposed to help with such tasks — and we have conducted some initial unsuccessful trials using the state predictive information bottleneck (SPIB) method — they remain an essentially unsolved problem in the enhanced sampling field.

Limitations of ligand-only affinity predictions

Although, on the overall dataset, random forest models trained on ligand features clearly outperform the structure-based methods, we noticed that the best-fit curves to the correlation plots in figure 6.6, shown as dashed lines, are too flat, compared to the expected gradient of 1. To study this property of our models further, we iterated across all substrate–substrate pairs in the dataset and calculated separate histograms of the experimental minus the predicted $\log_{10}(K_i)$ difference between the substrates grouped by the chemical similarity between the two substrates (figures 6.8–6.9c). In this analysis, we found pronounced tails at positive double-difference values — that is, our random forest models systematically underestimate activity differences. Moreover, this underprediction of differences is particularly prominent for chemically similar molecules. With Morgan fingerprint features, the larger block of 17028 ligand pairs with Tanimoto similarity between 0 and 0.2 showed an average ΔpK_i underprediction of 0.42, while the smaller number of 155 pairings of very similar molecules (Tanimoto score 0.8–1.0) had an average underprediction of 0.80. The 2D-pharmacophore features showed a similar behaviour with 0.52 and 0.62 average underpredictions for the same ligand-pair groups, amounting to somewhat more favourable behaviour for the most similar ligands but less favourable results for the least similar ligands. The underprediction of ΔpK_i, specifically for similar ligands, amounts to a crucial limitation, because activity cliffs, where small modifications in a molecule

effect large changes in activity, are often the cases of greatest interest in medicinal chemistry. (Stumpfe and Bajorath, 2012)

This inability to appropriately predict activity cliffs can be seen particularly well in the β-lactam antibiotic case (table 6.3). While the lower affinity of cloxacillin is reproduced by our models (in line with expectations, given that features corresponding to the amino terminus, which cloxacillin lacks, are clearly picked out), the predictions for cefadroxil and amoxicillin are very close to each other, despite the experimental difference of two orders of magnitude in K_i (with 2D-pharmacophore features, amoxicillin is even falsely predicted to bind tighter than cefadroxil). The regressor has clearly not learned to distinguish the two molecules based on the subtle structural differences in the β-lactam core. This deficiency may be a consequence of an imbalanced training dataset, an inappropriate training procedure that focuses on overall correlation instead of activity cliffs, or a combination thereof. Within the scope of this thesis, we could not address these issues further, and contend ourselves with noting that in a predictive setting, where the modification made to a molecule is often small, the ML models could be producing false negatives with respect to affinity gains.

Ligand	**-log$_{10}$(K_i)**		
	Experiment	Predicted (Morgan FP features)	Predicted (2D-pharmacophore features)
Cefadroxil	5.5	3.82 ± 0.11	3.90 ± 0.21
Amoxicillin	3.4	3.78 ± 0.38	4.38 ± 0.52
Cloxacillin	3.0	2.91 ± 0.13	2.85 ± 0.24

Table 6.3: Random forest regressor predictions of β-lactam antibiotic affinities at PepT2 compared to experimental data. The cefadroxil / amoxicillin activity cliff is not captured.

We also wished to investigate how well random forest regressors with ligand-only features could generalise across chemical space. In the 3-fold random cross-validation procedure, this cannot be studied because the training set for a given ligand will always contain a number of chemically similar molecules. To gain an idea, nonetheless, about how generalisable our ML models are, we trialled an alternative training–test split. On the left-hand side in figure 6.12, we show good results for the subset of 42 drugs in the dataset, using the same 3-fold cross validation procedure and Morgan fingerprint features as above. When, by contrast, we trained a model on the 142 natural and modified peptide molecules and used this model to predict affinities for the drugs, performance is much degraded. The best-fit curve is much flatter, indicating a very poor dynamic range for distinguishing good from bad substrates (even if some correlation is maintained at r = 0.55, ρ = 0.46). The affinity of most drugs is also underpredicted, leading to an order-of-magnitude increase in MSE to from 0.50 to 5.85. This result is not surprising, given that the drugs (mostly β-lactams), which are now all in the test set simultaneously, occupy a distinct region of chemical space from the training set. Considering future applications of ML models based on ligand-only features for drug design at POTs, these results urge caution when moving beyond the training-set chemical space, a general limitation of any ML model trained on ligand features alone. Broader, high-quality datasets will therefore be necessary.

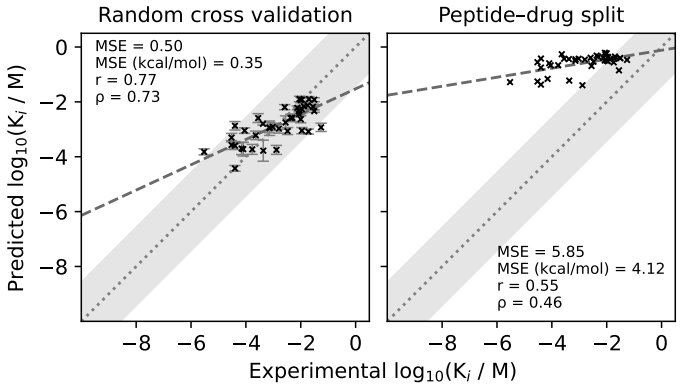


Figure 6.12: Random forest regressors predictions (Morgan fingerprint features) of 42 drugs in the data set, plotted against experimental $\log_{10}(K_i$ / M) values. On the left, good performance is shown with using random replicates of 3-fold cross validation for training–test splitting. On the right, performance is much degraded when using the natural and modified peptides part of the dataset for training to predict the subset of drugs.

6.3 Prospective application to drug design

Despite these limitations of the retrospective affinity predictions, we wished to explore the possibility of applying our models prospectively in the context of improving the oral bioavailability profile of known β-lactams that are not recognised by POTs, through making prodrugs that are POT substrates. After an initial trial to modify meropenem failed due to difficulties in purifying the resulting zwitterionic molecule, we agreed with experimental collaborators in the group of Chris Schofield to study the modification of penicillin G with a valyl ester (figure 6.13a), inspired by a successful similar modification of acyclovir into valacylovir as reviewed in section 1.2. Table 6.4 shows how all models predict a gain in affinity to PepT2 from this modification. We note that the gain is most pronounced in the ABFE values, and also that running a best-of-three-poses ABFE protocol, as discussed in section 6.2.4, increased the observed gain in affinity. Had we only run our simulations from the highest filtered docked pose, our ΔG predictions would have been $-3.6\,\mathrm{kcal\,mol^{-1}}$ for penicillin G and $-6.3\,\mathrm{kcal\,mol^{-1}}$ for val-pencillin G. While we would have drawn the same conclusion, this observation confirms a high degree of stochasticity in the quality of the docked poses that is not captured by docking scores or our filter function (as observed for cefadroxil above), and reiterates that one avenue for improvement of ABFE results could be increasing the sampling by using several distinct docked poses, as we discuss in 6.4.

Ligand	Docking	$-\Delta G_{bind}$ / kcal mol^{-1} (ABFE)	-log$_{10}$(K$_i$) (random forest)
Penicillin G	75.1	3.9	3.07
Val-penicillin G	80.6	16.7	3.51

Table 6.4: Results from docking, ABFE and random forest regressor (Morgan fingerprints) models concerning the attachment of a valyl ester on penicillin G. For ABFE, the three highest scoring docked poses were run, and the highest observed ABFE affinity is reported for each ligand. All models predict a gain in affinity to PepT2.

Encouraged by these affinity gains observed across the different models, our collaborators synthesised the val-penicillin G molecule by the scheme shown in figure 6.13a. We were, however, unable to test the inhibition of reference peptide transport by this compound, because — while being soluble in DMSO —

there was very poor solubility in aqueous buffer. In trials with cell-based transport assays by the same protocol as detailed in appendix B we determined that a maximum of around 10 % of DMSO in the assay buffer could be tolerated by the cells over the duration of the assay before losing uptake of the fluorescently labelled peptide. As shown in figure 6.13 b–c, adding water to the val-penicillin G initially dissolved in a small amount of DMSO to make a up a final DMSO concentration of 10 % led to a cloudy solution, indicating precipitation of the compound, while penicillin G dissolves well. We were also unable to improve the situation in further trials including adding stoichiometric amounts of acid and raising the DMSO content as far as 30 %, leading us to conclude that val-penicillin G is unsuitable for testing of activity at PepT2. We had chosen a valine amino acid for our ester because in the design of valacyclovir this side-chain demonstrated the best bioavailability. (Beauchamp et al., 1992) Given the poor solubility of the prodrug, a next step would be to trial other amino acids, although this was beyond the scope of work possible for this thesis owing to constraints on the synthetic collaborators' time. We note that this problem was not anticipated based on the simple logP predictor implemented in rdkit. While the attachment of the valine ester did increase the predicted logP from -0.47 to +0.48, the drastic loss of solubility in aqueous buffer was not expected from these calculations. Future trials of other amino acids will therefore likely need to be informed by more sophisticated models for solubility prediction.

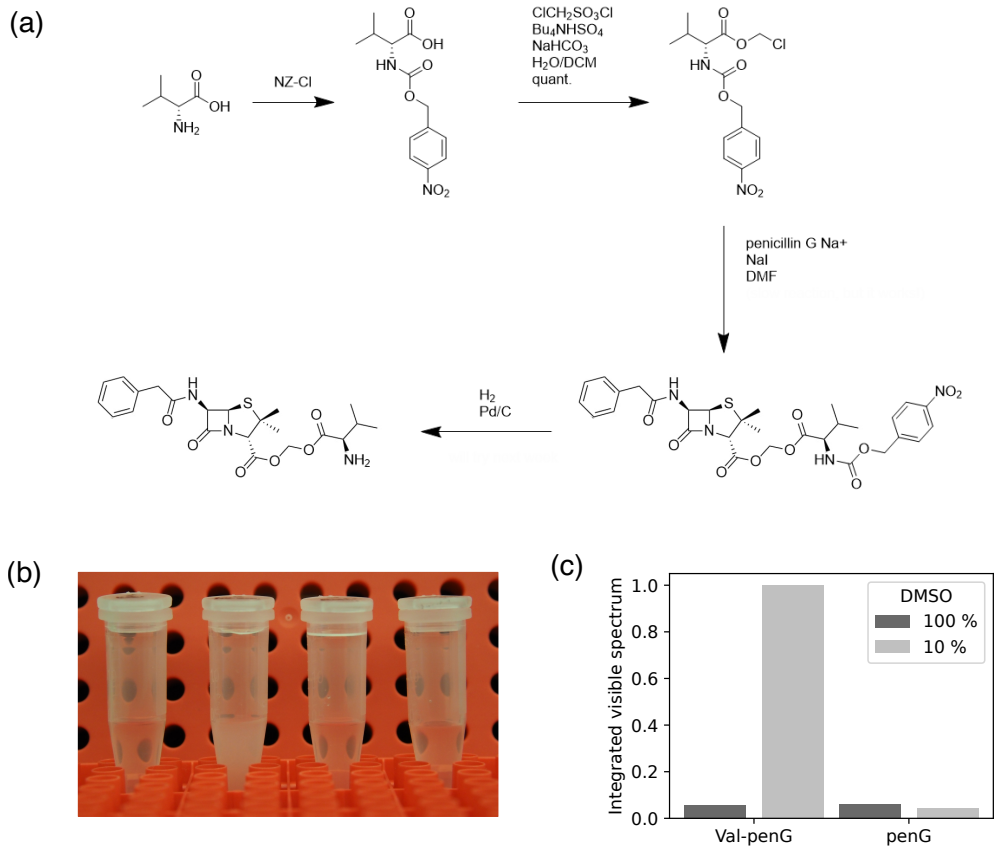


Figure 6.13: (a) Synthetic scheme for attaching a valine amino-acid ester to penicillin G (penG), devised and carried out by Dorian Munțean in the group of Chris Schofield. (b) Poor solubility of valine-penG in aqueous solvents. From left to right, all 2 mM prepared initially in DMSO and mixed with more DMSO or water: val-penG in 100 % DMSO, val-penG in 10 % DMSO (cloudy solution indicates insolubility), penG in 100 % DMSO, penG in 10 % DMSO. (c) Quantification of (val-)penG solution cloudiness as normalised integrated absorbance across the visible spectrum, measured from 380 –700 nm in 20 nm increments.

6.4 Conclusion and avenues for future exploration

The research presented in this chapter is by no means exhaustive of the possibilities that could be explored regarding structure or data-driven affinity prediction at POTs. Nonetheless, it highlights some key challenges faced by different conceptual approaches. A crucial factor limiting the accuracy and transferability of structure-based affinity prediction in this binding pocket is the initial prediction of high-quality protein–ligand binding poses. Here, we have — following initial exploration to establish a robust protocol — extensively trialled a state-of-the-art approach involving flexible ensemble docking using a modern docking programme, combined with either an empirical scoring function or with pose equilibration in MD and affinity prediction using ABFE. Improvements to the pipeline that could be considered in future work to help tackle the problem of finding the 'right' (that is, highest-affinity) binding pose include more exhaustive sampling by taking a larger number of structurally distinct docked poses through to MD and ABFE. This was also part of the strategy adopted by Parker et al. (2021) in their study concerning a small set of ligands. While more sampling was not feasible for the full dataset within the scope of this thesis, our results on cefadroxil and the valine-penicillin G modification suggest that this may make a difference by

removing a source of stochasticity and underprediction that appears to originate from in the quality of the docked poses. Various enhanced sampling methods such as replica exchange with solute tempering (REST2) (Wang et al., 2011) or binding-pose metadynamics (Fusani et al., 2020) could also help alleviate the situation. Furthermore, ML-based prediction of binding poses with programmes such as DiffDock (Corso et al., 2022) could improve the accuracy of initial ligand binding poses, and the co-structure prediction of proteins with ligands that is possible with Alphafold 3 (Abramson et al., 2024) and Umol (Bryant et al., 2024) may help alleviate the problem of subtle induced fit we have discussed. Even if this problem is addressed, be it with more sampling or better binding pose prediction, however, it remains unclear and to be determined in future work how good a correlation between ABFEs and transport inhibition constants is possible for this transporter, given that the experiments are measuring binding to a complex ensemble of protein states as opposed to a single conformation.

Regarding the cheminformatics and ML approach, we have trained random forest models on various ligand-only features to reproduce experimental K_i data for inhibition of reference peptide uptake through PepT2 in a cell-based assay. While we found a much higher correlation than with the structure-based methods, our models significantly underestimate crucial activity cliffs, hampering potential applications to the *in-silico* design of orally bioavailable drugs and prodrugs. Since our data set — while the largest currently available — is composed of several pooled studies and is potentially unbalanced regarding the included chemical space, a natural yet very expensive avenue of improvement could be the generation of an expanded experimental dataset. Targeted sub-sampling of the available dataset could also help with issues of balance and potential limitations from pooling data from separate studies. Additionally, work is ongoing in the field to characterise the performance of ML models based on ligand features for scenarios involving activity cliffs (Van Tilborg et al., 2022; Tamura et al., 2023), and dataset curation strategies and algorithms better suited to predicting activity cliffs may become available in the future. Lastly, we have shown how the ML models can suffer from transferability problems to new chemical space, making the extension of available datasets to larger regions of chemical space crucial for future applications in the drug discovery pipeline.

7

Conclusions and outlook

7.1 Summary of results

In this thesis, we have first summarised our current knowledge of the molecular mechanisms by which proton-coupled oligopeptide transporters (POTs) operate, as well as the factors underpinning their substrate promiscuity and the pharmaceutical relevance of these transporters in developing orally bioavailable medications (chapter 1). Having next introduced the theoretical background of molecular dynamics (MD) simulations and the advanced simulation techniques used throughout this thesis (chapter 2), we have described a new method, which we call MEMENTO, for tackling the hysteresis problem commonly encountered in enhanced sampling of conformational changes and benchmarked it on four biological example systems of increasing complexity (chapter 3). We then applied the MEMENTO method together with extensive replica-exchange umbrella sampling (REUS), constant-pH molecular dynamics (CpHMD) and absolute binding free energy (ABFE) simulations, as well as cell-based transport assays for experimental validation, to study the coupling between protonation-state changes, the presence of a dipeptide substrate and the alternating-access conformational changes of PepT2 (chapter 4). In a collaborative study combining cryo electron microscopy (cryo-EM) structure determination, biochemical assays and MD simulations, we subsequently investigated the binding of several β-lactam antibiotics to PepT2, identifying a key molecular mechanism by which substrate selectivity is achieved (chapter 5). Lastly, we have benchmarked structure- and machine learning (ML)-based methods for computationally predicting substrate affinities at PepT2, highlighting important limitations of the respective methods, and have started work towards the *in-silico* predictive design of POT-substrate prodrugs for improved oral bioavailability (chapter 6).

Regarding the formerly open research questions concerning POTs formulated in section 1.4, we present the following key conclusions from the work conducted for this thesis.

1. The protonation of the extracellular gating residues H87 and D342 drives the occluded (OCC) → outwards facing (OF)-state transition in PepT2, while the protonation of the intracellular gating residues E53 and E622 drives OCC → inwards facing (IF) opening. This strongly suggests a sequence of protonation events by which two protons moving through the transporter drive the full alternating-access conformational cycle.

2. H87 is part of an extracellular gating mechanism conserved between mammalian species and some (but not all) procaryotic POTs. The additional involvement of D342 in the PepT2 extracellular gate, which is not conserved in any bacterial POTs, further supports the notion that overall alternating access is more evolutionarily conserved than the detailed molecular gating mechanisms.

3. The intracellular gate-opening triggers in PepT2 are E53 and E622, the latter of which also has a role in substrate binding and release. Since they do not themselves seal the transporter cleft towards the intracellular side (a function accomplished by several hydrophobic residues), the intracellular gating mechanism is distinct from that of the extracellular gate. The universal conservation of these

residues across the tree of life indicates higher conservation of the proton coupling mechanism at the intracellular gate than at the extracellular gate.

4. The substrate couples to conformational changes in two ways: by subtly modulating the thermodynamics of the alternating-access transitions, and by controlling proton access to the ExxER motif.

5. The ExxER motif couples intracellular-gate opening, which is triggered by ExxER motif protonation, to substrate binding in a salt-bridge swap mechanism. The bound substrate engages R57, thus freeing up E56 to receive a proton that then drives the large-scale conformational change.

6. Good and bad substrates, such as in the cefadroxil vs amoxicillin example, are not distinguished by binding affinity to one protein conformation alone. Instead, the ability to engage E622 and the ExxER motif simultaneously, as required for progress through the transport cycle, provides a dynamic mechanism of discrimination. Inhibitors such as cloxacillin act by steric block of the binding site without picking up key interactions for transport.

7. Predicting the affinity of substrates to POTs comes with substantial challenges. The accuracy of structure-based methods such as ABFE simulations depends strongly on the quality of the initial predicted binding pose, and we were not able in this thesis to finalise a robust workflow to this end that can incorporate the subtle induced fit likely required for substrate engagement. Ligand-based ML approaches can give better correlations to experimental data, but struggle with capturing pharmaceutically important activity cliffs and chemical space not seen during training.

8. The attachment of amino acids to poorly bioavailable drugs to turn them into POT substrates remains a promising route for improving oral bioavailability, although poor solubility has hampered an initial trial we have made regarding computational predictive prodrug design.

7.2 Outlook

Notwithstanding the substantial advances made, through the work presented here, in our understanding of the molecular mechanisms by which POTs recognise and transport their various substrates, many details remain open to future work, and indeed some new questions have been raised by our research. A full mechanistic understanding of the various proton transfer processes proposed is still lacking, which would necessitate the future use of quantum mechanics / molecular mechanics (QM/MM) or reactive MD techniques. Such a study could also clarify the nature of potential further protonation sites, which may or may not include protonation of the substrate itself. It is also unknown how different substrates are accommodated throughout the full conformational cycle, and it would be worthwhile to explore how the findings presented here on PepT2 translate to the intestinal homologue PepT1, which — while having a broadly similar range of substrates — exhibits some differences in its pattern of substrate uptake. Concerning drug transport through POTs, a dynamical understanding is still lacking of how molecules that lack a negatively charged carboxyl group (such as valacyclovir) couple to proton transfer and conformational changes, and our proposed mechanism of discrimination between good and bad substrates needs to be elaborated in terms of the relative roles played by binding affinities and energetics of different protein–substrate conformations, as well as the contributions from previously described specificity pockets. The quantitative prediction of substrate affinities to POTs necessitates improvement in protocols to predict binding poses in flexible binding pockets like the PepT2 transport pocket, and better data sets are likely needed for

successful ML applications. Lastly, the modification of penicillin G (and similar molecules) into POT substrates by attaching amino acids requires addressing the solubility problem we have encountered in our trial.

In summary, by combining an array of established and novel computational methods with experimental validation and, collaboratively, structural biology and synthetic chemistry, this thesis provides a substantial contribution both to the fields of enhanced sampling of membrane protein conformational changes and, more specifically, POT biology. The results presented will be useful to the community for understanding coupling in alternating access mechanisms in general and benefit future mechanistic investigations as well as the design of orally bioavailable medicines harnessing active transport through POTs.

A

Detailed methods for chapter 3

All simulations were run in GROMACS (Abraham et al., 2015) in versions 2021.3/2021.4 (the slight version discrepancy is because of different installations on two compute clusters we used). For production simulations, we used the leap-frog integrator with a time-step of 2 fs, the v-rescale thermostat with stochastic term (Bussi et al., 2007) with a time constant of 0.5 ps and target temperature of 300 K (310 K for LeuT), the Parrinello-Rahman barostat (Parrinello and Rahman, 1981) with a time constant of 2 ps and target pressure of 1 bar, and a short-range cutoff of 1.2 nm. Where lipids were present, we employed two temperature coupling groups (membrane + protein / water + ions) and a semi-isotropic version of the barostat (x/y, z axes). Example GROMACS *.mdp files are provided with the PyMEMENTO package.

Before starting production simulations, we energy-minimised and then equilibrated all simulation boxes for 200 ps in the NVT ensemble at a time step of 1 fs, followed by 1 ns in the NPT ensemble at a time step of 2 fs (with the Berendsen barostat), both with Cα position restraints in place. This was done before the extra MEMENTO equilibrations as described above.

Umbrella sampling simulations were run with PLUMED (Tribello et al., 2014) in versions 2.7.2/2.7.3 (again, due to different cluster installations). We used replica exchange every 1000 steps (2 ps), and the WHAM algorithm (Kumar et al., 1992) in the implementation by Alan Grossfield. [1] Convergence was assessed by ensuring histogram overlap of neighbouring windows, visualising the potential of mean force (PMF) using different fractions of the data, and calculating the root mean squared deviation (RMSD) between PMFs incorporating successively more data.

Further details are given for the individual systems below. All simulation data is available at https://doi.org/10.5281/zenodo.7851906 in the form of key coordinate files, full PLUMED output files with collective variable (CV) projections and weighted histogram analysis method (WHAM) output free energy data.

A.1 Deca-alanine

We used the peptide building functionality of PyMOL to generate helical and extended conformations of the deca-alanine peptide, which we capped with ACE and NME residues. The simulations were run with the CHARMM 36 forcefield (Huang et al., 2017) (version July 2021). The peptide was solvated with 5237 solvent molecules in a cubic box of around 5.5 nm side length. Steered MD (SMD) was run starting from the helical state on the Cα end-to-end distance over 50 ns, with a restraint sliding from 1.446 to 2.666 nm and a force constant of 5000 kJ mol^{-1} nm^{-2}. REUS was done along the same CV, with restraint centers linearly interpolated between the terminal values and a force constant of 1000 kJ mol^{-1} nm^{-2} for 500 ns per window. For the SMD comparison, the starting frames were extracted from the SMD trajectory at even spacing in CV values. We conducted three replicates of both the MEMENTO and SMD-derived REUS. The total sampling time expended for deca-alanine was 72 µs.

[1] http://membrane.urmc.rochester.edu/?page_id=126

A.2 ADK

We obtained structures for adenylate kinase (ADK) in open (Müller et al., 1996) and closed (Müller and Schulz, 1992) states from the PDB (open: 4AKE, closed: 1AKE — with inhibitor). Since we could simulate all residues (1–214), we did not cap the termini. We solvated the protein with approximately 16,500 solvent molecules (precise number varies between replicates) and a NaCl concentration of $0.15\,\mathrm{M}$ in a cubic box of around $8.1\,\mathrm{nm}$ side length. The simulations were run using the AMBER ff14.sb forcefield (Maier et al., 2015); in the holo-state simulations, the AP5A inhibitor was parameterised using GAFF2. (Wang et al., 2004) At each of the apo end states, we ran $500\,\mathrm{ns}$ of unbiased MD, though the closed state opened spontaneously (as is expected from the PMF and the fact that the co-crystallised inhibitor was removed), so that our MEMENTO replicates could only be seeded from the initial closed structure and the 0, 250 and $500\,\mathrm{ns}$ frames of the unbiased open run. SMD was conducted in 3 replicates in the closed $\rightarrow$ open direction, and one replicate in the open $\rightarrow$ closed direction (since assessing the stability of the resulting conformation was impossible, given that even the native closed state opened spontaneously). We ran these for $150\,\mathrm{ns}$ each with a restraint on the Cα-RMSD to the respective target state that linearly increased to $5000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$, followed by $50\,\mathrm{ns}$ unbiased equilibration to judge stability of the obtained conformation. For 1D REUS, we ran three replicates for MEMENTO and SMD starting configurations each (the former prepared as described above, the latter as equally spaced frames from the three opening SMD runs) along the centre-of-mass distance between the ATP-binding LID (residues 123–159) and AMP-binding NMP (residues 31–73) domains. Restraint centers were averaged CV values over the MEMENTO equilibration runs for all boxes, with a force constant of $2000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$. 2D REUS ran along the LID–CORE (residues 1–30, 74–122 and 160–214) and NMP–CORE distances, with restraint centers again extracted from box equilibrations or the appropriate SMD frames, and a force constant of $1000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ along each CV. The 'alternative closed state' was obtained by clustering (using GROMACS simple linkage with default parameters) MEMENTO 1D REUS rep1 window 0 and SMD 1D REUS rep 1 window 2, as the most occupied cluster in each case. They were incorporated into 2D REUS by MEMENTO or SMD as described above.

We detail the amount of sampling collected on ADK in table A.1.

Type of run	Simulation time / ns
Unbiased MD	500 * 2
SMD	200 * 6
1D REUS MEMENTO	(259 + 234 + 258) * 24
1D REUS SMD	(372 + 240 + 245) * 24
2D REUS MEMENTO apo	(147 + 158) * 24
2D REUS SMD apo	(158 + 159 + 157 + 250 + 255) * 24
2D REUS MEMENTO holo	201 * 24
2D REUS MEMENTO holo, docked ligands	205 * 24
Total	$81\,\mu\mathrm{s}$

Table A.1: Overview of all sampling performed on ADK.

Type of run	Simulation time / ns
Unbiased MD	500 * 4
SMD	200 * 6
1D US MEMENTO	(150 + 142 + 140) * 24
1D US SMD (closing)	(148 + 141) * 24
1D US SMD (opening)	(147 + 130) * 24
1D US SMD (re-solvation)	(118 + 60) * 24
2D US MEMENTO apo	(137 + 123 + 101) * 24
2D US SMD (closing)	(110 + 158) * 24
2D US SMD (opening)	(179 + 101) * 24
2D US MEMENTO holo	(132 + 114 + 131) * 24
Total	62 µs

Table A.2: Overview of all sampling performed on P38α.

A.3 P38α

We obtained structures for the DFG-in (Wang et al., 1997) and DFG-out (Gill et al., 2005) states from the PDB (DFG-in: 1P38, DFG-out: 1W83 — with inhibitor). We used PyMOL to mutate two residues in 1P38 to match the sequence of 1W83: H48L and T263A. Note that in the PDB, entry 1P38 was superseded by 5UOJ in 2017, which does not model the DFG loop anymore. However, for consistency with exisiting simulation work on this system (e.g., Wang et al. (2020a)), we still use the 1P38 coordinates, which is validated by the fact that it is a stable conformation in unbiased MD. We simulated residues 4–354 capped with ACE and NME in the AMBER ff14.sb forcefield (Maier et al., 2015); in the holo-state simulations the pyridine-containing L11 inhibitor was parameterised using GAFF2. (Wang et al., 2004) We solvated the protein with approximately 27,300 solvent molecules (precise number varies between replicates) and a NaCl concentration of $0.15\,\mathrm{M}$ in a cubic box of around $9.6\,\mathrm{nm}$ side length. At each of the apo and holo end states (see note in the Adding ligands section), we simulated $500\,\mathrm{ns}$ of unbiased MD, the 0, 250 and $500\,\mathrm{ns}$ frames of which formed the input structures for the MEMENTO and SMD replicates. For bi-directional SMD, we used the RMSD of all Cα atoms to the respective target structure as CV, gradually increasing the force constant to $20\,000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ over $150\,\mathrm{ns}$, followed by $50\,\mathrm{ns}$ of unbiased relaxation. When proceeding to 1D REUS, we initially trialed a difference RMSD, $\mathrm{DRMSD} = \mathrm{RMSD}_{\text{(DFG-in)}} - \mathrm{RMSD}_{\text{(DFG-out)}}$, on all C$\alpha$ atoms, but found slightly better behaviour when keeping whole-protein aligning but restricting the RMSD calculation to the DFG loop (resides 166–176). We therefore used this CV for all 1D REUS, with a force constant of $4000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ and restraint centers averaged from MEMENTO box equilibrations or extracted from SMD frames (only reps 1 and 3 of the SMD were used for REUS, since rep 2 did not lead to meta-stable conformations in the vicinity of the target states). We also tried to re-solvate (using the GROMACS solvation tool) the conformations of 1D REUS SMD rep 1 at each intermediate, but stopped our simulations short of the sampling time used in other runs, since we observed no significant difference in the PMF. The 2D REUS CVs were derived as discussed in the previous section; we used a force constant of $2 \times 10^6\,\mathrm{kJ\,mol^{-1}}$ on each CV during REUS (note that the principle components (PCs) are dimensionless and have small numerical values when output by GROMACS).

We detail the amount of sampling collected on P38α in table A.2.

Type of run	Simulation time / ns
Unbiased MD	1000 * 2
SMD	600 * 6
1D US MEMENTO	(156 + 155 + 157) * 24
2D US MEMENTO	(495 + 514 + 517) * 24
2D US MEMENTO extra	510 * 24 + (523 + 499) * 16
2D US SMD	(507 + 511) * 24
2D US SMD extra	500 * 24
Total	118 µs

Table A.3: Overview of all sampling performed on LEUT.

A.4 LeuT

We obtained structures for the IF (Krishnamurthy and Gouaux, 2012) and outwards-facing OCC (Singh et al., 2008) states from the PDF (IF: 3TT3, OCC: 3F3E), using MODELLER to fix a missing loop in the OCC structure from sequence. We simulate here residues 11–507, with ACE and NME caps, embedded with the CHARMM-GUI membrane builder (Wu et al., 2014) in a bilayer of 344 POPE lipids, solvated with approximately 21,500 solvent molecules (precise number varies between replicates) at a NaCL concentration of 0.15 M in an orthorhombic box of around $10.6 * 10.6 * 9.8$ nm side lengths. We performed 1 µs of unbiased MD for each end-state, where the 0, 500 and 1000 ns frames served as the starting structures for MEMENTO and SMD replicates. SMD was run on the Cα-atom RMSD to the respective target state with a force constant linearly increasing to $10\,000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ over 250 ns, then held for further 250 ns and followed by unbiased relaxation for 100 ns. 1D REUS was set up with a distance RMSD CV, $\mathrm{DRMSD} = \mathrm{RMSD}_{(\mathrm{IF})} - \mathrm{RMSD}_{(\mathrm{OCC})}$, calculated on all C$\alpha$ atoms. The force constant was $20\,000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ and restraint centers were averaged from MEMENTO box equilibrations. The derivation of 2D CVs is described in section 3.2.4; we used a force constant of $5 \times 10^6\,\mathrm{kJ\,mol^{-1}}$ along each CV. We used the same MEMENTO starting frames as for 1D REUS, and SMD frames extracted at equal spacing (only reps 1 and 2 of the OCC $\rightarrow$ IF SMD were used for REUS, since the others did not lead to meta-stable conformations in the vicinity of the target states). To improve histogram overlap, we included additional windows in our 2D REUS:

1. a MEMENTO run between the starting and final conformations of the IF unbiased run for the 2D MEMENTO free energy surface (FES),
2. frames extracted from the IF unbiased run at equal CV spacing for the 2D SMD FES,
3. a MEMENTO run with 16 windows between the coordinates of windows 10 and 12 of MEMENTO rep 3, with higher force constant $2 \times 10^7\,\mathrm{kJ\,mol^{-1}}$, for the 2D-MEMENTO-FES
4. a MEMENTO run with 16 windows between the coordinates of windows 9 and 12 of MEMENTO rep 2, with higher force constant $2 \times 10^7\,\mathrm{kJ\,mol^{-1}}$, for the 2D-MEMENTO-FES.

We detail the amount of sampling collected on LeuT in table A.3.

B

Detailed methods for chapter 4

B.1 Data availability

Simulation and experimental data produced in this work is available at doi: 10.5281/zenodo.10561418. This includes key coordinate files and python scripts, as well as simulation trajectories projected onto the CVs of interest in plumed output format, and relevant processed files for PMF, pKa and ABFE calculations. Full trajectories will be shared upon reasonable request. The PyMEMENTO software is freely available at https://github.com/simonlichtinger/PyMEMENTO.

B.2 Definition of tip and base bundle CVs

For the analysis and interpretation of our unbiased MD runs, as well as for the use as a CV for initial metadynamics and umbrella sampling trials, we constructed the tip-CV and base-CV as centre-of-mass distances between the Cα atoms of the top and bottom 11 residues of the N-terminal and C-terminus bundles respectively, as illustrated in figure 4.5a. These residues were picked as the consensus of the DSSP analysis (Kabsch and Sander, 1983) of the PepT2 conformational states derived below, listed in table B.1.

Bundle	Residue numbers
N-terminal bundle tips	63–74, 79–90, 120–131, 143–154, 194–205, 217–228
N-terminal bundle bases	46–57, 93–104, 110–121, 161–172, 177–188, 227–238
C-terminal bundle tips	320–331, 341–352, 392–403, 609–620, 655–666, 671–682
C-terminal bundle bases	290–301, 359–370, 376–387, 626–637, 642–653, 686–697

Table B.1: Residue numbers used in the defintion of the tip-CV and base-CV.

B.3 MD setup and equilibration

We obtained protein coordinates from cryo-EM for the OF (7NQK) (Parker et al., 2021) and IF-partially-occluded (7PMY) (Killer et al., 2021) PepT2 conformations, as well as from alphafold 2 (Jumper et al., 2021) for the fully-open IF state. We used MODELLER (Šali and Blundell, 1993) to fit the rat PepT2 sequence (as used by Parker et al. (2021)) to the human PepT2 7PMY and alphafold models, using residues 43–409 (TM 1–9) linked as a continuous chain to residues 604–700 (TM 10–12), thereby truncating the extracellular domain. We scored 200 models with QMEANDisCo (Studer et al., 2020) and selected the highest-scoring protein model for embedding into a 3:1 POPE:POPG bilayer of target size $10 * 10$ nm (210/72 lipid molecules for IF and OCC, 218/72 for OF) with the CHARMM-GUI membrane builder. (Wu et al., 2014) We added ACE/NME capping residues using pymol and solvated the membrane system using GROMACS (Abraham et al., 2015) with approximately 21,000 solvent molecules (precise number varies between replicates) at a NaCl concentration of 0.15 M in an orthorhombic box of around $9.9 * 9.9 * 10.8$ nm side lengths. Topologies were generated using the AMBER ff14.SB (Maier et al., 2015) and slipids (Jämbeck and Lyubartsev, 2012) forcefields.

Using the GROMACS MD engine (Abraham et al., 2015) in versions 2021.3/2021.4 (the slight version discrepancy is because of different installations on two compute clusters we used), we energy-minimized the systems, assigned initial velocities, and equilibrated with Cα-atom restraints for 200 ps in the NVT ensemble with a leap-frog integrator (using the modified v-rescale thermostat with a stochastic term (Bussi et al., 2007) at 310 K throughout this chapter), then 1 ns in the NPT ensemble (with the berendsen barostat), followed by 20 ns of further Cα-restrained NPT equilibration (using the Parrinello-Rahman barostat (Parrinello and Rahman, 1981), as for all subsequent production runs).

We obtained Ala-Phe dipeptide-bound boxes by aligning the holo PepT1 cryo-EM structure (7PMW) onto our equilibrated PepT2 MD boxes, copying the ligand coordinates, and repeating the same equilibration protocol as before (where the peptide substrate Cα atoms were also restrained). The peptide ligand was parametrised using AMBER ff14.SB. (Maier et al., 2015)

B.4 Derivation of OF, OCC and IF conformational states

As shown in figure 4.2, the IF-partially-occluded structure (7PMY) does not behave well in MD (1 µs production runs from triplicate embeddings), since it either partially opens its extracellular gate (replicates 2–3) or partial helical unfolding in the intracellular gate occurs due to hydrophobic collapse (rep 1). This may be due to a variety of factors; one possibility is instability in the protein following the removal of the bound substrate in our simulations. In contrast (figure 4.2b), embedding replicates 1 and 3 of the alphafold IF state behave well. We picked the end-coordinates of replicate 3 as our IF state, due to the wider opening of its intracellular gate. We then sought to derive an OCC state via MD from replicate 1, see the paragraph below. We also note that embeddings from the OF cryo-EM structure (7NQK) remain stable in the OF conformation; we picked replicate 1 for our work.

To derive an OCC state from an IF box, we conducted 5 replicates of well-tempered metadynamics (Barducci et al., 2008) as implemented in PLUMED 2.7 (Tribello et al., 2014) along the base-bundle CV (see figure 4.5a), using 8 walkers, hill height 1 kJ mol^{-1} with sigma 0.022 nm deposited every 500 steps, and a bias factor of 100. From the resulting set of trajectories, we picked frames around the mark of 20 ns simulation time and a base-CV value of around 2.0 nm (we chose these values based on visual inspection of the trajectories, where we noticed that base-CV values significantly below 2.0 nm lead to artefacts such as partial unfolding of the ends of helices, as did continuing the metadynamics simulations for longer than necessary to obtain the desired states). We ran triplicate 100 ns-long unbiased MD from the obtained states for each of the five replicates, and found the OCC state obtained from the first replicate to be stably situated within the range of base-CV values observed in the OF-state trajectories. The micro-second-long unbiased MD runs as well as our 2D PMFs confirm that this protein conformation is a stable basin, and that different protonation states of key residues can drive its opening to either the IF or OF states. While this does not rule out the existence of different OCC states, it confirms the properties of the conformation we found as a functional OCC state.

B.5 Unbiased MD of the OCC state

Unbiased MD runs of the OCC state in different conditions were conducted in triplicates using the same simulation parameters as for the long Cα-restrained equilibrations described in the section on equilibration, but removing all restraints. The starting coordinates were — for the first replicate — the OCC state derived as described in the foregoing section, and the second and third replicates were initialised from the 500 ns and the final frame of the first replicate trajectory. Protonation-state changes and mutations were carried

out using PyMOL and GROMACS pdb2gmx independently for each replicate, followed by re-running the full equilibration protocol for all new boxes. Taken together, we conducted unbiased MD for 24 conditions, giving 72 µs of production sampling.

B.6 Metadynamics and steered MD

For our initial trials of enhanced sampling on PepT2 conformational changes — which showed hysteretic behaviour (see figure 4.8) — we attempted SMD and metadynamics for the OCC↔OF transition in the wild-type (WT), unprotonated state.

Two instances of 8-walker well-tempered metadynamics were run, starting from the OF and OCC states, biased along the tip-CV with hills of height $1\,\mathrm{kJ\,mol^{-1}}$ and sigma 0.0455 nm deposited every 500 steps, using a bias factor of 100. A harmonic flat-bottom restraint with boundaries of 2.0–3.0 nm and force constant $5 \times 10^4\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ was applied on the CV value. Sampling was run for 108 ns per walker for the simulations starting from OF, and 213 ns per walker starting from OCC.

To generate paths for umbrella sampling, SMD was run starting at OF towards OCC and vice versa with the heavy-atom RMSD to the target conformation as CV, using a harmonic potential centered to zero RMSD with a force constant sliding from 0 up to $2.5 \times 10^5\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ over 200 ns. The harmonic potential was then switched off over 2 ns, followed by 48 ns of unbiased MD. We then projected the SMD trajectories onto the tip-CV, picked 48 frames spaced equally along the CV and performed 1D REUS using a force constant of $3 \times 10^4\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$, for 92 ns per window for the OCC→OF derived boxes and 127 ns per window in the reverse direction. A total of 13.6 µs of MD time was thus expended on these trials.

B.7 MEMENTO path generation

We ran MEMENTO with 24 windows in triplicates for both the OCC↔OF and OCC↔IF transitions in different protein protonation and mutation states, and in the presence or absence of ligands. The apo and holo MEMENTO replicates were initialised from the 0 ns, 500 ns and 1000 ns frames of the first replicate of the 1 µs unbiased MD run for each conformational state (using always the unprotonated, WT condition, but apo/holo trajectories respectively). Protonation state changes and mutations were then carried out using the built-in functionality of PyMEMENTO, and equilibrated at each intermediate state for 90 ns. The total MD simulation time spent on equilibrations as part of the MEMENTO method across the 22 sampled conditions was 47 µs.

B.8 1D PMF calculations

Starting with the triplicate equilibrated MEMENTO boxes for the (all apo) OCC↔OF standard protonation and H87 & D342 protonated states, as well as OCC↔IF standard protonation and E53 protonated states, we ran 1D REUS with exchange every 1000 steps using PLUMED 2. (Tribello et al., 2014) The CVs were the tip- and base-CVs respectively, using a force constant of $4 \times 10^3\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$. The amount of sampling collected in each case is summarised in table B.2.

Condition	Simulation time / ns
OCC↔OF, standard prot	24 * (266 + 244 + 244)
OCC↔OF, H87 & D342 prot	24 * (327 + 244 + 244)
OCC↔IF, standard prot	24 * (242 + 248 + 246)
OCC↔IF, E53 prot	24 * (158 + 155 + 154)
Total	67 µs

Table B.2: Overview of all 1D PMF sampling.

B.9 2D CV derivation

Using the trajectory data from our 1D PMFs, we derived 2D CVs via a principle component analysis (PCA)-based approach we have previously described for LEUT in chapter 3. In short, we pooled all sampling collected in 1D REUS runs along the tip-CV for apo OCC↔OF (and equivalently for OCC↔IF along the base-CV; the same procedure was taken for these trajectories, and we will only explicitly write about OCC↔OF in the following paragraph). Using GROMACS tools, we ran PCA of the Cα positions of residues contained in the transmembrane region (see table B.1 above). The first PC accounts for 50% of the variance; adding an extra 15 CVs increases coverage to 78% of the variance (comparable to our results on LEUT). The first PC (see figure 4.10) describes the gating motions of the respective conformational changes, behaving similarly to the tip CV (or base CV) — expectedly so, given it was the CV used in our 1D REUS. To explain differences between replicates (see figure 4.9a and c), we used differential evolution (Storn and Price, 1997) as implemented in scipy (Virtanen et al., 2020) to maximise an entropy-like metric of distances between MEMENTO path frames for linear combinations of the PCs 2–16:

$$\frac{2}{N_{\text{rep}} * (N_{\text{rep}} - 1)} \sum_{i=1}^{N_{\text{rep}}} \sum_{j=i+1}^{N_{\text{rep}}} \sum_{n}^{N_{\text{windows}}} log(\sqrt{\sum_{\text{PC}} (\mathbf{X}(n,i) - \mathbf{X}(n,j))^2}), \tag{B.1}$$

where N_{rep} is the number of replicates and by $\mathbf{X}(n,i) - \mathbf{X}(n,j)$ we denote the distance between two conformational frames in different replicates i and j, evaluated in a projection along a given combination of principal components. The result was termed PC 2 henceforth for simplicity and used as the second CV in 2D REUS (figure 4.10).

B.10 2D PMF calculations

Using the same equilibrated MEMENTO paths as above and the 2D CVs we derived, we calculated 2D PMFs of the OCC↔OF and OCC↔IF transitions in several protonation / mutation states, with and without Ala-Phe substrate bound, using 2D REUS. As shown in figure 4.16, we found that a lower force constant of 2×10^6 kJ mol^{-1} nm^{-2} leads to good histogram overlap in the lower-lying regions of the PMF, but has poor overlap near the transition region. In turn, a higher force constant of 1×10^7 kJ mol^{-1} nm^{-2} gives good window overlap in the transition region while not sampling broadly enough in large basins. Therefore, for each condition and MEMENTO replicate, we ran windows at both force constants and included them all in the WHAM analysis, thus ensuring sufficient sampling through the CV space. Replica-exchange was run within the 24 windows corresponding to each MEMENTO replicate–force constant combination, and a total of 144 windows contribute to each 2D PMF.

The sampling collected in all conditions is detailed in table B.3, aiming for between 180 and 210 ns per window though exact amounts differ with heterogeneous hardware and slightly different box sizes.

Condition	Simulation time / ns
OCC↔OF, standard prot	24 * (233 + 195 + 195 + 208 + 208 + 207)
OCC↔OF, H87 prot	24 * (230 + 200 + 188 + 201 + 228 + 229)
OCC↔OF, D342 prot	24 * (204 + 232 + 204 + 224 + 229 + 207)
OCC↔OF, H87&D342 prot	24 * (196 + 196 + 195 + 194 + 194 + 201)
OCC↔OF, D342A	24 * (182 + 183 + 206 + 181 + 183 + 183)
OCC↔OF, H87A	24 * (176 + 182 + 183 + 184 + 182 + 177)
OCC↔OF, E53 prot	24 * (185 + 180 + 209 + 179 + 218 + 213)
OCC↔OF, E56 prot	24 * (194 + 194 + 194 + 192 + 193 + 193)
OCC↔OF, R206D&D342R	24 * (227 + 213 + 183 + 222 + 211 + 180)
OCC↔OF, holo, standard prot	24 * (200 + 196 + 195 + 200 + 194 + 201)
OCC↔OF, holo, E53 prot	24 * (185 + 180 + 210 + 179 + 218 + 213)
OCC↔OF, holo, E56 prot	24 * (194 + 194 + 194 + 192 + 193 + 193)
OCC↔OF, holo, H87&D342 prot	24 * (198 + 191 + 198 + 190 + 182 + 201)
OCC↔IF, standard prot	24 * (197 + 216 + 203 + 192 + 197 + 199)
OCC↔IF, E53 prot	24 * (197 + 208 + 199 + 192 + 193 + 195)
OCC↔IF, E622 prot	24 * (199 + 201 + 200 + 196 + 217 + 197)
OCC↔IF, E53&E622 prot	24 * (193 + 237 + 240 + 187 + 250 + 200)
OCC↔IF, holo, standard prot	24 * (190 + 195 + 191 + 196 + 192 + 197)
OCC↔IF, holo, E53 prot	24 * (189 + 196 + 197 + 199 + 189 + 198)
OCC↔IF, holo, E622 prot	24 * (194 + 191 + 206 + 197 + 198 + 199)
OCC↔IF, holo, E53&E622 prot	24 * (197 + 198 + 180 + 195 + 190 + 204)
Total	598 µs

Table B.3: Overview of all 2D PMF sampling.

B.11 Absolute binding free energies

To probe the affinity of the Ala-Phe substrate to the PepT2 OF and IF conformations in different protonation states, we conducted ABFE simulations (Aldeghi et al., 2016, 2018) in GROMACS, using an equilibrium approach. For this, we changed our mdp files to use the stochastic dynamics integrator (doubling as a thermostat) and set the relevant free-energy flags, including soft-core van-der-Waals interactions (alpha = 0.5, power = 1, sigma = 0.3) and the couple-intramol=yes flag for consistency with larger ligands in other work. Our lambda-protocol was to first add Boresch restraints (Boresch et al., 2003) (for the complex thermodynamic leg only, ligand side was calculated using the analytic formula; through values 0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0), then annihilate coulumb interactions (even 0.1 spacings) followed by vdw interactions (even 0.05 spacings). We equilibrated for 200 ps of NVT and 1 ns of NPT at each lambda window, and ran production simulations with replica-exchange attempts every 1000 steps for 30 ns per window on the complex thermodynamic leg and 100 ns on the ligand-only leg.

The Boresch restraints for ABFE simulations were obtained using MDRestraintsGenerator (Alibay et al., 2022) by running the restraint finding algorithm over the ≈200 ns 2D REUS trajectory at the relevant conformation and protonation state, and we used the trajectory frame closest to the restraint centre as input for subsequent ABFE. This was done for each of the triplicate MEMENTO runs, giving 3 candidate ligand binding poses. For the pose that was found to have the highest single-replicate affinity, four replicates of unbiased, restraint-free 200 ns-long equilibrations were also started from the frames, processed with MDRestraintsGenerator and used to make replicates for ABFE runs to give an error estimate as mean ±

standard deviation. By following this protocol for the OF, OF E56 prot, IF and IF E622 prot conditions, we sampled for a total of 4 conditions $*$ 7 boxes per condition $*$ 44 windows $*$ 31.2 ns = $38\,\mu s$.

B.12 Constant-pH MD

To probe the substrate-dependence of the E53 and E56 pKa values, we ran CpHMD using the hybrid solvent approach with discrete protonation states (Swails et al., 2014) as implemented in the AMBER software. (D.A. Case et al., 2021) We took the MEMENTO starting frames from above as triplicate initial coordinates of OF and OCC states in the presence and absence of Ala-Phe. We then used tleap and in-house scripts to convert our boxes to the AMBER constant pH forcefield fork for protein, substrate and solutes and to the lipid21 force field (Dickson et al., 2022) for the membrane. We prepared CpHMD simulations as in the tutorial by Swails (2014), and ran them for $1\,\mu s$ at 8 pH replica windows (ph 0–7), in the NVT ensemble with a langevin thermostat (at $310\,K$ as before), attempting protonation state changes every 100 steps, running 100 steps of relaxation dynamics for every exchange, and attempting replica exchange every 1000 steps. Analysis was performed using the cphstats programme and in-house scripts for fitting titration curves. In figure 4.22, analysis is performed per replicate, reporting mean $\pm$ standard deviation for each condition and residue. In figures 4.23–4.24 we show pKa values estimated over simulation time from $10\,ns$ chunks of all CpHMD runs. We also analyse this data in terms of histograms of the chunk-estimated pKa values, pooling all data for each condition and residue.

A total of 4 conditions $*$ 3 replicates $*$ 8 windows $*$ 1 μs = $96\,\mu s$ of sampling were thus collected.

B.13 Cell-based transport assays

Transport assays were carried out using a modified version of the protocol by Parker et al. (2021). Hela cells were cultured in DMEM + GlutaMAX medium, supplemented with 10% FBS. 12-well plates were prepared by seeding 9×10^4 cells per well in $1\,mL$ of medium, and transfected after 24h with $0.8\,\mu g$ of PepT2 constructs in pEF5-FRT-eGFP vector (or $0.5\,\mu g$ of insert vector + $0.3\,\mu g$ of empty vector, where specified), with $1.6\,\mu g$ of fugene transfection reagent. The medium was exchanged 24h post transfection, and assays were carried out 40h post transfection. Cells were washed 2 times with $\approx 0.6\,mL$ of assay buffer (20 mM HEPES pH 7.5, 120 mM NaCl, 2 mM $MgSO_4$ and 25 mM glucose), then incubated with $0.3\,mL$ assay buffer containing 20 mM β-ala-lys-AMCA substrate for 15 min. The cells were then washed 3 times with $\approx 0.6\,mL$ of assay buffer, and incubated with $0.25\,mL$ of lysis buffer (20 mM Trist pH 7.5 + 0.2% Triton x-100) for 5 min. The fluorescence ($340\,nm$ excitation, $460\,nm$ read-out) of $0.15\,mL$ of the lysate was normalised by the protein amount in each well (as determined from BCA assay of $20\,\mu L$ lysate). We removed two outliers from the WT ($0.8\,\mu g$) transport assay dataset, giving n=46. The data was then scaled to the mean WT ($0.8\,\mu g$) transport level as 100%.

B.14 Protein expression controls

For comparing PepT2 WT and mutant expression levels, Hela cells were seeded in 6-well plates at 1.8×10^5 cells per well in $2\,mL$ of medium. Transfection was after 24h with $1.6\,\mu g$ of PepT2-constructs in pEF5-FRT-eGFP vector (or $1.0\,\mu g$ of insert vector + $0.6\,\mu g$ of empty vector, where specified) and $3.2\,\mu g$ of fugene; the medium was exchanged 24h post transfection. The cells were washed 3 times with $\approx 0.6\,mL$ of PBS, harvested using 0.1% trypsin, pelleted, re-suspended in $100\,\mu L$ PBS with protease inhibitor and lysed through 3x freeze-thawing. The lysates from 3 wells were pooled for each mutant to increase between-sample consistency. $4.5\,\mu L$ of each sample were loaded onto a 10% SDS-PAGE gel, and western

blot was performed using an anti-GFP antibody. The membrane was then stripped and developed again with an anti-β-actin antibody to control for gel loading.

B.15 Protein localisation controls

To confirm the plasma membrane localisation of PepT2 WT and mutants, we seeded 1.8×10^5 Hela cells per well in 2 mL of medium in a 6-well plate with added coverslips. Transfection was — as for the expression controls — 24h after seeding. 20h post transfection, the cells were washed 3 times with $\approx$0.6 mL of PBS, fixed with PFA for 10 min at room temperature, washed 3 times, incubated with PBS + 50 mM NH_4Cl for 10 min, washed 3 times, incubated with PBS + 0.1 % Triton x-100 for 5 min, washed 3 times, stained with DAPI, washed 5 times, and mounted on slides with ImmuMount. Images were recorded in the GFP and DAPI channels.

C

Detailed methods for chapter 5

C.1 Unbiased MD

The protein models from cryo-EM were patched at the missing extracellular domain (ECD) loop using MODELLER (Šali and Blundell, 1993), including residues 43–409 and 604–700 as a continuous chain in the processed model (as done in chapter 4). We scored 200 models with QMEANDisCo (Studer et al., 2020) and selected the highest-scoring protein model for embedding into a 3:1 POPE:POPG bilayer of target size $10 * 10$ nm (213/71 lipid molecules) with the CHARMM-GUI membrane builder. (Wu et al., 2014) Terminal groups were patched with ACE/NME residues in pymol, and the E53 / E56 residue protonation states were assigned as required in the GROMACS 2021 (Abraham et al., 2015) pbd2gmx tool. The protein and lipids were parametrised in the AMBER ff.14sb (Maier et al., 2015) and slipids (Jämbeck and Lyubartsev, 2012) forcefields respectively. The ligand parameters were obtained using the OpenFF 2.0 (SAGE) forcefield. (Boothroyd et al., 2023) Gromacs was then used to solvate the system with 0.15 M NaCl and 23,000 TIP3P water molecules. The box size after energy minimization was $9.9 * 9.9 * 11.3$ nm. From these boxes, we ran six replicates of equilibration and production MD at temperature 310 K and pressure 1 bar. The thermostat was v-rescale (Bussi et al., 2007) (separate temperature coupling groups for membrane and solvent); the barostat was berendsen for equilibration and Parrinello-Rahman for production. (Parrinello and Rahman, 1980) The equilibration protocol was: 200 ps NVT, 1 ns NPT (Cα restraints and ligand heavy atoms), 20 ns NPT (Cα restraints only). The production simulations without restraints were then run for 1 µs.

C.2 ABFE simulations

We ran ABFE simulations (see section 2.3) for cefadroxil, amoxicillin and cloxacillin (pose 1 only) binding in the E56 unprotonated and protonated protein conditions. We derived Boresch restraints (Boresch et al., 2003) from the last 200 ns of each of the 6-replicate unbiased simulations and picked the frame closest to the restraint centre as ABFE starting frames, using MDRestraintsGenerator. (Alibay et al., 2022) For cefadroxil and amoxicillin, we further made 4 replicate starting poses from the unbiased replicate with the highest affinity by running 4 times additional 200 ns-long equilibrations and deriving new sets of Boresch restraints from them (we report mean $\pm$ standard deviation among these 5 replicates). For cloxacillin, we report the (single) highest binding affinity of the unbiased E56 unprotonated and protonated replicates, respectively. Our lambda-protocol was to first add Boresch restraints (for the complex thermodynamic leg only, ligand side was calculated using the analytic formula; through values 0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0), then annihilate coulumb interactions (even 0.1 spacings) followed by vdw interactions (even 0.05 spacings). This gives 44 windows for the complex and 31 windows for the ligand sides of the cycle. At each lambda window, ran energy minimization, 200 ps NVT equilibration (310 K, stochastic dynamics integrator), 1 ns NPT equilibration (310 K, stochastic dynamics integrator, 1 bar, berendsen), then 30 ns (complex windows) / 100 ns (ligand windows) production (Parrinello-Rahman) with replica exchange. We analysed the simulations using alchemlyb (https://github.com/alchemistry/alchemlyb) with the mbar esitmator. (Shirts and Chodera, 2008)

D

Detailed methods for chapter 6

D.1 Reference dataset preparation

The dataset contained in a PDF version of Biegel et al. (2006) was parsed, annotating the ligand names with SMILES strings semi-manually with in-house scripts. Amino and carboxyl chemical groups were treated as charged by default. Of affinities reported at PepT2, the data presented for [^{14}C]Gly-Sar uptake inhibition as K_i values at SKPT cell monolayers was most comprehensive. The dataset encompasses 184 compounds, including 51 dipeptides with natural amino acids, 30 tripeptides with natural amino acids, 61 peptide derivatives with modified backbone and/or side chains, as well as 35 β-lactam and 7 non-β-lactam drugs. In a few cases where several affinity values were reported from separate studies, we opted for the lower K_i by default. We used rdkit 2022.3.5 (`https://www.rdkit.org`) to generate 3D conformers from the SMILES strings.

D.2 Ensemble docking

We extracted protein configurations from triplicate 1 µs-long unbiased MD runs of OF PepT2 without ligand as well as with bound Ala-Phe and cefadroxil, as previously prepared (see appendices B and C for details), using frames spaced by 200 ns and disregarding the first 200 ns as equilibration. This yielded an ensemble of 45 protein configurations. We used the GOLD docking software (version 2020.3) to dock the ligand set into all receptor structures, defining the binding site as a sphere of radius 1.5 nm around the phenolic oxygen atom of Y61 and treating the side chains of R57, Y61, Y94, D317 and E622 as fully flexible. Up to 100 poses were generated per protein conformation (allowing for early termination with default settings), and the default ChemPLP fitness function was used.

For each ligand, we extracted both the highest overall docking score, and the highest score after filtering the docked poses by canonical binding site interactions. For this, we used rdkit to identify formally charged atoms in the ligands. We defined an additive score, where a positive ligand atom within a cutoff of 4 Å of E622 counts 2 and next to N192 and N348 counts 1, and a negative atom next to R57 counts 2 as well. The highest filtered docking score is then the highest docking score of those docked poses which have the highest filtered score; the procedure slightly improved correlation with experimental data (see figure 6.4) by removing outliers that had high docking scores in non-canonical binding poses.

D.3 ABFE calculations

For each docked ligand, we converted the docked pose with the highest filtered score back to a simulation box, by integrating the ligand position and side-chain rearrangements of flexible residues into the membrane-embedded protein boxes from which the respective protein configuration was taken. As before, the protein was parameterised using AMBER 14.sb (Maier et al., 2015), the lipids using slipids (Jämbeck and Lyubartsev, 2012), and solvation with TIP3P water molecules was performed using GROMACS (Abraham et al., 2015) tools. The ligands were parameterised using the OpenFF toolkit and the OpenFF 2.0 (SAGE)

force field. (Boothroyd et al., 2023)

For each pose thus embedded, we ran 200 ps NVT and 1 ns NPT equilibrations with Cα-position restraints followed by 50 ns-long unbiased MD (setup details were the same as described in appendix B). From these trajectories, we derived Boresch restraints using MDRestraintsGenerator (Alibay et al., 2022), picked the frames closest to the restraint centers and followed the same ABFE protocol as above (appendix C), except for shortening the simulation time per window to 10 ns for the complex windows and 50 ns for the ligand windows. Where required, we introduced an appropriate number of alchemical co-ions in a charge-transfer procedure, as suggested by Wu and Biggin (2022) for membrane proteins. No replicates were run due to constraints on computational time. Out of the 184 ligands, for 4 ligands the Boresch restraints were too numerically unstable to compute ABFE values (indicating particularly unstable binding poses); we removed these from the dataset to yield 180 ligands included in our analysis. The total sampling expended on the complex boxes in this way was 90 µs.

For the penicillin G and valine-penicillin G ligands, we adopted the same docking and ABFE protocols, except for also running ABFEs for the second- and third-best filtered docked poses, as it was feasible for the smaller number of ligands in this case. We report the highest ΔG value among the tested poses.

D.4 Random forest models

Using the dataset of 184 compounds described above, we derived 3 sets of features with rdkit: (1) a set of all 64 implemented low-dimensional molecular descriptors, (2) 1024-bit Morgan fingerprints (with 2 iterations), and (3) 2048-bit 2- and 3-point 2d-pharmacophores (with default rdkit feature definitions and 0–2, 2–5 and 5–8 bond bins). We split the dataset using a k-fold cross-validation procedure with a random 2:1 training:test split, running 5 repeats to give us mean $\pm$ standard deviation for each compound. For the training of random forest models, we used the implementation in scikit-learn (Pedregosa et al., 2011), version 1.2.2. We optimised hyperparameters through 2-dimensional grid search (see figure 6.7), where we kept the minimum number of samples per split constant at 2 and the minimum number of samples per leaf at 1 to find that performance did not improve when using a number of estimators larger than 100 and a maximum depth of larger than 10, for all sets of features. Similarly, we found that at number of estimators 100 and maximum depth 10, performance degrades when increasing the number of samples per split and number of samples per leaf from their minimum values at 2 and 1, respectively. The number of features used in searching for a split is defined as the square root of the total number of features; in initial trials we found no benefit in moving towards a higher number of included features, and degraded performance in going significantly lower.

For the interpretation of Morgan fingerprint bits picked out as important by our model, we used built-in rdkit functionality to print the first instances of a given fingerprint bit being triggered by a molecular fragment in our dataset. To interpret the 2D-pharmacophore bits, we picked the Asp-Phe dipeptide as an example and show, using an in-house script, which atoms contribute to a given pharmacophore bit being triggered.

Bibliography

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. SoftwareX. 2015 Sep; 1–2:19–25. doi: 10.1016/j.softx.2015.06.001.

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung CC, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Žemgulytė A, Arvaniti E, Beattie C, et al. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. Nature. 2024 May; doi: 10.1038/s41586-024-07487-w.

Aduri NG, Montefiori M, Khalil R, Gajhede M, Jørgensen FS, Mirza O. Molecular Dynamics Simulations Reveal the Proton:Peptide Coupling Mechanism in the Bacterial Proton-Coupled Oligopeptide Transporter YbgH. ACS Omega. 2019 Jan; 4(1). doi: 10.1021/acsomega.8b02131.

Aduri NG, Prabhala BK, Ernst HA, Jørgensen FS, Olsen L, Mirza O. Salt Bridge Swapping in the EXXERFXYY Motif of Proton-coupled Oligopeptide Transporters. Journal of Biological Chemistry. 2015 Dec; 290(50). doi: 10.1074/jbc.M115.675603.

Ahmad K, Rizzi A, Capelli R, Mandelli D, Lyu W, Carloni P. Enhanced-Sampling Simulations for the Estimation of Ligand Binding Kinetics: Current Status and Perspective. Frontiers in Molecular Biosciences. 2022 Jun; 9:899805. doi: 10.3389/fmolb.2022.899805.

Aho N, Buslaev P, Jansen A, Bauer P, Groenhof G, Hess B. Scalable Constant pH Molecular Dynamics in GROMACS. Journal of Chemical Theory and Computation. 2022 Oct; 18(10):6148–6160. doi: 10.1021/acs.jctc.2c00516.

Aldeghi M, Bluck JP, Biggin PC. Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide. In: *Computational Drug Discovery and Design, Methods in Molecular Biology*, vol. 1762; 2018.doi: 10.1007/978-1-4939-7756-7_11.

Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Accurate Calculation of the Absolute Free Energy of Binding for Drug Molecules. Chemical Science. 2016; 7(1):207–218. doi: 10.1039/C5SC02678D.

Alder BJ, Wainwright TE. Phase Transition for a Hard Sphere System. The Journal of Chemical Physics. 1957 Nov; 27(5):1208–1209. doi: 10.1063/1.1743957.

Alibay I, Magarkar A, Seeliger D, Biggin PC. Evaluating the Use of Absolute Binding Free Energy in the Fragment Optimisation Process. Communications Chemistry. 2022 Sep; 5(1):105. doi: 10.1038/s42004-022-00721-4.

Alqahtani MS, Kazi M, Alsenaidy MA, Ahmad MZ. Advances in Oral Drug Delivery. Frontiers in Pharmacology. 2021 Feb; 12:618411. doi: 10.3389/fphar.2021.618411.

Amaro RE, Baudry J, Chodera J, Demir Ö, McCammon JA, Miao Y, Smith JC. Ensemble Docking in Drug Discovery. Biophysical Journal. 2018 May; 114(10):2271–2278. doi: 10.1016/j.bpj.2018.02.038.

Andersen HC. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. The Journal of Chemical Physics. 1980 Feb; 72(4):2384–2393. doi: 10.1063/1.439486.

Apostolakis J, Ferrara P, Caflisch A. Calculation of Conformational Transitions and Barriers in Solvated Systems: Application to the Alanine Dipeptide in Water. The Journal of Chemical Physics. 1999 Jan; 110(4):2099–2108. doi: 10.1063/1.477819.

Arakawa H, Yamada H, Arai K, Kawanishi T, Nitta N, Shibata S, Matsumoto E, Yano K, Kato Y, Kumamoto T, Aoki I, Ogihara T. Possible Utility of Peptide-Transporter-Targeting [19F]Dipeptides for Visualization of the Biodistribution of Cancers by Nuclear Magnetic Resonance Imaging. International Journal of Pharmaceutics. 2020; 586(May):119575. doi: 10.1016/j.ijpharm.2020.119575.

Arora K, Brooks CL. Large-Scale Allosteric Conformational Transitions of Adenylate Kinase Appear to Involve a Population-Shift Mechanism. Proceedings of the National Academy of Sciences. 2007 Nov; 104(47):18496–18501. doi: 10.1073/pnas.0706443104.

Bailey PD, Boyd CAR, Bronk JR, Collier ID, Meredith D, Morgan KM, Temple CS. How to Make Drugs Orally Active: A Substrate Template for Peptide Transporter PepT1. Angewandte Chemie International Edition. 2000; 39(3):505–508. doi: 10.1002/(SICI)1521-3773(20000204)39:3<505::AID-ANIE505>3.0.CO;2-B.

Barducci A, Bussi G, Parrinello M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. Physical Review Letters. 2008 Jan; 100(2):020603. doi: 10.1103/PhysRevLett.100.020603.

Batista MRB, Watts A, José Costa-Filho A. Exploring Conformational Transitions and Free-Energy Profiles of Proton-Coupled Oligopeptide Transporters. Journal of Chemical Theory and Computation. 2019 Nov; 15(11). doi: 10.1021/acs.jctc.9b00524.

Bayly CI, Cieplak P, Cornell W, Kollman PA. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. The Journal of Physical Chemistry. 1993 Oct; 97(40). doi: 10.1021/j100142a004.

Beale JH, Parker JL, Samsudin F, Barrett AL, Senan A, Bird LE, Scott D, Owens RJ, Sansom MSP, Tucker SJ, Meredith D, Fowler PW, Newstead S. Crystal Structures of the Extracellular Domain from PepT1 and PepT2 Provide Novel Insights into Mammalian Peptide Transport. Structure. 2015 Oct; 23(10). doi: 10.1016/j.str.2015.07.016.

Beauchamp LM, Orr GF, De Miranda P, Burnette T, Krenitsky TA. Amino Acid Ester Prodrugs of Acyclovir. Antiviral Chemistry and Chemotherapy. 1992; 3(3):157–164. doi: 10.1177/095632029200300305.

Beckstein O, Becksteinlab/GromacsWrapper: Release 0.8.2, Tempe, AZ; 2021. doi: 10.5281/ZENODO.5498364. Zenodo.

Beckstein O, Denning EJ, Perilla JR, Woolf TB. Zipping and Unzipping of Adenylate Kinase: Atomistic Insights into the Ensemble of Open-Closed Transitions. Journal of Molecular Biology. 2009 Nov; 394(1):160–176. doi: 10.1016/j.jmb.2009.09.009.

Beckstein O, Naughton F. General Principles of Secondary Active Transporter Function. Biophysics Reviews. 2022 Mar; 3(1):011307. doi: 10.1063/5.0047967.

Bender BJ, Gahbauer S, Luttens A, Lyu J, Webb CM, Stein RM, Fink EA, Balius TE, Carlsson J, Irwin JJ, Shoichet BK. A Practical Guide to Large-Scale Docking. Nature Protocols. 2021 Oct; 16(10):4799–4832. doi: 10.1038/s41596-021-00597-z.

Bennett CH. Efficient Estimation of Free Energy Differences from Monte Carlo Data. Journal of Computational Physics. 1976 Oct; 22(2). doi: 10.1016/0021-9991(76)90078-4.

Berendsen HJC, Postma JPM, Van Gunsteren WF, DiNola A, Haak JR. Molecular Dynamics with Coupling to an External Bath. The Journal of Chemical Physics. 1984 Oct; 81(8):3684–3690. doi: 10.1063/1.448118.

Beutler TC, Mark AE, Van Schaik RC, Gerber PR, Van Gunsteren WF. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. Chemical Physics Letters. 1994 Jun; 222(6):529–539. doi: 10.1016/0009-2614(94)00397-1.

Bhakat S. Collective Variable Discovery in the Age of Machine Learning: Reality, Hype and Everything in Between. RSC Advances. 2022; 12(38):25010–25024. doi: 10.1039/D2RA03660F.

Biegel A, Knütter I, Hartrodt B, Gebauer S, Theis S, Luckner P, Kottra G, Rastetter M, Zebisch K, Thondorf I, Daniel H, Neubert K, Brandsch M. The Renal Type H+/Peptide Symporter PEPT2: Structure-Affinity Relationships. Amino Acids. 2006 Aug; 31(2). doi: 10.1007/s00726-006-0331-0.

Biegel A, Gebauer S, Hartrodt B, Brandsch M, Neubert K, Thondorf I. Three-Dimensional Quantitative Structure-Activity Relationship Analyses of β-Lactam Antibiotics and Tripeptides as Substrates of the Mammalian H^+/Peptide Cotransporter PEPT1. Journal of Medicinal Chemistry. 2005 Jun; 48(13):4410–4419. doi: 10.1021/jm048982w.

Bleiziffer P, Schaller K, Riniker S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. Journal of Chemical Information and Modeling. 2018 Mar; 58(3):579–590. doi: 10.1021/acs.jcim.7b00663.

Boggavarapu R, Jeckelmann JM, Harder D, Ucurum Z, Fotiadis D. Role of Electrostatic Interactions for Ligand Recognition and Specificity of Peptide Transporters. BMC Biology. 2015 Dec; 13(1). doi: 10.1186/s12915-015-0167-8.

Boothroyd S, Behara PK, Madin OC, Hahn DF, Jang H, Gapsys V, Wagner JR, Horton JT, Dotson DL, Thompson MW, Maat J, Gokey T, Wang LP, Cole DJ, Gilson MK, Chodera JD, Bayly CI, Shirts MR, Mobley DL. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. Journal of Chemical Theory and Computation. 2023 Jun; 19(11):3251–3275. doi: 10.1021/acs.jctc.3c00039.

Boresch S, Tettinger F, Leitgeb M, Karplus M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. The Journal of Physical Chemistry B. 2003 Sep; 107(35). doi: 10.1021/jp0217839.

Brandsch M. Drug Transport via the Intestinal Peptide Transporter PepT1. Current Opinion in Pharmacology. 2013 Dec; 13(6). doi: 10.1016/j.coph.2013.08.004.

Brandsch M, Knütter I, Thunecke F, Hartrodt B, Born I, Börner V, Hirche F, Fischer G, Neubert K. Decisive Structural Determinants for the Interaction of Proline Derivatives with the Intestinal H^+/Peptide Symporter. European Journal of Biochemistry. 1999 Dec; 266(2):502–508. doi: 10.1046/j.1432-1327.1999.00885.x.

Brandsch M, Thunecke F, Küllertz G, Schutkowski M, Fischer G, Neubert K. Evidence for the Absolute Conformational Specificity of the Intestinal H+/Peptide Symporter, PEPT1. Journal of Biological Chemistry. 1998 Feb; 273(7):3861–3864. doi: 10.1074/jbc.273.7.3861.

Branduardi D, Gervasio FL, Parrinello M. From A to B in Free Energy Space. The Journal of Chemical Physics. 2007 Feb; 126(5). doi: 10.1063/1.2432340.

Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. doi: 10.1023/A:1010933404324.

Bryant P, Kelkar A, Guljas A, Clementi C, Noé F. Structure Prediction of Protein-Ligand Complexes from Sequence Information with Umol. Nature Communications. 2024 May; 15(1):4536. doi: 10.1038/s41467-024-48837-6.

Buslaev P, Aho N, Jansen A, Bauer P, Hess B, Groenhof G. Best Practices in Constant pH MD Simulations: Accuracy and Sampling. Journal of Chemical Theory and Computation. 2022 Oct; 18(10):6134–6147. doi: 10.1021/acs.jctc.2c00517.

Bussi G, Donadio D, Parrinello M. Canonical Sampling through Velocity Rescaling. The Journal of Chemical Physics. 2007 Jan; 126(1):014101. doi: 10.1063/1.2408420.

Bussi G, Laio A. Using Metadynamics to Explore Complex Free-Energy Landscapes. Nature Reviews Physics. 2020 Apr; 2(4). doi: 10.1038/s42254-020-0153-0.

Casalino L, Dommer AC, Gaieb Z, Barros EP, Sztain T, Ahn SH, Trifan A, Brace A, Bogetti AT, Clyde A, Ma H, Lee H, Turilli M, Khalid S, Chong LT, Simmerling C, Hardy DJ, Maia JD, Phillips JC, Kurth T, et al. AI-driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics. The International Journal of High Performance Computing Applications. 2021 Sep; 35(5):432–451. doi: 10.1177/10943420211006452.

Castellana NE, Lushnikov A, Rotkiewicz P, Sefcovic N, Pevzner PA, Godzik A, Vyatkina K. MORPH-PRO: A Novel Algorithm and Web Server for Protein Morphing. Algorithms for Molecular Biology. 2013 Jan; 8(1):19. doi: 10.1186/1748-7188-8-19.

Chan MC, Selvam B, Young HJ, Procko E, Shukla D. The Substrate Import Mechanism of the Human Serotonin Transporter. Biophysical Journal. 2022 Mar; 121(5):715–730. doi: 10.1016/j.bpj.2022.01.024.

Chan MC, Shukla D. Markov State Modeling of Membrane Transport Proteins. Journal of Structural Biology. 2021 Dec; 213(4):107800. doi: 10.1016/j.jsb.2021.107800.

Chen M. Collective Variable-Based Enhanced Sampling and Machine Learning. The European Physical Journal B. 2021 Oct; 94(10):211. doi: 10.1140/epjb/s10051-021-00220-w.

Chen XZ, Steel A, Hediger MA. Functional Roles of Histidine and Tyrosine Residues in the H+-Peptide Transporter PepT1. Biochemical and Biophysical Research Communications. 2000; 272(3):726–730. doi: 10.1006/bbrc.2000.2851.

Chen XZ, Zhu T, Smith DE, Hediger MA. Stoichiometry and Kinetics of the High-affinity H+-Coupled Peptide Transporter PepT2. Journal of Biological Chemistry. 1999 Jan; 274(5). doi: 10.1074/jbc.274.5.2773.

Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical Free Energy Methods for Drug Discovery: Progress and Challenges. Current Opinion in Structural Biology. 2011 Apr; 21(2):150–160. doi: 10.1016/j.sbi.2011.01.011.

Colas C, Masuda M, Sugio K, Miyauchi S, Hu Y, Smith DE, Schlessinger A. Chemical Modulation of the Human Oligopeptide Transporter 1, hPepT1. Molecular Pharmaceutics. 2017 Dec; 14(12). doi: 10.1021/acs.molpharmaceut.7b00775.

Corey RA, Vickery ON, Sansom MSP, Stansfeld PJ. Insights into Membrane Protein-Lipid Interactions from Free Energy Calculations. Journal of Chemical Theory and Computation. 2019; 15(10):5727–5736. doi: 10.1021/acs.jctc.9b00548.

Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv. 2022; doi: 10.48550/ARXIV.2210.01776.

Cournia Z, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. Journal of Chemical Information and Modeling. 2017 Dec; 57(12). doi: 10.1021/acs.jcim.7b00564.

Cournia Z, Chipot C. Applications of Free-Energy Calculations to Biomolecular Processes. A Collection. Journal of Chemical Information and Modeling. 2024; 64:2129–2131. doi: 10.1021/acs.jcim.4c00349.

Crooks GE. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. Physical Review E. 1999 Sep; 60(3):2721–2726. doi: 10.1103/PhysRevE.60.2721.

D A Case, H M Aktulga, K Belfon, I Y Ben-Shalom, S R Brozell, D S Cerutti, T E Cheatham, III, G A Cisneros, V W D Cruzeiro, T A Darden, R E Duke, G Giambasu, M K Gilson, H Gohlke, A W Goetz, R , Harris, S Izadi, S A Izmailov, C Jin, K Kasavajhala, M C Kaymak, E King, A Kovalenko, T Kurtzman,, T S Lee, S LeGrand, P Li, C Lin, J Liu, T Luchko, R Luo, M Machado, V Man, M Manathunga, K M , Merz, Y Miao, O Mikhailovskii, G Monard, H Nguyen, K A O'Hearn, A Onufriev, F Pan, S Pantano,, R Qi, A Rahnamoun, D R Roe, A Roitberg, C Sagui, S Schott-Verdugo, J Shen, C L Simmerling, N R , Skrynnikov, J Smith, J Swails, R C Walker, J Wang, H Wei, R M Wolf, X Wu, Y Xue, D M York, S , Zhao, and P A Kollman, Amber 2021; 2021. University of California, San Francisco.

Dama JF, Parrinello M, Voth GA. Well-Tempered Metadynamics Converges Asymptotically. Physical Review Letters. 2014 Jun; 112(24). doi: 10.1103/PhysRevLett.112.240602.

Daniel H, Kottra G. The Proton Oligopeptide Cotransporter Family SLC15 in Physiology and Pharmacology. Pfluegers Archiv European Journal of Physiology. 2004 Feb; 447(5):610–618. doi: 10.1007/s00424-003-1101-4.

Daniel H, Spanier B, Kottra G, Weitz D. From Bacteria to Man: Archaic Proton-Dependent Peptide Transporters at Work. Physiology. 2006 Apr; 21(2). doi: 10.1152/physiol.00054.2005.

Darden T, York D, Pedersen L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems. The Journal of Chemical Physics. 1993 Jun; 98(12):10089–10092. doi: 10.1063/1.464397.

De Miguel E, Jackson G. The Nature of the Calculation of the Pressure in Molecular Simulations of Continuous Models from Volume Perturbations. The Journal of Chemical Physics. 2006 Oct; 125(16):164109. doi: 10.1063/1.2363381.

del Alamo D, Sala D, Mchaourab HS, Meiler J. Sampling Alternative Conformational States of Transporters and Receptors with AlphaFold2. eLife. 2022 Mar; 11:e75751. doi: 10.7554/eLife.75751.

Dellago C, Bolhuis PG. Transition Path Sampling Simulations of Biological Systems. In: Reiher M, editor. *Atomistic Approaches in Modern Biology*, vol. 268 Springer Berlin Heidelberg; 2007.p. 291–317. doi: 10.1007/128_085.

Dickson CJ, Madej BD, Skjevik ÅA, Betz RM, Teigen K, Gould IR, Walker RC. Lipid14: The Amber Lipid Force Field. Journal of Chemical Theory and Computation. 2014 Feb; 10(2):865–879. doi: 10.1021/ct4010307.

Dickson CJ, Walker RC, Gould IR. Lipid21: Complex Lipid Membrane Simulations with AMBER. Journal of Chemical Theory and Computation. 2022 Mar; 18(3):1726–1736. doi: 10.1021/acs.jctc.1c01217.

Doki S, Kato HE, Solcan N, Iwaki M, Koyama M, Hattori M, Iwase N, Tsukazaki T, Sugita Y, Kandori H, Newstead S, Ishitani R, Nureki O. Structural Basis for Dynamic Mechanism of Proton-Coupled Symport by the Peptide Transporter POT. Proceedings of the National Academy of Sciences. 2013 Jul; 110(28). doi: 10.1073/pnas.1301079110.

Döring F, Will J, Amasheh S, Clauss W, Ahlbrecht H, Daniel H. Minimal Molecular Determinants of Substrates for Recognition by the Intestinal Peptide Transporter. Journal of Biological Chemistry. 1998 Sep; 273(36). doi: 10.1074/jbc.273.36.23211.

Drew D, Boudker O. Shared Molecular Mechanisms of Membrane Transporters. Annual Review of Biochemistry. 2016 Jun; 85(1):543–572. doi: 10.1146/annurev-biochem-060815-014520.

Drew D, North RA, Nagarathinam K, Tanabe M. Structures and General Transport Mechanisms by the Major Facilitator Superfamily (MFS). Chemical Reviews. 2021 May; 121(9):5289–5335. doi: 10.1021/acs.chemrev.0c00983.

Durrant JD, Kochanek SE, Casalino L, Ieong PU, Dommer AC, Amaro RE. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. ACS Central Science. 2020 Feb; 6(2):189–196. doi: 10.1021/acscentsci.9b01071.

Dyla M, Kjærgaard M, Poulsen H, Nissen P. Structure and Mechanism of P-Type ATPase Ion Pumps. Annual Review of Biochemistry. 2020 Jun; 89(1):583–603. doi: 10.1146/annurev-biochem-010611-112801.

Ekins S, Johnston JS, Bahadduri P, D'Souza VM, Ray A, Chang C, Swaan PW. In Vitro and Pharmacophore-Based Discovery of Novel hPEPT1 Inhibitors. Pharmaceutical Research. 2005 Apr; 22(4):512–517. doi: 10.1007/s11095-005-2505-y.

Ewald PP. Die Berechnung Optischer Und Elektrostatischer Gitterpotentiale. Annalen der Physik. 1921 Jan; 369(3):253–287. doi: 10.1002/andp.19213690304.

Farrell DW, Speranskiy K, Thorpe MF. Generating Stereochemically Acceptable Protein Pathways. Proteins: Structure, Function, and Bioinformatics. 2010 Nov; 78(14):2908–2921. doi: 10.1002/prot.22810.

Fei YJ, Kanai Y, Nussberger S, Ganapathy V, Leibach FH, Romero MF, Singh SK, Boron WF, Hediger MA. Expression Cloning of a Mammalian Proton-Coupled Oligopeptide Transporter. Nature. 1994 Apr; 368(6471). doi: 10.1038/368563a0.

Fei YJ, Liu W, Prasad PD, Kekuda R, Oblak TG, Ganapathy V, Leibach FH. Identification of the Histidyl Residue Obligatory for the Catalytic Activity of the Human H^+/Peptide Cotransporters PEPT1 and PEPT2. Biochemistry. 1997 Jan; 36(2):452–460. doi: 10.1021/bi962058p.

Fei YJ, Nara E, Liu JC, Boyd CAR, Ganapathy V, Leibach FH. Preferential Recognition of Zwitterionic Dipeptides as Transportable Substrates by the High-Affinity Peptide Transporter PEPT2. Biochimica et Biophysica Acta (BBA) - Biomembranes. 1999 May; 1418(2):344–351. doi: 10.1016/S0005-2736(99)00046-2.

Fiorin G, Klein ML, Hénin J. Using Collective Variables to Drive Molecular Dynamics Simulations. Molecular Physics. 2013 Dec; 111(22-23). doi: 10.1080/00268976.2013.813594.

Foley D, Bailey P, Pieri M, Meredith D. Targeting Ketone Drugs towards Transport by the Intestinal Peptide Transporter, PepT1. Organic & Biomolecular Chemistry. 2009; 7(6):1064. doi: 10.1039/b818606e.

Foley DW, Bermudez I, Bailey PD, Meredith D. A Cyclosporine Derivative Is a Substrate of the Oligopeptide Transporter PepT1. MedChemComm. 2016; 7(5):999–1002. doi: 10.1039/C5MD00575B.

Foley DW, Pathak RB, Phillips TR, Wilson GL, Bailey PD, Pieri M, Senan A, Meredith D. Thiodipeptides Targeting the Intestinal Oligopeptide Transporter as a General Approach to Improving Oral Drug Delivery. European Journal of Medicinal Chemistry. 2018 Aug; 156. doi: 10.1016/j.ejmech.2018.06.064.

Formoso E, Limongelli V, Parrinello M. Energetics and Structural Characterization of the Large-Scale Functional Motion of Adenylate Kinase. Scientific Reports. 2015 Jul; 5(1):8425. doi: 10.1038/srep08425.

Forrest LR. (Pseudo-)Symmetrical Transport. Science. 2013 Jan; 339(6118):399–401. doi: 10.1126/science.1228465.

Forrest LR. Structural Symmetry in Membrane Proteins. Annual Review of Biophysics. 2015 Jun; 44(1):311–337. doi: 10.1146/annurev-biophys-051013-023008.

Forrest LR, Krämer R, Ziegler C. The Structural Basis of Secondary Active Transport Mechanisms. Biochimica et Biophysica Acta (BBA) - Bioenergetics. 2011 Feb; 1807(2):167–188. doi: 10.1016/j.bbabio.2010.10.014.

Forrest LR, Rudnick G. The Rocking Bundle: A Mechanism for Ion-Coupled Solute Flux by Symmetrical Transporters. Physiology. 2009 Dec; 24(6):377–386. doi: 10.1152/physiol.00030.2009.

Forrest LR, Zhang YW, Jacobs MT, Gesmonde J, Xie L, Honig BH, Rudnick G. Mechanism for Alternating Access in Neurotransmitter Transporters. Proceedings of the National Academy of Sciences. 2008 Jul; 105(30):10338–10343. doi: 10.1073/pnas.0804659105.

Fowler PW, Orwick-Rydmark M, Radestock S, Solcan N, Dijkman PM, Lyons JA, Kwok J, Caffrey M, Watts A, Forrest LR, Newstead S. Gating Topology of the Proton-Coupled Oligopeptide Symporters. Structure. 2015 Feb; 23(2). doi: 10.1016/j.str.2014.12.012.

Fujita T, Morishita Y, Ito H, Kuribayashi D, Yamamoto A, Muranishi S. Enhancement of the Small Intestinal Uptake of Phenylalanyl-Glycine via a H+/Oligopeptide Transport System by Chemical Modification with Fatty Acids. Life Sciences. 1997 Nov; 61(25):2455–2465. doi: 10.1016/S0024-3205(97)00980-6.

Furuta T. Structural Dynamics of ABC Transporters: Molecular Simulation Studies. Biochemical Society Transactions. 2021 Feb; 49(1):405–414. doi: 10.1042/BST20200710.

Fusani L, Palmer DS, Somers DO, Wall ID. Exploring Ligand Stability in Protein Crystal Structures Using Binding Pose Metadynamics. Journal of Chemical Information and Modeling. 2020; 60(3):1528–1539. doi: 10.1021/acs.jcim.9b00843.

Ganapathy ME, Huang W, Wang H, Ganapathy V, Leibach FH. Valacyclovir: A Substrate for the Intestinal and Renal Peptide Transporters PEPT1 and PEPT2. Biochemical and Biophysical Research Communications. 1998 May; 246(2). doi: 10.1006/bbrc.1998.8628.

Gapsys V, Pérez-Benito L, Aldeghi M, Seeliger D, van Vlijmen H, Tresadern G, de Groot BL. Large Scale Relative Protein Ligand Binding Affinities Using Non-Equilibrium Alchemy. Chemical Science. 2020; 11(4). doi: 10.1039/C9SC03754C.

Gapsys V, Yildirim A, Aldeghi M, Khalak Y, van der Spoel D, de Groot BL. Accurate Absolute Free Energies for Ligand–Protein Binding Based on Non-Equilibrium Approaches. Communications Chemistry. 2021 Dec; 4(1). doi: 10.1038/s42004-021-00498-y.

Garaeva AA, Slotboom DJ. Elevator-Type Mechanisms of Membrane Transport. Biochemical Society Transactions. 2020 Jun; 48(3):1227–1241. doi: 10.1042/BST20200290.

Gebauer S, Knütter I, Hartrodt B, Brandsch M, Neubert K, Thondorf I. Three-Dimensional Quantitative Structure-Activity Relationship Analyses of Peptide Substrates of the Mammalian H^+/Peptide Cotransporter PEPT1. Journal of Medicinal Chemistry. 2003 Dec; 46(26):5725–5734. doi: 10.1021/jm030976x.

Gill AL, Frederickson M, Cleasby A, Woodhead SJ, Carr MG, Woodhead AJ, Walker MT, Congreve MS, Devine LA, Tisi D, O'Reilly M, Seavers LCA, Davis DJ, Curry J, Anthony R, Padova A, Murray CW, Carr RAE, Jhoti H. Identification of Novel P38α MAP Kinase Inhibitors Using Fragment-Based Lead Generation. Journal of Medicinal Chemistry. 2005 Jan; 48(2):414–426. doi: 10.1021/jm049575n.

Goga N, Rzepiela AJ, De Vries AH, Marrink SJ, Berendsen HJC. Efficient Algorithms for Langevin and DPD Dynamics. Journal of Chemical Theory and Computation. 2012 Oct; 8(10):3637–3649. doi: 10.1021/ct3000876.

Gowers R, Linke M, Barnoud J, Reddy T, Melo M, Seyler S, Domański J, Dotson D, Buchoux S, Kenney I, Beckstein O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In: *Python in Science Conference* Python in Science Conference, Austin, Texas: Python in Science Conference; 2016. p. 98–105. doi: 10.25080/Majora-629e541a-00e.

Griffiths DJ. Introduction to Quantum Mechanics. 2nd ed ed. Upper Saddle River, NJ: Pearson Prentice Hall; 2005.

Grouleff J, Søndergaard S, Koldsø H, Schiøtt B. Properties of an Inward-Facing State of LeuT: Conformational Stability and Substrate Release. Biophysical Journal. 2015 Mar; 108(6):1390–1399. doi: 10.1016/j.bpj.2015.02.010.

Grubmüller H, Heymann B, Tavan P. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. Science. 1996 Feb; 271(5251):997–999. doi: 10.1126/science.271.5251.997.

Guettou F, Quistgaard EM, Raba M, Moberg P, Löw C, Nordlund P. Selectivity Mechanism of a Bacterial Homolog of the Human Drug-Peptide Transporters PepT1 and PepT2. Nature Structural & Molecular Biology. 2014 Aug; 21(8). doi: 10.1038/nsmb.2860.

Guettou F, Quistgaard EM, Trésaugues L, Moberg P, Jegerschöld C, Zhu L, Jong AJO, Nordlund P, Löw C. Structural Insights into Substrate Recognition in Proton-dependent Oligopeptide Transporters. EMBO reports. 2013 Sep; 14(9). doi: 10.1038/embor.2013.107.

Gur M, Zomot E, Cheng MH, Bahar I. Energy Landscape of LeuT from Molecular Simulations. The Journal of Chemical Physics. 2015 Dec; 143(24):243134. doi: 10.1063/1.4936133.

Halder R, Manna RN, Chakraborty S, Jana B. Modulation of the Conformational Dynamics of Apo-Adenylate Kinase through a π–Cation Interaction. The Journal of Physical Chemistry B. 2017 Jun; 121(23):5699–5708. doi: 10.1021/acs.jpcb.7b01736.

Han Hk, de Vrueh RLA, Rhie JK, Covitz KMY, Smith PL, Lee CP, Oh DM, Sadee W, Amidon1 GL. 5'-Amino Acid Esters of Antiviral Nucleosides, Acyclovir, and AZT Are Absorbed by the Intestinal PEPT1 Peptide Transporter. Pharmaceutical Research. 1998; 15(8). doi: 10.1023/A:1011919319810.

Harpole TJ, Delemotte L. Conformational Landscapes of Membrane Proteins Delineated by Enhanced Sampling Molecular Dynamics Simulations. Biochimica et Biophysica Acta - Biomembranes. 2018 Apr; 1860(4):909–926. doi: 10.1016/j.bbamem.2017.10.033.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, et al. Array Programming with NumPy. Nature. 2020 Sep; 585(7825):357–362. doi: 10.1038/s41586-020-2649-2.

Hazel A, Chipot C, Gumbart JC. Thermodynamics of Deca-Alanine Folding in Water. Journal of Chemical Theory and Computation. 2014; 10(7):2836–2844. doi: 10.1021/ct5002076.

Henderson JA, Lui R, Harris JA, Huang Y, De Oliviera VM, Shen J. A Guide to the Continuous Constant pH Molecular Dynamics Methods in Amber and CHARMM [Article v1.0]. Living Journal of Computational Molecular Science. 2022; 4(1). doi: 10.33011/livecoms.4.1.1563.

Henderson RK, Fendler K, Poolman B. Coupling Efficiency of Secondary Active Transporters. Current Opinion in Biotechnology. 2019 Aug; 58:62–71. doi: 10.1016/j.copbio.2018.11.005.

Hénin J, Lelièvre T, Shirts MR, Valsson O, Delemotte L. Enhanced Sampling Methods for Molecular Dynamics Simulations. arXiv. 2022; doi: 10.48550/ARXIV.2202.04164.

Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A Linear Constraint Solver for Molecular Simulations. Journal of Computational Chemistry. 1997 Sep; 18(12):1463–1472. doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

Hollingsworth SA, Dror RO. Molecular Dynamics Simulation for All. Neuron. 2018 Sep; 99(6):1129–1143. doi: 10.1016/j.neuron.2018.08.011.

Hoover WG. Canonical Dynamics: Equilibrium Phase-Space Distributions. Physical Review A. 1985 Mar; 31(3):1695–1697. doi: 10.1103/PhysRevA.31.1695.

Huang H, Ozkirimli E, Post CB. Comparison of Three Perturbation Molecular Dynamics Methods for Modeling Conformational Transitions. Journal of Chemical Theory and Computation. 2009 May; 5(5):1304–1314. doi: 10.1021/ct9000153.

Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell AD. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. Nature Methods. 2017 Jan; 14(1):71–73. doi: 10.1038/nmeth.4067.

Hukushima K, Nemoto K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. Journal of the Physical Society of Japan. 1996 Jun; 65(6):1604–1608. doi: 10.1143/JPSJ.65.1604.

Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007; 9(3):90–95. doi: 10.1109/MCSE.2007.55.

Immadisetty K, Hettige J, Moradi M. What Can and Cannot Be Learned from Molecular Dynamics Simulations of Bacterial Proton-Coupled Oligopeptide Transporter GkPOT? The Journal of Physical Chemistry B. 2017 Apr; 121(15). doi: 10.1021/acs.jpcb.6b09733.

Immadisetty K, Moradi M. Mechanistic Picture for Chemomechanical Coupling in a Bacterial Proton-Coupled Oligopeptide Transporter from Streptococcus Thermophilus. The Journal of Physical Chemistry B. 2021; 125(34):9738–9750. doi: 10.1021/acs.jpcb.1c03982.

Ito K, Hikida A, Kawai S, Lan VTT, Motoyama T, Kitagawa S, Yoshikawa Y, Kato R, Kawarasaki Y. Analysing the Substrate Multispecificity of a Proton-Coupled Oligopeptide Transporter Using a Dipeptide Library. Nature Communications. 2013 Dec; 4(1). doi: 10.1038/ncomms3502.

Izrailev S, Stepaniants S, Isralewitz B, Kosztin D, Lu H, Molnar F, Wriggers W, Schulten K. Steered Molecular Dynamics. In: Griebel M, Keyes DE, Nieminen RM, Roose D, Schlick T, Deuflhard P, Hermans J, Leimkuhler B, Mark AE, Reich S, Skeel RD, editors. *Computational Molecular Dynamics: Challenges, Methods, Ideas*, vol. 4 Berlin, Heidelberg: Springer Berlin Heidelberg; 1999.p. 39–65. doi: 10.1007/978-3-642-58360-5_2.

Jackson MB. On the Time Scale and Time Course of Protein Conformational Changes. The Journal of Chemical Physics. 1993 Nov; 99(9):7253–7259. doi: 10.1063/1.465418.

Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. Journal of Computational Chemistry. 2000; 21(2):132–146.

Jakalian A, Jack DB, Bayly CI. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. Journal of Computational Chemistry. 2002; 23(16):1623–1641. doi: 10.1002/jcc.10128.

Jämbeck JPM, Lyubartsev AP. Derivation and Systematic Validation of a Refined All-Atom Force Field for Phosphatidylcholine Lipids. The Journal of Physical Chemistry B. 2012 Mar; 116(10). doi: 10.1021/jp212503e.

Jana B, Adkar BV, Biswas R, Bagchi B. Dynamic Coupling between the LID and NMP Domain Motions in the Catalytic Conversion of ATP and AMP to ADP by Adenylate Kinase. The Journal of Chemical Physics. 2011 Jan; 134(3):035101. doi: 10.1063/1.3516588.

Jarzynski C. Equilibrium Free-Energy Differences from Nonequilibrium Measurements: A Master-Equation Approach. Physical Review E. 1997 Nov; 56(5):5018–5035. doi: 10.1103/PhysRevE.56.5018.

Jensen JM, Ernst HA, Wang X, Hald H, Ditta AC, Ismat F, Rahman M, Mirza O. Functional Investigation of Conserved Membrane-Embedded Glutamate Residues in the Proton-Coupled Peptide Transporter YjdL. Protein & Peptide Letters. 2012 Mar; 19(3):282–287. doi: 10.2174/092986612799363109.

Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and Validation of a Genetic Algorithm for Flexible Docking. Journal of Molecular Biology. 1997 Apr; 267(3):727–748. doi: 10.1006/jmbi.1996.0897.

Jørgensen ME, Olsen CE, Geiger D, Mirza O, Halkier BA, Nour-Eldin HH. A Functional EXXEK Motif Is Essential for Proton Coupling and Active Glucosinolate Transport by NPF2.11. Plant and Cell Physiology. 2015 Dec; 56(12). doi: 10.1093/pcp/pcv145.

Jorgensen WL, Buckner JK, Boudon S, Tirado-Rives J. Efficient Computation of Absolute Free Energies of Binding by Computer Simulations. Application to the Methane Dimer in Water. The Journal of Chemical Physics. 1988 Sep; 89(6):3742–3746. doi: 10.1063/1.454895.

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of Simple Potential Functions for Simulating Liquid Water. The Journal of Chemical Physics. 1983 Jul; 79(2). doi: 10.1063/1.445869.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, et al. Highly Accurate Protein Structure Prediction with AlphaFold. Nature. 2021 Aug; 596(7873):583–589. doi: 10.1038/s41586-021-03819-2.

Juraszek J, Vreede J, Bolhuis PG. Transition Path Sampling of Protein Conformational Changes. Chemical Physics. 2012 Mar; 396:30–44. doi: 10.1016/j.chemphys.2011.04.032.

Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. Biopolymers. 1983 Dec; 22(12):2577–2637. doi: 10.1002/bip.360221211.

Kanada R, Terayama K, Tokuhisa A, Matsumoto S, Okuno Y. Enhanced Conformational Sampling with an Adaptive Coarse-Grained Elastic Network Model Using Short-Time All-Atom Molecular Dynamics. Journal of Chemical Theory and Computation. 2022 Apr; 18(4):2062–2074. doi: 10.1021/acs.jctc.1c01074.

Kandt C, Ash WL, Peter Tieleman D. Setting up and Running Molecular Dynamics Simulations of Membrane Proteins. Methods. 2007 Apr; 41(4):475–488. doi: 10.1016/j.ymeth.2006.08.006.

Kapoor K, Thangapandian S, Tajkhorshid E. Extended-Ensemble Docking to Probe Dynamic Variation of Ligand Binding Sites during Large-Scale Structural Changes of Proteins. Chemical Science. 2022; 13(14):4150–4169. doi: 10.1039/D2SC00841F.

Kashefolgheta S, Vila Verde A. Developing Force Fields When Experimental Data Is Sparse: AMBER/GAFF-compatible Parameters for Inorganic and Alkyl Oxoanions. Physical Chemistry Chemical Physics. 2017; 19(31):20593–20607. doi: 10.1039/C7CP02557B.

Kazmier K, Sharma S, Quick M, Islam SM, Roux B, Weinstein H, Javitch JA, Mchaourab HS. Conformational Dynamics of Ligand-Dependent Alternating Access in LeuT. Nature Structural & Molecular Biology. 2014 May; 21(5):472–479. doi: 10.1038/nsmb.2816.

Khalak Y, Tresadern G, Aldeghi M, Baumann HM, Mobley DL, De Groot BL, Gapsys V. Alchemical Absolute Protein–Ligand Binding Free Energies for Drug Design. Chemical Science. 2021; 12(41):13958–13971. doi: 10.1039/D1SC03472C.

Killer M, Wald J, Pieprzyk J, Marlovits TC, Löw C. Structural Snapshots of Human PepT1 and PepT2 Reveal Mechanistic Insights into Substrate and Drug Transport across Epithelial Membranes. Science Advances. 2021 Nov; 7(45). doi: 10.1126/sciadv.abk3259.

King E, Aitchison E, Li H, Luo R. Recent Developments in Free Energy Calculations for Drug Discovery. Frontiers in Molecular Biosciences. 2021 Aug; 8:712085. doi: 10.3389/fmolb.2021.712085.

Kirillova S, Cortés J, Stefaniu A, Siméon T. An NMA-guided Path Planning Approach for Computing Large-Amplitude Conformational Changes in Proteins. Proteins: Structure, Function, and Bioinformatics. 2007 Jul; 70(1):131–143. doi: 10.1002/prot.21570.

Klauda JB, Venable RM, Freites JA, O'Connor JW, Tobias DJ, Mondragon-Ramirez C, Vorobyov I, MacKerell AD, Pastor RW. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. The Journal of Physical Chemistry B. 2010 Jun; 114(23):7830–7843. doi: 10.1021/jp101759q.

Klesse G, Rao S, Sansom MSP, Tucker SJ. CHAP: A Versatile Tool for the Structural and Functional Annotation of Ion Channel Pores. Journal of Molecular Biology. 2019 Aug; 431(17):3353–3365. doi: 10.1016/j.jmb.2019.06.003.

Knütter I, Hartrodt B, Theis S, Foltz M, Rastetter M, Daniel H, Neubert K, Brandsch M. Analysis of the Transport Properties of Side Chain Modified Dipeptides at the Mammalian Peptide Transporter PEPT1. European Journal of Pharmaceutical Sciences. 2004 Jan; 21(1):61–67. doi: 10.1016/S0928-0987(03)00141-6.

Korkut A, Hendrickson WA. Computation of Conformational Transitions in Proteins by Virtual Atom Molecular Mechanics as Validated in Application to Adenylate Kinase. Proceedings of the National Academy of Sciences. 2009 Sep; 106(37):15673–15678. doi: 10.1073/pnas.0907684106.

Kotov V, Killer M, Jungnickel KEJ, Lei J, Finocchio G, Steinke J, Bartels K, Strauss J, Dupeux F, Humm AS, Cornaciu I, Márquez JA, Pardon E, Steyaert J, Löw C. Plasticity of the Binding Pocket in Peptide Transporters Underpins Promiscuous Substrate Recognition. Cell Reports. 2023 Aug; 42(8):112831. doi: 10.1016/j.celrep.2023.112831.

Krebs WG. SURVEY AND SUMMARY: The Morph Server: A Standardized System for Analyzing and Visualizing Macromolecular Motions in a Database Framework. Nucleic Acids Research. 2000 Apr; 28(8):1665–1675. doi: 10.1093/nar/28.8.1665.

Krishnamurthy H, Gouaux E. X-Ray Structures of LeuT in Substrate-Free Outward-Open and Apo Inward-Open States. Nature. 2012 Jan; 481(7382):469–474. doi: 10.1038/nature10737.

Kubitzki MB, de Groot BL. The Atomistic Mechanism of Conformational Transition in Adenylate Kinase: A TEE-REX Molecular Dynamics Study. Structure. 2008 Aug; 16(8):1175–1182. doi: 10.1016/j.str.2008.04.013.

Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. Journal of Computational Chemistry. 1992 Oct; 13(8):1011–1021. doi: 10.1002/jcc.540130812.

Laio A, Parrinello M. Escaping Free-Energy Minima. Proceedings of the National Academy of Sciences. 2002 Oct; 99(20):12562–12566. doi: 10.1073/pnas.202427399.

Leach AR. Molecular Modelling: Principles and Applications. 2nd ed ed. Pearson Education; 2001.

Lee BH, Park SW, Jo S, Ki Kim M. Protein Conformational Transitions Explored by a Morphing Approach Based on Normal Mode Analysis in Internal Coordinates. PLoS ONE. 2021; 16(11 November):1–21. doi: 10.1371/journal.pone.0258818.

Léran S, Varala K, Boyer JC, Chiurazzi M, Crawford N, Daniel-Vedele F, David L, Dickstein R, Fernandez E, Forde B, Gassmann W, Geiger D, Gojon A, Gong JM, Halkier BA, Harris JM, Hedrich R, Limami AM, Rentsch D, Seo M, et al. A Unified Nomenclature of NITRATE TRANSPORTER 1/PEPTIDE TRANSPORTER Family Members in Plants. Trends in Plant Science. 2014 Jan; 19(1):5–9. doi: 10.1016/j.tplants.2013.08.008.

Li C, Yue Z, Newstead S, Voth GA. Proton Coupling and the Multiscale Kinetic Mechanism of a Peptide Transporter. Biophysical Journal. 2022 Jun; 121(12):2266–2278. doi: 10.1016/j.bpj.2022.05.029.

Li D, Liu MS, Ji B. Mapping the Dynamics Landscape of Conformational Transitions in Enzyme: The Adenylate Kinase Case. Biophysical Journal. 2015 Aug; 109(3):647–660. doi: 10.1016/j.bpj.2015.06.059.

Li Y, Gong H. Identifying a Feasible Transition Pathway between Two Conformational States for a Protein. Journal of Chemical Theory and Computation. 2022 Jul; 18(7):4529–4543. doi: 10.1021/acs.jctc.2c00390.

Liao Q. Enhanced Sampling and Free Energy Calculations for Protein Simulations. In: *Progress in Molecular Biology and Translational Science*, vol. 170 Elsevier; 2020.p. 177–213. doi: 10.1016/bs.pmbts.2020.01.006.

Lichtinger SM, Biggin PC. Tackling Hysteresis in Conformational Sampling: How to Be Forgetful with MEMENTO. Journal of Chemical Theory and Computation. 2023 Jun; 19(12):3705–3720. doi: 10.1021/acs.jctc.3c00140.

Lichtinger SM, Parker JL, Newstead S, Biggin PC, The Mechanism of Mammalian Proton-Coupled Peptide Transporters. eLife reviewed preprint; 2024. doi: 10.7554/eLife.96507.2.

Limongelli V, Bonomi M, Parrinello M. Funnel Metadynamics as Accurate Binding Free-Energy Method. Proceedings of the National Academy of Sciences. 2013 Apr; 110(16). doi: 10.1073/pnas.1303186110.

Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved Side-chain Torsion Potentials for the Amber ff99SB Protein Force Field. Proteins: Structure, Function, and Bioinformatics. 2010 Jun; 78(8):1950–1958. doi: 10.1002/prot.22711.

Lo YC, Rensi SE, Torng W, Altman RB. Machine Learning in Chemoinformatics and Drug Discovery. Drug Discovery Today. 2018 Aug; 23(8):1538–1546. doi: 10.1016/j.drudis.2018.05.010.

Lopes PEM, Guvench O, MacKerell AD. Current Status of Protein Force Fields for Molecular Dynamics Simulations. In: Kukol A, editor. *Molecular Modeling of Proteins*, vol. 1215 New York, NY: Springer New York; 2015.p. 47–71. doi: 10.1007/978-1-4939-1465-4_3.

López-Blanco JR, Chacón P. New Generation of Elastic Network Models. Current Opinion in Structural Biology. 2016 Apr; 37:46–53. doi: 10.1016/j.sbi.2015.11.013.

Lu J, Scheerer D, Haran G, Li W, Wang W. Role of Repeated Conformational Transitions in Substrate Binding of Adenylate Kinase. The Journal of Physical Chemistry B. 2022 Oct; 126(41):8188–8201. doi: 10.1021/acs.jpcb.2c05497.

Lu Q, Wang J. Single Molecule Conformational Dynamics of Adenylate Kinase: Energy Landscape, Structural Correlations, and Transition State Ensembles. Journal of the American Chemical Society. 2008 Apr; 130(14):4772–4783. doi: 10.1021/ja0780481.

Luckner P, Brandsch M. Interaction of 31 β-Lactam Antibiotics with the H+/Peptide Symporter PEPT2: Analysis of Affinity Constants and Comparison with PEPT1. European Journal of Pharmaceutics and Biopharmaceutics. 2005 Jan; 59(1). doi: 10.1016/j.ejpb.2004.07.008.

Lyons JA, Parker JL, Solcan N, Brinth A, Li D, Shah ST, Caffrey M, Newstead S. Structural Basis for Polyspecificity in the <scp>POT</Scp> Family of Proton-coupled Oligopeptide Transporters. EMBO reports. 2014 Aug; 15(8). doi: 10.15252/embr.201338403.

MacKerell AD, Feig M, Brooks CL. Improved Treatment of the Protein Backbone in Empirical Force Fields. Journal of the American Chemical Society. 2004 Jan; 126(3):698–699. doi: 10.1021/ja036959e.

Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. Journal of Chemical Theory and Computation. 2015 Aug; 11(8):3696–3713. doi: 10.1021/acs.jctc.5b00255.

Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G. String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. The Journal of Chemical Physics. 2006 Jul; 125(2):024106. doi: 10.1063/1.2212942.

Marber MS, Rose B, Wang Y. The P38 Mitogen-Activated Protein Kinase Pathway—A Potential Target for Intervention in Infarction, Hypertrophy, and Heart Failure. Journal of Molecular and Cellular Cardiology. 2011 Oct; 51(4):485–490. doi: 10.1016/j.yjmcc.2010.10.021.

Martinez Molledo M, Quistgaard EM, Flayhan A, Pieprzyk J, Löw C. Multispecific Substrate Recognition in a Proton-Dependent Oligopeptide Transporter. Structure. 2018 Mar; 26(3). doi: 10.1016/j.str.2018.01.005.

Martins De Oliveira V, Liu R, Shen J. Constant pH Molecular Dynamics Simulations: Current Status and Recent Applications. Current Opinion in Structural Biology. 2022 Dec; 77:102498. doi: 10.1016/j.sbi.2022.102498.

Marx D, Hutter J. Ab Initio Molecular Dynamics Basic Theory and Advanced Methods. [online-ausg.] ed. Cambridge: Cambridge University Press; 2009.

Matsunaga Y, Fujisaki H, Terada T, Furuta T, Moritsugu K, Kidera A. Minimum Free Energy Path of Ligand-Induced Transition in Adenylate Kinase. PLoS Computational Biology. 2012 Jun; 8(6):e1002555. doi: 10.1371/journal.pcbi.1002555.

Matsunaga Y, Komuro Y, Kobayashi C, Jung J, Mori T, Sugita Y. Dimensionality of Collective Variables for Describing Conformational Changes of a Multi-Domain Protein. The Journal of Physical Chemistry Letters. 2016 Apr; 7(8):1446–1451. doi: 10.1021/acs.jpclett.6b00317.

McCammon JA, Gelin BR, Karplus M. Dynamics of Folded Proteins. Nature. 1977 Jun; 267(5612):585–590. doi: 10.1038/267585a0.

McComas SE, Reichenbach T, Mitrovic D, Alleva C, Bonaccorsi M, Delemotte L, Drew D. Determinants of Sugar-Induced Influx in the Mammalian Fructose Transporter GLUT5. eLife. 2023 Jul; 12:e84808. doi: 10.7554/eLife.84808.

McKinney W. Data Structures for Statistical Computing in Python. In: *Python in Science Conference* Python in Science Conference, Austin, Texas; 2010. p. 56–61. doi: 10.25080/Majora-92bf1922-00a.

McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, Sunseri J, Koes DR. GNINA 1.0: Molecular Docking with Deep Learning. Journal of Cheminformatics. 2021 Dec; 13(1):43. doi: 10.1186/s13321-021-00522-2.

Mehdi S, Smith Z, Herron L, Zou Z, Tiwary P. Enhanced Sampling with Machine Learning. Annual Review of Physical Chemistry. 2024 Apr; 75(1):annurev–physchem–083122–125941. doi: 10.1146/annurev-physchem-083122-125941.

Mehdi S, Wang D, Pant S, Tiwary P. Accelerating All-Atom Simulations and Gaining Mechanistic Understanding of Biophysical Systems through State Predictive Information Bottleneck. Journal of Chemical Theory and Computation. 2022 May; 18(5):3231–3238. doi: 10.1021/acs.jctc.2c00058.

Meli R, Morris GM, Biggin PC. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. Frontiers in Bioinformatics. 2022 Jun; 2:885983. doi: 10.3389/fbinf.2022.885983.

Meredith D, Temple CS, Guha N, Sword CJ, Boyd CAR, Collier ID, Morgan KM, Bailey PD. Modified Amino Acids and Peptides as Substrates for the Intestinal Peptide Transporter PepT1. European Journal of Biochemistry. 2000 Jun; 267(12):3723–3728. doi: 10.1046/j.1432-1327.2000.01405.x.

Meshkin H, Zhu F. Toward Convergence in Free Energy Calculations for Protein Conformational Changes: A Case Study on the Thin Gate of Mhp1 Transporter. Journal of Chemical Theory and Computation. 2021 Oct; 17(10):6583–6596. doi: 10.1021/acs.jctc.1c00585.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics. 1953 Jun; 21(6):1087–1092. doi: 10.1063/1.1699114.

Mey ASJS, Allen BK, Bruce Macdonald HE, Chodera JD, Hahn DF, Kuhn M, Michel J, Mobley DL, Naden LN, Prasad S, Rizzi A, Scheen J, Shirts MR, Tresadern G, Xu H. Best Practices for Alchemical Free Energy Calculations [Article v1.0]. Living Journal of Computational Molecular Science. 2020; 2(1). doi: 10.33011/livecoms.2.1.18378.

Minhas GS, Newstead S. Structural Basis for Prodrug Recognition by the SLC15 Family of Proton-Coupled Peptide Transporters. Proceedings of the National Academy of Sciences. 2019 Jan; 116(3). doi: 10.1073/pnas.1813715116.

Mitchell P. Osmochemistry of Solute Translocation. Research in Microbiology. 1990; 141(3):286–289. doi: 10.1016/0923-2508(90)90002-8.

Mitchell P. A General Theory of Membrane Transport from Studies of Bacteria. Nature. 1957; 180(4577):134–136.

Mobley DL, Bannan CC, Rizzi A, Bayly CI, Chodera JD, Lim VT, Lim NM, Beauchamp KA, Slochower DR, Shirts MR, Gilson MK, Eastman PK. Escaping Atom Types in Force Fields Using Direct Chemical Perception. Journal of Chemical Theory and Computation. 2018 Nov; 14(11):6076–6092. doi: 10.1021/acs.jctc.8b00640.

Moradi M, Enkavi G, Tajkhorshid E. Atomic-Level Characterization of Transport Cycle Thermodynamics in the Glycerol-3-Phosphate:Phosphate Antiporter. Nature Communications. 2015 Dec; 6(1):8393. doi: 10.1038/ncomms9393.

Müller CW, Schulz GE. Structure of the Complex between Adenylate Kinase from Escherichia Coli and the Inhibitor Ap5A Refined at 1.9 Å Resolution. Journal of Molecular Biology. 1992 Mar; 224(1):159–177. doi: 10.1016/0022-2836(92)90582-5.

Müller C, Schlauderer G, Reinstein J, Schulz G. Adenylate Kinase Motions during Catalysis: An Energetic Counterweight Balancing Substrate Binding. Structure. 1996 Feb; 4(2):147–156. doi: 10.1016/S0969-2126(96)00018-4.

Müller J, Keiser M, Drozdzik M, Oswald S. Expression, Regulation and Function of Intestinal Drug Transporters: An Update. Biological Chemistry. 2017 Feb; 398(2):175–192. doi: 10.1515/hsz-2016-0259.

Nerenberg PS, Head-Gordon T. New Developments in Force Fields for Biomolecular Simulations. Current Opinion in Structural Biology. 2018 Apr; 49:129–138. doi: 10.1016/j.sbi.2018.02.002.

Newstead S. Molecular Insights into Proton Coupled Peptide Transport in the PTR Family of Oligopeptide Transporters. Biochimica et Biophysica Acta (BBA) - General Subjects. 2015 Mar; 1850(3). doi: 10.1016/j.bbagen.2014.05.011.

Newstead S. Recent Advances in Understanding Proton Coupled Peptide Transport via the POT Family. Current Opinion in Structural Biology. 2017 Aug; 45. doi: 10.1016/j.sbi.2016.10.018.

Newstead S, Drew D, Cameron AD, Postis VLG, Xia X, Fowler PW, Ingram JC, Carpenter EP, Sansom MSP, McPherson MJ, Baldwin SA, Iwata S. Crystal Structure of a Prokaryotic Homologue of the Mammalian Oligopeptide-Proton Symporters, PepT1 and PepT2. The EMBO Journal. 2011 Jan; 30(2). doi: 10.1038/emboj.2010.309.

Noé F, Clementi C. Collective Variables for the Study of Long-Time Kinetics from Molecular Trajectories: Theory and Methods. Current Opinion in Structural Biology. 2017 Apr; 43:141–147. doi: 10.1016/j.sbi.2017.02.006.

Nosé S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. Molecular Physics. 1984 Jun; 52(2):255–268. doi: 10.1080/00268978400101201.

Omori A, Fujisawa Y, Sasaki S, Shimono K, Kikukawa T, Miyauchi S. Protonation State of a Histidine Residue in Human Oligopeptide Transporter 1 (hPEPT1) Regulates hPEPT1-Mediated Efflux Activity. Biological and Pharmaceutical Bulletin. 2021 May; 44(5). doi: 10.1248/bpb.b20-01013.

Orellana L. Large-Scale Conformational Changes and Protein Function: Breaking the in Silico Barrier. Frontiers in Molecular Biosciences. 2019 Nov; 6:117. doi: 10.3389/fmolb.2019.00117.

Ovchinnikov V, Karplus M. Analysis and Elimination of a Bias in Targeted Molecular Dynamics Simulations of Conformational Transitions: Application to Calmodulin. The Journal of Physical Chemistry B. 2012 Jul; 116(29):8584–8603. doi: 10.1021/jp212634z.

Pan AC, Sezer D, Roux B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. The Journal of Physical Chemistry B. 2008 Mar; 112(11):3432–3440. doi: 10.1021/jp0777059.

Pan AC, Weinreich TM, Shan Y, Scarpazza DP, Shaw DE. Assessing the Accuracy of Two Enhanced Sampling Methods Using EGFR Kinase Transition Pathways: The Influence of Collective Variable Choice. Journal of Chemical Theory and Computation. 2014 Jul; 10(7):2860–2865. doi: 10.1021/ct500223p.

Pande VS, Beauchamp K, Bowman GR. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. Methods. 2010 Sep; 52(1):99–105. doi: 10.1016/j.ymeth.2010.06.002.

Parker JL, Deme JC, Lichtinger S, Kuteyi G, Biggin PC, Lea SM, Newstead S, Structural Basis for Antibiotic Transport and Inhibition in PepT2, the Mammalian Proton-Coupled Peptide Transporter. Research Square; 2024. doi: 10.21203/rs.3.rs-4435259/v1.

Parker JL, Deme JC, Wu Z, Kuteyi G, Huo J, Owens RJ, Biggin PC, Lea SM, Newstead S. Cryo-EM Structure of PepT2 Reveals Structural Basis for Proton-Coupled Peptide and Prodrug Transport in Mammals. Science Advances. 2021 Aug; 7(35). doi: 10.1126/sciadv.abh3355.

Parker JL, Li C, Brinth A, Wang Z, Vogeley L, Solcan N, Ledderboge-Vucinic G, Swanson JMJ, Caffrey M, Voth GA, Newstead S. Proton Movement and Coupling in the POT Family of Peptide Transporters. Proceedings of the National Academy of Sciences. 2017 Dec; 114(50). doi: 10.1073/pnas.1710727114.

Parker JL, Mindell JA, Newstead S. Thermodynamic Evidence for a Dual Transport Mechanism in a POT Peptide Transporter. eLife. 2014 Dec; 3. doi: 10.7554/eLife.04273.

Parker JL, Newstead S. Molecular Basis of Nitrate Uptake by the Plant Nitrate Transporter NRT1.1. Nature. 2014 Mar; 507(7490). doi: 10.1038/nature13116.

Parrinello M, Rahman A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. Physical Review Letters. 1980 Oct; 45(14). doi: 10.1103/PhysRevLett.45.1196.

Parrinello M, Rahman A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. Journal of Applied Physics. 1981 Dec; 52(12):7182–7190. doi: 10.1063/1.328693.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–2830.

Peng C, Wang J, Shi Y, Xu Z, Zhu W. Increasing the Sampling Efficiency of Protein Conformational Change by Combining a Modified Replica Exchange Molecular Dynamics and Normal Mode Analysis. Journal of Chemical Theory and Computation. 2021; 17(1):13–28. doi: 10.1021/acs.jctc.0c00592.

Piana S, Laio A. A Bias-Exchange Approach to Protein Folding. The Journal of Physical Chemistry B. 2007 May; 111(17):4553–4559. doi: 10.1021/jp067873l.

Pieri M, Gan C, Bailey P, Meredith D. The Transmembrane Tyrosines Y56, Y91 and Y167 Play Important Roles in Determining the Affinity and Transport Rate of the Rabbit Proton-Coupled Peptide Transporter PepT1. The International Journal of Biochemistry & Cell Biology. 2009 Nov; 41(11). doi: 10.1016/j.biocel.2009.04.014.

Pinzi L, Rastelli G. Molecular Docking: Shifting Paradigms in Drug Discovery. International Journal of Molecular Sciences. 2019 Sep; 20(18):4331. doi: 10.3390/ijms20184331.

Pizzagalli MD, Bensimon A, Superti-Furga G. A Guide to Plasma Membrane Solute Carrier Proteins. The FEBS Journal. 2021 May; 288(9):2784–2835. doi: 10.1111/febs.15531.

Pomarici ND, Mehdi S, Quoika PK, Lee S, Loeffler JR, Liedl KR, Tiwary P, Fernández-Quintero ML. Learning High-Dimensional Reaction Coordinates of Fast-Folding Proteins Using State Predictive Information Bottleneck and Bias Exchange Metadynamics. Biophysics; 2023.

Poroca DR, Pelis RM, Chappe VM. ClC Channels and Transporters: Structure, Physiological Functions, and Implications in Human Chloride Channelopathies. Frontiers in Pharmacology. 2017 Mar; 8. doi: 10.3389/fphar.2017.00151.

Pravin MA, Singh SK. Enhanced Sampling and Free Energy Methods to Study Protein Folding and Dynamics. In: Saudagar P, Tripathi T, editors. *Protein Folding Dynamics and Stability* Singapore: Springer Nature Singapore; 2023.p. 165–179. doi: 10.1007/978-981-99-2079-2_9.

Radestock S, Forrest LR. The Alternating-Access Mechanism of MFS Transporters Arises from Inverted-Topology Repeats. Journal of Molecular Biology. 2011 Apr; 407(5). doi: 10.1016/j.jmb.2011.02.008.

Rahman A. Correlations in the Motion of Atoms in Liquid Argon. Physical Review. 1964 Oct; 136(2A):A405–A411. doi: 10.1103/PhysRev.136.A405.

Raiteri P, Laio A, Gervasio FL, Micheletti C, Parrinello M. Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics †. The Journal of Physical Chemistry B. 2006 Mar; 110(8). doi: 10.1021/jp054359r.

Raniolo S, Limongelli V. Ligand Binding Free-Energy Calculations with Funnel Metadynamics. Nature Protocols. 2020 Sep; 15(9). doi: 10.1038/s41596-020-0342-4.

Refaeilzadeh P, Tang L, Liu H. Cross-Validation. In: Liu L, Özsu MT, editors. *Encyclopedia of Database Systems* Boston, MA: Springer US; 2009.p. 532–538. doi: 10.1007/978-0-387-39940-9_565.

Rocklin GJ, Mobley DL, Dill KA, Hünenberger PH. Calculating the Binding Free Energies of Charged Species Based on Explicit-Solvent Simulations Employing Lattice-Sum Methods: An Accurate Correction Scheme for Electrostatic Finite-Size Effects. The Journal of Chemical Physics. 2013 Nov; 139(18). doi: 10.1063/1.4826261.

Rodríguez-Pérez R, Miljković F, Bajorath J. Machine Learning in Chemoinformatics and Medicinal Chemistry. Annual Review of Biomedical Data Science. 2022 Aug; 5(1):43–65. doi: 10.1146/annurev-biodatasci-122120-124216.

Rogers D, Hahn M. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling. 2010 May; 50(5):742–754. doi: 10.1021/ci100050t.

Roser P, Weisner J, Simard JR, Rauh D, Drescher M. Direct Monitoring of the Conformational Equilibria of the Activation Loop in the Mitogen-Activated Protein Kinase P38α. Chemical Communications. 2018; 54(85):12057–12060. doi: 10.1039/C8CC06128A.

Roy K, Kar S, Das RN. Introduction to 3D-QSAR. In: *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* Elsevier; 2015.p. 291–317. doi: 10.1016/B978-0-12-801505-6.00008-9.

Rubio-Aliaga I, Boll M, Daniel H. Cloning and Characterization of the Gene Encoding the Mouse Peptide Transporter PEPT2. Biochemical and Biophysical Research Communications. 2000 Sep; 276(2):734–741. doi: 10.1006/bbrc.2000.3546.

Šali A, Blundell TL. Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology. 1993 Dec; 234(3):779–815. doi: 10.1006/jmbi.1993.1626.

Samsudin F, Parker JL, Sansom MSP, Newstead S, Fowler PW. Accurate Prediction of Ligand Affinities for a Proton-Dependent Oligopeptide Transporter. Cell Chemical Biology. 2016 Feb; 23(2). doi: 10.1016/j.chembiol.2015.11.015.

Sauve S, Williamson J, Polasa A, Moradi M. Ins and Outs of Rocker Switch Mechanism in Major Facilitator Superfamily of Transporters. Membranes. 2023 Apr; 13(5):462. doi: 10.3390/membranes13050462.

Schauperl M, Nerenberg PS, Jang H, Wang LP, Bayly CI, Mobley DL, Gilson MK. Non-Bonded Force Field Model with Advanced Restrained Electrostatic Potential Charges (RESP2). Communications Chemistry. 2020 Apr; 3(1):44. doi: 10.1038/s42004-020-0291-4.

Schlitter J, Engels M, Krüger P, Jacoby E, Wollmer A. Targeted Molecular Dynamics Simulation of Conformational Change-Application to the T - R Transition in Insulin. Molecular Simulation. 1993 Aug; 10(2-6):291–308. doi: 10.1080/08927029308022170.

Selvam B, Chiang N, Shukla D, Energetics of Substrate Transport in Proton-Dependent Oligopeptide Transporters; 2024. doi: 10.1101/2024.05.01.592129.

Selvam B, Mittal S, Shukla D. Free Energy Landscape of the Complete Transport Cycle in a Key Bacterial Transporter. ACS Central Science. 2018 Sep; 4(9). doi: 10.1021/acscentsci.8b00330.

Senn HM, Thiel W. QM/MM Methods for Biomolecular Systems. Angewandte Chemie International Edition. 2009 Feb; 48(7):1198–1229. doi: 10.1002/anie.200802019.

Seo S, Kim MK. KOSMOS: A Universal Morph Server for Nucleic Acids, Proteins and Their Complexes. Nucleic Acids Research. 2012 Jul; 40(W1):W531–W536. doi: 10.1093/nar/gks525.

Shao Q. Enhanced Conformational Sampling Technique Provides an Energy Landscape View of Large-Scale Protein Conformational Transitions. Physical Chemistry Chemical Physics. 2016; 18(42):29170–29182. doi: 10.1039/C6CP05634B.

Shaw DE, Adams PJ, Azaria A, Bank JA, Batson B, Bell A, Bergdorf M, Bhatt J, Butts JA, Correia T, Dirks RM, Dror RO, Eastwood MP, Edwards B, Even A, Feldmann P, Fenn M, Fenton CH, Forte A, Gagliardo J, et al. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* St. Louis Missouri: ACM; 2021. p. 1–11. doi: 10.1145/3458817.3487397.

Shen J, Hu M, Fan X, Ren Z, Portioli C, Yan X, Rong M, Zhou M. Extracellular Domain of PepT1 Interacts with TM1 to Facilitate Substrate Transport. Structure. 2022 Jul; 30(7):1035–1041.e3. doi: 10.1016/j.str.2022.04.011.

Shi Y. Common Folds and Transport Mechanisms of Secondary Active Transporters. Annual Review of Biophysics. 2013 May; 42(1):51–72. doi: 10.1146/annurev-biophys-083012-130429.

Shirts MR, Chodera JD. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. The Journal of Chemical Physics. 2008 Sep; 129(12). doi: 10.1063/1.2978177.

Shirts MR, Pande VS. Comparison of Efficiency and Bias of Free Energies Computed by Exponential Averaging, the Bennett Acceptance Ratio, and Thermodynamic Integration. The Journal of Chemical Physics. 2005 Apr; 122(14):144107. doi: 10.1063/1.1873592.

Sidky H, Chen W, Ferguson AL. Machine Learning for Collective Variable Discovery and Enhanced Sampling in Biomolecular Simulation. Molecular Physics. 2020 Mar; 118(5):e1737742. doi: 10.1080/00268976.2020.1737742.

Singh SK, Piscitelli CL, Yamashita A, Gouaux E. A Competitive Inhibitor Traps LeuT in an Open-to-Out Conformation. Science. 2008 Dec; 322(5908):1655–1661. doi: 10.1126/science.1166777.

Smith D, Artursson P, Avdeef A, Di L, Ecker GF, Faller B, Houston JB, Kansy M, Kerns EH, Krämer SD, Lennernäs H, Van De Waterbeemd H, Sugano K, Testa B. Passive Lipoidal Diffusion and Carrier-Mediated Cell Uptake Are Both Important Mechanisms of Membrane Permeation in Drug Disposition. Molecular Pharmaceutics. 2014 Jun; 11(6):1727–1738. doi: 10.1021/mp400713v.

Solcan N, Kwok J, Fowler PW, Cameron AD, Drew D, Iwata S, Newstead S. Alternating Access Mechanism in the POT Family of Oligopeptide Transporters. The EMBO Journal. 2012 Aug; 31(16). doi: 10.1038/emboj.2012.157.

Song HD, Zhu F. Conformational Dynamics of a Ligand-Free Adenylate Kinase. PLoS ONE. 2013 Jul; 8(7):e68023. doi: 10.1371/journal.pone.0068023.

Song X, Lorenzi PL, Landowski CP, Vig BS, Hilfinger JM, Amidon GL. Amino Acid Ester Prodrugs of the Anticancer Agent Gemcitabine: Synthesis, Bioconversion, Metabolic Bioevasion, and hPEPT1-mediated Transport. Molecular Pharmaceutics. 2005; 2(2):157–167. doi: 10.1021/mp049888e.

Song X, Vig BS, Lorenzi PL, Drach JC, Townsend LB, Amidon GL. Amino Acid Ester Prodrugs of the Antiviral Agent 2-Bromo-5,6-Dichloro-1- (β-D-ribofuranosyl)Benzimidazole as Potential Substrates of hPEPT1 Transporter. Journal of Medicinal Chemistry. 2005; 48(4):1274–1277. doi: 10.1021/jm049450i.

Stauffer M, Jeckelmann JM, Ilgü H, Ucurum Z, Boggavarapu R, Fotiadis D. Peptide Transporter Structure Reveals Binding and Action Mechanism of a Potent PEPT1 and PEPT2 Inhibitor. Communications Chemistry. 2022 Dec; 5(1):23. doi: 10.1038/s42004-022-00636-0.

Steel A, Nussberger S, Romero MF, Boron WF, Boyd CA, Hediger MA. Stoichiometry and pH Dependence of the Rabbit Proton-Dependent Oligopeptide Transporter PepT1. The Journal of Physiology. 1997 Feb; 498(3). doi: 10.1113/jphysiol.1997.sp021883.

Steinhauser M, Hiermaier S. A Review of Computational Methods in Materials Science: Examples from Shock-Wave and Polymer Physics. International Journal of Molecular Sciences. 2009 Dec; 10(12):5135–5216. doi: 10.3390/ijms10125135.

Stillwell W. Membrane Transport. In: *An Introduction to Biological Membranes* Elsevier; 2016.p. 423–451. doi: 10.1016/B978-0-444-63772-7.00019-1.

Storn R, Price K. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. Journal of Global Optimization. 1997; 11(4):341–359. doi: 10.1023/A:1008202821328.

Strathern P. Mendeleyev's Dream: The Quest for the Elements. 1st u.s. ed ed. New York: Thomas Dunne Books; 2001.

Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo—Distance Constraints Applied on Model Quality Estimation. Bioinformatics. 2020 Mar; 36(6):1765–1771. doi: 10.1093/bioinformatics/btz828.

Stumpfe D, Bajorath J. Exploring Activity Cliffs in Medicinal Chemistry: Miniperspective. Journal of Medicinal Chemistry. 2012 Apr; 55(7):2932–2942. doi: 10.1021/jm201706b.

Sugano K, Kansy M, Artursson P, Avdeef A, Bendels S, Di L, Ecker GF, Faller B, Fischer H, Gerebtzoff G, Lennernaes H, Senner F. Coexistence of Passive and Carrier-Mediated Processes in Drug Transport. Nature Reviews Drug Discovery. 2010 Aug; 9(8). doi: 10.1038/nrd3187.

Sugita Y, Kitao A, Okamoto Y. Multidimensional Replica-Exchange Method for Free-Energy Calculations. The Journal of Chemical Physics. 2000 Oct; 113(15):6042–6051. doi: 10.1063/1.1308516.

Sugita Y, Okamoto Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. Chemical Physics Letters. 1999 Nov; 314(1-2):141–151. doi: 10.1016/S0009-2614(99)01123-9.

Sun J, Bankston JR, Payandeh J, Hinds TR, Zagotta WN, Zheng N. Crystal Structure of the Plant Dual-Affinity Nitrate Transporter NRT1.1. Nature. 2014 Mar; 507(7490). doi: 10.1038/nature13074.

Sun Y, Gan W, Lei M, Jiang W, Cheng M, He J, Sun Q, Liu W, Hu L, Jin Y. PEPT1-mediated Prodrug Strategy for Oral Delivery of Peramivir. Asian Journal of Pharmaceutical Sciences. 2018 Nov; 13(6). doi: 10.1016/j.ajps.2018.05.008.

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences. 2003 Nov; 43(6):1947–1958. doi: 10.1021/ci034160g.

Swaan PW, Tukker JJ. Molecular Determinants of Recognition for the Intestinal Peptide Carrier. Journal of Pharmaceutical Sciences. 1997 May; 86(5):596–602. doi: 10.1021/js960359e.

Swails JM, Explicit Solvent CpHMD; 2014.

Swails JM, York DM, Roitberg AE. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. Journal of Chemical Theory and Computation. 2014; 10(3):1341–1352. doi: 10.1021/ct401042b.

Tamura K, Lee CP, Smith PL, Borchardt RT. [No Title Found]. Pharmaceutical Research. 1996; 13(11):1663–1667. doi: 10.1023/A:1016436606183.

Tamura S, Miyao T, Bajorath J. Large-Scale Prediction of Activity Cliffs Using Machine and Deep Learning Methods of Increasing Complexity. Journal of Cheminformatics. 2023 Jan; 15(1):4. doi: 10.1186/s13321-022-00676-7.

Tao W, Zhao D, Sun M, Wang Z, Lin B, Bao Y, Li Y, He Z, Sun Y, Sun J. Intestinal Absorption and Activation of Decitabine Amino Acid Ester Prodrugs Mediated by Peptide Transporter PEPT1 and Enterocyte Enzymes. International Journal of Pharmaceutics. 2018 Apr; 541(1-2). doi: 10.1016/j.ijpharm.2018.02.033.

Tateoka R, Abe H, Miyauchi S, Shuto S, Matsuda A, Kobayashi M, Miyazaki K, Kamo N. Significance of Substrate Hydrophobicity for Recognition by an Oligopeptide Transporter (PEPT1). Bioconjugate Chemistry. 2001 Jul; 12(4):485–492. doi: 10.1021/bc000135u.

Tembre BL, Mc Cammon JA. Ligand-Receptor Interactions. Computers & Chemistry. 1984 Jan; 8(4):281–283. doi: 10.1016/0097-8485(84)85020-2.

Temple CS, Stewart AK, Meredith D, Lister NA, Morgan KM, Collier ID, Vaughan-Jones RD, Boyd CAR, Bailey PD, Bronk JR. Peptide Mimics as Substrates for the Intestinal Peptide Transporter. Journal of Biological Chemistry. 1998 Jan; 273(1):20–22. doi: 10.1074/jbc.273.1.20.

Terada T, Saito H, Mukai M, Inui Ki. Identification of the Histidine Residues Involved in Substrate Recognition by a Rat H^+ /Peptide Cotransporter, PEPT1. FEBS Letters. 1996 Sep; 394(2):196–200. doi: 10.1016/0014-5793(96)00952-0.

Terada T, Sawada K, Irie M, Saito H, Hashimoto Y, Inui Ki. Structural Requirements for Determining the Substrate Affinity of Peptide Transporters PEPT1 and PEPT2. Pflügers Archiv - European Journal of Physiology. 2000 Sep; 440(5). doi: 10.1007/s004240000339.

Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Migues AN, Bickel J, Wang Y, Pincay J, Wu Q, Simmerling C. Ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. Journal of Chemical Theory and Computation. 2020; 16(1):528–552. doi: 10.1021/acs.jctc.9b00591.

Tiwary P, Parrinello M. From Metadynamics to Dynamics. Physical Review Letters. 2013 Dec; 111(23):230602. doi: 10.1103/PhysRevLett.111.230602.

Tokonami S, Onose M, Nakasone Y, Terazima M. Slow Conformational Changes of Blue Light Sensor BLUF Proteins in Milliseconds. Journal of the American Chemical Society. 2022 Mar; 144(9):4080–4090. doi: 10.1021/jacs.1c13121.

Torrie GM, Valleau JP. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. Journal of Computational Physics. 1977 Feb; 23(2):187–199. doi: 10.1016/0021-9991(77)90121-8.

Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. PLUMED 2: New Feathers for an Old Bird. Computer Physics Communications. 2014 Feb; 185(2):604–613. doi: 10.1016/j.cpc.2013.09.018.

Tuckerman ME. Statistical Mechanics: Theory and Molecular Simulation. 2nd ed ed. Oxford Graduate Texts, Oxford: Oxford university press; 2023.

Tummino PJ, Copeland RA. Residence Time of Receptor-Ligand Complexes and Its Effect on Biological Function. Biochemistry. 2008 May; 47(20):5481–5492. doi: 10.1021/bi8002023.

Uchiyama T, Kulkarni AA, Davies DL, Lee VHL. Biophysical Evidence for His ^{57} as a Proton-Binding Site in the Mammalian Intestinal Transporter hPepT1. Pharmaceutical Research. 2003 Dec; 20(12):1911–1916. doi: 10.1023/B:PHAM.0000008036.05892.e9.

Ural-Blimke Y, Flayhan A, Strauss J, Rantos V, Bartels K, Nielsen R, Pardon E, Steyaert J, Kosinski J, Quistgaard EM, Löw C. Structure of Prototypic Peptide Transporter DtpA from *E. Coli* in Complex with Valganciclovir Provides Insights into Drug Binding of Human PepT1. Journal of the American Chemical Society. 2019 Feb; 141(6). doi: 10.1021/jacs.8b11343.

Våbenø J, Nielsen CU, Steffansen B, Lejon T, Sylte I, Jørgensen FS, Luthman K. Conformational Restrictions in Ligand Binding to the Human Intestinal Di-/Tripeptide Transporter: Implications for Design of hPEPT1 Targeted Prodrugs. Bioorganic & Medicinal Chemistry. 2005 Mar; 13(6):1977–1988. doi: 10.1016/j.bmc.2005.01.019.

Vale N, Ferreira A, Matos J, Fresco P, Gouveia MJ. Amino Acids in the Development of Prodrugs. Molecules. 2018; 23(9). doi: 10.3390/molecules23092318.

Van Der Spoel D. Systematic Design of Biomolecular Force Fields. Current Opinion in Structural Biology. 2021 Apr; 67:18–24. doi: 10.1016/j.sbi.2020.08.006.

Van Tilborg D, Alenicheva A, Grisoni F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. Journal of Chemical Information and Modeling. 2022 Dec; 62(23):5938–5951. doi: 10.1021/acs.jcim.2c01073.

Vani BP, Aranganathan A, Tiwary P. Exploring Kinase Asp-Phe-Gly (DFG) Loop Conformational Stability with AlphaFold2-RAVE. Journal of Chemical Information and Modeling. 2023 Nov; p. acs.jcim.3c01436. doi: 10.1021/acs.jcim.3c01436.

Vani BP, Aranganathan A, Wang D, Tiwary P. AlphaFold2-RAVE: From Sequence to Boltzmann Ranking. Journal of Chemical Theory and Computation. 2023 Jul; 19(14):4351–4354. doi: 10.1021/acs.jctc.3c00290.

Veevers R, Hayward S. Morphing and Docking Visualisation of Biomolecular Structures Using Multi-Dimensional Scaling. Journal of Molecular Graphics and Modelling. 2018 Jun; 82:108–116. doi: 10.1016/j.jmgm.2018.04.013.

Vergara-Jaque A, Fenollar-Ferrer C, Kaufmann D, Forrest LR. Repeat-Swap Homology Modeling of Secondary Active Transporters: Updated Protocol and Prediction of Elevator-Type Mechanisms. Frontiers in Pharmacology. 2015 Sep; 6. doi: 10.3389/fphar.2015.00183.

Verlet L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Physical Review. 1967 Jul; 159(1):98–103. doi: 10.1103/PhysRev.159.98.

Vig BS, Stouch TR, Timoszyk JK, Quan Y, Wall DA, Smith RL, Faria TN. Human PEPT1 Pharmacophore Distinguishes between Dipeptide Transport and Binding. Journal of Medicinal Chemistry. 2006 Jun; 49(12):3636–3644. doi: 10.1021/jm0511029.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020 Mar; 17(3):261–272. doi: 10.1038/s41592-019-0686-2.

Vogtherr M, Saxena K, Hoelder S, Grimme S, Betz M, Schieborr U, Pescatore B, Robin M, Delarbre L, Langer T, Wendt KU, Schwalbe H. NMR Characterization of Kinase P38 Dynamics in Free and Ligand-Bound Forms. Angewandte Chemie International Edition. 2006 Jan; 45(6):993–997. doi: 10.1002/anie.200502770.

Wang A, Zhang D, Li Y, Zhang Z, Li G. Large-Scale Biomolecular Conformational Transitions Explored by a Combined Elastic Network Model and Enhanced Sampling Molecular Dynamics. The Journal of Physical Chemistry Letters. 2020 Jan; 11(1):325–332. doi: 10.1021/acs.jpclett.9b03399.

Wang D, Tiwary P. State Predictive Information Bottleneck. The Journal of Chemical Physics. 2021 Apr; 154(13):134111. doi: 10.1063/5.0038198.

Wang J, Peng C, Yu Y, Chen Z, Xu Z, Cai T, Shao Q, Shi J, Zhu W. Exploring Conformational Change of Adenylate Kinase by Replica Exchange Molecular Dynamic Simulation. Biophysical Journal. 2020 Mar; 118(5):1009–1018. doi: 10.1016/j.bpj.2020.01.001.

Wang J, Shao Q, Xu Z, Liu Y, Yang Z, Cossins BP, Jiang H, Chen K, Shi J, Zhu W. Exploring Transition Pathway and Free-Energy Profile of Large-Scale Protein Conformational Change by Combining Normal Mode Analysis and Umbrella Sampling Molecular Dynamics. The Journal of Physical Chemistry B. 2014 Jan; 118(1):134–143. doi: 10.1021/jp4105129.

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and Testing of a General Amber Force Field. Journal of Computational Chemistry. 2004 Jul; 25(9):1157–1174. doi: 10.1002/jcc.20035.

Wang L, Friesner RA, Berne BJ. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). Journal of Physical Chemistry B. 2011; 115(30):9431–9438. doi: 10.1021/jp204407d.

Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. Journal of the American Chemical Society. 2015 Feb; 137(7):2695–2703. doi: 10.1021/ja512751q.

Wang Y, Gan L, Wang E, Wang J. Exploring the Dynamic Functional Landscape of Adenylate Kinase Modulated by Substrates. Journal of Chemical Theory and Computation. 2013 Jan; 9(1):84–95. doi: 10.1021/ct300720s.

Wang Y, Pulido I, Takaba K, Kaminow B, Scheen J, Wang L, Chodera JD. EspalomaCharge: Machine Learning-Enabled Ultra-Fast Partial Charge Assignment. arXiv. 2023; doi: 10.48550/ARXIV.2302.06758.

Wang Z, Harkins PC, Ulevitch RJ, Han J, Cobb MH, Goldsmith EJ. The Structure of Mitogen-Activated Protein Kinase P38 at 2.1-Å Resolution. Proceedings of the National Academy of Sciences. 1997 Mar; 94(6):2327–2332. doi: 10.1073/pnas.94.6.2327.

Warren MT, Biggs CI, Bissoyi A, Gibson MI, Sosso GC, Data-Driven Discovery of Potent Small Molecule Ice Recrystallisation Inhibitors; 2023. doi: 10.26434/chemrxiv-2023-gvvk0.

Weiss DR, Levitt M. Can Morphing Methods Predict Intermediate Structures? Journal of Molecular Biology. 2009 Jan; 385(2):665–674. doi: 10.1016/j.jmb.2008.10.064.

Wells S, Menor S, Hespenheide B, Thorpe MF. Constrained Geometric Simulation of Diffusive Motion in Proteins. Physical Biology. 2005 Nov; 2(4):S127–S136. doi: 10.1088/1478-3975/2/4/S07.

Wilkens S. Structure and Mechanism of ABC Transporters. F1000Prime Reports. 2015 Feb; 7. doi: 10.12703/P7-14.

Wu EL, Cheng X, Jo S, Rui H, Song KC, Dávila-Contreras EM, Qi Y, Lee J, Monje-Galvan V, Venable RM, Klauda JB, Im W. CHARMM-GUI *Membrane Builder* toward Realistic Biological Membrane Simulations. Journal of Computational Chemistry. 2014 Oct; 35(27):1997–2004. doi: 10.1002/jcc.23702.

Wu Z, Biggin PC. Correction Schemes for Absolute Binding Free Energies Involving Lipid Bilayers. Journal of Chemical Theory and Computation. 2022 Mar; p. acs.jctc.1c01251. doi: 10.1021/acs.jctc.1c01251.

Ye C, Ding C, Ma R, Wang J, Zhang Z. Electrostatic Interactions Determine Entrance/Release Order of Substrates in the Catalytic Cycle of Adenylate Kinase. Proteins: Structure, Function, and Bioinformatics. 2019 Apr; 87(4):337–347. doi: 10.1002/prot.25655.

You W, Tang Z, Chang CeA. Potential Mean Force from Umbrella Sampling Simulations: What Can We Learn and What Is Missed? Journal of Chemical Theory and Computation. 2019 Apr; 15(4):2433–2443. doi: 10.1021/acs.jctc.8b01142.

Yuan Y, Zhu Q, Song R, Ma J, Dong H. A Two-Ended Data-Driven Accelerated Sampling Method for Exploring the Transition Pathways between Two Known States of Protein. Journal of Chemical Theory and Computation. 2020 Jul; 16(7):4631–4640. doi: 10.1021/acs.jctc.9b01184.

Zeller F, Zacharias M. Substrate Binding Specifically Modulates Domain Arrangements in Adenylate Kinase. Biophysical Journal. 2015 Nov; 109(9):1978–1985. doi: 10.1016/j.bpj.2015.08.049.

Zhao Y, Mao G, Liu M, Zhang L, Wang X, Zhang XC. Crystal Structure of the E. Coli Peptide Transporter YbgH. Structure. 2014 Aug; 22(8). doi: 10.1016/j.str.2014.06.008.

Zheng Y, Cui Q. Multiple Pathways and Time Scales for Conformational Transitions in Apo-Adenylate Kinase. Journal of Chemical Theory and Computation. 2018 Mar; 14(3):1716–1726. doi: 10.1021/acs.jctc.7b01064.

Zosimos of Panoplis. Cheirokmeta; 300.

Zwanzig RW. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. The Journal of Chemical Physics. 1954 Aug; 22(8):1420–1426. doi: 10.1063/1.1740409.