THE UNIVERSITY OF CHICAGO

MULTISCALE SIMULATION METHODS AND THEIR APPLICATIONS FOR CONDENSED PHASE REACTIVE SYSTEMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
DA TENG

CHICAGO, ILLINOIS
AUGUST 2024

Copyright © 2024 by Da Teng

All Rights Reserved

To my family, who instilled in me a deep-rooted love for truth and science

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES vii

ABSTRACT viii

1 INTRODUCTION 1
1.1 Proton Transporters and Secondary Gating 1
1.2 The Sampling Problem in Computer Simulations 3
References 5

2 THE C-TERMINUS OF THE MULTI-DRUG EFFLUX PUMP EMRE PREVENTS PROTON LEAK BY GATING TRANSPORT 8
2.1 Introduction 8
2.2 Results 11
2.2.1 The C-terminus Controls a Water Wire into the Primary Binding Site 11
2.2.2 Structural Basis for the C-terminus Gating 15
2.2.3 The Potential of Mean Force of Proton Transport Supports a Hydrophobic Bottleneck 16
2.3 Methods 18
2.3.1 Molecular Dynamics 18
2.3.2 Water Path Length Calculation 18
2.3.3 Umbrella sampling with MS-RMD 19
2.4 Discussions 19
References 21

3 DETERMINING THE RATE LIMITING STEP OF PROTON LEAKING IN EMRE 25
3.1 Introduction 25
3.2 Results 29
3.2.1 Proton Hopping between the Two E14 is Fast 29
3.2.2 EmrE Occludes Itself at Low pH 32
3.2.3 Hydration of the E25 pathway is possible upon a hydrogen bond breaking 33
3.3 Methods 36
3.3.1 Simulation Software 36
3.3.2 MS-RMD Umbrella Sampling with PT CV 36
3.3.3 Rate Calculation with Transition State Theory 37
3.3.4 Low pH Structure of EmrE and Metadynamics on Hydration CV . . 38
3.4 Discussions 38
References 41

4 QM/CG-MM: SYSTEMATIC EMBEDDING OF QUANTUM MECHANICAL SYSTEMS IN A COARSE-GRAINED ENVIRONMENT WITH ACCURATE ELECTROSTATICS . . . 45
4.1 Introduction . . . 46
4.2 Theory and Methods . . . 49
4.2.1 Solvent Effects: S_N2 Reactions as an Illustrative Example . . . 49
4.2.2 Multiscale Coarse-graining (MS-CG) . . . 50
4.2.3 Electrostatic coupling in QM/MM and QM/CG-MM . . . 51
4.2.4 Description of Solvent Polarization Calls for Explicit Electrostatics Coupling . . . 52
4.2.5 The QM/CG-MM Coupling . . . 54
4.2.6 Implementations and Simulation Details . . . 55
4.3 Results . . . 56
4.3.1 Monopole Approximation Fails to Capture Essential Electrostatics . . 56
4.3.2 Electrostatic Fitting Suggests a Two-Bead or a Three-Bead Mapping Scheme . . . 59
4.3.3 QM/CG-MM Correctly Reproduces Solvent Electrostatic Effects . . . 63
4.3.4 The QM/CG-MM Model is Transferable to Other Reactive Systems . 63
4.3.5 QM/CG-MM Speeds up Sampling through Faster Solvent Relaxation 65
4.4 Discussion . . . 68
4.5 Supporting Information . . . 70
4.5.1 Simulation Details . . . 70
4.5.2 QM/CG-MM parametrization . . . 77
4.5.3 Free energy sampling . . . 78
4.5.4 Discussion on the center-of-charge mapping . . . 82
References . . . 83

A APPENDIX: EXPERIMENTAL RESULTS ON THE LEAKAGE OF Δ107-EMRE 89
A.1 Acknowledgements . . . 89
A.2 Results . . . 89
A.2.1 EmrE is properly folded and functional when the C-terminal tail is truncated . . . 89
A.2.2 Truncation of the C-terminal tail enhances proton leak through EmrE 91
A.2.3 Re-assessing protonation state by NMR . . . 95
A.2.4 Proton leak through Δ107-EmrE does not synergize with harmane . . 98
References . . . 100

LIST OF FIGURES

2.1 Model of coupled antiport and uncoupled proton leak through EmrE 9
2.2 The C-terminus tail caps the water wire from the open side 12
2.3 The structural basis of the C-terminus gating 14
2.4 Hydrogen bonding of D84 and the C-terminal tail 16

3.1 Partial transport network of EmrE 28
3.2 PMF for proton transport between the two E14s 29
3.3 In fully protonated EmrE, the open side complete closes by hydrogen bonding . 32
3.4 PMF for hydration of the PT pathways 35

4.1 QM/CG-MM umbrella sampling of the chloride-methyl chloride system 57
4.2 One-site acetone model 58
4.3 Mapping of the acetone molecule 60
4.4 Resulting CG models 62
4.5 Reaction PMF for the bromide-methyl chloride system 63
4.6 Kemp decarboxylation reaction in acetone 66
4.7 The convergence of QM/CG-MM 67
4.8 Fitting bonded parameters 75
4.9 Integrating 2-D PMF into 1-D 81
4.10 Result of CoC mapping 83

A.1 C-terminal tail truncation does not impair the ability of EmrE to confer resistance to toxic compounds 90
A.2 C-terminal tail truncation enhances proton leak 92
A.3 pH titration of TPP^+-bound Δ107-EmrE supports the possibility of secondary gating 96
A.4 Intrinsic leak in Δ107-EmrE does not synergize with harmane-induced leak . . . 99

LIST OF TABLES

4.1 Nonbonded forcefield parameters used in present work 72

ABSTRACT

Multiscale simulation methods utilize microscopic information, such as atomic interactions, to infer macroscopic thermodynamic or kinetic properties. For example, free energy sampling connects dynamics with free energy curves, enabling the estimation of reaction rates and equilibrium constants, while coarse-graining methods pre-average interactions in finer-grain systems to build coarse-grained models, allowing longer simulations of larger systems. A key challenge of these multiscale methods is their reliance on accurate microscopic molecular dynamics statistics and the need for ergodic sampling of the phase space. Despite advances in computational power enabling unprecedented levels of brute force sampling, the curse of dimensionality dictates that such efforts become exponentially more difficult as system size increases. This limitation is particularly relevant for systems like membrane proteins, which are challenging to study due to the limited temporal and spatial resolution of experimental techniques, yet their biological significance necessitates a deeper understanding for the sake of drug discovery and bioengineering.

This study first demonstrates the multiscale simulation of EmrE, a secondary active transporter that effluxes toxins from *E. coli* using proton motive force. EmrE, a member of the Small Multidrug Resistant (SMR) transporter family, exhibits transport cycles that deviate significantly from the canonical alternating access model while preventing substrate leakage. Our simulations show that at higher pH levels, the C-terminal of EmrE acts as a secondary gate, controlling a water wire into the binding site. In the tail truncation mutant Δ107-EmrE, a hydrated water wire not previously observed in WT-EmrE is found. Further, we demonstrate that this gate allosterically controls a hydrophobic gate, preventing water wire formation. This secondary gating hypothesis is corroborated by high pH structural analysis of EmrE, revealing a hydrogen bonding network occluding the binding site. A 2D potential of mean force (PMF) analysis, using hydrogen bond formation and channel hydration as collective variables, shows that breaking this hydrogen bond leads to EmrE

hydration, facilitating subsequent proton transport.

While cleverly choosing collective variables can aid in sampling specific slow degrees of freedom, unbiased slow degrees of freedom are often abundant and difficult to identify. Coarse-graining is a well-known method for accelerating system dynamics and handling longer timescale sampling. We developed the Quantum Mechanics with Coarse-Grained Molecular Mechanics (QM/CG-MM) method to introduce reactivity into coarse-grained models, extending previous methodologies to account for electrostatic coupling in polar environments. Demonstrations with a chloride-methyl chloride S_N2 reaction in acetone, a system sensitive to solvent polarity, show that the QM/CG-MM method accurately replicates the potential of mean force and reaction barriers compared to atomistic simulations. The method's generalizability is validated across multiple reactive systems, achieving consistent agreement with atomistic models and proportional sampling speed-up relative to the acceleration of solvent rotational dynamics in the CG system. This advancement highlights the potential of QM/CG-MM to provide accurate and computationally efficient simulations for complex chemical systems.

CHAPTER 1

INTRODUCTION

1.1 Proton Transporters and Secondary Gating

Membrane transport proteins are a crucial class of membrane proteins essential for various biological functions. They regulate the flux of substances across the cell membrane and are involved in a wide variety of physiological processes. Traditionally, these membrane proteins are classified into channels and transporters. Channels facilitate passive transport along the downhill electrochemical gradient, while transporters enable "uphill" transport by coupling it to an energy source. Compared to channels, transporters typically have more complex working mechanisms, as they need to ensure coupled transport without leaking either the energy source or the substrate being transported. Transporters can be further categorized into two classes based on their mode of operation. The first class, known as Primary Transporters, harnesses energy from light or chemical reactions to drive the transport of their substrates. Notable examples include the F_0F_1-ATPase and Complex I in the respiratory chain. The second class, called Secondary Transporters, utilizes the electrochemical free energy of another ion, usually sodium or proton, to drive the transport of substrates.[1,2]

Conventionally, secondary transporters are divided into symporters and antiporters based on the direction of transport. Symporters move the driving ion and substrate in the same direction, whereas antiporters transport them in opposite directions. The alternating access model effectively explains the coupled transport mechanism of these transporters. This model posits that the transporter has two conformations: outward-facing (OF) and inward-facing (IF), where only one side of the membrane has access to the binding site at any given time. Upon binding of either the ion or the substrate, the protein occludes its pore and alternates its conformation between OF and IF states. Notably, alternating access of the *apo* states is prevented to avoid leakages by diffusion through the transporter.

The binary categorization of secondary transporters into symporters and antiporters has been challenged by many emerging systems. While there are perfect examples of systems that strictly follow the alternating access model with a step-by-step transport cycle like an internal combustion engine,[3] many systems do not conform to this model. For instance, some transporters, such as ClC-ec1—a proton/chloride antiporter—exhibit a 2.2:1 non-integer stoichiometry, which cannot be explained by the alternating access model[4,5] In such cases, a kinetic model becomes useful to describe their transport behavior, where the net transport flux is the cumulative effect of all possible cycles. This approach also implies that the transport mode can be altered by changing the substrate concentration or the rate of certain fundamental steps. Consequently, an antiporter can be modified into a symporter or even a uniporter that only moves one of its substrates.[6,7]

An interesting system called EmrE has been studied long as a minimum model for proton-coupled transport.[8] EmrE is a member of the Small Multirug Resistant (SMR) transporters, using proton motive force to efflux toxins from the cell. With only 110 residues in either of its two monomers, it exhibits unexpected colmepexity in the transport mechanism. It does not satisfy either of the two criteria for coupled transport in the alternating access model, showing no competitive binding and alternates its access whatever the substrate binding state is.[9] Meanwhile, it is still capable of preventing proton leakage.

Beyond the alternating access model, there is a growing recognition that many transporters, such as EmrE, may have more than one "gate" to regulate coupled transport.[10] In the case of EmrE, our research indicates that the C-terminal residues play a crucial role in coordinating interactions responsible for this secondary gating within the transport pathway. Experimental data (see Appendix) and molecular dynamics simulations (Chapter 2) demonstrate that the C-terminal tail of EmrE is essential for preventing uncoupled proton leak in the absence of a substrate. The significance of this finding is underscored by the conservation of the C-terminal residue H110 across the SMR family, including clinically relevant

homologs from Pseudomonas aeruginosa and Staphylococcus aureus. This insight also addresses a longstanding inconsistency in previous mechanistic models of the small multidrug resistance (SMR) transporter.

Kinetic model analysis has shown that EmrE's alternative transport mode can be triggered by altering substrate on/off rates and proton on/off rates.[6,7] However, the validity of such modeling remains uncertain and requires a quantitative understanding of the fundamental proton transport steps. Additionally, understanding the structural details of this regulation is essential for future drug development. In Chapter 3, using multiscale reactive molecular dynamics (MS-RMD), we studied several proton transport steps of EmrE and identified the fast and slow steps in its proton transport cycle.[11] Our findings indicate that the slow step involves the proton leaving, which faces a high hydration barrier when EmrE is fully protonated. This result aligns well with modeling predictions that a fast proton on rate and a slow off rate will lead to coupled antiport.

1.2 The Sampling Problem in Computer Simulations

Computer simulations of molecular systems nowadays are running at unimaginable scales and speeds to those trailblazers who started the field. In 1975, Michael Levitt and Arieh Warshel published the first ever molecular dynamics (MD) simulation of a protein called Bovine Pancreatic Trypsin Inhibitor (BPTI), with only 58 residues.[12] In today's standard, their protein model is actually a "coarse-grained" model that only contains the C^{α} atoms and the centroid of the side chain. In the conclusion, yet inaugural, remarks, they wrote:

> *The general concept of using a simple model based on effective time-averaged forces when the detailed forces are too complicated has many potential applications; for example, the formation of protein quarternary structure and multienzyme complexes, virus assembly and so on. At each level of complexity, forces would be time averaged over those substructures that are relatively fixed or seem to play a less important role*

in the assembly. Such a hierarchical approach might eventually lead to an understanding and simulation of very complicated biological assembly processes.

(Michael Levitt and Arieh Warshel, 1975)

We can now proudly assert that, after nearly 50 years, we are capable of simulating these processes at atomistic resolution with sufficient confidence. However, the depressing fact from statistical mechanics is still demanding advancement in novel theories and algorithms. As the degrees of freedom of a system increase, the possibilities in the configurational space grow exponentially. The physically meaningful regions of this vast space are tiny compared to what is mathematically possible. Without prior knowledge, it is intractable to locate these sparse, physically possible configurations. State-of-the-art long-range electrostatic calculations now achieve a scaling of $O(N \log N)$, and distributed high-performance computing can effectively reduce the cost of calculating other interactions to nearly linear.[13] Despite these advancements, the mismatch between our linear computational power and the exponentially growing sampling problem remains the real bottleneck in solving larger and more complex problems.

In our previous work with EmrE, we identified a slow degree of freedom (DoF) that significantly alters the hydration landscape. We discovered that breaking a hydrogen bond between E25 and T56 drastically reduces the hydration barrier for proton transport. However, identifying such additional slow DoF requires deep biological insight into the specific system. Without prior experiments highlighting the critical role of E25, it would be nearly impossible to identify this DoF.[14] This lack of prior knowledge would result in insufficient sampling and questionable results. Our sampling capacity also limits the number of slow DoFs we can explicitly sample. For instance, if a one-dimensional (1D) umbrella sampling simulation consists of 50 windows, adding another degree of freedom will increase the computational effort by a factor of 50. This exponential growth in computational demand highlights the challenges we face in accurately sampling multiple slow DoFs, necessitating further

advancements in sampling techniques and computational efficiency.[15]

Coarse-graining is known to expedite the dynamics of a system without compromising too much detail, enabling simulations on longer timescales.[16] By pre-averaging the forces, as Levitt and Warshel did in their work, the free energy landscape is smoothened, allowing for friction-free movement through it.[17] Therefore, incorporating a reactive system into a coarse-grained model can significantly expedite sampling, potentially eliminating the need to bias additional degrees of freedom (DoFs) while still achieving sufficient sampling. In Chapter 4, we introduced a method called Quantum Mechanics/Coarse-Grained Molecular Mechanics (QM/CG-MM). This approach extends the well-known Quantum Mechanics/Molecular Mechanics (QM/MM) method to allow embedding within coarse-grained environments.[18] Since QM/MM is expensive, any theoretical method that reduces the need for extensive QM calculations while maintaining the same level of sampling is highly valuable.

Specifically, we expanded on previous work that primarily addressed reactions in non-polar solvents by accurately treating their electrostatic interactions.[19,20] Our treatment is not only highly accurate but also transferable, meaning the coarse-grained (CG) parameters we developed for one reactive system can be seamlessly integrated into another. This greatly expands the applicability of the model and allows the reusing of existing CG models. We also demonstrated that the expedited sampling is roughly proportional to the acceleration of rotational dynamics of the solvent, confirming that the faster convergence arises from the accelerated dynamics. This validation underscores the effectiveness of our approach in improving sampling efficiency without compromising accuracy.

References

(1) Beckstein, O.; Naughton, F. General Principles of Secondary Active Transporter Function. *Biophys. Rev.* **2022**, *3*, 011307.

(2) Ashcroft, F.; Gadsby, D.; Miller, C. Introduction. The Blurred Boundary between Channels and Transporters. *Philos Trans R Soc Lond B Biol Sci* **2009**, *364*, 145–147.

(3) Abramson, J.; Smirnova, I.; Kasho, V.; Verner, G.; Iwata, S.; Kaback, H. The Lactose Permease of Escherichia Coli: Overall Structure, the Sugar-Binding Site and the Alternating Access Model for Transport. *FEBS Letters* **2003**, *555*, 96–101.

(4) Mayes, H. B.; Lee, S.; White, A. D.; Voth, G. A.; Swanson, J. M. J. Multiscale Kinetic Modeling Reveals an Ensemble of Cl^-/H^+ Exchange Pathways in ClC-ec1 Antiporter. *J. Am. Chem. Soc.* **2018**, *140*, 1793–1804.

(5) Yue, Z.; Bernardi, A.; Li, C.; Mironenko, A. V.; Swanson, J. M. J. Toward a Multipathway Perspective: pH-Dependent Kinetic Selection of Competing Pathways and the Role of the Internal Glutamate in Cl^-/H^+ Antiporters. *J Phys Chem B* **2021**, *125*, 7975–7984.

(6) Hussey, G. A.; Thomas, N. E.; Henzler-Wildman, K. A. Highly Coupled Transport Can Be Achieved in Free-Exchange Transport Models. *Journal of General Physiology* **2020**, *152*.

(7) Awtrey, N.; Beckstein, O. Kinetic Diagram Analysis: A Python Library for Calculating Steady-State Observables of Kinetic Systems Analytically. *bioRxiv* **2024**, 2024.05.27.596119.

(8) Burata, O. E.; Yeh, T. J.; Macdonald, C. B.; Stockbridge, R. B. Still Rocking in the Structural Era: A Molecular Overview of the Small Multidrug Resistance (SMR) Transporter Family. *J. Biol. Chem.* **2022**, *298*, 102482.

(9) Robinson, A. E.; Thomas, N. E.; Morrison, E. A.; Balthazor, B. M.; Henzler-Wildman, K. A. New Free-Exchange Model of EmrE Transport. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, E10083–E10091.

(10) Peyton J. Spreacker; Nathan E. Thomas; Will F. Beeninga; Merissa Brousseau; Colin J. Porter; Kylie M. Hibbs; Katherine A. Henzler-Wildman Activating Alternative Transport Modes in a Multidrug Resistance Efflux Pump to Confer Chemical Susceptibility | Nature Communications. *Nat. Commun.* **2022**, *13*, 7655.

(11) Kaiser, S.; Yue, Z.; Peng, Y.; Nguyen, T.; Chen, S.; Teng, D.; Voth, G. Molecular Dynamics Simulation of Complex Reactivity with the Rapid Approach for Proton Transport and Other Reactions (RAPTOR) Software Package. *J. Phys. Chem. B* **2024**,

(12) Levitt, M.; Warshel, A. Computer Simulation of Protein Folding. *Nature* **1975**, *253*, 694–698.

(13) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(14) Glaubitz, C.; Gröger, A.; Gottschalk, K.; Spooner, P.; Watts, A.; Schuldiner, S.; Kessler, H. ^{31}P-CP–MAS NMR Studies on TPP^+ Bound to the Ion-Coupled Multidrug Transport Protein EmrE. *FEBS Letters* **2000**, *480*, 127–131.

(15) Beckett, D.; Voth, G. A. Unveiling the Catalytic Mechanism of GTP Hydrolysis in Microtubules. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, e2305899120.

(16) Jin, J.; Pak, A. J.; Durumeric, A. E. P.; Loose, T. D.; Voth, G. A. Bottom-up Coarse-Graining: Principles and Perspectives. *J. Chem. Theory Comput.* **2022**, *18*, 5759–5791.

(17) Jin, J.; Pak, A. J.; Voth, G. A. Understanding Missing Entropy in Coarse-Grained Systems: Addressing Issues of Representability and Transferability. *J. Phys. Chem. Lett.* **2019**, *10*, 4549–4557.

(18) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–49.

(19) Sinitskiy, A. V.; Voth, G. A. Quantum Mechanics/Coarse-Grained Molecular Mechanics (QM/CG-MM). *J. Chem. Phys.* **2018**, *148*, 014102.

(20) Mironenko, A. V.; Voth, G. A. Density Functional Theory-Based Quantum Mechanics/Coarse-Grained Molecular Mechanics: Theory and Implementation. *J. Chem. Theory Comput.* **2020**, *16*, 6329–6342.

CHAPTER 2

THE C-TERMINUS OF THE MULTI-DRUG EFFLUX PUMP EMRE PREVENTS PROTON LEAK BY GATING TRANSPORT

These works are in collaboration with our experimantal collaborator, Ms. Merissa Brousseau, Dr. Nathan Thomas, and Prof. Katherine Henzler-Wildman from the University of Wisconsin, Madison. They demonstated the proton leaking behavior of a tail truncation mutant of EmrE named Δ107-EmrE. Their experimental results are included in the Appendix as the basis of our computational research.

2.1 Introduction

Antibiotic resistance can result from various mechanisms, including the active removal of drugs by versatile multidrug efflux pumps.[1] Within these pumps, the small multi-drug resistance (SMR) family is widespread among bacteria, showing diverse substrate preferences.[2–4] *Escherichia coli*'s EmrE is the most extensively studied SMR transporter, providing resistance against a wide range of toxic compounds through secondary active transport.[5] These transporters utilize proton motive force (PMF) for efficient drug export, making EmrE a prime model for understanding proton-coupled transport mechanisms.

EmrE exhibits electrogenic transport for tetraphenyl-phosphonium (TPP^+) and electroneutral transport for methyl viologen (MV^{2+}), consistent with a $2H^+$:1 drug antiport mechanism.[6] Early mechanistic models concentrated on the essential states and transitions required for this stoichiometric antiport. Recent NMR studies of EmrE's protonation and alternating access revealed additional states and transitions that play significant roles at near-physiological conditions.[7] These findings expanded the mechanistic model to include pathways for alternative transport activities like symport, drug-uniport, and proton-uniport (leak) (Figure 2.1). Different environmental conditions or small molecule substrates can alter

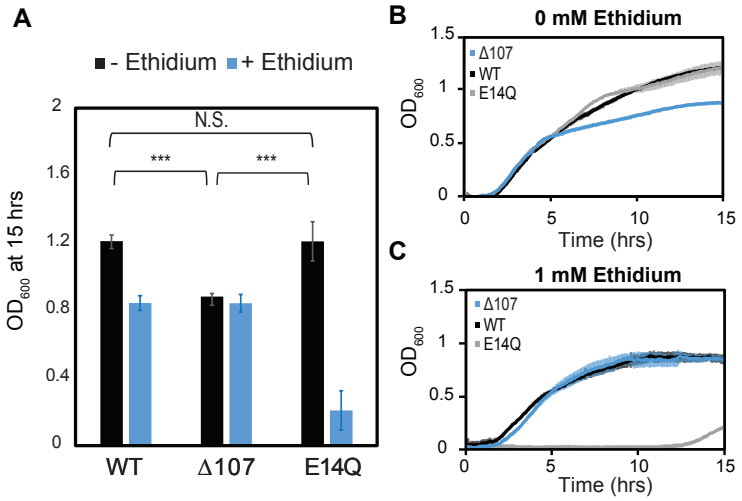


Figure 2.1: **Model of coupled antiport and uncoupled proton leak through EmrE.** All of the drug- and proton-bound states that are reasonably populated at near physiological temperature and pH and the transitions between these states observed by NMR leads to a model for EmrE transport that allows for both coupled antiport (A, orange) and proton leak (A, red solid line). In WT-EmrE, the C-terminal tail on the open face acts as a secondary gate, minimizing proton leak in the absence of substrate (B, top). Drug binding to a secondary binding site near the tail opens the gate and allows proton entry and exit to the primary binding site near E14, and either coupled antiport (A, orange) or drug-gated proton leak (A, red dashed line) depending on the substrate. Truncation of the C-terminal tail in Δ107-EmrE removes this secondary gate and allows uncoupled proton leak in the absence of substrate (B, bottom).

the transport behavior by shifting the rates of microscopic steps.[7,8] This has been confirmed experimentally, as the small molecule substrate harmane causes uncontrolled proton leak through EmrE, negatively impacting *E. coli* growth and NADH production (Figure 2.1).[9] However, how EmrE avoids catastrophic leakage remains an unresolved question.

Measurements directly monitoring proton release upon drug binding indicate that drug-induced deprotonation occurs at the C-terminal histidine (H110), along with the essential glutamate-14 residues that define the primary drug and proton binding site.[10] Furthermore, the C-terminus in one protomer of the homodimer is highly sensitive to the specific drug bound in the primary site.[11] Early solid-state ^{31}P NMR experiments suggested the presence of a second, lower affinity TPP^+ binding site near the acidic loop residues E25 and D84,[12] which are likely in close spatial proximity to the C-terminal tail in this small transporter. Based on these findings, we propose a secondary gating model where the C-terminal tail inhibits proton release until drug binding at a peripheral site on the transporter surface displaces the tail.[10]

Unfortunately, the available structures of EmrE do not offer high-resolution insights into the conformation of the C-terminal tail and adjacent loop regions. Early CryoEM and crystallographic structures revealed the unique asymmetric arrangement of the transmembrane helices and antiparallel topology of the EmrE homodimer, but they lacked resolution and density in the loops and tails.[13,14] Recent higher-resolution crystal structures and NMR structures have improved our understanding of substrate binding within the transport pore.[15–17] However, the crystal structures utilized a monobody that necessitated mutation of three residues in the TM1–TM2 loop (E25N, W13I, V34M), including E25, and had limited or missing density for other loops on the transporter's open face and the C-terminal tail after residue 104. In the NMR structures, distance restraints were confined to substrate-protein distances within the transmembrane helices lining the primary binding site, with only chemical-shift-derived backbone torsion angles restraining the loops and tail. As a result,

there is no structural model for the last 5 residues of EmrE, and experimental data on the loops are limited. However, these regions likely play a crucial role in gating access to the central binding site defined by residue E14. In such cases, molecular dynamics (MD) simulations have proven invaluable, providing the only atomic resolution model of the C-terminal tail from MD simulations.[18]

Our collaborator has studied a C-terminal deletion mutant of EmrE truncated after residue 106, denoted Δ107-EmrE to directly determine the regulatory role of the C-terminal tail.(See Appendix) Comparisons of growth and resistance phenotypes, alternating access rates, and transport activities of Δ107- and WT-EmrE confirm the importance of the C-terminal tail in regulating tightly coupled antiport and minimizing proton leak. They have discovered that the truncation of the C-terminal tail enhances proton leak through EmrE. Detailed results from their experiments can be found in Appendix.

Here, based on these experimental findings, we are using molecular dynamics simulations to give structural and mechanistic insights into this phenomenon. Our simulations on these two systems suggest differences in water structure and hydrogen bonding patterns when the C-terminus is truncated. This gives a possible secondary site and the structural basis for this regulation.

2.2 Results

2.2.1 The C-terminus Controls a Water Wire into the Primary Binding Site

To investigate how the C-terminal tail of EmrE may interact with other regions of EmrE and gate access into the transport pore, we carried out MD simulations of substrate-free WT-EmrE and Δ107-EmrE in a DMPC lipid bilayer. The protonation states were set to simulate a pH between 7.0 and 8.0, where only $E14^{A}$ with the higher pK_a is protonated. We used an NMR structure determined with TPP^+ (PDB: 7JK8) as the initial structural

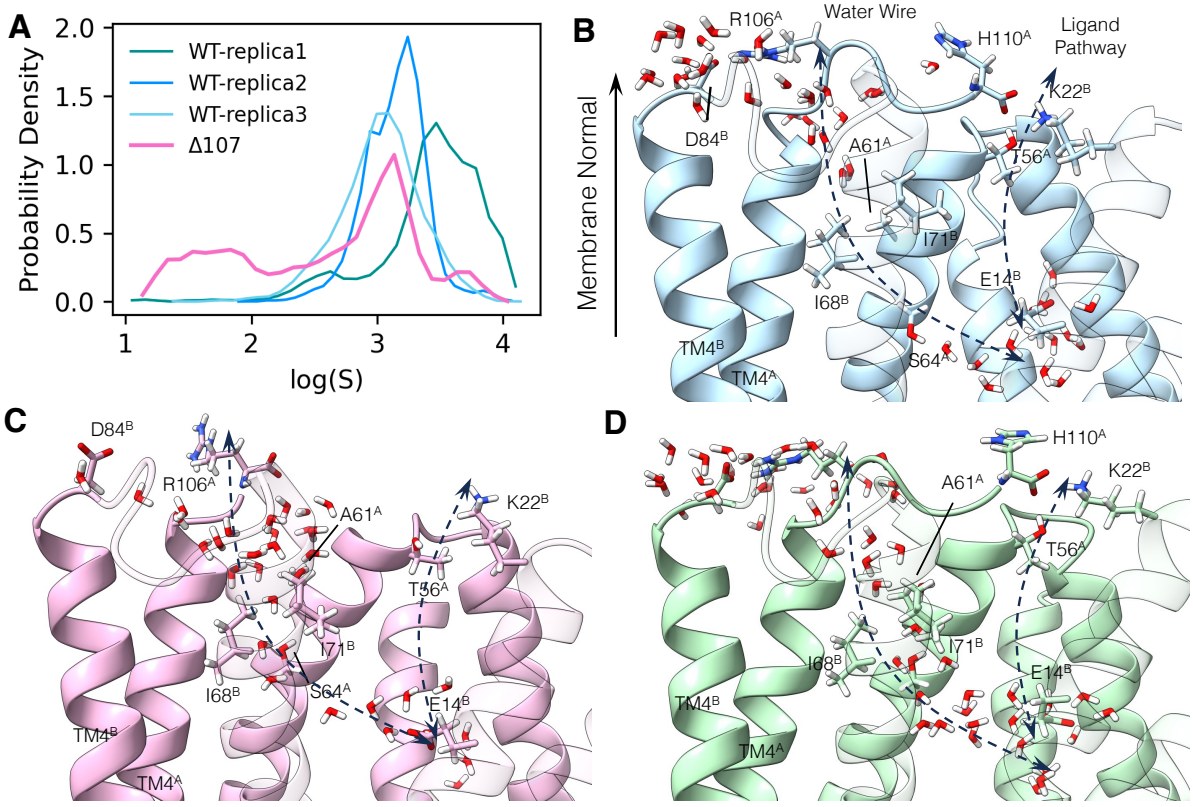


Figure 2.2: **The C-terminus tail caps the water wire from the open side.** (A) Logarithm of the minimum water distance $\log(S)$ histogram. (B-D) The following panels illustrate a few snapshots in the simulation. The membrane normal vector points to the open side of EmrE. Two dash line arrows show the ligand pathway and the water chain respectively. TM1 to TM3 in subunit B is shown transparently to show what is at the interface between the two subunits. (B) Dry snapshot of WT-EmrE. (C) Wet snapshot of Δ107-EmrE (D) A rare event snapshot when WT-EmrE is hydrated. These color codes are the same as in Figure 2.3.

model.[17] Since this NMR structure does not include the C-terminal tail, we modeled it with CHARMM-GUI.[19] The systems were first equilibrated at constant temperature 310 K and 1 bar pressure for 400 ns, while position restraints were gradually released. After this equilibration, the RMSD of the protein compared with the initial structure plateaued. The production simulations were run for another 1000 ns and MD trajectories were output every 0.1 ns. With the same starting structure, we ran three parallel replicas to ensure consistency.

In the Δ107-EmrE MD simulations, we identified a water chain (aka "water wire") not seen in WT simulations that connects E14 at the primary binding site to bulk water. It enters the protein from the open side near $R106^{A}$, passing through the triad of $A61^{A}$, $I68^{B}$, and $I71^{B}$, and then comes into the primary binding site (Figure 2.2C). To quantitatively understand the connectivity of this water wire over time, we calculated the shortest water path length S for each frame in the trajectory.[20] This water path length is defined to be equal to the number of water molecules in the chain when the water molecules are perfectly connected. When some adjacent water molecules are too far away, this length increases dramatically. Therefore, the shorter this distance is, the better water molecules are connected. The logarithm of this distance $\log(S)$ is plotted for all simulation systems (Figure 2.2A). For Δ107-EmrE, there is a leaking state characterized by a much smaller $\log(S)$, where the water wire is very well connected, and for WT-EmrE, this $\log(S)$ is consistently large.

Previously, despite knowledge about proton release from the protein, the microscopic pathway for such proton transport remained largely unknown. A critical prerequisite of proton transport is a water wire, either transient or long-lasting, that allows protons to transport through the Grotthuss hopping mechanism.[21] The existence of this water chain in Δ107-EmrE is consistent with the enhanced proton leak observed experimentally. This newly found water wire starts very close to the C-terminus and is distinct from the ligand entry path.[22]

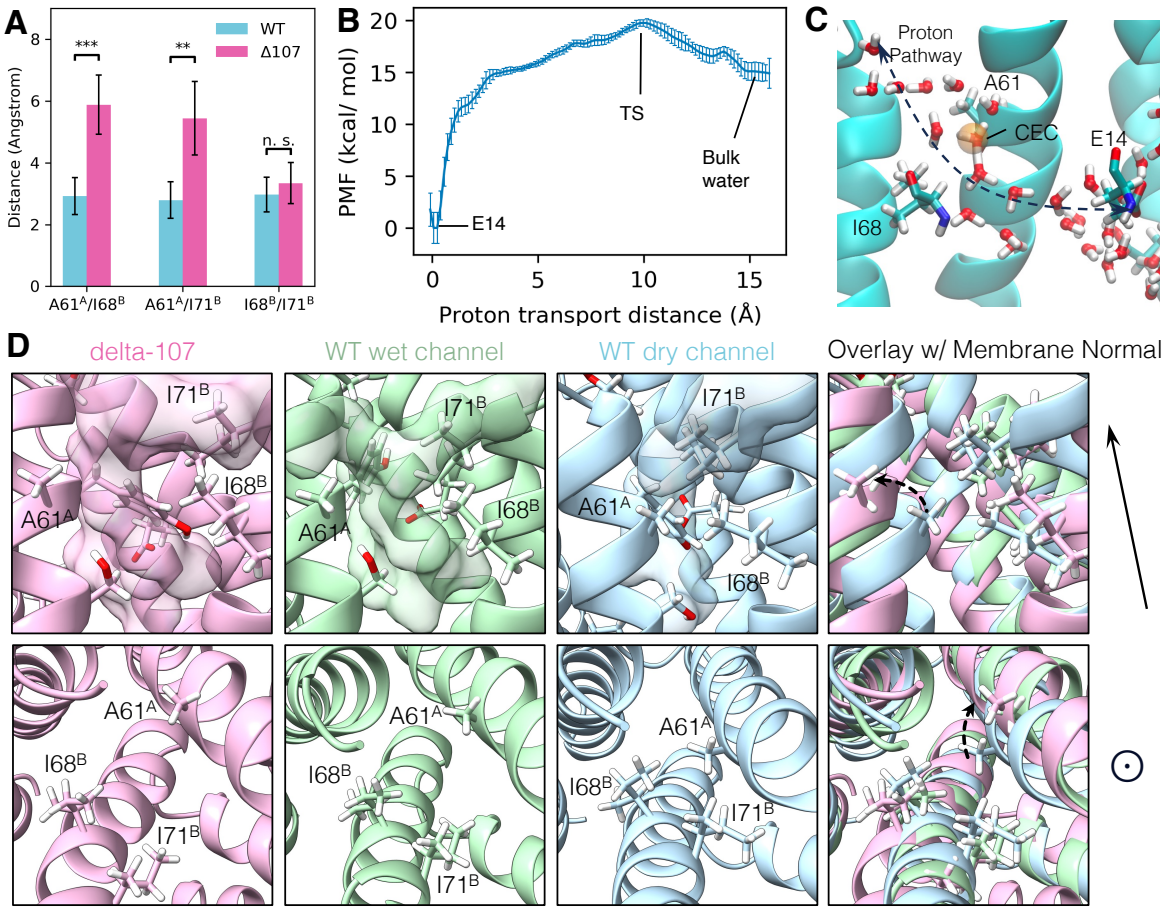


Figure 2.3: **The structural basis of the C-terminus gating.** (A) The minimum distance between side chain hydrogens between $A61^A$, $I68^B$, and $I71^B$. In WT-EmrE, the sidechain of $A61^A$ is significantly closer to I68B and I71B, while the distance between $I68^B$ and $I71^B$ is not significantly changing. The error bars show the standard deviation along the trajectory. All p-values were calculated from a two-sided t-test. (**) $p < 0.01$ (***) $p < 0.001$ (B) The proton transport PMF, as a function of the distance between the CEC and the donor E14 on the direction of transport (See Eq. 2.1 in Methods). (C) A snapshot of the transition state. The orange sphere is the proton center of the excess charge (CEC) (D) Conformations of $A61^A$, $I68^B$, $I71^B$ triad from two different angles. The membrane normal showed at the right points to the open side of EmrE. The upper panels are from a side view, and the lower panels are from a top-down angle, where the vector points outside from the paper. The transparent surface in the upper panels shows the water wire. Unlabeled residues were $E14^B$, $Y60^A$, and $S64^A$.

2.2.2 Structural Basis for the C-terminus Gating

In the initial structure of WT-EmrE, the C-terminus is floating in the water. After equilibration, we observed the tail come closer and interact with the protein in all three replicas. The tail has two notable interactions with other part of the protein. One is a salt bridge between D84 and R106. The other is involves the carbonyl group of the C-terminus, which forms hydrogen bonds with T56, and sometimes salt bridges with K22 (Figure 2.4). When these interactions occur, the tail comes in close proximity to residues near the water chain. Examining characteristic snapshots from the simulation trajectories illustrates the effect of the C-terminal tail on water wire formation. Figure 2.2B shows the worst-hydrated snapshot of WT-EmrE determined by the highest S. Figure 2.2C shows the best-hydrated snapshot of Δ107-EmrE determined by the lowest S. Despite the overall dryness in the channel of WT-EmrE, there were a few rare moments where a transient water wire formed, characterized by a sudden drop in water path length, and Figure 2.2D shows the best-hydrated snapshot from that simulation. In WT-EmrE, the water wire is broken at the triad of $A61^{A}$, $I68^{B}$, and $I71^{B}$ (Figure 2.2B, 2.3D), suggesting that these three hydrophobic residues may act as a bottleneck for the water wire.

To test this hypothesis with statistically meaningful results, we calculated the distances between the closest pairs of side-chain hydrogens among these three residues for the whole trajectory. For WT-EmrE, the minimum hydrogen-hydrogen distances of $A61^{A}$–$I68^{B}$ and $A61^{A}$–$I71^{B}$ are 2.9 ± 0.6 Å and 2.8 ± 0.6 Å respectively, and for Δ107-EmrE, these distances increased to 5.9 ± 1.0 Å and 5.4 ± 1.2 Å (Figure 2.3A). The $I68^{B}$–$I71^{B}$ distance does not differ significantly. These increased distances suggest that $A61^{A}$ is moving away in the Δ107-variant, opening a pore for the water wire to form. For comparison, the diameter of a water molecule is measured to be around 2.7 Å.[23] This means a triangle larger than 5.4 Å in sides is required for a water molecule to fit inside. Additionally, the most hydrated snapshot in WT-EmrE showed $A61^{A}$ takes a conformation more similar to Δ107-EmrE and

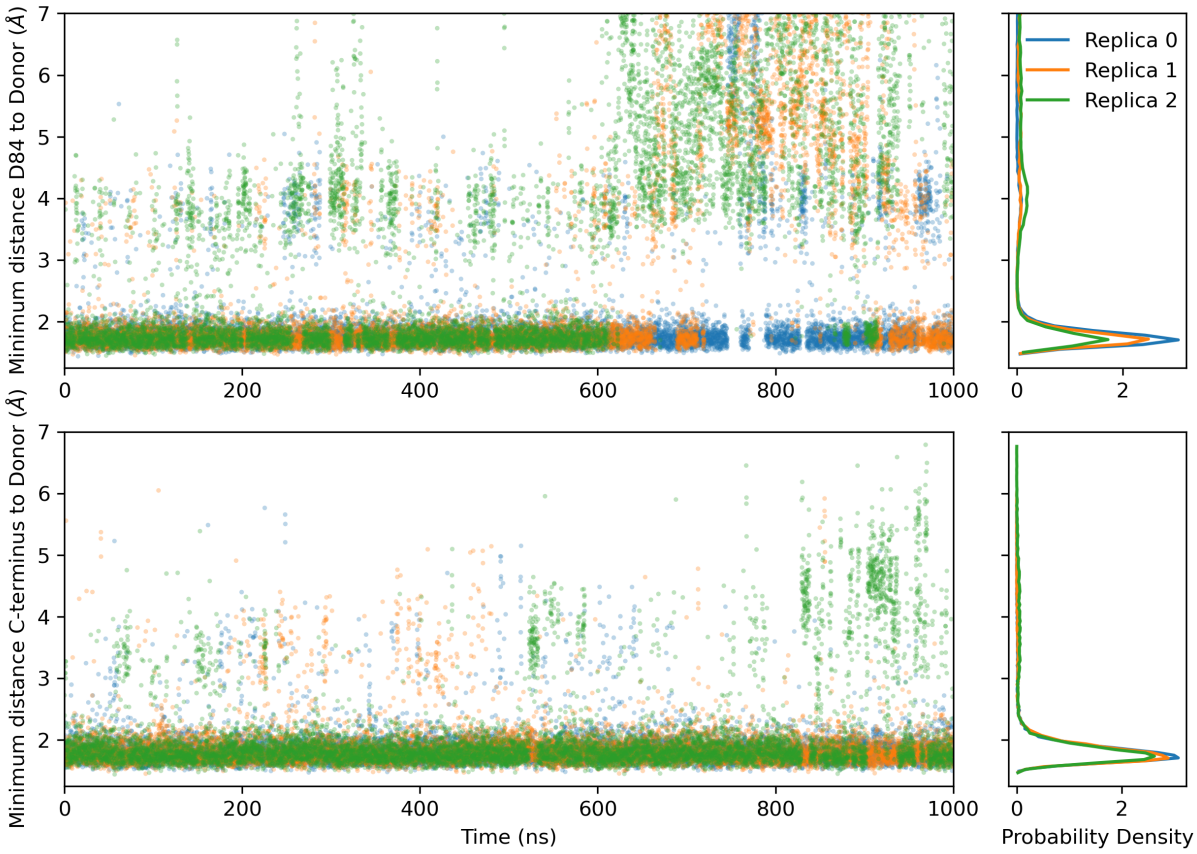


Figure 2.4: **Hydrogen bonding of D84 and the C-terminal tail.** The hydrogen bonding of D84 and the carboxylate group of the C-terminal can be characterized by the distances between these oxygens and their nearest donor hydrogen. The left two panels show the time series of these distances, and the right two panels shows the probability density. The donor of the C-terminus is the side chain of T56. While most of the donor for D84 is R106, there is a small probability density of H-bonding to S105 near 4 Å.

very different from dry WT-EmrE. This confirms the role of $A61^{A}$ rotation in controlling the water wire formation.

2.2.3 The Potential of Mean Force of Proton Transport Supports a Hydrophobic Bottleneck

The potential of mean force (PMF) for explicit proton transport shows the free energy change of the system as a function of a particular collective variable (CV) or reaction coordinate.

It can provide additional information beyond structural snapshots of a reaction, which are only incomplete samples of the ensemble. For example, the existence of a water chain does not necessarily mean good proton conductance, but by contrast, the PMF for explicit proton transport (including Grotthuss proton shuttling) can provide key thermodynamic and kinetic information about the PT process.[20,24] However, free energy sampling involving proton transport is also intrinsically complicated. First, proton transport involves chemical bond breaking and formation is beyond the capability of classical MD, so expensive quantum mechanics/molecular mechanics (QM/MM) may be required. Second, when an excess proton is solvated in the water, the net positive excess charge defect arising from the presence of the excess proton can be delocalized in the water network. It is not possible to define which proton is exactly the "excess proton" as Grotthuss shuttling dynamically rearranges these definitions. To address these issue, we have developed a method called Multiscale Reactive MD (MS-RMD).[25] It can model bond forming and breaking involving excess proton shuttling at a computational cost near classical MD. In this approach, one can also conveniently define a "center of excess charge" (CEC) to describe the location of the excess positive charge defect.

We carried out umbrella sampling with MS-RMD that describes the proton transport from E14B to the bulk water in WT-EmrE. The CV "x" we chose is defined similarly as in Ref. 26, which is the distance between the glutamate oxygen to the CEC, mapped along a vector that aligns with the proton transport direction. The resulting PMF shown in Figure 2.2B indicates a deep well near $x = 0\,\text{Å}$ where the proton is on the glutamate, and a transition state near $x = 10\,\text{Å}$. We took a conformational snapshot from the transition state as shown in Figure 2.2C. The location of the CEC is marked with a transparent orange sphere, and it is in close proximity to the pore defined by I68 and A61. This result supports the hypothesis that the bottleneck of the proton transport is this hydrophobic gate.

2.3 Methods

2.3.1 Molecular Dynamics

All MD simulations were conducted with GROMACS 2020.4.[27] Simulation inputs were generated by CHARMM-GUI membrane bilayer builder.[19] The protein was solvated by 162 DMPC molecules, and 40 mM NaCl was added to the water to neutralize the system. The system was coupled to a Nose-Hoover thermostat[28,29] and a Parrinello-Rahman barostat[30], at 310 K and 1 bar, respectively. The system was minimized, then equilibrated with position constraints gradually releasing, as by the default setting of CHARMM-GUI. Then, the system was further equilibrated for 400 ns without constraint, where the RMSD plateaued, and the box sizes were stable. For the WT Δ107-EmrE simulation, we observed lipids penetrating the protein in this 400 ns equilibration, so we added another 100 ns simulation with backbone constraints to further equilibrate the membrane before releasing these constraints again. Figures were rendered with ChimeraX. The hydrogen distances were analyzed with PLUMED,[31] by computing the softmin of all hydrogen-hydrogen distances between two sidechains with $\beta = 500$.

2.3.2 Water Path Length Calculation

The water path length calculation was implemented in an in-house modification of PLUMED. The algorithm can be briefly described as follows: Each water oxygen is considered a node in a graph, and the distance for each edge connecting two node is determined by a function that is close to 1 when the oxygen-oxygen distance is smaller than a r_0 and grows rapidly when it is larger than r_0. This r_0 is set to be 3 Å, which is the typical distance between the oxygens of hydrogen-bonded water. A more detailed discussion can be found in reference 20. Then, for each frame in the trajectory, the shortest path is found for the graph. We used the two oxygens of $E14^B$ at the starting point and the midpoint of C_α of $R102^A$ and $G57^A$

as the destination.

2.3.3 Umbrella sampling with MS-RMD

The simulations were run with the LAMMPS MD engine and the umbrella sampling was carried out as implemented in PLUMED.[31,32] The codes were co-compiled with a plug-in developed in house called RAPTOR to model proton transport reactions.[25]

The starting structure was taken from the classical MD simulation of WT-EmrE. A water molecule was protonated at the mouth of the channel and steered MD was then used to create initial configurations at different CV values. The CV is defined as

$$x = \mathbf{d}_{\mathrm{O-C}} \cdot \mathbf{e}_{PT} \tag{2.1}$$

where $\mathbf{d}_{\mathrm{O-C}}$ is a vector pointing from the closer glutamate oxygen to the CEC, and $\mathbf{e}_{PT}$ is a unit vector of the direction of proton transport. A total of 43 umbrella windows spanning from $x = 0.0\,\text{Å}$ to $15.0\,\text{Å}$ were used, with a varying restraint force constant of 80 to 15 $2.4\,\text{kcal/mol/Å}^2$. Each umbrella window was equilibrated for 1 ns and then the production run was for 2 ns. The PMF was reconstructed with the weighted histogram analysis method (WHAM)[33]. In these simulations, the C_α of the residues at least 10 Å away from the path and those in TM1-2 were restrained to its initial coordinate with a $2.4\,\text{kcal/mol/Å}^2$ harmonic potential to ensure the bias force does not unrealistically distort the protein conformation.

2.4 Discussions

Traditionally, transporters have been categorized as antiporters, symporters, or uniporters. However, there's a growing understanding that uniport, symport, and antiport merely represent extremes of a unified transport model encompassing all conceivable binding and conformational states and their transitions.[34] EmrE, one of the smallest and most extensively

studied multidrug resistance transporters, has emerged as a model system for investigating coupled transport.[5] Despite initial expectations that this diminutive transporter would have a straightforward mechanism and elucidate the minimal requirements for coupled transport, EmrE has proven unexpectedly complex, revealing unforeseen aspects of membrane protein topology and transport mechanisms.

The free exchange model, an expansion of a universal 8-state transport model to incorporate EmrE's ability to bind two protons at the two E14 residues in the core of the homodimer,[7] largely fits the available data. It encompasses all observed states and transitions identified by NMR and explains EmrE's capability to confer resistance to some substrates while making it susceptible to others.[9] However, this model predicts rapid proton leakage through WT-EmrE, whereas experimental data indicates a minor proton leakage of a smaller magnitude and similar timescale to coupled transport. Coupled with earlier findings emphasizing the significance of the C-terminal tail,[10] the experimental data and MD simulations presented here endorse a regulatory role for the C-terminal tail as part of a secondary gate that minimizes proton leakage in the absence of substrate and can be opened upon binding a drug-substrate.

Previous NMR investigations[9,10,12] prompted the notion that the C-terminal tail functions as a secondary gate, blocking the primary binding site defined by E14 in the absence of drug-substrate. Upon drug binding to a peripheral site, this gate opens, allowing the release of protons from E14. This model elucidates the observed connection between the C-terminal tail, drug-binding events, and protonation at the primary site,[10] as well as the correlation between proton off-rate and substrate on-rate in earlier stopped-flow studies of EmrE.[35]

Meanwhile, unlike the transport of other ions, proton transport requires water as a medium. Due to the lack of structural information and experimental technique, very few is known about the water formation in this flexible protein. With the latest experimental structure available, we are able to use the tail-truncated mutant, Δ107-EmrE, to gain more

understanding on the proton transport mechanism of this protein. Through many replicas of simulation, we have found out the key interactions involving D84, T56, and several residues on the tail that keeps it locked in place.

Comparison of MD simulations between WT and Δ107-EmrE reveals that the C-terminus controls a water wire into the primary binding site. Surprisingly, this water wire is not the most straightforward way of connecting the primary binding site to the outside. Instead of going straight outside, it enters the protein from the TM3–TM4 loop. This finding offers a structural explanation for how the tail regulates access to the primary EmrE binding site at E14 and controls proton entry and exit from that site, crucial for preventing proton leakage.

References

(1) Munita, J. M.; Arias, C. A. Mechanisms of Antibiotic Resistance. *Microbiol. Spectr.* **2016**, *4*, 10.1128/microbiolspec.vmbf–0016–2015.

(2) Brown, M. H.; Skurray, R. A. Staphylococcal Multidrug Efflux Protein QacA. *J. Mol. Microbiol. Biotechnol.* **2001**, *3*, 163–170.

(3) Pérez-Varela, M.; Corral, J.; Aranda, J.; Barbé, J. Roles of Efflux Pumps from Different Superfamilies in the Surface-Associated Motility and Virulence of Acinetobacter Baumannii ATCC 17978. *Antimicrob. Agents Chemother.* **2019**, *63*, 10.1128/aac.02190–18.

(4) Schuldiner, S. EmrE, a Model for Studying Evolution and Mechanism of Ion-Coupled Transporters. *Biochimica et Biophysica Acta - Proteins and Proteomics* **2009**, *1794*, 748–762.

(5) Yerushalmi, H.; Schuldiner, S. A Model for Coupling of H^+ and Substrate Fluxes Based on "Time-Sharing" of a Common Binding Site. *Biochemistry* **2000**, *39*, 14711–14719.

(6) Rotem, D.; Schuldiner, S. EmrE, a Multidrug Transporter from *Escherichia Coli*, Transports Monovalent and Divalent Substrates with the Same Stoichiometry*. *J. Biol. Chem.* **2004**, *279*, 48787–48793.

(7) Robinson, A. E.; Thomas, N. E.; Morrison, E. A.; Balthazor, B. M.; Henzler-Wildman, K. A. New Free-Exchange Model of EmrE Transport. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, E10083–E10091.

(8) Hussey, G. A.; Thomas, N. E.; Henzler-Wildman, K. A. Highly Coupled Transport Can Be Achieved in Free-Exchange Transport Models. *Journal of General Physiology* **2020**, *152*.

(9) Peyton J. Spreacker; Nathan E. Thomas; Will F. Beeninga; Merissa Brousseau; Colin J. Porter; Kylie M. Hibbs; Katherine A. Henzler-Wildman Activating Alternative Transport Modes in a Multidrug Resistance Efflux Pump to Confer Chemical Susceptibility | Nature Communications. *Nat. Commun.* **2022**, *13*, 7655.

(10) Thomas, N. E.; Wu, C.; Morrison, E. A.; Robinson, A. E.; Werner, J. P.; Henzler-Wildman, K. A. The C Terminus of the Bacterial Multidrug Transporter EmrE Couples Drug Binding to Proton Release. *Journal of Biological Chemistry* **2018**, *293*, 19137–19147.

(11) Morrison, E. A.; Henzler-Wildman, K. A. Transported Substrate Determines Exchange Rate in the Multidrug Resistance Transporter EmrE. *Journal of Biological Chemistry* **2014**, *289*, 6825–6836.

(12) Glaubitz, C.; Gröger, A.; Gottschalk, K.; Spooner, P.; Watts, A.; Schuldiner, S.; Kessler, H. ^{31}P-CP–MAS NMR Studies on TPP^+ Bound to the Ion-Coupled Multidrug Transport Protein EmrE. *FEBS Letters* **2000**, *480*, 127–131.

(13) Fleishman, S. J.; Harrington, S. E.; Enosh, A.; Halperin, D.; Tate, C. G.; Ben-Tal, N. Quasi-Symmetry in the Cryo-EM Structure of EmrE Provides the Key to Modeling Its Transmembrane Domain. *J. Mol. Biol.* **2006**, *364*, 54–67.

(14) Chen, Y.-J.; Pornillos, O.; Lieu, S.; Ma, C.; Chen, A. P.; Chang, G. X-Ray Structure of EmrE Supports Dual Topology Model. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18999–19004.

(15) Kermani, A. A.; Burata, O. E.; Ben Koff, B.; Koide, A.; Koide, S.; Stockbridge, R. B. Crystal Structures of Bacterial Small Multidrug Resistance Transporter EmrE in Complex with Structurally Diverse Substrates. *eLife* **2022**, *11*, 1–26.

(16) Shcherbakov, A. A.; Spreacker, P. J.; Dregni, A. J.; Henzler-Wildman, K. A.; Hong, M. High-pH Structure of EmrE Reveals the Mechanism of Proton-Coupled Substrate Transport. *Nat. Commun.* **2022**, *13*, 991.

(17) Shcherbakov, A. A.; Hisao, G.; Mandala, V. S.; Thomas, N. E.; Soltani, M.; Salter, E. A.; Davis, J. H.; Henzler-Wildman, K. A.; Hong, M. Structure and Dynamics of the Drug-Bound Bacterial Transporter EmrE in Lipid Bilayers. *Nat. Commun.* **2021**, *12*, 1–13.

(18) Vermaas, J. V.; Rempe, S. B.; Tajkhorshid, E. Electrostatic Lock in the Transport Cycle of the Multidrug Resistance Transporter EmrE. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, E7502–E7511.

(19) Lee, J. et al. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **2016**, *12*, 405–413.

(20) Li, C.; Voth, G. A. A Quantitative Paradigm for Water-Assisted Proton Transport through Proteins and Other Confined Spaces. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2113141118.

(21) Agmon, N. The Grotthuss Mechanism. *Chem. Phys. Lett.* **1995**, *244*, 456–462.

(22) Jurasz, J.; Bagiński, M.; Czub, J.; Wieczór, M. Molecular Mechanism of Proton-Coupled Ligand Translocation by the Bacterial Efflux Pump EmrE. *PLoS Computational Biology* **2021**, *17*, 1–19.

(23) Schatzberg, P. Molecular Diameter of Water from Solubility and Diffusion Measurements. *J. Phys. Chem.* **1967**, *71*, 4569–4570.

(24) Ilan, B.; Tajkhorshid, E.; Schulten, K.; Voth, G. A. The Mechanism of Proton Exclusion in Aquaporin Channels. *Proteins: Structure, Function, and Bioinformatics* **2004**, *55*, 223–228.

(25) Kaiser, S.; Yue, Z.; Peng, Y.; Nguyen, T.; Chen, S.; Teng, D.; Voth, G. Molecular Dynamics Simulation of Complex Reactivity with the Rapid Approach for Proton Transport and Other Reactions (RAPTOR) Software Package. *J. Phys. Chem. B* **2024**,

(26) Li, C.; Voth, G. A. Accurate and Transferable Reactive Molecular Dynamics Models from Constrained Density Functional Theory. *J. Phys. Chem. B* **2021**, *125*, 10471–10480.

(27) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(28) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *The Journal of Chemical Physics* **1984**, *81*, 511–519.

(29) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(30) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(31) Bonomi, M. et al. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16*, 670–673.

(32) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.;

Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.

(33) Grossfield, Alan WHAM: The Weighted Histogram Analysis Method.

(34) Beckstein, O.; Naughton, F. General Principles of Secondary Active Transporter Function. *Biophys. Rev.* **2022**, *3*, 011307.

(35) Adam, Y.; Tayer, N.; Rotem, D.; Schreiber, G.; Schuldiner, S. The Fast Release of Sticky Protons: Kinetics of Substrate Binding and Proton Release in a Multidrug Transporter. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17989–17994.

CHAPTER 3

DETERMINING THE RATE LIMITING STEP OF PROTON LEAKING IN EMRE

3.1 Introduction

EmrE is a membrane protein belonging to the Small Multidrug Resistant (SMR) Transporters family found in *Escherichia coli*. As a secondary active transporter, it utilizes the electrochemical gradient of proton across its inner membrane to efflux toxins from the cell. EmrE has garnered significant attention due to its relevance to drug resistance and the unexpected complexity it exhibits despite being one of the smallest ion-coupled transporters.[1] EmrE consists of only 110 residues in each subunit, forming a homodimer with an unusual dual topology.[2,3] These two subunits are identical in sequence but possess different tertiary structures and are antiparallel to each other. Under its working conditions, EmrE antiports protons and drug molecules at an approximately 2:1 stoichiometry.[4] The substrates it effluxes are primarily aromatic and cationic, with sizes ranging from a few hundred to a few thousand Daltons.[5] A primary binding site was identified as two glutamates near the center of the protein. These are the two E14s from each subunit and are found to bind either proton or the cationic substrate.[6,7]

One of the most intriguing questions about EmrE is its unusual gating mechanism. For many secondary transporters, a canonical alternating access model can explain the coupled transport. Specifically for antiporters, this model requires that (a) the protein binds the two ligands or ions competitively, and (b) it only alternates access upon binding. These two conditions together prevent channel-like uncoupled transport, or leakage, of any of its substrates down its electrochemical gradient.[8] However, EmrE has been found in violation of both of these conditions. First, when EmrE is saturated with tetraphenylphosphonium (TPP^+), one of its most-studied toxic substrates, it can still bind protons.[9] This indicates that substrate

binding to EmrE is not exclusive. Second, alternating rate measurements have shown that EmrE will alternate access regardless of whether a substrate or proton is bound.[10] Instead of allowing for alternating access, substrate binding slows it down. Without these two conditions, one would expect proton leakage without the presence of the substrate. However, in reality, protons do not leak rapidly through EmrE under physiological conditions.[9] Otherwise, it will quickly dissipate the proton gradient established in the electron transport chain, eventually preventing ATP synthesis and kills the cell.

Interestingly, while it is still unclear why EmrE does not leak proton, a few conditions have been found where proton leaking can be triggered under certain conditions. Some substrates, such as harmane or crown ether, can trigger proton leaking through EmrE without introducing much conformation change.[11] In another case, a C-terminal deletion variant (Δ107-EmrE) missing the tail from S107 to the C-terminus residue H110 was found to enhance proton leak, while this effect does not synergize with the effect of harmane. These findings, together with the non-stoichiometric antiport of drugs and protons, strongly suggest that the transport of EmrE may be regulated kinetically. Under a kinetic model framework called "free exchange model," changes in the rates of some elementary steps in the kinetic network or concentration of some reactants can switch EmrE's behavior from antiporting to alternative modes, such as proton leaking or drug uniport.[12,13] Specifically, to understand the proton leaking behavior, it is essential to comprehend the elementary proton transport (PT) steps in the EmrE working cycle. While there are numerous experimental measurements of alternating rates and drug on/off rates, very few exist for proton on/off rates.[14] Direct measurements are intrinsically difficult due to the limited time and spatial resolution of experimental techniques, making it challenging to pinpoint transient proton translocations.

The relatively small size of EmrE has enabled extensive molecular dynamics (MD) studies.[15–19] MD can sample the conformation of a protein at atomistic resolution for a few microseconds, providing a great complement to NMR studies. These works have offered

structural details on the hydrogen bonding near the primary binding site and the ligand entry process. Despite these successes in structural studies, MD studies on proton transport (PT) processes in proteins are difficult and consequently rare. Unlike other ions that move through translocation, proton transport follows the Grotthuss shuttling mechanism, which involves simultaneously switching a covalent bond into a hydrogen bond and a hydrogen bond into a new covalent bond.[20] When the proton is solvated in water, the resulting net positive excess charge defect is delocalized in the water network. Defining a proper hydronium molecule in such a scenario is impossible, as Grotthuss shuttling dynamically rearranges these bonds.[21,22] Treating a delocalized positive charge defect as a hydronium ion can give wrong kinetic or thermodynamic properties. Therefore, modeling PT processes is beyond the capability of classical MD.

Nevertheless, structural insights provided by classical MD can help understand PT processes to some extent. Classical MD can identify water chains (also known as water wires) that are necessary as mediums for proton transport.[15] However, a lasting water wire is not a necessary condition for effective proton conductance. Hydration events sometimes have a large free energy barrier in a dry, hydrophobic pore and are, therefore, transient. Unbiased classical simulations often can only sample conformations near the system's lowest free energy state. The same is true for experimentally resolved structures. Excess solvated protons can "grow" a transient water wire within a hydrophobic environment to facilitate proton transport. Although such water wire formation is entropically unfavorable, the much stronger hydrogen bonding due to the excess charge can offset this penalty by enthalpy. We have shown in the case of ClC-ec1 and a carbon nanotube that proton can grow its own water wires and permeate through a dry, hydrophobic channel.[22–24] In these simulations, enhanced sampling methods on a hydration collective variable (CV) will be needed to sample these hydration events sufficiently.

To address these challenges—specifically the sampling of hydration events and the explicit

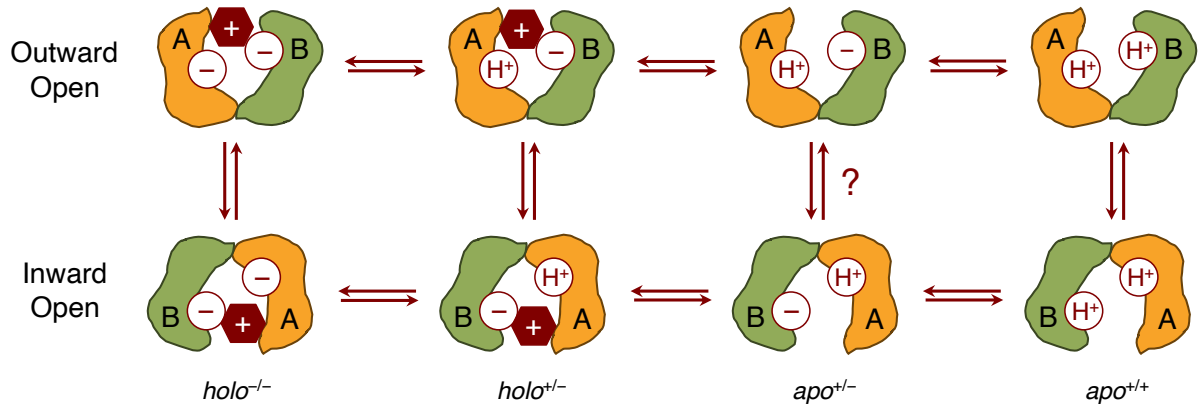


Figure 3.1: **Partial transport network of EmrE.** The free exchange model of EmrE without the two completely deprotonated apo state ($apo^{-/-}$). The *holo* states refers to those bound with a toxin substrate, and apo for those without. The protonation states are designated by subscripts of the protonation state of $E14^{A}$ and $E14^{B}$ respectively and in that order.

treatment of Grotthuss shuttling—our group has developed two methods: a differentiable collective variable (CV) to measure water connectivity,[23] and Multiscale Reactive Molecular Dynamics (MS-RMD), previously known as MS-EVB, to explicitly model proton transfer (PT) processes.[24] Our water connectivity CV can calculate a value ranging from 0 to 1 to measure the water connectivity along a specific path on the fly with any simulation. MS-RMD allows bond formations and breakings during the simulation at a cost comparable to classical molecular dynamics (MD), and it is orders of magnitude faster than first-principle methods such as QM/MM. Additionally, MS-RMD provides a definition for "center of excess charge" (CEC) to describe the location of the delocalized solvated proton, which can be used to monitor the proton transport progress. MS-RMD has been proven successful in simulating water and many proton transporters.[25–27]

In this chapter, we are using the above-described computational methods to review a few key proton transport steps in the proton transport cycle (Figure 3.1). First, we will show that proton moving between the two E14s is faster than alternating access and will not slow down alternating access of EmrE at singly protonated state of EmrE. Then, we will show

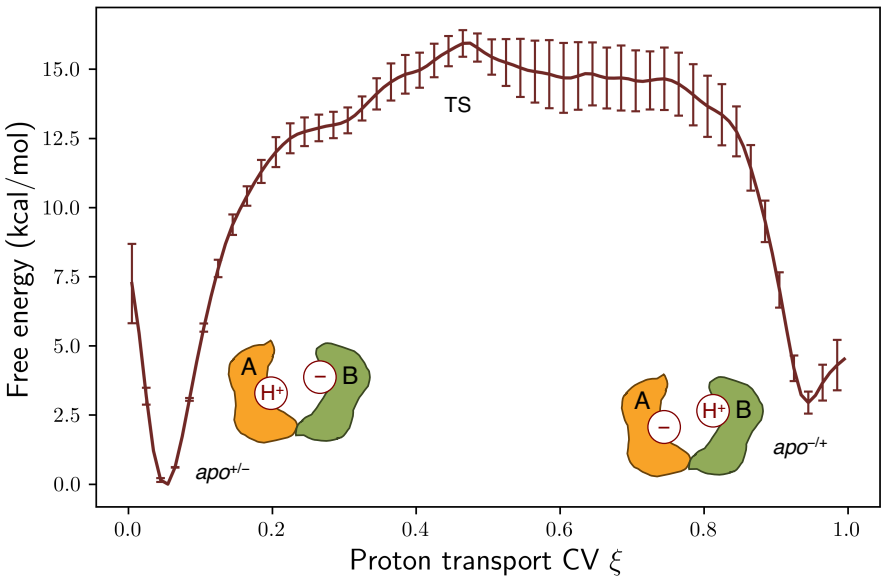


Figure 3.2: **PMF for proton transport between the two E14s.**

that the hydration of a proton pathway will be a slow step for proton leaking.

3.2 Results

3.2.1 Proton Hopping between the Two E14 is Fast

Since EmrE has a dual topology, the two E14 residues at the primary binding site exhibit distinct properties, despite being the same glutamate in the sequence. We designate the subunit with their C- and N-terminals both on the open side as subunit A, and the other one as B. These two glutamates have different pKa due to the asymmetric topology. At 25 °C, the pKa of the E14 on subunit A ($E14^A$) is approximately 8.5, and the other one on subunit B ($E14^B$) has a pKa of around 6.8.[28] These two glutamates are electrostatically uncoupled, meaning the protonation state of one does not affect the pKa of the other.[29] Given that the pK_a of E. coli cytoplasm is around 7.5, a large percentage of EmrE will be singly protonated at the E14A.

This dual topology and asymmetric pKa leads to an interesting question: how will the

proton react to alternating access in this singly protonated $apo^{+/-}$ EmrE? During alternating access, the pKa values of the two glutamates will be exchanged. This lowers the pKa of the previously protonated E14 and destabilizes the protonation.[30] It is unknown whether the proton will leave for the new E14A during alternating access, a phenomenon known as "proton hopping." The possibility of proton hopping has significant implications for the working cycle of EmrE. If it does not occur, alternating access of a correctly protonated EmrE ($apo^{+/-}$) will exchange with an incorrectly pronated one ($apo^{-/+}$). This will impose a free energy penalty for alternating access of 2.9 kcal/mol, making it an asymmetrical transition (Figure 3.1, step with a question mark).[29] Based on this measurement, a "hop-free" mechanism was proposed based on this measurement to explain coupled transport. If $apo^{+/-}$ state can only exchange with $apo^{-/+}$ state, alternating access for singly protonated EmrE species will be very slow. Consequently, the only way for alternating access to happen will be through the binding of a drug molecule, which would further deprotonate EmrE and recover the symmetry in the protonation state. This is supported by observations that the alternating rate is slower at higher pH where EmrE is singly protonated. They concluded that the rate-limiting step for proton leaking should be the alternating access of singly protonated EmrE.

However, a few findings do not align with the hop-free mechanism. Firstly, while alternating access is slower at higher pH, proton leaking was observed to be faster at higher pH. If alternating access is the rate-limiting step for proton leaking, one would expect more leaking at lower pH, contrasting the leaking measurements. Secondly, the measured $\sim$2.9 kcal/mol free energy difference only affects the thermodynamic equilibrium. It does not provide any information about the kinetics of the process, or how fast is the transition between these two states. A thermodynamically unfavorable process can still quickly establish an equilibrium between reactants and products if it has a low free energy barrier. In such case, the proton hopping thermodynamic barrier will not prevent alternating access. However, no

measurements have been done on the rates of the proton hopping so far.

To gain information about the kinetics of the process, we used MS-RMD with free energy sampling methods to calculate the potential of mean force (PMF) of the PT process between the two glutamates. The PMF can provide information about not only the differences between the starting and end states but also about the intermediate states along the reaction process. The reaction process for this proton hopping is defined with a single CV, which, in this case, is defined by the distance between the CEC and the starting glutamate, divided by the distance between the two glutamates (see Methods). This CV can describe the PT progress within a range of 0 to 1, for 0 being the starting state ($apo^{+/-}$, proton on E14A) and 1 being the end state ($apo^{-/+}$, proton on E14B). Starting from an $apo^{+/-}$ structure taken from Chapter 2, we used umbrella sampling to reconstruct the PMF (see Methods). The resulting PMF (Figure 3.2) clearly shows the two protonation states. The free energy difference between the two states is $2.9 \pm 0.4\,\mathrm{kcal/mol}$, which agrees well with experimental data. Additionally, the transition state (TS) has a height of $15.9 \pm 0.5\,\mathrm{kcal/mol}$.

Using transition state theory (TST), we can derive an upper bound for the reaction rate from the PMF, ignoring possible recrossing (see Methods). The forward rate is calculated to be $1.5 \times 10^3\,\mathrm{s}^{-1}$ (turnover time 0.66 ms), and the backward rate is $1.4 \times 10^5\,\mathrm{s}^{-1}$ (turnover time 7.1 μs). The forward rate is still approximately 15 times faster than the alternating rate of EmrE at this pH ($\sim 100\,\mathrm{s}^{-1}$), and the backward reaction is orders of magnitude faster. This high backward rate indicates that any incorrectly protonated EmrE can readily convert to the more favorable protonation state. The forward rate also suggests that, although slow, a protonation equilibrium can be established between the two E14 residues faster than the alternating access.

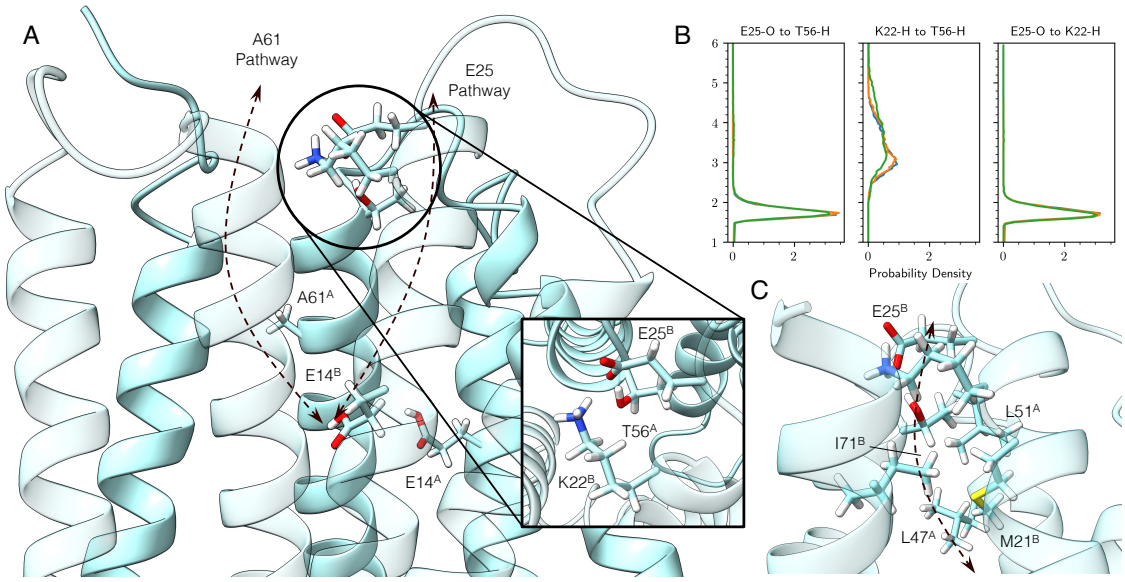


Figure 3.3: **In fully protonated EmrE, the open side complete closes by hydrogen bonding.** (A) EmrE conformation when fully protonated at E14 and H110. The dashed lines show the E25 pathway and the A61 pathway. The hydrogen bonding pattern is shown in the small window. (B) The distance between the donor-acceptor pair is plotted as the probability density. (C) Hydrophobic residues line along the E25 pathway.

3.2.2 EmrE Occludes Itself at Low pH

Both experimental[31] and computational studies14 have shown that when both E14s are protonated, a conformation change closes the mouth of the protein. Specifically, simulation with the faRM EmrE structure shows that the protonation at E14A shuts down a water pathway that connects E14 and the outside from E25.[15] Proton transport requires a water wire as a medium. This leaves an important question unanswered: how does the proton leave the protein when both E14s are protonated?

We first conducted a classical MD simulation of EmrE at low pH. The C-terminus residue H110 was found to be in close proximity to the mouth of the protein, so its protonation state needs to be treated carefully. NMR titrations have determined that this H110 has a pKa of around 7.0 at 25°, which is slightly higher than the pKa of $E14^B$ at 6.8 at the same temperature.[32] Therefore, when both E14s are protonated, H110s are also likely protonated.

Therefore, starting from the singly protonated structure above, we protonated all E14s and H110s to simulate a low pH scenario. A total of three replicas were run for this system. We first equilibrated the structure for 400 to 800 ns to wait for its RMSD against the starting structure to stabilize. These equilibrations were followed by 2.5 microseconds for each production simulation, where trajectories are kept every 1 ns. In all three replicas, we confirmed the observation in the faRM structure simulation that the pore is consistently dry.[15] In addition, we observed a pattern at the mouth of the protein that seals this opening (Figure 3.3A). E25 forms a consistent hydrogen bond with T56 and a salt bridge with K22 within the same helix. The probability density of the donor-acceptor pair between E25—T56 and E25—K22 is a sharp, narrow peak, while the distribution for T56—K22 is flatter and wider (Figure 3.3B). This suggests that T56 and K22 do not interact directly but are tethered together by E25. Several nearby hydrophobic residues (L51, I71, and M21) were also packed together to block the hydration of this water pathway (Figure 3.3C).

3.2.3 Hydration of the E25 pathway is possible upon a hydrogen bond breaking

The lack of hydration in experimental and computational structures suggests we need to know the PT path before investigating other aspects of the transport. We have mentioned previously that a dry pore in experimental structures or MD trajectories does not necessarily mean prohibited proton transport, as hydrations are often transient with a high free energy barrier, so they are hard to observe in experimental structures and unbiased MD simulations, as they only reflect the lowest energy conformations. Two pathways are possible for this proton transport. The first is the path directly connecting the E14s to the outside through E25 (Figure 3.3A). This path (referred to as the E25 pathway) was the most direct path and also conjured to be the ligand pathway. The other pathway (referred to as the A61 pathway) is the one we previously reported for proton leaking in Δ107-EmrE, which spanned from

D84 via a hydrophobic gate at A61 and enters the primary binding site near E14 (Figure 3.3A). We want to know if these two pathways can be hydrated at this occluded state.

We conducted well-tempered metadynamics with the water connectivity as the CV to get the free energy profile for the hydration of these two pathways, respectively referred to as the E25 pathway and the A61 pathway. When biasing the water connectivity in the E25 pathway simulation, large conformation changes was observed near the E25 area. Due to the bias added, E25 and the loop region next to it swung back and forth against the other subunit. To ensure sufficient sampling in those additional degrees of freedom other than the water connectivity, this slow degree of freedom needs to be explicitly addressed. Consequently, a conformational change CV was added to the metadynamics simulation, for which we picked the minimum distance between the donor and receptor in the E25–T56 hydrogen bond. A distance upper bound restraint is also added to this CV, so they do not go over 7.0 Å and tear apart the structure.

Figure 3.4 shows the resulting 2D free energy profile. The two axes correspond to the hydration and distance of the hydrogen bond donor-acceptor pair. In the PMF, the lower left basin corresponds to the state we get in the classical unbiased simulations. It shows that EmrE at this protonation state likes to stay dry and closed, with a water connectivity of 0.27 and a hydrogen bond distance of 1.7 Å. More importantly, hydration is only possible when this hydrogen bond is broken. At high water connectivity of 0.8, the free energy is significantly higher ($\sim 4\,\mathrm{kcal/mol}$, or $6.6\,kT$ at 310 K) for an unbroken hydrogen bond (1.7 Å) than a broken one (6.0 Å). This hydrogen bond between E25 and T56 functions like a gate that controls the water connection into E14. At this mouth region of the protein, the key interaction that glues these two subunits together is the hydrophobic interaction. Without opening this mouth part, it is much more difficult to hydrate this pathway. We then integrated along the conformation change CV and converted it into a 1-D PMF on water connectivity CV to compare with the A61 pathway. The result shows that it is almost

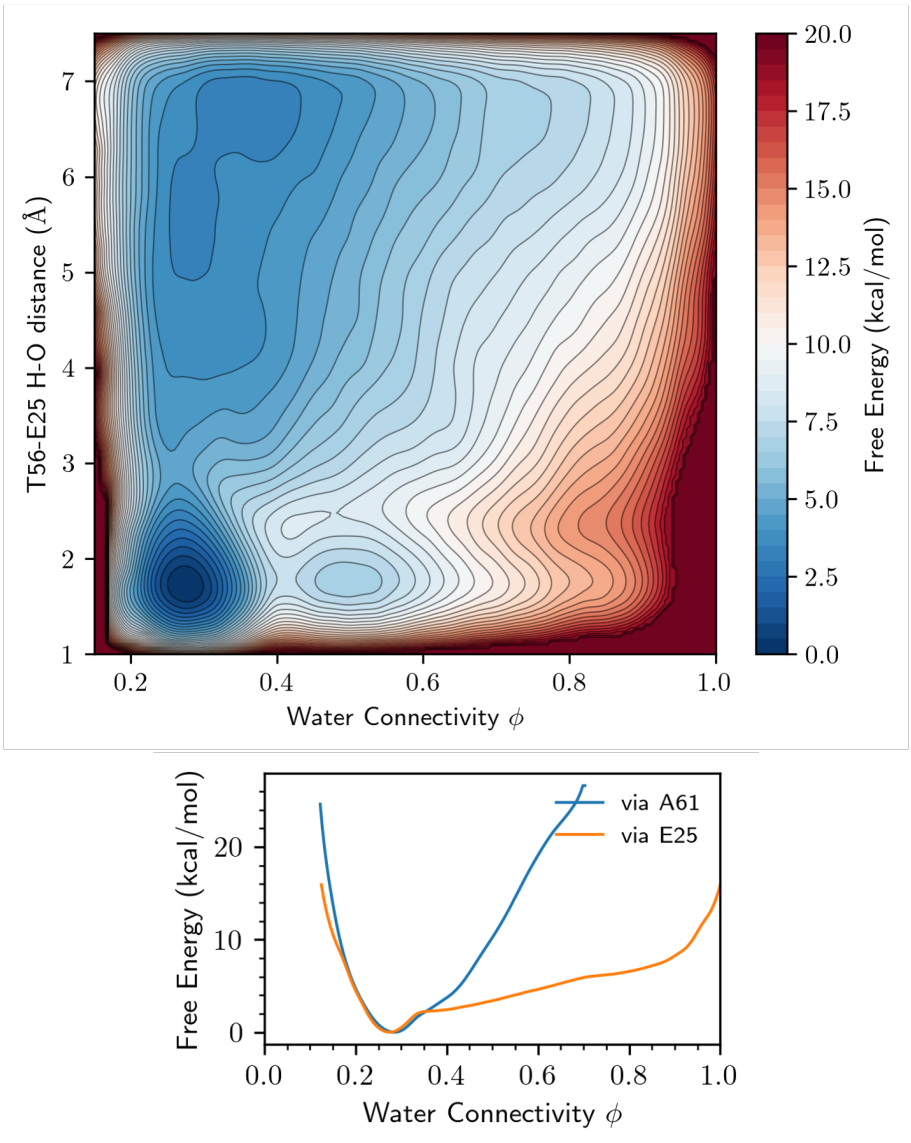


Figure 3.4: **PMF for hydration of the PT pathways.** The 2D PMF for hydration of the E25 pathway, and the integrated PMF with a comparison with the A61 pathway.

impossible to hydrate the A61 pathway under this protonation state. There is an unrealistic free energy barrier to hydrate it at this protonation state.

3.3 Methods

3.3.1 Simulation Software

Our classical simulations are carried out in GROMACS 2020.4 version with CHARMM36m force field and cation-pi interactions.[33] The classical simulations have an integration step of 2 fs using verlet integration. Van der Waals interactions and are cutoff with a switching function from 1.0 nm to 1.2 nm. Short-range electrostatic interactions have a cutoff of 1.2 nm and the long-range part is done with PME. A Nose-Hoover thermostat was coupled to the system at 310 K.[34,35] For NPT simulations, the Parrinello-Rahman barostat was also used with a reference pressure of 1.0 bar.[36] This GROMACS is patched with PLUMED 2.8 to facilitate enhanced sampling.The PLUMED 2.8 we used has an in-house modification to allow for calculations of the water connectivity CV.[37]

MS-RMD simulations are carried out with LAMMPS patched with interfaces to PLUMED 2.8 and RAPTOR. RAPTOR is a plugin for LAMMPS developed by the Voth group to allow reactive simulation of proton reactions. These simulations are done in NVT ensembles at $T = 310\,\mathrm{K}$. The dynamics are integrated at a step of 1 fs. More details on the method and its implementation is available in reference 24.

3.3.2 MS-RMD Umbrella Sampling with PT CV

We started with a structure taken from the previous chapter and constructed an MS-RMD model with the glutamate parameters in reference 38. The proton transport CV was defined as

$$\xi = \frac{\mathbf{e}_{\mathrm{PT}} \cdot (\mathbf{r}_{\mathrm{CEC}} - \mathbf{r}_A)}{\mathbf{e}_{\mathrm{PT}} \cdot (\mathbf{r}_B - \mathbf{r}_A)} \tag{3.1}$$

Where $\mathbf{r}_{\mathrm{CEC}} - \mathbf{r}_A$ is the vector pointing from the nearest oxygen of E14A to the CEC, and $\mathbf{r}_B - \mathbf{r}_A$ is the vector pointing from the nearest pair of E14B to E14A. These distance minimums are evaluated as implemented in PLUMED with a switching factor $\beta = 200$. The vector $\mathbf{e}_{PT}$ is a unit vector pointing at the direction of proton transport.

We aligned a classical simulation trajectory we acquired from the same simulations to get our starting structure, and calculated the vector pointing from the δ-carbon of the E14A side chain to that of E14B. This vector is averaged over the trajectory and normalized to acquire $\mathbf{e}_{PT}$.

Umbrella sampling was carried out for 37 windows from $\xi = 0.050$ to 0.095, with a spacing of 0.025. The spring constant for all windows are 2500 kcal/mol. We restrained the position of all CA atoms with a spring constant of $2.4\,\mathrm{kcal/mol/\AA^2}$. The time series of CV are output every 1 fs, and the PMF is constructed from these time series with weighted histogram analysis method (WHAM).[39] The error bar is acquired from a block analysis of 5 equally divided blocks. The resulting total simulation time is 31.5 ns.

3.3.3 Rate Calculation with Transition State Theory

According to transition state theory, the rate of a reaction can be approximated by

$$k = \sqrt{\frac{\langle v^2 \rangle}{2\pi}} \int_0^{x^\ddagger} \exp\left[-\beta(F(x) - F(x^\ddagger))\right] \mathrm{d}x \tag{3.2}$$

Where $x^\ddagger$ is the location of the transition state, and $F(x)$ is the PMF. The $\langle v^2 \rangle$ term in the prefactor is the mean square velocity along the proton transport CV. Since this CV is nonlinear, it was approximated from the umbrella sampling simulation trajectory.

3.3.4 Low pH Structure of EmrE and Metadynamics on Hydration CV

We used the same structure above and constructed a DMPC bilayer simulation system with CHARMM-GUI. All E14 and H110 were protonated to simulate a low-pH environment. Sodium chloride of 0.04 M was added to the simulation box to neutralize the charges. These structures were first minimized and equilibrated as in the default steps provided by CHARMM-GUI. After the initial minimization and equilibration where the distance restraint was gradually released, each replica was further equilibrated with 400 to 800 ns depending on its RMSD to the initial structure. Then, all restraint was released, and the system was simulated for another 2500 ns. The trajectory was saved every 1 ns for the production simulation.

The well-tempered metadynamics simulation used an initial structure taken from the 1000 ns snapshot of replica 1 of the classical simulation. A Gaussian of 0.6 kcal/mol was deposited every 2 ps. For the water connectivity dimension, the width of the Gaussian was 0.02, and for the distance dimension, the width was 0.3 Å (for 2D case only). The bias factor for the simulation was set as $\beta = 35$. Tiwary-Parrinello reweighting was used to reconstruct the PMF with the histogram from last 250 ns of the simulation.[40] On the distances, an upper wall at 7.0 Å with a spring constant of 200 kcal/mol/Å^2 was added to prevent unrealistic conformations. The total time for the metadynamics is 1000 ns.

3.4 Discussions

In this paper, we presented two major results. First, we showed that proton hopping can happen at a faster time scale than alternating access. The proton will be able to move with the alternating access and stay with the E14A of that opening orientation, contrasting what is hypothesized in reference 29. This suggests that this thermodynamic barrier of 2.9 kcal/mol will not be hindering EmrE from alternating access when asymmetrically protonated. Additionally, the proton leaking measurements showed an opposite trend as alternating rates.

Overall, these findings do not support the alternating access as the rate-limiting step. The possibility of this step also underscores the necessity of a kinetic model treatment, as this additional step will branch out multiple transport pathways at this singly protonated apo state (Figure 3.1).

The alternating access model is a charming theory that can explain the coupled transport behavior of many transporters.[8,41] In such a model, a transporter cycles through a defined and exclusive path to ensure the tightly coupled transport of its substrates. It is particularly appealing for its analogy that proteins are little molecular machines with precision and definite mechanisms just like human-made ones. For example, the well-studied LacY is one of the best examples of transporters strictly following the alternating access model.[42] However, there is a growing appreciation that many transporters are regulated kinetically. Instead of following a fixed path of partial reactions (transport loops), there are multiple possible pathways between states, and the net transport is the result of all these contributing pathways.[12] Kinetic networks are susceptible to regulations by changing the substrate concentrations or the rates of some fundamental steps.11, 12 Therefore, for those transporters involving proton, pH changes can shift their gear not only by changing proton concentration but also allosterically impacting the proton transport pathway. For example, a chloride/proton antiporter, ClC-ec1, has a non-integer stoichiometry and a pH-dependent transport rate.[43,44] These behaviors are best explained through a kinetic model network instead of a fixed path model. Specifically, for EmrE, which also has a pH-dependent transport behavior and non-integer stoichiometry, a kinetic model is very needed to model its behavior.[9,12,13]

Second, we showed that proton binding allosterically affects the formation of a water wire, which resulted in an increase in the proton transport free energy barrier. Protonation of EmrE at E14 is known to rearrange the hydrogen bonding network near the primary binding site by changing the local electrostatic interaction pattern. It tethers TM1 with

TM3, resulting in tighter packing of the primary binding site and its vicinity. Allosterically, this will tilt the C-terminal end of TM1, where K22 and E25 are located. We have further confirmed on a microscopic level that the allosteric effect of the protonation involves a new hydrogen bonding pattern around E25, and it gates a water wire into the primary binding site. This hydrogen bonding pattern with E25 we reported is new to the low pH EmrE and was previously not found in simulations representing higher pH where only E14A and no H110 were protonated. In this high-pH system, T56 was found to be hydrogen bonding with the C-terminus carboxylate group. At low pH environments, the protonation of H110 significantly changes the electrostatic environment near the C-terminus tail. This additional positive charge neutralizes the negative charge in the carboxylate C-terminal and makes the net charge zero. An electrostatic dipole (zero net charge but has separation of positive and negative charge) has a much shorter interaction range than that of a monopole (has non-zero net charge). This can significantly change the conformation preference of this flexible end of EmrE. Given its important function in gating transport, we believe the H110 and the C-terminal will act like a pH sensor for the protein.

There has long been a hypothesis about a peripheral binding site in EmrE.[10,32,45] It has been found that even on E14Q mutants with abolished primary binding sites, EmrE still undergoes pH-dependent conformation change and will still react to TPP^+. They proposed that either E25 or D84, the other two carboxylate groups, or the H110, can form a peripheral binding site that senses pH change or substrate presence. Our results provided more insight by showing the gating role of E25. In our 2D PMF of hydration (Figure 3.3), we showed that breaking the hydrogen bond between T56 and E25 strongly favors the hydration of the channel directly connecting E14 and E25, allowing proton transport. This provides a possible mechanism for drug molecules to open the gate. Since most EmrE substrates are cationic, when they are bound to this peripheral site, these positively charged substrates will break or weaken the hydrogen bonding between E25 and T56, allowing for a water wire to form.

Meanwhile, large chemical shift perturbation has been found near D84 between harmane-bound WT-EmrE and E14Q-EmrE, indicating harmane binding to EmrE near D84. The proximity of the A61 pathway to D84 also indicates that the harmane-introduced proton leak may be through a similar mechanism that controls the A61 water pathway.

References

(1) Burata, O. E.; Yeh, T. J.; Macdonald, C. B.; Stockbridge, R. B. Still Rocking in the Structural Era: A Molecular Overview of the Small Multidrug Resistance (SMR) Transporter Family. *J. Biol. Chem.* **2022**, *298*, 102482.

(2) Morrison, E. A.; DeKoster, G. T.; Dutta, S.; Vafabakhsh, R.; Clarkson, M. W.; Bahl, A.; Kern, D.; Ha, T.; Henzler-Wildman, K. A. Antiparallel EmrE Exports Drugs by Exchanging between Asymmetric Structures. *Nature* **2012**, *481*, 45–50.

(3) Rapp, M.; Granseth, E.; Seppälä, S.; von Heijne, G. Identification and Evolution of Dual-Topology Membrane Proteins. *Nat Struct Mol Biol* **2006**, *13*, 112–116.

(4) Rotem, D.; Schuldiner, S. EmrE, a Multidrug Transporter from *Escherichia Coli*, Transports Monovalent and Divalent Substrates with the Same Stoichiometry*. *J. Biol. Chem.* **2004**, *279*, 48787–48793.

(5) Morrison, E. A.; Henzler-Wildman, K. A. Transported Substrate Determines Exchange Rate in the Multidrug Resistance Transporter EmrE. *Journal of Biological Chemistry* **2014**, *289*, 6825–6836.

(6) Yerushalmi, H.; Schuldiner, S. A Model for Coupling of H^+ and Substrate Fluxes Based on "Time-Sharing" of a Common Binding Site. *Biochemistry* **2000**, *39*, 14711–14719.

(7) Yerushalmi, H.; Schuldiner, S. A Common Binding Site for Substrates and Protons in EmrE, an Ion-Coupled Multidrug Transporter. *FEBS Lett* **2000**, *476*, 93–97.

(8) Mitchell, P. A General Theory of Membrane Transport From Studies of Bacteria. *Nature* **1957**, *180*, 134–136.

(9) Robinson, A. E.; Thomas, N. E.; Morrison, E. A.; Balthazor, B. M.; Henzler-Wildman, K. A. New Free-Exchange Model of EmrE Transport. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, E10083–E10091.

(10) Gayen, A.; Leninger, M.; Traaseth, N. J. Protonation of a Glutamate Residue Modulates the Dynamics of the Drug Transporter EmrE. *Nature Chemical Biology* **2016**, *12*, 141–145.

(11) Peyton J. Spreacker; Nathan E. Thomas; Will F. Beeninga; Merissa Brousseau; Colin J. Porter; Kylie M. Hibbs; Katherine A. Henzler-Wildman Activating Alternative Transport Modes in a Multidrug Resistance Efflux Pump to Confer Chemical Susceptibility | Nature Communications. *Nat. Commun.* **2022**, *13*, 7655.

(12) Hussey, G. A.; Thomas, N. E.; Henzler-Wildman, K. A. Highly Coupled Transport Can Be Achieved in Free-Exchange Transport Models. *Journal of General Physiology* **2020**, *152*.

(13) Awtrey, N.; Beckstein, O. Kinetic Diagram Analysis: A Python Library for Calculating Steady-State Observables of Kinetic Systems Analytically. *bioRxiv* **2024**, 2024.05.27.596119.

(14) Adam, Y.; Tayer, N.; Rotem, D.; Schreiber, G.; Schuldiner, S. The Fast Release of Sticky Protons: Kinetics of Substrate Binding and Proton Release in a Multidrug Transporter. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17989–17994.

(15) Vermaas, J. V.; Rempe, S. B.; Tajkhorshid, E. Electrostatic Lock in the Transport Cycle of the Multidrug Resistance Transporter EmrE. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, E7502–E7511.

(16) Ovchinnikov, V.; Stone, T. A.; Deber, C. M.; Karplus, M. Structure of the Emre Multidrug Transporter and Its Use for Inhibitor Peptide Design. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, E7932–E7941.

(17) Jurasz, J.; Bagiński, M.; Czub, J.; Wieczór, M. Molecular Mechanism of Proton-Coupled Ligand Translocation by the Bacterial Efflux Pump EmrE. *PLoS Computational Biology* **2021**, *17*, 1–19.

(18) Li, J.; Her, A. S.; Besch, A.; Ramirez-Cordero, B.; Crames, M.; Banigan, J. R.; Mueller, C.; Marsiglia, W. M.; Zhang, Y.; Traaseth, N. J. Dynamics Underlie the Drug Recognition Mechanism by the Efflux Transporter EmrE. *Nat Commun* **2024**, *15*, 4537.

(19) Padariya, M.; Kalathiya, U.; Baginski, M. Structural and Dynamic Insights on the EmrE Protein with TPP^+ and Related Substrates through Molecular Dynamics Simulations. *Chemistry and Physics of Lipids* **2018**, *212*, 1–11.

(20) Agmon, N. The Grotthuss Mechanism. *Chem. Phys. Lett.* **1995**, *244*, 456–462.

(21) Swanson, J. M. J.; Maupin, C. M.; Chen, H.; Petersen, M. K.; Xu, J.; Wu, Y.; Voth, G. A. Proton Solvation and Transport in Aqueous and Biomolecular Systems: Insights from Computer Simulations. *J. Phys. Chem. B* **2007**, *111*, 4300–4314.

(22) Peng, Y.; Swanson, J. M. J.; Kang, S.-g.; Zhou, R.; Voth, G. A. Hydrated Excess Protons Can Create Their Own Water Wires. *J. Phys. Chem. B* **2015**, *119*, 9212–9218.

(23) Li, C.; Voth, G. A. A Quantitative Paradigm for Water-Assisted Proton Transport through Proteins and Other Confined Spaces. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2113141118.

(24) Kaiser, S.; Yue, Z.; Peng, Y.; Nguyen, T.; Chen, S.; Teng, D.; Voth, G. Molecular Dynamics Simulation of Complex Reactivity with the Rapid Approach for Proton Transport and Other Reactions (RAPTOR) Software Package. *J. Phys. Chem. B* **2024**,

(25) Wang, Z.; Swanson, J. M.; Voth, G. A. Modulating the Chemical Transport Properties of a Transmembrane Antiporter via Alternative Anion Flux. *Journal of the American Chemical Society* **2018**, *140*, 16535–16543.

(26) Kratochvil, H. T.; Watkins, L. C.; Mravic, M.; Thomaston, J. L.; Nicoludis, J. M.; Somberg, N. H.; Liu, L.; Hong, M.; Voth, G. A.; DeGrado, W. F. Transient Water Wires Mediate Selective Proton Transport in Designed Channel Proteins. *Nat. Chem.* **2023**, *15*, 1012–1021.

(27) Watkins, L. C.; DeGrado, W. F.; Voth, G. A. Multiscale Simulation of an Influenza A M2 Channel Mutant Reveals Key Features of Its Markedly Different Proton Transport Behavior. *J. Am. Chem. Soc.* **2022**, *144*, 769–776.

(28) Morrison, E. A.; Robinson, A. E.; Liu, Y.; Henzler-Wildman, K. A. Asymmetric Protonation of EmrE. *Journal of General Physiology* **2015**, *146*, 445–461.

(29) Li, J.; Sae Her, A.; Traaseth, N. J. Asymmetric Protonation of Glutamate Residues Drives a Preferred Transport Pathway in EmrE. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2110790118.

(30) Morrison, E. A.; Henzler-Wildman, K. A. Reconstitution of Integral Membrane Proteins into Isotropic Bicelles with Improved Sample Stability and Expanded Lipid Composition Profile. *Biochim. Biophys. Acta.* **2012**, *1818*, 814–820.

(31) Dastvan, R.; Fischer, A. W.; Mishra, S.; Meiler, J.; McHaourab, H. S. Protonation-Dependent Conformational Dynamics of the Multidrug Transporter EmrE. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*, 1220–1225.

(32) Thomas, N. E.; Wu, C.; Morrison, E. A.; Robinson, A. E.; Werner, J. P.; Henzler-Wildman, K. A. The C Terminus of the Bacterial Multidrug Transporter EmrE Couples Drug Binding to Proton Release. *Journal of Biological Chemistry* **2018**, *293*, 19137–19147.

(33) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(34) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *The Journal of Chemical Physics* **1984**, *81*, 511–519.

(35) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(36) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(37) Bonomi, M. et al. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16*, 670–673.

(38) Li, C.; Voth, G. A. Accurate and Transferable Reactive Molecular Dynamics Models from Constrained Density Functional Theory. *J. Phys. Chem. B* **2021**, *125*, 10471–10480.

(39) Grossfield, Alan WHAM: The Weighted Histogram Analysis Method.

(40) Tiwary, P.; Parrinello, M. A Time-Independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B* **2015**, *119*, 736–742.

(41) Jardetzky, O. Simple Allosteric Model for Membrane Pumps. *Nature* **1966**, *211*, 969–970.

(42) Abramson, J.; Smirnova, I.; Kasho, V.; Verner, G.; Iwata, S.; Kaback, H. The Lactose Permease of Escherichia Coli: Overall Structure, the Sugar-Binding Site and the Alternating Access Model for Transport. *FEBS Letters* **2003**, *555*, 96–101.

(43) Yue, Z.; Bernardi, A.; Li, C.; Mironenko, A. V.; Swanson, J. M. J. Toward a Multipathway Perspective: pH-Dependent Kinetic Selection of Competing Pathways and the Role of the Internal Glutamate in Cl^-/H^+ Antiporters. *J Phys Chem B* **2021**, *125*, 7975–7984.

(44) Mayes, H. B.; Lee, S.; White, A. D.; Voth, G. A.; Swanson, J. M. J. Multiscale Kinetic Modeling Reveals an Ensemble of Cl^-/H^+ Exchange Pathways in ClC-ec1 Antiporter. *J. Am. Chem. Soc.* **2018**, *140*, 1793–1804.

(45) Glaubitz, C.; Gröger, A.; Gottschalk, K.; Spooner, P.; Watts, A.; Schuldiner, S.; Kessler, H. ^{31}P-CP–MAS NMR Studies on TPP^+ Bound to the Ion-Coupled Multidrug Transport Protein EmrE. *FEBS Letters* **2000**, *480*, 127–131.

CHAPTER 4

QM/CG-MM: SYSTEMATIC EMBEDDING OF QUANTUM MECHANICAL SYSTEMS IN A COARSE-GRAINED ENVIRONMENT WITH ACCURATE ELECTROSTATICS

Quantum Mechanics/Molecular Mechanics (QM/MM) can describe chemical reactions in molecular dynamics (MD) simulations at a much lower cost than ab initio MD. Still, it is prohibitively expensive for many systems of interest because such systems usually require long simulations for sufficient statistical sampling. Additional MM degrees of freedom are often slow and numerous but secondary in interest. Coarse-graining (CG) is well known to be able to speed up sampling through both reduction in simulation cost and the ability to accelerate the dynamics. Therefore, embedding a QM system in a CG environment can be a promising way of expediting sampling without compromising the information about the QM subsystem. Sinitskiy and Voth first proposed the theory of Quantum Mechanics/Coarse-grained Molecular Mechanics (QM/CG-MM) with a bottom-up CG mapping. Mironenko and Voth subsequently introduced the DFT-QM/CG-MM formalism to couple a Density Functional Theory (DFT) treated QM system and to an apolar environment. Here, we present a more complete theory that addresses MM environments with significant polarity by explicitly accounting for the electrostatic coupling. We demonstrate our QM/CG-MM method with a chloride-methyl chloride S_N2 reaction system in acetone, which is sensitive to solvent polarity. The method accurately recapitulates the potential of mean force for the substitution reaction, and the reaction barrier from the best model agrees with the atomistic simulations within sampling error. These models also have generalizability. In two other reactive systems that they have not been trained on, the QM/CG-MM model still achieves the same level of agreement with the atomistic QM/MM models. Finally, we show that in these examples the speed-up in the sampling is proportional to the acceleration of the

rotational dynamics of the solvent in the CG system.

4.1 Introduction

A central conundrum in the field of theoretical and computational chemistry is the trade-off between computational efficiency and methodological accuracy, with the former often winning the competition due to a desire by the user to obtain rapid results and publish papers. As such, interest in simulating chemical reactions in larger systems has led to hybrid methods like Quantum Mechanics/Molecular Mechanics (QM/MM).[1–3] In QM/MM, a quantum mechanically treated subsystem (the QM region) is embedded in an environment modeled by molecular mechanics (the MM region). Treating the reactive region with a QM method allows for a treatment of bond breaking and formation, while representing its surroundings at the MM level of theory significantly reduces the number of electronic degrees of freedom being treated and also brings in environmental influences on the QM region such as a protein and/or solvent environment. Consequently, QM/MM enables significant speed-up in simulation scenarios where only a relatively small part of the system of primary interest is at the expensive QM level. Since its origination, QM/MM has found widespread use in, e.g., biological and material systems, enabling simulations of enzymatic reactions[2] and proton transporters,[4,5] among many other applications.[3]

Despite the numerous advantages of QM/MM simulations, they are still computationally expensive relative to standard classical (empirical forcefield) MD, especially for large systems (with QM/MM often a factor of 100 to 1000 times slower for reasonably accurate QM methods used in the QM zone). This is primarily due to the requirement of long simulation trajectories to obtain statistically meaningful results. For example, in QM/MM MD sampling of a chemical reaction in a protein system, the additional MM degrees of freedom are often slow, plentiful, and require considerable sampling to converge the results statistically (e.g., a rection free energy barrier: a recent example is found in ref 6). The time scales

associated with these slower modes can range from several nanoseconds to several milliseconds, leading to a substantial demand for CPU hours to obtain sufficient sampling. It is worth noting, however, that these additional MM degrees of freedom are often secondary to the main focus of interest in the reactive free energy sampling. Therefore, a strategy to reduce the total number of required QM calculations that can accelerate both the QM/MM MD equilibration and subsequent statistical sampling will significantly enhance simulation efficiency and therefore the ultimate simulation outcome, provided such a strategy is carried out carefully and systematically (i.e., not *ad hoc*).

Coarse-graining (CG) is a well-established technique known for accelerating the sampling of molecular mechanics systems.[7,8] Bottom-up CG methods map several atoms into a single CG "bead" and then derive the effective interaction between these beads, significantly reducing the system's degrees of freedom. These CG interactions are often conceived as "pre-averaged" atomistic interactions, averaging over the less important higher frequency forces in the system. As a result, the CG energy landscape is smoother than the atomistic one, and the overall dynamics of the system are expedited.[7,9,10] In addition to this, the reduction in degrees of freedom also saves computation effort. The overall speed-up can be up to four orders of magnitude.[11–13] Given these advantages, coarse-graining the MM environment for a QM/MM simulation can allow for more rapid sampling while losing very little information obtained about the QM subsystem.

So far, several attempts have been made to incorporate coarse-grained MM models into QM/MM simulations. Much of the early work is more similar to implicit solvent models built from the bottom up and therefore less incapable of retaining information about the MM system.[14,15] In 2015, Sokkar et al. proposed a QM/AA/CG method that sandwiches an AA layer between the QM subsystem and the CG environment, avoiding direct contact between QM and CG systems.[16] In 2016, Shen at al. proposed a more sophisticated QM/AA+CG method, applying a switching function to allow the representation to switch smoothly from

AA to CG as its distance to the QM region increases.[17] However, both methods circumvented the issue of directly embedding QM atoms into the CG environment by adding a layer of atomistic resolution molecules to buffer the QM-CG contacts. Consequently, sampling from these methods will still be limited by the slow dynamics of the immediate surrounding environment. As QM calculations typically consume a substantial portion of the computation time, and the MM part only costs around 10% to 40% of the total time,[18] if these methods cannot accelerate the effective sampling of the environment and reduce the number of QM calculations needed, the achievable speed-up will be capped by this limited percentage of MM calculation time.

The concept of Multiscale Coarse-graining (MS-CG) was initially proposed by Izvekov and Voth in 2005.[19,20] Since its introduction, MS-CG has spurred significant theoretical advancements and a wide range of practical applications.[7] In this present work, we chose to model the MM environment with MS-CG primarily due to its two substantial advantages. Firstly, MS-CG models are good at retaining microscopic information since they are constructed bottom-up by integrating information from AA simulations. In contrast, top-down CG models are developed by fitting more macroscopic properties and cannot guarantee correct microscopic properties, such as pair-wise radial distribution functions (RDF). In the context of QM/CG-MM, an accurate recapitulation of microscopic structural details is crucial for achieving a correct embedding. Secondly, the application of MS-CG is not restricted by the availability of pre-existing CG models, providing much greater flexibility. Theoretically, an MS-CG model can be constructed using any classical simulation trajectory, so the availability of CG models will therefore not limit the applicability of MS-CG.

To date, two papers from our group have been devoted to directly coupling a QM system to an MS-CG environment, aiming to couple the two resolutions directly without compromising structural information. In 2018, Sinitskiy and Voth introduced a formal framework for embedding QM systems into MS-CG models.[21] Building upon this foundation, Mironenko

and Voth subsequently recast the framework in the density functional theory (DFT) formalism and demonstrated the feasibility of the resulting DFT-QM/CG-MM approach in a simplified scenario.[18] Their work specifically focused on apolar environments, such as a CCl_4 solvent coarse-grained into charge-neutral beads. However, for a complete QM/CG-MM theory, accurately modeling electrostatic coupling between QM and MM regions becomes crucial. This study therefore extends the previous DFT-QM/CG-MM method and addresses the electrostatic coupling. This present more complete method can therefore model polar solvents and effectively recapitulate their polarization effect on the QM subsystem. Furthermore, we demonstrate here the method's generalizability. In this framework, QM/CG-MM models can be constructed upon existing MS-CG models with minimal extra parameterization, eliminating the need for lengthy QM/MM simulations for model construction. Additionally, we demonstrate the transferability of our approach, showcasing how models constructed for one QM system can be applied to another without reparameterization. This characteristic further enhances the practicality and efficiency of the QM/CG-MM method, as it enables the reuse of models and reduces the computational burden associated with model construction.

4.2 Theory and Methods

4.2.1 Solvent Effects: S_N2 Reactions as an Illustrative Example

The reaction rate and thermodynamics of S_N2 reactions are highly sensitive to their environment[22]. For example, for a chloride-methyl chloride reaction, the rate is measured 2500 times faster in acetone than in water at 25°C[23]. Specifically, the S_N2 reaction of a halide ion and another methyl halide is arguably one of the most studied systems for solvent effects on reactions, both experimentally[22,23] and theoretically[24–31]. This work is therefore no exception.

The high susceptibility of this system and ones like it to solvent effects makes it an ex-

cellent benchmark for evaluating how well our QM/CG-MM model can accurately capture these electrostatic coupling effects compared to all-atom QM/MM, which come mostly from solvent reorganization. The negatively charged chloride ion will be solvated by surrounding solvent molecules by reorienting their dipoles, effectively screening the electrostatic interaction between the chloride ion and the methyl chloride molecule. Additionally, solvent reorganization contributes substantially to the reaction free energy barrier. As the reaction happens, the solvents must reorient their dipoles, resulting in a significant energy penalty. To assess the capability of our QM/CG-MM model for accurately capturing these solvent effects, a direct comparison can be made between the reaction potential of mean force (PMF) obtained from our QM/CG-MM simulations and those obtained from all-atom QM/MM simulations. A collective variable (CV) is often used in free energy sampling to describe the progress of a reaction. An intuitive and effective collective variable we used for this system is the difference between two C−Cl distances, $r_c = r_{\mathrm{CCl}} - r_{\mathrm{CCl'}}$. This CV is symmetric and has an easy-to-identify transition state at $r_c = 0$.

In our work, we specifically chose acetone as the solvent. It is relatively easy to model using force fields compared to protic solvents like water. We also have prior knowledge of the feasibility of coarse-graining this particular solvent.

4.2.2 Multiscale Coarse-graining (MS-CG)

In MS-CG, force matching is the most straightforward approach for obtaining the CG interactions, particularly for smaller molecules like acetone. The force matching method minimizes the ensemble average of the least square residuals between forces derived from atomistic trajectories and those generated by the CG potentials. This procedure ensures that the CG model reproduces the average forces experienced by the atoms in the atomistic simulations. This present work employed an open-source implementation of MS-CG known as OpenMSCG[32]. It was developed by the Voth group and ships as a Python package. It can

be imported as a Python class or invoked via command-line interfaces. OpenMSCG offers the tools and algorithms to build a CG model and parametrize its interactions. Detailed information and the source code for OpenMSCG can be found on the Voth group website at `https://software.rcc.uchicago.edu/mscg/`.

4.2.3 Electrostatic coupling in QM/MM and QM/CG-MM

QM/MM uses an *ab initio* theory to solve the QM region and a classical force field for MM regions. The key is how to embed the QM subsystem into the MM environment. The most widely used QM/MM embedding scheme today is additive QM/MM, which states that the total energy can be decomposed into the energy of two subsystems plus a coupling term:

$$E^{\text{total}} = E^{\text{QM}} + E^{\text{MM}} + E^{\text{QM/MM}}. \tag{4.1}$$

There are several ways to model this coupling term $E^{\text{QM/MM}}$, , among which the most commonly used today is electrostatic coupling.[3] In the simplest scenario where there is no bond crossing the QM/MM boundary, this QM/MM part of the Hamiltonian usually has two terms in it:

$$E^{\text{QM/MM}} = \sum_{a \in \text{MM}} q_a \int \frac{\rho(\mathbf{r})}{|\mathbf{r} - \mathbf{r}_a|} \mathrm{d}\mathbf{r} + E^{\text{LJ}}, \tag{4.2}$$

where E^{LJ} is the Lennard-Jones energy between QM and MM atoms, which can be modeled by different functional forms. This term compensates for the inability of QM charges to polarize the MM atoms when using a non-polarizable force field, thus avoiding collisions between QM and MM atoms.

The more important first term is the Coulomb energy between the QM and MM regions. In this term, q_a is the classical charge assigned to each MM atom by the force field, and $\rho(r)$ is the electron density in the QM region. While this equation is straightforward, determining this QM electron density consumes most of the computation effort. In our DFT-QM/CG-

MM framework, this is solved from the Kohn-Sham equation with an external field. In the Kohn-Sham formalism, the QM atoms cannot "see" the MM atoms directly. The impact of the MM environment is introduced through an external electrostatic field V_{es}.

$$\left(\underbrace{-\frac{1}{2}\nabla^2 + v_N(\mathbf{r}) + \phi[\mathbf{r};\rho] + \mu_{xc}[\rho]}_{\hat{H}_0} + V_{es}(\mathbf{r}) \right) \psi(\mathbf{r}) = \varepsilon\psi(\mathbf{r}) \tag{4.3}$$

Here, $\hat{H}_0$ is the Hamiltonian for the system in vacuum, where $v_N(\mathbf{r})$ is the electrostatic potential from nuclei; $\phi[\mathbf{r};\rho]$ is the Hartree potential. In QM/MM electrostatic coupling, this extra $V_{es}(\mathbf{r})$ term will be approximated by MM parameters.

Classical force fields assign each atom a charge q_i, and each atom has a covalent radius R_i. To avoid the potential singularity of $1/r$-style potentials from point charges and to improve the convergence, a common practice is to replace the point charge with a Gaussian-distributed charge with a finite width, usually its covalent radius R_i. The electrostatic coupling in many software packages, such as CPMD with EGO QM/MM interface and CP2K used in our work, are all implemented this way.[33,34] The resulting potential from this Gaussian charge, (Eq. 4.4) below, compared to that from point charge, has an additional error function term that eliminates the singular point at $r = |r - r_a| = 0$ and will converge asymptotically to $1/r$ at large r such that

$$V_{es}(\mathbf{r}) = \sum_{a \in \mathrm{MM}} \frac{q_a}{|\mathbf{r} - \mathbf{r}_a|} \operatorname{erf}\left(\frac{|\mathbf{r} - \mathbf{r}_a|}{R_a} \right), \tag{4.4}$$

4.2.4 Description of Solvent Polarization Calls for Explicit Electrostatics Coupling

Mironenko and Voth first implemented a DFT-QM/CG-MM methodology by recapitulating the QM-MM coupling through force-matching the non-QM forces obtained from a QM/MM

simulation[18]. These non-QM forces include all interactions within the MM subsystem and those across the QM-MM boundary. Force-matching these forces will give an effective QM-CG potential and the CG-CG potential. In subsequent CG resolution simulations, the QM problem is solved in a vacuum, and then the obtained potentials are added to restore the coupling between the QM and MM regions. This approach demonstrated the ability to accurately reproduce pairwise radial (RDF) and three-body distribution functions (3DF), as well as the reaction PMF for radical reactions occurring in apolar solvents such as carbon tetrachloride.

An important assumption underlying this method is that a pairwise potential can effectively capture the impact of adding a V_{es} term to the Kohn-Sham equation. This approximation proves to be highly accurate for apolar, neutral solvents. In the case of symmetric molecules like the CCl_4, their electrostatic fields are also very short-ranged. Consequently, the V_{es} term tends to be relatively small and smooth within the QM region. As a result, the QM-CG potential primarily needs to account for the Lennard-Jones interactions, which is to the strengths of MS-CG. However, in scenarios where the coupling effects are prominently electrostatic in nature, the assumption of solely capturing the coupling through the QM-CG potential requires careful reevaluation.

One example where the assumption stated above is problematic is when the charge distribution in the QM region undergoes significant changes during a reaction. In the case of the S_N2 reaction example, the atomic charge for the chloride atom changes substantially during the substitution process. Consequently, the electrostatic interactions between the chloride and MM atoms also change proportionally. If all the CG-chloride interactions were to be packed into a single, tabulated potential, it would not be able to adapt to the evolving QM charge distributions. Such scenarios necessitate an explicit treatment of electrostatic coupling that considers these changing electrostatic interactions between the QM and MM regions. Instead of packing the effects from V_{es} into a uniform, pairwise potential, reproduc-

ing the V_{es} will be needed.

4.2.5 The QM/CG-MM Coupling

As discussed in the previous paragraphs, the two most important factors of a correct coupling are (a) the model of V_{es}, which determines how the QM region feels the MM region, and (b) the Lennard-Jones term, which gives the correct structure of the QM/CG boundary. In this section, we will describe how these two couplings can be modeled.

Potential matching has already been widely used to determine atomistic charges from QM calculations[35–39]. These efforts try to reproduce the electrostatic potential a molecule generates, often obtained from QM calculations, with atomistic models by assigning them radii and charges. We will apply it similarly to reproduce a molecule's electrostatic potential from MM forcefields by assigning each CG bead a radius R_i and a charge Q_i. We denote the set of parameters to be fit as $\xi = \{R_i, Q_i\}$. Then, the objective is to find the best set of parameters $\hat{\xi}$ that generates the electrostatic potential with the minimum RMSD compared to the atomistic potential V_{es}^{AA}. Given an ensemble of atomistic conformations $\{\mathbf{r}^n\}$ and its corresponding CG conformations, we have

$$\hat{\xi} = \arg\min_{\xi} \left\langle \frac{1}{|B|} \int_B \left[V_{es}^{CG}(\mathbf{x};\xi) - V_{es}^{AA}(\mathbf{x}) \right]^2 \mathrm{d}^3\mathbf{x} \right\rangle_{\mathbf{r}^n} \tag{4.5}$$

Here, $\langle \cdots \rangle_{\mathbf{r}^n}$ denotes the ensemble average over solvent configurations. The integration over $\mathbf{x}$ is done within the spatial box of QM region B with its volume denoted as $|B|$. When implementing this RMSD calculation, the integral is replaced with a discrete sum over grid points over the QM box. While this fitting algorithm best reproduces the electrostatic field, there is no guarantee that the CG charge it gives will equal the sum of atomistic charges that make up the CG bead.

The Lennard-Jones parameters are acquired similarly. We used a 12-6 Lennard-Jones

interaction similar to many atomistic force fields. They are defined by two parameters: a radius σ and an intensity ε.

$$V_{LJ}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{4.6}$$

The optimal set of L-J parameters $\hat{\chi}$ is given by minimizing the ensemble average of the RMSD.

$$\hat{\chi} = \arg\min_{\chi} \left\langle \sum_{A} \left[V_{LJ}^{CG}(\mathbf{x}_A; \chi, \mathbf{r}^n) - V_{LJ}^{AA}(\mathbf{x}_A; \mathbf{r}^n) \right]^2 \right\rangle_{\mathbf{r}^n} \tag{4.7}$$

Again, $\langle \cdots \rangle_{\mathbf{r}^n}$ denotes the ensemble average. To approximate this ensemble average, the coordinates $\mathbf{x}_A$ and $\mathbf{r}^n$ are sampled from a classical simulation.

The optimization for both types of parameters is done first using a Nelder-Mead algorithm[40], followed by a conjugate gradient algorithm. When the landscape of the objective function is smooth, the Nelder-Mead algorithm gives a reasonably accurate result with very few function evaluations. We then use conjugate gradient optimization to improve it further.

4.2.6 Implementations and Simulation Details

We ran our QM/MM and QM/CG-MM simulations with CP2K[41]. CP2K is a widely used, open-source software capable of a series of quantum mechanical and QM/MM calculations. We have implemented new features for CP2K for our QM/CG-MM models. Namely, there are two. First, we implemented a feature to allow using a LAMMPS-style tabulated energy file for non-bonded interactions. In MS-CG, the CG-CG interactions are modeled by spline-based potential energy functions. This feature was already implemented in Mironenko and Voth's proof-of-concept work[18]. These codes added a section called `TABPOT` in CP2K to load an external table file as the pairwise potential. Second, all electrostatic and L-J interactions between CG beads are already included in the CG potential. At the same time, we still need to assign them a charge and corresponding L-J parameters for correct QM-CG coupling.

This will lead to double counting these two interactions within the MM subsystem. We implemented a switch to turn off the electrostatics within the MM regions, which can be done by stating "`DO_ELECTROSTATICS FALSE`" in the MM section of the input. We also implemented the `TABPOT` section in a way that when a tabulated potential is assigned to two atom types, the L-J between them will be overridden. As of the publication of this work, these features are now part of the CP2K package and are available on GitHub. More details are available in the CP2K manuals.

All reactive systems are solvated in 500 acetone molecules with PACKMOL[42]. The force field parameters for both the acetone and the reactive molecule were generated by LigParGen web server[43]. The system was first equilibrated with 40 ps of NPT simulation at 300 K and 1 bar with Nosé-Hoover thermostat-barostat as in implemented in LAMMPS[44]. The resulting cubic simulation box has an average edge length of 41.5 Å. All following calculations were done in an NVT ensemble at 300 K with this box size fixed. Free energy samplings were done with an open-source package PLUMED2[45] co-compiled with CP2K. For consistency, all QM calculation setups are kept the same throughout the work. All other additional simulation details can be found in the Supporting Information.

4.3 Results

4.3.1 Monopole Approximation Fails to Capture Essential Electrostatics

We first calculated the single-point energies for the chloride-methyl chloride system in the gas phase (Figure 4.1A, red). At each CV value, we constrained the CV with a big harmonic potential $k = 2$ a.u. and allowed the conformation to relax by minimizing the energy until convergence. Without solvent screening, the leaving and attacking halides will still feel the electrostatic attraction from the rest of the system at a far distance. This results in increasing single-point energy at larger CVs on both sides. Such electrostatic attraction between the

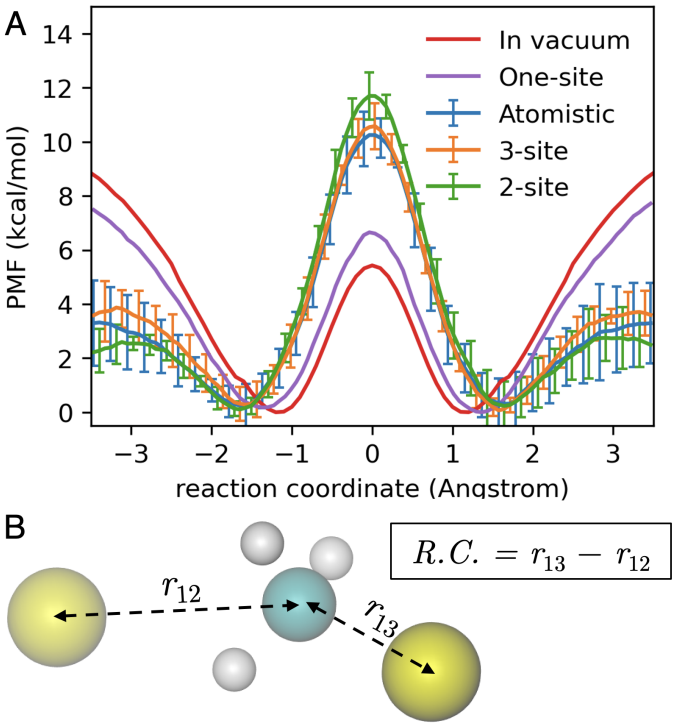


Figure 4.1: **QM/CG-MM umbrella sampling of the chloride-methyl chloride system.** (A) The reaction PMF or single point energy of the reactive system. (B) The schematic plot of the chloride-methyl chloride system and the definition of the reaction coordinate (also referred to as CV). Error bars are from a block analysis of 5 blocks, shown as 95% CI.

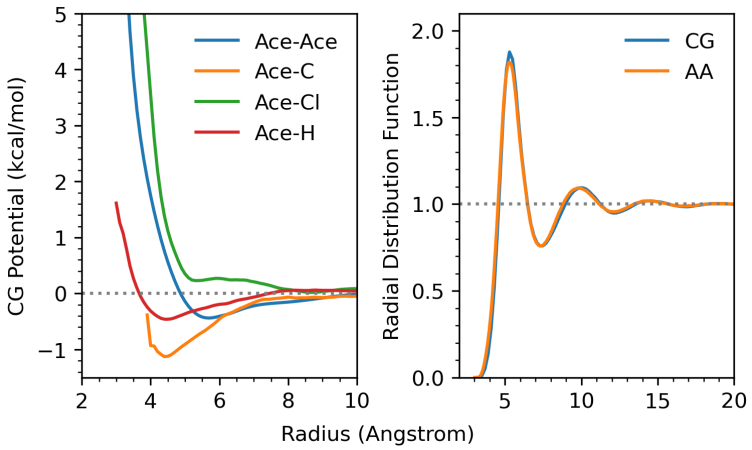


Figure 4.2: **One-site acetone model.** Left: CG potential between beads. Right: RDF between beads.

ion and the methyl chloride dipole creates a minimum in the PMF before the substitution happens.

A QM/MM umbrella sampling approach was employed to determine the PMF of the reaction in acetone (Figure 4.1A, blue). We used 20 windows spanning the CV from $-3.5\,\text{Å}$ to $+3.5\,\text{Å}$ each simulated for 40 ps. The timeseries of CV were analyzed with Weighted Histogram Analysis Method (WHAM) to reconstruct the PMF.[46] The PMF was obtained through a block analysis of 5 blocks to gain statistical insight. The resulting barrier height is $10.3 \pm 1.1\,\text{kcal/mol}$.

Two significant differences were observed when comparing the PMF obtained in acetone with the corresponding vacuum case. First, the free energy barrier in the solvent's presence was considerably higher than in the vacuum case. This can be attributed to the necessity for solvent molecules to reorient their dipoles during the nucleophilic attack, thus leading to a higher energy barrier. Secondly, as the CV value increases, the PMF curve flattens. This behavior can be explained by the solvent's ability to screen the charge of the leaving group, resulting in a reduced electrostatic interaction between the reactants.

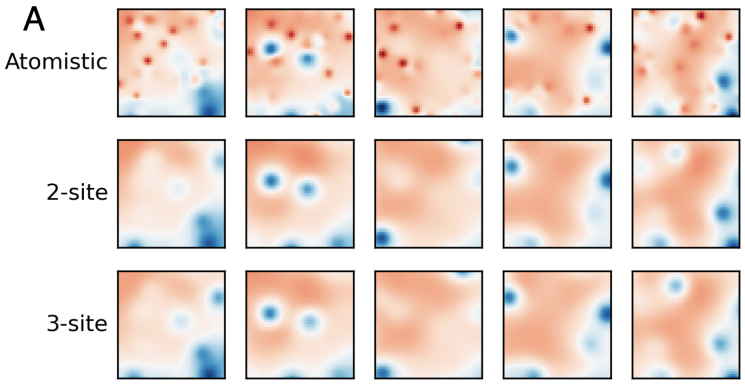
A one-site CG model was then built using force matching the non-QM force as described previously[18]. As explained in the theory section, this previous method requires the beads

to be charge-neutral. Consequently, each acetone molecule has to be mapped into one single bead to ensure charge neutrality. A short 1 ns classical trajectory with only acetone was mapped into CG resolution and force-matched to get the acetone-acetone interaction. Then, the non-QM force was obtained from a 40 ps unbiased atomistic QM/MM simulation. We fixed the acetone-acetone interaction and applied force-matching to acquire the undetermined acetone-QM atoms interaction (Figure 4.2, left). To validate the model, we first ran an acetone-only classical CG simulation of 5 ns to validate the pairwise distributions (Figure 4.2, right). The result shows a nearly perfectly overlapping RDF for the solvents.

A reaction PMF was then obtained by umbrella sampling along the CV with the identical simulation setups from the atomistic simulation. The results (Figure 4.1A, purple) show that the monopole approximation only captured an inadequate degree of the two effects we are interested in, *i.e.*, the barrier height increase due to solvent reorientation and the flattened asymptotic behavior due to charge screening. The barrier is slightly higher than the vacuum case; meanwhile, the position of the meta-stable state is closer to the QM/MM one, and at CV= ± 3.5 Å, the solvation reduces the PMF by roughly 1 kcal/mol. However, this PMF curve is more like the gas phase curve than the QM/MM one in liquid acetone. This means that the effects of polar solvents on the QM system cannot be well-incorporated into QM-CG pair potentials with a one site CG solvent.

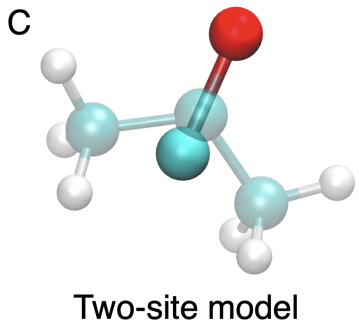
4.3.2 Electrostatic Fitting Suggests a Two-Bead or a Three-Bead Mapping Scheme

Next, we used the potential matching described in the theory section to recapitulate the electrostatic potentials (V_{es}) in the QM box by assigning CG beads charges and radius. We randomly picked 100 snapshots from the previous atomistic QM/MM umbrella sampling simulations, and calculated the (V_{es}) in their QM box with CP2K using its Gaussian Expansion of Electrostatics Potential (GEEP) method[33]. The first row of Figure 4.3A shows



B

	2-site model		3-site model	
Element	O	CG	O	CG
Radius (Å)	0.759	1.998	0.778	1.764
Charge (e)	−0.384	+0.384	−0.382	+0.191
σ (Å)	2.96	3.505	2.96	3.542
ε (kcal/mol)	0.21	1.88	0.21	0.414



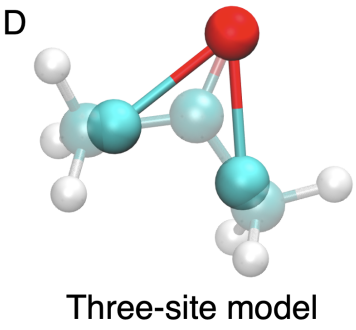


Figure 4.3: **Mapping of the acetone molecule.** (A) A few 2D tomographies of the electrostatic potential in the QM region in atomistic and two different CG resolutions. Blue means negative value, and red represents positive. (B) Coupling parameters from fitting. (C) The two-site CG mapping scheme. Everything else but the oxygen atom was mapped into a single bead. (D) The three-site CG mapping scheme. The $CH_3{-}C{-}CH_3$ part was split into two equal beads, breaking the carbonyl carbon atom into two beads.

five slices of 2D tomographies of the QM box, colored by the electrostatic potentials at these slices. One of the most significant patterns in this map is the dark blue dots from oxygen atoms. While the potential from the positive charges spreads all over the space, that from the negative charge is concentrated around one atom. This pattern suggested that the oxygen atom as a negative charge center better remains at atomistic resolution, while the rest of the molecule can be coarse-grained.

Two possible mapping schemes can be surmised from these observations. The two-site model (Figure 4.3C) maps everything but the oxygen into one bead while preserving the oxygen at atomistic resolution. The three-site model (Figure 4.3D) maps the oxygen into an individual bead and splits the rest of the molecule into two equivalent beads. This mapping splits the carbonyl carbon, but it preserves the symmetry of acetone better. Each model has two kinds of CG beads (O and CG, respectively), and each bead has two parameters used in QM/MM coupling: their charge q_i and radius R_i. The fidelity of these CG beads-generated electrostatic field can be evaluated by their RMSD compared to the atomistic reference with (Eq. 4.5). The parameters were optimized by minimizing this RMSD on these 100 snapshots of training sets, with an additional physical constraint that the molecule's net charge has to be zero. The resulting parameters we got are shown in Figure 4.3B.

The corresponding electrostatic fields generated by the CG beads are shown in The the last two rows of Figure 4.3A. Both models captured the negative charge center while averaging out the positive potentials generated by the hydrogen atoms. The three-site model has a slightly better RMSD (10.2 kcal/mol) than the two-site model (11.4 kcal/mol). The charge assigned to oxygen atoms was slightly smaller than in the atomistic force field ($-0.402e$). The large radius fit for the CG beads corresponds to a more spread-out charge distribution. Lennard-Jones parameters were also fit for these beads using the method in the theory section. To decrease the degrees of freedom in the fitting, the L-J parameters of the oxygen were left the same as those in atomistic resolution. We assume arithmetic mixing for the L-J

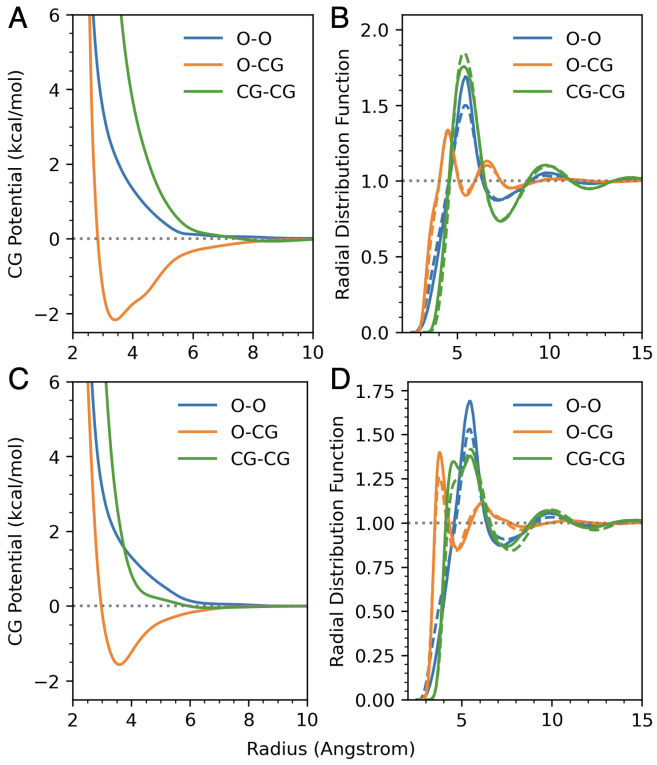


Figure 4.4: **Resulting CG models.** CG potentials and radial distribution functions, respectively. The dashed lines showed atomistic simulation RDF (reference), and the solid line is from CG simulation. (A–B) Two-site model. (C–D) Three-site model.

parameters, where two different atom kinds, i and j, interact with a $\sigma = (\sigma_i + \sigma_j)/2$ and $\varepsilon = \sqrt{\varepsilon_i \varepsilon_j}$. The resulting L-J parameters for the CG beads are listed in Figure 4.3B as well.

The CG potentials between the beads were acquired from the same classical trajectory we used in building the one-site model. It was mapped accordingly and force-matched to obtain CG-CG interactions (Figure 4.4A, C). A subsequent classical, solvent-only simulation was run to validate the model. The resulting RDFs are shown in the solid line of Figure 4.4B and D, while the RDFs from the mapped atomistic trajectory are in dashed lines. The RDFs have similar peak locations with their atomistic references, but the heights of the peaks vary slightly.

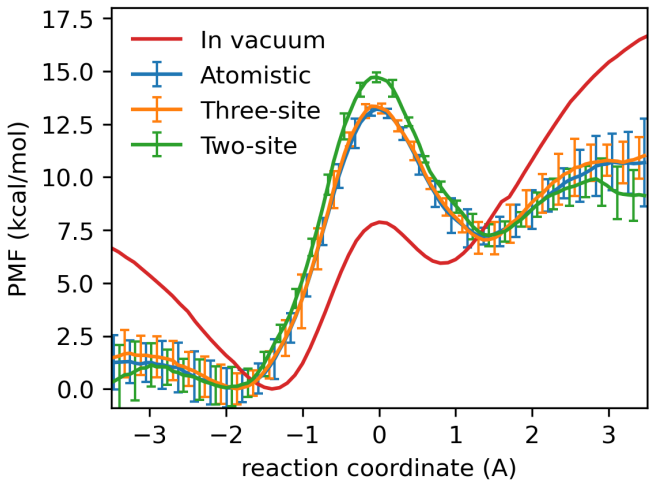


Figure 4.5: **Reaction PMF for the bromide-methyl chloride system.**

4.3.3 QM/CG-MM Correctly Reproduces Solvent Electrostatic Effects

Following the parametrization of the CG models, umbrella sampling was used to help reconstruct the PMFs for this reaction. The resulting PMFs are shown in Figure 4.1 as the orange and green lines. These results show much more similarity to the all-atom QM/MM PMF than the one-site CG model. The position of the ion-dipole complex is almost identical to the blue "Atomistic" QM/MM curve. The transition state barrier and 95% CI given by the two-site model is 11.7 ± 0.9 kcal/mol, while for the three-site model, they are 10.6 ± 0.9 kcal/mol barrier. These numbers are in good agreement with the atomistic QM/MM barrier of 10.3 ± 1.1 kcal/mol. The two-site CG model gave a slightly higher barrier, likely due to an overpolarization tendency.

4.3.4 The QM/CG-MM Model is Transferable to Other Reactive Systems

After testing the chloride-methyl chloride system, we ported the same set of electrostatic and Lennard-Jones parameters to other reactive systems. We wanted to see how these parameters are generalizable to systems for which they have yet be trained. A similar but asymmetric system, the bromide methyl-chloride system, was chosen. One chloride was

substituted with bromide, while all other settings were kept the same. A gas phase PMF (Figure 4.5 red) was obtained first, followed by an atomistic QM/MM umbrella sampling in solvent (blue). Similar to the chloride-methyl chloride case, we observed both a shift in the location of the semi-stable complex and a flattened-out tail. The barrier is also much higher when solvent is present. Then, two QM/CG-MM umbrella sampling simulations were run with the two-site model (green) and the three-site model (orange), respectively. The results showed that despite the significant solvent effect compared to the gas phase case, the QM/CG-MM model recapitulates the reaction PMF well. Similar to the symmetric chloride-methyl chloride system, the three-site model performs slightly better than the two-site model in terms of the reaction barrier. Still, the location of the energy minima agrees equally well.

Perhaps unsurprisingly, a model fit with one S_N2 reaction will perform well in another similar S_N2 reaction. We therefore investigated a different reaction type to demonstrate further generalizability of QM/CG-MM. Previous QM/MM calculations on a Kemp decarboxylation reaction have found its barrier height highly sensitive to the solvent.[47] We reexamined this system with QM/MM and QM/CG-MM in acetone. The progress for this reaction is usually described by two CVs, the C−C distance and the O−N distance (Figure 4.6C). With the same computational setup, we used 2D well-tempered metadynamics with a bias factor $\gamma = 40$ to find the PMF as a function of these two CVs.[48] Only the three-site model was tested for this system due to its better performance in the previous tests. We used four walkers and ran the two systems for the same wall clock time, resulting in an average simulation time of 110 ps for QM/MM and 135 ps for QM/CG-MM, an average of 23% speed-up.

The resulting free energy surface in Figure 4.6A showed good similarity. The results have the same reactant basin and the same dissociation direction. We then applied the string method to obtain a minimum free energy path (shown in black dashed lines)[49]. These two paths have a quite similar shape and location of the transition state. These PMFs show

that O−N bond first stretches significantly with a much smaller change in C−C distance to reach the transition state; after that, the CO_2 leaves simultaneously as the O−N distance increases.

Nevertheless, the minimum free energy path in the 2D PMF only tells half of the story and an entropic correction is needed when reducing the 2D surface to a 1D curve. Those degrees of freedom orthogonal to the reaction path contribute to the 1D PMF through entropic effects[50]. For example, the bottleneck given by the QM/CG-MM is slightly "narrower" than the QM/MM one. This will raise the QM/CG-MM transition state free energy. We identified those regions contributing to the path[51], and integrated them with the method described in the Supporting Information. The resulting 1D PMF along the path is shown in Figure 4.6B. These two methods resulted in nearly identical 1D PMFs, again showcasing the good generalizability of the QM/CG-MM models.

4.3.5 QM/CG-MM Speeds up Sampling through Faster Solvent Relaxation

We have demonstrated in the Kemp decarboxylation reactions that, in terms of computational cost, QM/CG-MM is 23% faster than QM/MM for this system. However, we have stated earlier that the main objective for QM/CG-MM is to speed up the effective dynamics of the MM subsystem, allowing faster equilibration and convergence of sampling, in addition to the reduction of per-step computation cost. We now test this idea with the data we acquired in the bromide-methyl chloride system.

To better quantify how much speed-up we can get, several reaction PMFs are calculated from the same umbrella sampling trajectories but only using the first few picoseconds. We calculated the PMF from the first 0.5, 1, 2, 4, 8, and 16 ps of the trajectory and the RMSD of these preliminary PMFs against the PMF from the block analysis of the entire 40 ps simulation (Figure 4.7A). After only 4.0 ps of simulation, the two-site CG model can give similar PMFs to the full simulation length, while the three-site model converges slower but

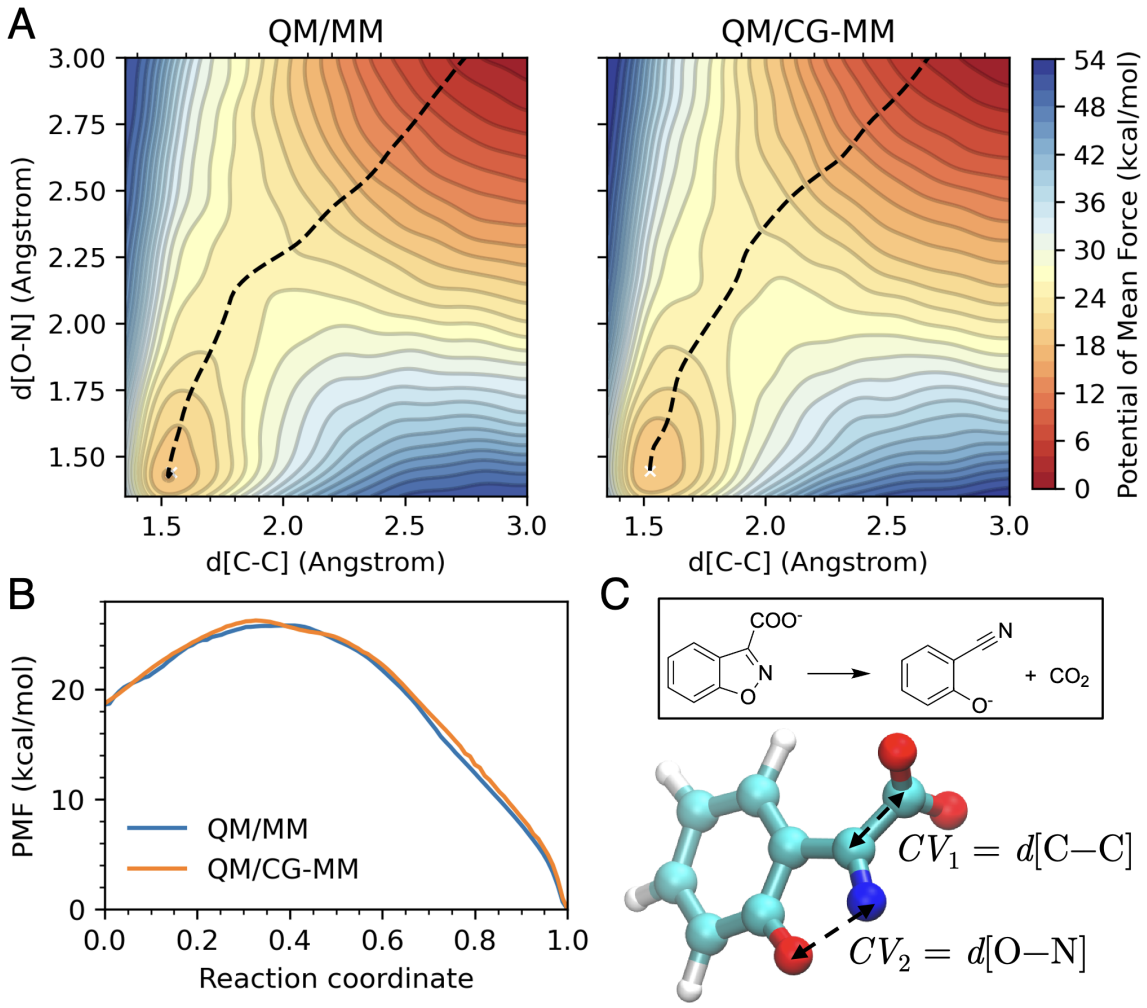


Figure 4.6: **Kemp decarboxylation reaction in acetone.** (A) The 2-D reaction PMF as a function of the two CVs. (B) The 1-D reaction PMF integrated from the 2-D PMF, following the minimum free energy path. (C) The reaction scheme and the definition of the two CVs.

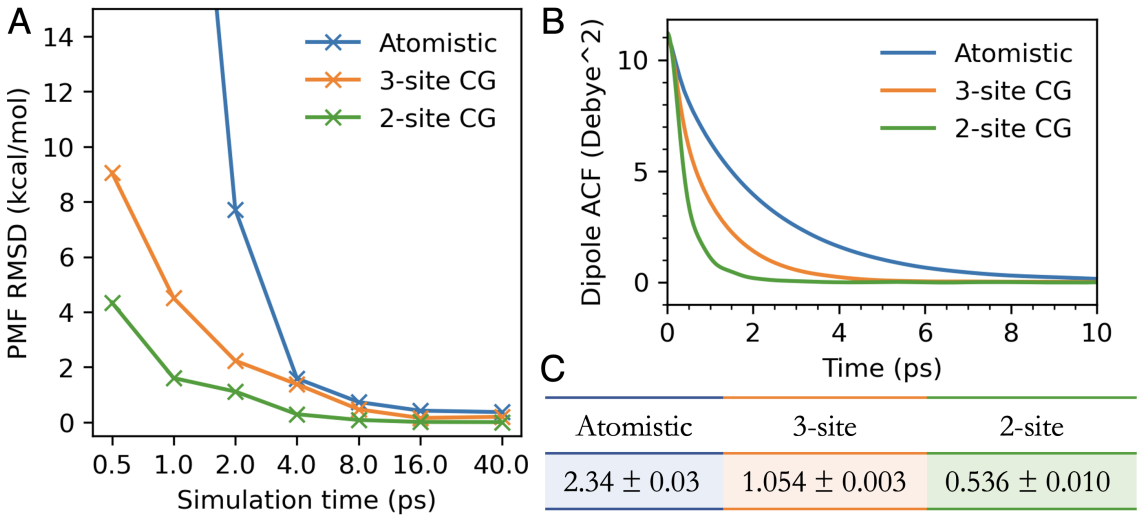


Figure 4.7: **The convergence of QM/CG-MM.** (A) The PMF from truncated trajectories. (B) The dipole auto-correlation function for different acetone models in a pure solvent simulation. (C) Dielectric relaxation time (in ps) from a linear fit of $\ln C(t)$ over t. The shown uncertainty is 95% CI.

still faster than the atomistic simulation (note that the 3-site CG model is nearly atomistic in its resolution).

The time scale for the PMF convergence falls into that of acetone's dielectric relaxation time, which is the characteristic time of the dipole auto-correlation function (ACF). A 1965 experiment estimated this relaxation time to be 2.8 ps at 303 K with the Cole-Cole equation.[52,53] The following calculation also suggests this sampling speed-up is due to the faster dipole relaxation. Ideally, the dipole ACF decays exponentially over time.

$$\langle \mu(0) \cdot \mu(t) \rangle = C(t) = C(0) \cdot e^{-t/\tau} \tag{4.8}$$

where τ is the dielectric relaxation time. We calculated the dipole ACF for pure acetone in three models (Figure 4.7B). Fitting the natural logarithm of the dipole ACF with time will give this relaxation time (Figure 4.7C). From the atomistic to the 3-site coarse-grained model, we have seen a 2.2x faster rotational dynamics and another 2x speed up when using the 2-site model. This speed-up of dynamics agrees with the speed-up of sampling. From

Figure 7A, 1 ps of sampling with the 2-site model, which is approximately 2τ, has similar accuracy with 2 ps of 3-site sampling and 4 ps of atomistic sampling.

4.4 Discussion

There is no doubt that free energy sampling is extremely helpful in modeling chemical and biophysical processes. Usually, this begins by defining a set of CVs and then using free energy sampling of some sort to determine the PMF as a function of these CVs. This PMF can be interpreted to be related to many useful macroscopic observables, such as reaction rate, reaction mechanisms, and the reaction thermodynamics. However, free energy sampling is costly for large and complex systems. While several enhanced sampling methods have emerged to accelerate the exploration along a CV[48,54–56], they may also leave additional degrees of freedom (DoF) poorly sampled. For systems of substantial complexity like proteins, the timescale of these DoF spans from femtoseconds to several seconds[57]. Even when some enhanced sampling methods are adept at handling tens of the slowest DoF[56], ergodic sampling of other slow DoF still demands large simulation time. For example, when a metadynamics simulation is near "convergence," the system can rapidly traverse those biased CVs. However, this leaves very little time for the slower, unbiased DoF to relax, resulting in questionable convergence. Computer hardware advancements[58], as well as faster QM calculations and reactive MD[59], have pushed free energy sampling to a new level; still, any method that can better treat the problem of fully ergodic sampling in large systems is still very much needed.

In this paper, we presented a systematic approach for integrating a QM reactive system into a CG environment. We observed significant improvement in sampling efficiency, primarily from the acceleration of the additional degrees of freedom, and also non-negligibly from reduced QM/MM coupling and MM computation cost. In our particular systems, this speedup is roughly proportional to the acceleration in the solvent's dielectric relaxation time. Additionally, this speedup can also be expanded by any enhanced sampling method that can

cleverly sample the CV space. In addition, unlike other existing methods, our method has a direct QM-CG interface. By assigning a charge and a radius to each CG bead, our QM/CG-MM approach can accurately capture the influence of MM electrostatics on the QM system with a reduced number of MM degrees of freedom.

The QM/CG-MM method proposed in this work offers several notable features and advantages. Most importantly, the QM/CG-MM method has demonstrated good transferability. We illustrated the portability of our QM/CG-MM acetone CG model to other reactive QM systems, although it was initially parametrized using a simple symmetric S_N2 reaction. In the bromide-methyl chloride example, the QM/CG-MM solvent accurately reproduced the QM/MM PMF, exhibiting good agreement for the PMF reaction barrier height and reaction thermodynamics. Moreover, the method also exhibited excellent accuracy in sampling the Kemp decarboxylation, a distinct type of organic reaction very different from the S_N2 reaction. The transferability also means these electrostatic parameters depend only on the underlying MM force field and may be independent of other environmental parameters like temperature. Furthermore, the parametrization of the QM/CG-MM model does not require extensive QM/MM simulations. It only involves fitting the MM electrostatics and Lennard-Jones parameters using some representative conformations. In our study, we selected a modest number of 100 frames from our QM/MM simulations and achieved a satisfactory outcome. Moreover, it is reasonable to expect this approach to be further optimized. As we previously discussed, the transferability of QM/CG-MM parameters allows for the possibility of performing electrostatic fitting solely from classical simulation trajectories, thereby reducing the overall cost of parametrization. This advance holds the potential for even more efficient application of the method.

Lastly, fitting the coupling parameters is independent of obtaining the CG-CG potentials for the MM region and does not require prior knowledge of the CG system. Consequently, a reaction system can be seamlessly integrated into any pre-existing CG model without re-

parametrizing the system's CG portion. This feature significantly expands the applicability of the QM/CG-MM approach, allowing it to be readily employed in a wide range of systems.

This work has also inspired more understanding of the representability problem of coarse-graining. Usually, the CG model construction is determined by the properties one wants to reproduce. In most cases, kinetic information from atomistic simulations can be incorporated into the mapping to ensure atoms moving together are mapped into the same CG bead[60]. Microscopically, the Henderson theorem ensures that pairwise interactions can uniquely determine the RDF of a system[61]. Reproducing observables other than that will require additional terms added to the CG interactions, such as three-body terms, often resulting in a trade-off between computational tractability and compatibility[62]. In a previous paper, acetone was found to be best coarse-grained with a "center-of-charge" (COC) mapping for reproducing radial and angle distribution functions[63]. The COC mapping maps all positively charged atoms into one bead and those negatively charged into another, connected with a bond. It guarantees the invariance of dipole moments between AA and CG representations and, therefore, the long-range electrostatic behavior. In our particular case, the COC mapping that best preserved the liquid structure may not work well, as it cannot capture local electrostatics patterns near the inhomogeneous QM/MM boundary (see Supplementary Information). We used potential matching to acquire these CG coupling parameters so that they best fit our purpose, although these mappings may not be optimal for reproducing other observables such as RDFs in the CG region.

4.5 Supporting Information

4.5.1 Simulation Details

All simulations were done in an NVT ensemble at 300 K unless otherwise noted. All classical simulations were carried out with LAMMPS[44], with the timestep set at 1.0 fs. All QM/MM

and QM/CG-MM simulations were done with CP2K[41], with a timestep at 0.5 fs. A Nose-Hoover thermostat was used with both simulation software. In LAMMPS, the T_{damp} was set as 100 K. In CP2K, the time constant was set as 200 fs.

All QM calculations were carried out using density functional theory (DFT) with a mixed Gaussian and plane wave basis as implemented in CP2K[64]. We assumed the ground state to be a singlet state, therefore using restricted methods. The Gaussian part was calculated with a double-zeta basis set with polarization added (DZVP). Goedecker-Teter-Hutter pseudopotentials were used to model the effects of core electrons[65]. The exchange-correlation interactions were modeled by the PBE functional[66]. The SCF problem was solved by the Broyden minimizer. Both inner and outer SCF loops had a convergence threshold of 10^{-5}.

For all our systems, the system has a total charge of -1. This will be considered explicitly in the DFT calculations. Instead of adding a counter ion, we decided to let the net charge of our simulation box be non-zero. A detailed discussion on this choice and their respective effects is available in reference[67]. The authors of ref[67] concluded that (a) the effect of a non-zero net charge is equivalent to a uniform background charge over the simulation box due to the omission of the $k = 0$ term in Ewald summations, and that (b) the artifact of such treatment depends on the homogeneity of the simulation system's dielectric properties. In our system, where most of the system is homogenous acetone solvents, we believe such artifact is negligible. Additionally, a previous study of the same reaction system in water has ignored the finite size effect of this non-neutral simulation box[31]. Lastly, our most important argument in this paper is the consistency of QM/MM and QM/CG-MM treatments. As long as the QM/CG-MM system can recapitulate what is observed in QM/MM, these artifacts are secondary to our considerations.

System setup

The initial configuration and force field parameters for both the reactive system and the acetone molecule were generated by LigParGen web server[43]. Then, the chloride methyl-chloride system was solvated in 500 acetone molecules with PACKMOL[42]. The system was first equilibrated classically with an NPT ensemble using LAMMPS, with Nose-Hoover thermostat and barostat for 1 ns at 300 K and 1 bar[68,69]. The average box size was taken from the last half of the simulation to be 41.5 Å. This box size was kept the same throughout the work and for all reactive systems. Therefore, for other systems, we simply solvated the reactive system within a 41.5 Å big cube box and minimized it, skipping the NPT simulation.

Classical simulation

The input was prepared by psfgen, a VMD package, and charmm2lammps. A configuration was taken from the NPT simulation, first minimized, and ran classically in an NVT ensemble for 1 ns. The real space interaction is modeled with a charmm-style cutoff of 8 Å and 12 Å. Long-range electrostatics was done with PPPM and a k-space cutoff of 10^{-6}. A trajectory of force and coordinates was output every 0.1 ps.

Table 4.1: **Nonbonded forcefield parameters used in present work.**

Atom	charge	$-\varepsilon$ (kcal/mol)	$r_{\min}/2$	Note
CA	−0.3040	0.0660	1.964309	methyl carbon
CO	0.3110	0.0700	1.992370	carbonyl carbon
Oac	−0.4020	0.2100	1.661244	carbonyl oxygen
Hac	0.1165	0.1165	1.403078	methyl hydrogen
C	−0.1519	0.0660	1.964309	reactive system carbon
H	0.1098	0.0300	1.403078	reactive system hydrogen
CL	−0.1773	0.3000	1.908185	reactive system chloride
BR	N/A	0.4700	1.947472	reactive bromide

Force field parameters Here's the exact force field parameters we used for the simulation. The nonbonded parameters are listed in Table 4.1. Note that another chloride of exactly

charge -1 was placed in the classical system as well. The bromide was never used in any classical simulation, so its classical charge was not used. The bonded interactions are listed below in CHARMM PRM format:

```
BOND
CA   CO   317.000 1.522
Oac  CO   570.000 1.229
Hac  CA   340.000 1.090

ANGLE
CA  CO  Oac 80.000 120.400
CA  CO  CA  70.000 116.000
CO  CA  Hac 35.000 109.500
Hac CA  Hac 33.000 107.800
Oac CO  CA  80.000 120.400

DIHEDRAL
Hac CA CO CA  0.00000 1 0.00000
Hac CA CO CA  0.00000 2 180.00000
Hac CA CO CA  0.13750 3 0.00000
Hac CA CO CA  0.00000 4 180.00000

IMPROPER
CO  CA CA Oac 10.50000 2 0.000000
```

CG model construction, RDF computation

The CG mapping and force matching was done with OpenMSCG, an open-source software developed by Voth group. The LAMMPS trajectories were mapped to CG resolutions with different schemes with "`cgmap`" command.

Two-site model The two-site model maps the carbonyl oxygen to its own bead, while the rest of the whole molecule is mapped to its center of mass. During the mapping, the force on each atom is also added to the bead it maps to. The mass of each bead is the sum of the masses of its components.

Three-site model The three-site model maps the carbonyl oxygen to its own bead, while splitting the rest of the molecule into two equivalent beads. This will involve splitting the carbonyl carbon into two halves. Hence, each CG bead contains one methyl group and half of a carbon. The CG site is also located at the center of mass of its components. Forces are mapped similarly as the two-site model, but the force on the carbonyl carbon was also split into two halves.

CG model construction The RDF was calculated with OpenMSCG for all pairwise interactions with "`cgrem`" command, with a cutoff of 20 Å using 100 bins. This RDF also provides information about the range of these interactions. The forces were force matched to CG potentials with the "`cgfm`" command. Non-bonded potentials were represented with 4th-order splines, and bonded potentials (bonds, angles) were represented with 3rd-order splines. All non-bonded interaction was cut off at 20 Å. The lower cutoffs were determined by their RDF where it falls below 0.01, respectively.

These bond and angle potentials were fit to parabolas to get harmonic interaction parameters. A linear mapping on the force gives very similar results to a polynomial fit on the potential (Figure 4.8). For the 2-site mapping, the bond length is determined to be 1.800 Å

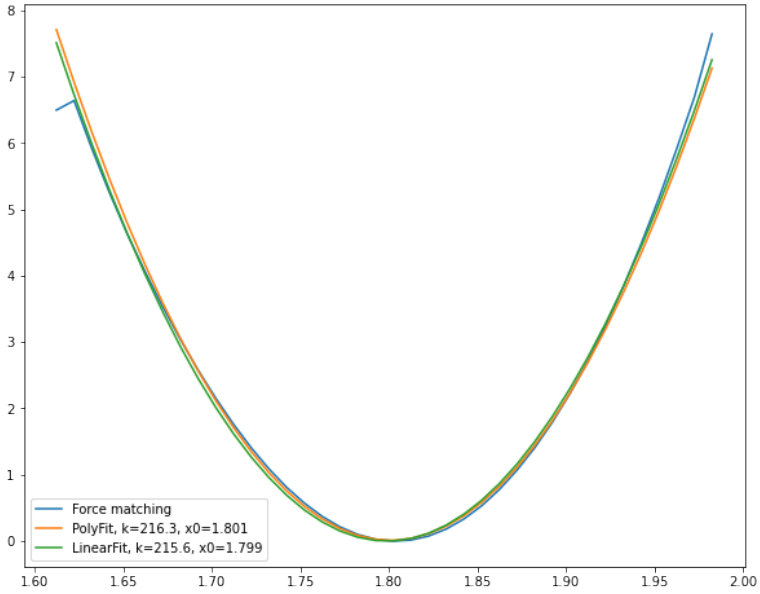


Figure 4.8: **Fitting bonded parameters.** Example of a comparison between a linear fit of force and a 2-order fit of potential.

and

$$\frac{1}{2}k = 216\,\text{kcal/mol/Å}^2.$$

For the 3-site model, the two bonds are identical, with a bond length 2.050 Å and

$$\frac{1}{2}k = 131.8\,\text{kcal/mol/Å}^2.$$

The angle has a $\theta_0 = 56.75°$ and

$$\frac{1}{2}k = 228.8\,\text{kcal/mol/rad}^2.$$

Single point energy calculation

The single point energy was calculated with the above DFT formalism. The system consists of only the chloride and the methyl-chloride, a total of 6 atoms. We scanned the CV spanning from -3.5 Å to $+3.5$ Å with 0.1 Å spacing. Each point was calculated by restaining the CV

with a huge harmonic potential $k = 2\,\text{a.u.}(\text{hartree}/\text{Å}^2)$, and minimizing the energy of the system until convergence.

QM/MM simulations

All QM/MM simulations were done by CP2K in its Gaussian electrostatic coupling scheme[33]. The QM box used was a cube with 12 Å in dimension, centered each simulation step at the center of the reactive system.

The covalent radius of the atoms will determine the distribution width of the Gaussian charge. For hydrogens, we used 0.315 Å; for methyl carbons we used 0.76 Å; for the carbonyl carbon, 0.73 Å, and for the oxygen 0.66 Å. The electrostatic potentials were expanded into Gaussians and a residue using CP2K's internal Gaussian Expansion of Electrostatic Potential (GEEP) library using 10 Gaussians.

One-site model

First, the CG–CG interaction was force-matched from the classical simulation above in a similar way, resulting in a tabulated potential. The one-site model RDF was then computed from a classical simulation of 5 ns, with only this tabulated potential.

The QM-CG interaction was modeled by running an unbiased QM/MM simulation for 100 ps. The trajectory of coordinates and non-QM forces were output every 2.5 fs for a total of 40000 frames. Then, these non-QM forces were mapped to CG resolution and force-matched, with the CG–CG interaction as prior, to get the QM–CG interaction. These interactions were then input to CP2K with the "`TABPOT`" feature we implemented.

The PMF was calculated with an QM/MM umbrella sampling done by co-compiled PLUMED2 and CP2K.[41,45] We used a total of 15 windows, spanning from -3.5 Å to $+3.5$ Å with 0.5 Å spacing. The force constant for the harmonic restraint is set to be $k = 20\,\text{kcal}/\text{mol}/\text{Å}^2$ for all windows. The resulting time series of CV was analyzed by

Weighted Histogram Analysis Method (WHAM) to reconstruct the reaction PMF.

4.5.2 QM/CG-MM parametrization

Potential matching on electrostatics

Potential matching minimizes the RMSD between the electrostatic fields in the QM box, generated by charged CG beads and the atomistic representations.

First, 100 random frames were selected from the QM/MM umbrella sampling trajectories. These atomistic coordinates were mapped to CG resolution with different CG mapping schemes. Then, we used CP2K to calculate the MM electrostatic potential in the QM box, evaluated by GEEP algorithm on a real space grid[33]. This is a native feature of CP2K and can be activated by its `POTENTIAL` keyword `QMMM%PRINT` section. The result is a grid on a $48 \times 48 \times 48$ real space grid over the 12 Å cube. The definition of the RMSD over this grid is shown in the main text in the Method section.

We then used the Nelder-Mead algorithm as implemented in the SciPy package to optimize the parameters. A constraint is implemented to ensure the charge neutrality of CG molecules. For the two-bead model, this is $q(\mathrm{O}) = -q(\mathrm{CG})$; for the three-bead model, this is $q(\mathrm{O}) = -2q(\mathrm{CG})$. For each function evaluation of the RMSD, a CP2K process is called to evaluate the electrostatic field of this area. The final RMSD is the average over all the 100 conformations. Since this program is intrinsically parallelizable, it is highly efficient with multithreading mechanisms.

The final parameters will go through another conjugate gradient optimization. The reason for this two-staged optimization is that the Nelder-Mead algorithm requires far fewer function evaluations than gradient-based methods. The final conjugate gradient minimization will ensure it arrives at a minimum. The RMSD landscape for this optimization problem is mostly convex, so the Nelder-Mead algorithm can often ensure a very good fitting.

Lennard-Jones fitting

For Lennard-Jones fitting, we also fit the RMSD of LJ energy generated from atomistic simulations and CG simulations. We picked a total of 979 conformations from the QM/MM simulation. These conformations are similarly mapped to CG resolution. We calculated the LJ energy each QM atom feels when embedded in the atomistic and CG environments, and hence their RMSD. The only two free parameters are the σ and ε for the CG bead. This optimization is mostly non-linear. We only used the Nelder-Mead algorithm to optimize it, as gradient-based algorithms are sometimes non-convergent depending on the initial guess and learning rate.

4.5.3 Free energy sampling

Umbrella sampling

Block analysis The PMFs are calculated from a block analysis to gauge the uncertainty of the free energies. For each umbrella sampling, we cut the trajectories into 5 parts at 8 ps each, which is still more than 3 times longer than the rotational timescale of acetone molecules. Weighted Histogram Analysis Method (WHAM) was used to build a PMF for each of the 5 blocks, and the PMFs were aligned by minimizing RMSD. Then, the average and standard deviations were calculated. The 95% CIs shown were based on t-distributions with two-sided $p = 0.05$. We used 100 points through -3.5 to $+3.5$ Å for the grid.

Umbrella windows Initially, when running umbrella sampling for the one-site acetone models, windows were set at 0.5 Å spacing from -3.5 Å to $+3.5$ Å, with 20 kcal/mol/Å^2 spring constant. However, when using the same setup for other simulations, the histogram does not overlap well around the big peak near $r = 0$. For the other simulations, some additional windows were added.

Atomistic QM/MM We added four windows at $\pm 0.25, \pm 0.75$ Å, and adjusted the spring constant for those windows within (including, same below) ± 0.5 Å to 40 kcal/mol/Å^2.

QM/CG-MM Compared to QM/MM windows described above we added four additional windows at ± 0.1 and ± 0.40 Å. Spring constants were set at 40 kcal/mol/Å^2 for windows within ± 0.75 Å and 80 kcal/mol/Å^2 for windows within ± 0.40 Å.

These setups were the same between the chloride-methyl chloride and bromide-methyl chloride system.

2D WT-MetaD

Since the kemp-decarboxylation is best described with 2 CVs, we used the 2D well-tempered metadynamics instead of 2D umbrella sampling to save computation efforts. PLUMED was also used for this simulation. For every 100 steps (50 fs), A Gaussian potential is deposited with 0.5 kcal/mol in height and 0.1 Å in width in both dimensions. Additional wall potentials were added for those CVs greater than 3.0 Å and smaller than 1.35 Å with a spring constant of 150 kcal/mol/Å^2 to constrain the system in our studied CV region. We used 4 walkers for each system and they are communicated every 100 steps (50 fs).

Speed-up calculation

Convergence (Figure 4.7A) We picked the first t picoseconds from the umbrella sampling trajectories and ran WHAM on these shortened trajectories, where t is picked as $0.5, 1.0, 2.0, 4.0, 8.0, 16.0$ and 40.0 ps, yielding a PMF for each t. These PMF $V(r)$ are compared with the PMF from block analysis $V_0(r)$ described above. The RMSD σ is calculated as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[V(r_i) - V_0(r_i)]^2}, \tag{4.9}$$

where r_i is the points the PMF is calculated at, and $N = 100$ is the total number of points.

Note that the WHAM results from 40 ps of trajectory may not be the same as the average from the block analysis; therefore, the RMSD is exactly not 0 for all systems at $t = 40\,\text{ps}$.

Autocorrelation function (Figure 4.7B,C) Using PLUMED, we calculated the dipole moments of each acetone molecule from the classical trajectory without the reactive system in it. Then, we used a short self-written code to compute the following

$$C(t) = \frac{T-t}{\Delta t} \sum_{t_0=0}^{T-t} \frac{1}{N} \sum_{i=1}^{N} \mu_i(t_0) \cdot \mu_i(t_0 + t), \tag{4.10}$$

where $N = 500$ is the number of acetone molecules in the box; T is the total simulation time, Δt is the output frequency. Therefore, $(T-t)/\Delta t$ is the total number of frames that were summed.

The rotational time was from the inversed slope of a linear fit of $\ln C(t)$ to t, with a cutoff when $C(t)$ falls lower than 10^{-10}. The error was calculated assuming a two-sided t-distribution with $p = 0.05$.

Intregration of 2D PMF into 1D

Math basis A path is defined by $r(t) = (x(t), y(t))$ where t is the path CV in $[0, 1]$. Parrinello published a paper discussing the mapping of an arbitrary point $R(x, y)$ not on the curve $r(t)$ to a path CV value t[51].

They used the following equation:

$$\xi(R) = \lim_{\gamma \to +\infty} \frac{\displaystyle\int_0^1 t \exp(-\gamma ||R - r(t)||^2)\, \mathrm{d}t}{\displaystyle\int_0^1 \exp(-\gamma ||R - r(t)||^2)\, \mathrm{d}t} \tag{4.11}$$

This is like an "ensemble average" of t, weighed by a temperature factor γ and a harmonic spring-style potential energy function $||R - r(t)||^2$. Basically, only those $r(t)$ closest to R

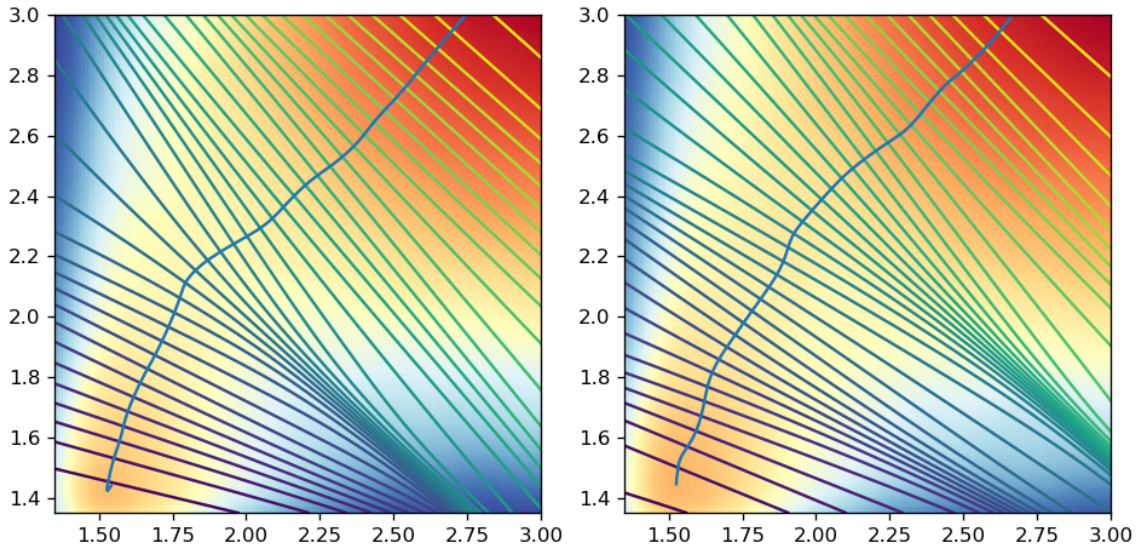


Figure 4.9: **Integrating 2-D PMF into 1-D** The equal-CV contour for 2-D PMF. Left: QM/MM. Right: QM/CG-MM.

will contribute significantly to this integral. Therefore, this gives a continuous and derivable function that can map any point to a path CV. Then a limit of $\gamma \to +\infty$ is taken, making it like a 0 K situation, where theoretically only the closest t contributes.

In numerical computations, the accuracy of this limit is constrained by the representability of a 64-bit double-precision float. Therefore, this exponential should not be smaller than 10^{-300}, or e^{-690}. If $\max ||R - r(t)||^2 < 3$, a practical choice of γ should not exceed 230, where we used 50 in our work.

The PMF comes by 200 by 200 points grids. For each point, we calculated the path CV it should map to. We got an equal-CV contour as Figure 4.9.

Integration The integral to get 1-D PMF from a 2-D one looks like the following

$$V(\xi_0) = -kT \ln \int e^{-\beta v(R)} \delta[\xi(R) - \xi_0] \, \mathrm{d}x\mathrm{d}y \tag{4.12}$$

Here $R = R(x, y)$ is a point on the 2-D CV space, and $\xi(R)$ is the above mapping from equation (4.11), which was pre-computed on those grid points. The partition function in

equation (4.12), after taking the logarithm, is a constant and is omitted.

The resulting path CV space ($[0, 1]$) is binned into 100 parts for numerical computation. For each grid point $R_0(x_0, y_0)$ on the xy-space, the exponential of the integral in (4.12) is evaluated and summed to the bin where the CV on that point $\xi_0 = \xi(R_0)$ falls into. So for those path CV value ξ with many grid points mapped to it, the sum will be bigger and its negative logarithm smaller, reflecting an entropic effect.

4.5.4 Discussion on the center-of-charge mapping

Center-of-Charge (CoC) mapping is a mapping scheme proposed by Cao and Voth in 2015 in reference[63]. The idea of CoC mapping is that it is exact in terms of long-range electrostatics. For an index i that loops over all atoms, p loops over all positively charged atoms, and n for negatively charged atoms, we have the dipole moment $\mathbf{d}$ as

$$\mathbf{d} = \sum_i q_i \mathbf{r}_i = \sum_p q_p \mathbf{r}_p + \sum_n q_n \mathbf{r}_n. \tag{4.13}$$

While in the CoC mapping, the center of positive charge (COPC) bead is defined by

$$\mathbf{R}_p = \sum_p \frac{q_p}{Q} \mathbf{r}_p,$$

where Q is the sum of all positive charges, and $-Q$ will be the sum of all negative charges assuming a neutral molecule. Hence we also have the location of the center of negative charge (CONC)

$$\mathbf{R}_n = \sum_p \frac{q_n}{-Q} \mathbf{r}_n.$$

Consequently, equation (4.13) becomes

$$\mathbf{d} = Q\mathbf{R}_p + (-Q)\mathbf{R}_n. \tag{4.14}$$

This is exactly the dipole for the CoC coarse-grained CG molecule. In ref[63], they had a couple of successful examples how this works for modeling pairwise interactions.

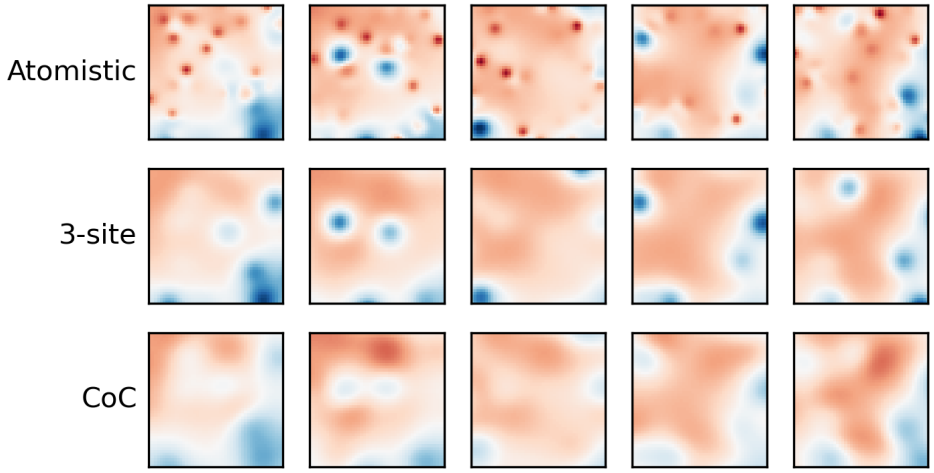


Figure 4.10: **Result of CoC mapping**. The first two rows are identical with the main text. The last row shows the electrostatic map with the center-of-charge (CoC) mapping.

As we've discussed in the main text, preserving good long-range interactions may not work the best in the context of QM/MM. Many QM/MM coupling is short-ranged. While CoC acetone model is a great success, it may not work well with the electrostatic fitting. We tried to redo the electrostatic fitting with CoC acetone and found it is insufficient to represent the electrostatic field, overpolarizing it. The RMSD is 16.5 kcal/mol, much larger compared to that of 3-site mapping (10.2 kcal/mol) and 2-site mapping (11.4 kcal/mol)

References

(1) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–49.

(2) Magalhães, R. P.; Fernandes, H. S.; Sousa, S. F. Modelling Enzymatic Mechanisms with QM/MM Approaches: Current Status and Future Challenges. *Isr. J. Chem.* **2020**, *60*, 655–666.

(3) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

(4) Liang, R.; Li, H.; Swanson, J. M.; Voth, G. A. Multiscale Simulation Reveals a Multifaceted Mechanism of Proton Permeation through the Influenza A M2 Proton Channel. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, 9396–9401.

(5) Liang; Swanson, J.; Wikstrom, M.; Voth, G. A. Understanding the Essential Proton Pumping Kinetic Gates and Decoupling Mutations in Cytochrome c Oxidase. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*.

(6) Beckett, D.; Voth, G. A. Unveiling the Catalytic Mechanism of GTP Hydrolysis in Microtubules. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, e2305899120.

(7) Jin, J.; Pak, A. J.; Durumeric, A. E. P.; Loose, T. D.; Voth, G. A. Bottom-up Coarse-Graining: Principles and Perspectives. *J. Chem. Theory Comput.* **2022**, *18*, 5759–5791.

(8) Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The Power of Coarse Graining in Biomolecular Simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 225–248.

(9) Izvekov, S.; Voth, G. A. Modeling Real Dynamics in the Coarse-Grained Representation of Condensed Phase Systems. *J. Chem. Phys.* **2006**, *125*, 151101.

(10) Hijón, C.; Español, P.; Vanden-Eijnden, E.; Delgado-Buscalioni, R. Mori–Zwanzig Formalism as a Practical Computational Tool. *Faraday Discuss.* **2009**, *144*, 301–322.

(11) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews* **2016**, *116*, 7898–7936.

(12) Liwo, A.; Khalili, M.; Scheraga, H. A. Ab Initio Simulations of Protein-Folding Pathways by Molecular Dynamics with the United-Residue Model of Polypeptide Chains. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362–2367.

(13) Gupta, M.; Pak, A. J.; Voth, G. A. Critical Mechanistic Features of HIV-1 Viral Capsid Assembly. *Science Advances* **2023**, *9*, eadd7434.

(14) Theel, K. L.; Wen, S.; Beran, G. J. Communication: Constructing an Implicit Quantum Mechanical/Molecular Mechanics Solvent Model by Coarse-Graining Explicit Solvent. *J. Chem. Phys.* **2013**, *139*.

(15) Nocito, D.; Beran, G. J. Averaged Condensed Phase Model for Simulating Molecules in Complex Environments. *J. Chem. Theory Comput.* **2017**, *13*, 1117–1129.

(16) Sokkar, P.; Boulanger, E.; Thiel, W.; Sanchez-Garcia, E. Hybrid Quantum Mechanics/Molecular Mechanics/Coarse Grained Modeling: A Triple-Resolution Approach for Biomolecular Systems. *J. Chem. Theory Comput.* **2015**, *11*, 1809–1818.

(17) Shen, L.; Yang, W. Quantum Mechanics/Molecular Mechanics Method Combined with Hybrid All-Atom and Coarse-Grained Model: Theory and Application on Redox Potential Calculations. *J. Chem. Theory Comput.* **2016**, *12*, 2017–2027.

(18) Mironenko, A. V.; Voth, G. A. Density Functional Theory-Based Quantum Mechanics/Coarse-Grained Molecular Mechanics: Theory and Implementation. *J. Chem. Theory Comput.* **2020**, *16*, 6329–6342.

(19) Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.

(20) Izvekov, S.; Voth, G. A. Multiscale Coarse Graining of Liquid-State Systems. *J. Chem. Phys.* **2005**, *123*.

(21) Sinitskiy, A. V.; Voth, G. A. Quantum Mechanics/Coarse-Grained Molecular Mechanics (QM/CG-MM). *J. Chem. Phys.* **2018**, *148*, 014102.

(22) Tanaka, K.; Mackay, G. I.; Payzant, J. D.; Bohme, D. K. Gas-Phase Reactions of Anions with Halogenated Methanes at 297±2 K. *Can. J. Chem.* **1976**, *54*, 1643–1659.

(23) Parker, A. J. Solvation of Ions. Part II. Dipolar Aprotic Solvents as Media for Nucleophilic Substitution Reactions at a Saturated Carbon Atom. *J. Chem. Soc.* **1961**, 1328–1337.

(24) Sola, M.; Lledos, A.; Duran, M.; Bertran, J.; Abboud, J. L. M. Analysis of Solvent Effects on the Menshutkin Reaction. *J. Am. Chem. Soc.* **1991**, *113*, 2873–2879.

(25) Morokuma, K. Potential Energy Surface of the S_N2 Reaction in Hydrated Clusters. *J. Am. Chem. Soc.* **1982**, *104*, 3732–3733.

(26) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. S_N2 Reaction Profiles in the Gas Phase and Aqueous Solution. *J. Am. Chem. Soc.* **1984**, *106*, 3049–3050.

(27) Garver, J. M.; Fang, Y.-r.; Eyet, N.; Villano, S. M.; Bierbaum, V. M.; Westaway, K. C. A Direct Comparison of Reactivity and Mechanism in the Gas Phase and in Solution. *J. Am. Chem. Soc.* **2010**, *132*, 3808–3814.

(28) Kubelka, J.; Bickelhaupt, F. M. Activation Strain Analysis of S_N2 Reactions at C, N, O, and F Centers. *J. Phys. Chem. A* **2017**, *121*, 885–891.

(29) Szabó, I.; Czakó, G. Revealing a Double-Inversion Mechanism for the $F^- + CH_3Cl$ S_N2 Reaction. *Nat. Commun.* **2015**, *6*, 5972.

(30) Hamlin, T. A.; van Beek, B.; Wolters, L. P.; Bickelhaupt, F. M. Nucleophilic Substitution in Solution: Activation Strain Analysis of Weak and Strong Solvent Effects. *Chem. Eur. J.* **2018**, *24*, 5927–5938.

(31) Tirado-Rives, J.; Jorgensen, W. L. QM/MM Calculations for the Cl^- + CH_3Cl S_N2 Reaction in Water Using CM5 Charges and Density Functional Theory. *J. Phys. Chem. A* **2019**, *123*, 5713–5717.

(32) Peng, Y.; Pak, A. J.; Durumeric, A. E. P.; Sahrmann, P. G.; Mani, S.; Jin, J.; Loose, T. D.; Beiter, J.; Voth, G. A. OpenMSCG: A Software Tool for Bottom-Up Coarse-Graining. *J. Phys. Chem. B* **2023**, *127*, 8537–8550.

(33) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. An Efficient Real Space Multigrid QM/MM Electrostatic Coupling. *Journal of Chemical Theory and Computation* **2005**, *1*, 1176–1184.

(34) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: Cambridge, 2009.

(35) Momany, F. A. Determination of Partial Atomic Charges from Ab Initio Molecular Electrostatic Potentials. Application to Formamide, Methanol, and Formic Acid. *J. Phys. Chem.* **1978**, *82*, 592–601.

(36) Cox, S. R.; Williams, D. E. Representation of the Molecular Electrostatic Potential by a Net Atomic Charge Model. *J. Comput. Chem.* **1981**, *2*, 304–323.

(37) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **1990**, *11*, 361–373.

(38) Wiberg, K. B.; Rablen, P. R. Comparison of Atomic Charges Derived via Different Procedures. *J. Comput. Chem.* **1993**, *14*, 1504–1518.

(39) Verstraelen, T.; Pauwels, E.; De Proft, F.; Van Speybroeck, V.; Geerlings, P.; Waroquier, M. Assessment of Atomic Charge Models for Gas-Phase Computations on Polypeptides. *J. Chem. Theory Comput.* **2012**, *8*, 661–676.

(40) Nelder, J. A.; Mead, R. A Simplex Method for Function Minimization. *Comp. J.* **1965**, *7*, 308–313.

(41) Kühne, T. D. et al. CP2K: An Electronic Structure and Molecular Dynamics Software Package - Quickstep: Efficient and Accurate Electronic Structure Calculations. *J. Chem. Phys.* **2020**, *152*, 194103.

(42) Martinez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *Journal of Computational Chemistry* **2009**, *30*, 2157–2164.

(43) Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen Web Server: An Automatic OPLS-AA Parameter Generator for Organic Ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336.

(44) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.

(45) Bonomi, M. et al. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16*, 670–673.

(46) Grossfield, Alan WHAM: The Weighted Histogram Analysis Method.

(47) Acevedo, O.; Jorgensen, W. L. *Chapter 14 Solvent Effects on Organic Reactions from QM/MM Simulations*; Annual Reports in Computational Chemistry; 2006; pp 263–278.

(48) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(49) E, Weinan; Ren, Weiqing; Vanden-Eijnden, Eric String Method for the Study of Rare Events. *Phys. Rev. B* **2002**, *66*, 052301.

(50) Liu, Y.; Li, C.; Voth, G. A. Generalized Transition State Theory Treatment of Water-Assisted Proton Transport Processes in Proteins. *J. Phys. Chem. B* **2022**, *126*, 10452–10459.

(51) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in Free Energy Space. *J. Chem. Phys.* **2007**, *126*, 054103.

(52) Shakhparonov, M. I.; Akhadov, Ya. Yu. Dielectric Properties and Molecular Structure of Water-Acetone Solutions. *J. Struct. Chem.* **1965**, *6*, 15–19.

(53) Cole, K. S.; Cole, R. H. Dispersion and Absorption in Dielectrics I. Alternating Current Characteristics. *J. Chem. Phys.* **1941**, *9*, 341–351.

(54) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.

(55) Zuckerman, D. M.; Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu. Rev. Biophys.* **2017**, *46*, 43–57.

(56) Wang, D.; Wang, Y.; Chang, J.; Zhang, L.; Wang, H.; E., W. Efficient Sampling of High-Dimensional Free Energy Landscapes Using Adaptive Reinforced Dynamics. *Nat Comput Sci* **2022**, *2*, 20–29.

(57) McCammon, J. A. Protein Dynamics. *Rep. Prog. Phys.* **1984**, *47*, 1.

(58) Shaw, D. E. et al. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. New York, NY, USA, 2021; pp 1–11.

(59) Kaiser, S.; Yue, Z.; Peng, Y.; Nguyen, T.; Chen, S.; Teng, D.; Voth, G. Molecular Dynamics Simulation of Complex Reactivity with the Rapid Approach for Proton Transport and Other Reactions (RAPTOR) Software Package. *J. Phys. Chem. B* **2024**,

(60) Wu, J.; Xue, W.; Voth, G. A. K-Means Clustering Coarse-Graining (KMC-CG): A Next Generation Methodology for Determining Optimal Coarse-Grained Mappings of Large Biomolecules. *J. Chem. Theory Comput.* **2023**, *19*, 8987–8997.

(61) Henderson, R. L. A Uniqueness Theorem for Fluid Pair Correlation Functions. *Phys. Lett. A* **1974**, *49*, 197–198.

(62) Wagner, J. W.; Dama, J. F.; Durumeric, A. E.; Voth, G. A. On the Representability Problem and the Physical Meaning of Coarse-Grained Models. *J. Chem. Phys.* **2016**, *145*, 044108.

(63) Cao, Z.; Voth, G. A. The Multiscale Coarse-Graining Method. XI. Accurate Interactions Based on the Centers of Charge of Coarse-Grained Sites. *Journal of Chemical Physics* **2015**, *143*, 1–11.

(64) Vandevondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and Accurate Density Functional Calculations Using a Mixed Gaussian and Plane Waves Approach. *Computer Physics Communications* **2005**, *167*, 103–128.

(65) Goedecker, S.; Teter, M.; Hutter, J. Separable Dual-space Gaussian Pseudopotentials. *Phys. Rev. B* **1996**, *54*, 1703–1710.

(66) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(67) Hub, J. S.; de Groot, B. L.; Grubmüller, H.; Groenhof, G. Quantifying Artifacts in Ewald Simulations of Inhomogeneous Systems with a Net Charge. *J Chem Theory Comput* **2014**, *10*, 381–390.

(68) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *The Journal of Chemical Physics* **1984**, *81*, 511–519.

(69) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

CHAPTER A

APPENDIX: EXPERIMENTAL RESULTS ON THE LEAKAGE OF Δ107-EMRE

A.1 Acknowledgements

This appendix shows the experimental results from our collaborator, who demonstated the leaking properties of the tail truncation mutant Δ107-EmrE we studied in Chapter 2. This part is written by our collaborator, Ms. Merissa Brousseau, Dr. Nathan Thomas, and Prof. Katherine Henzler-Wildman from the University of Wisconsin, Madison. These results are included in this dissertation to provide the necessary background for our computational work.

A.2 Results

A.2.1 EmrE is properly folded and functional when the C-terminal tail is truncated

We first assessed whether C-terminal tail truncation affected the ability of EmrE to confer resistance to toxic substrates, the well-established primary function of this transporter. Growth assays of MG1655-Δ*emrE E. coli* cells expressing WT-, Δ107-, or E14Q-EmrE show that all strains grow well in the absence of toxic compounds (Figure A.1B). In these assays, uninduced leaky expression from a low-copy number plasmid (p15 origin) with a pTrc promoter keeps transporter expression relatively low.[1] In the presence of ethidium bromide, a substrate commonly used to assess the activity of EmrE and other multidrug efflux pumps, functional transporter is required for survival (Figure A.1C). This confirms the known resistance activity of WT-EmrE, that E14Q-EmrE is non-functional, and that Δ107-EmrE is

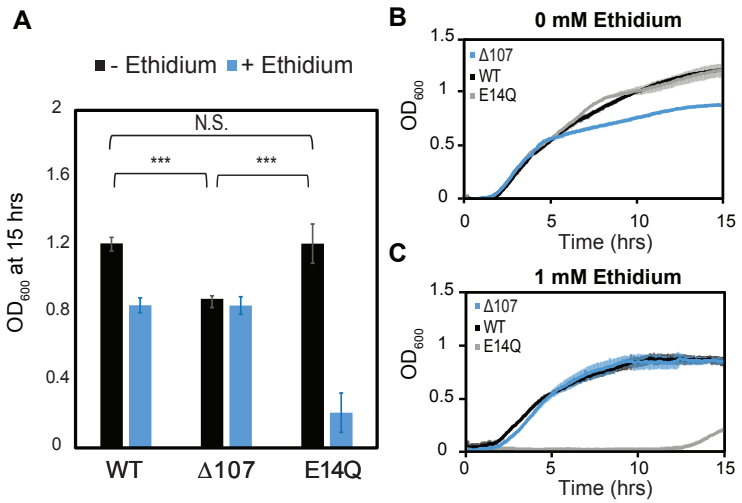


Figure A.1: **C-terminal tail truncation does not impair the ability of EmrE to confer resistance to toxic compounds.** (A-C) WT-, E14Q-, or Δ107-EmrE was heterologously expressed in MG1655-ΔEmrE E. coli using a plasmid with p15 origin and pTrc promoter without induction to minimize any growth defect due to expression. *In vivo* growth assays were monitored by OD700 to allow consistent monitoring in the absence (B) or presence of (C) ethidium bromide. Growth at 15 hours (A) shows identical growth for WT-EmrE and Δ107-EmrE in the presence of ethidium, while E14Q-EmrE is severely impaired (A,B). There is a 20% reduction in growth for Δ107-EmrE relative to WT-EmrE or non-functional EmrE ($p < 0.001$), but this does not prevent the mutant from transporting ethidium out of the cell and thus conferring resistance (A,C).

properly expressed to the inner membrane, folded and functionally able to confer resistance to toxic compounds in a manner comparable to the WT transporter.

A.2.2 Truncation of the C-terminal tail enhances proton leak through EmrE

There is a small but reproducible growth defect for cells expressing Δ107-EmrE in the absence of exogenous substrate (Figure A.1A, B). This defect becomes apparent around 5 hours, the point at which available fermentable sugars in LB media are depleted, increasing dependence on the proton motive force (PMF) for energy production[2]. A similar time dependent growth inhibition is observed for WT-EmrE in the presence of harmane, and we have previously shown that this substrate triggers uncoupled proton leak through EmrE.[1] Thus, while Δ107-EmrE competently performs the proton-coupled drug antiport necessary to confer resistance to toxic substrates, it is detrimental to *E. coli* in the absence of known small molecule substrate in a manner suggestive of proton leak.

To directly test this hypothesis, we measured proton leak in proteoliposomes using the pH-sensitive dye pyranine. Pyranine was encapsulated inside proteoliposomes at pH 6.5, and the liposomes were then diluted 100-fold into pH 7.5 buffer. If protons leak out of the liposome down the proton concentration gradient, the internal pH will rise and pyranine fluorescence will increase. We compared the fluorescence, normalized to time zero (just after dilution), of proteoliposomes with WT-, Δ107-, or E14Q-EmrE. WT-EmrE proteoliposomes show a gradual increase in internal pH over time (Figure A.2A, solid black), which is faster than the pH change for E14Q-EmrE proteoliposomes (Figure A.2A, solid gray). This is consistent with a small amount of proton leak through WT-EmrE and a role for E14 in mediating leak. Δ107-EmrE proteoliposomes show a much faster rise in internal pH in this assay (Figure A.2A, solid red), consistent with enhanced proton leak upon truncation of the C-terminal tail. As a positive control for rapid proton leak, we repeated the experiment with the protonophore CCCP in the external buffer. In this case, WT-, E14Q- and Δ107-

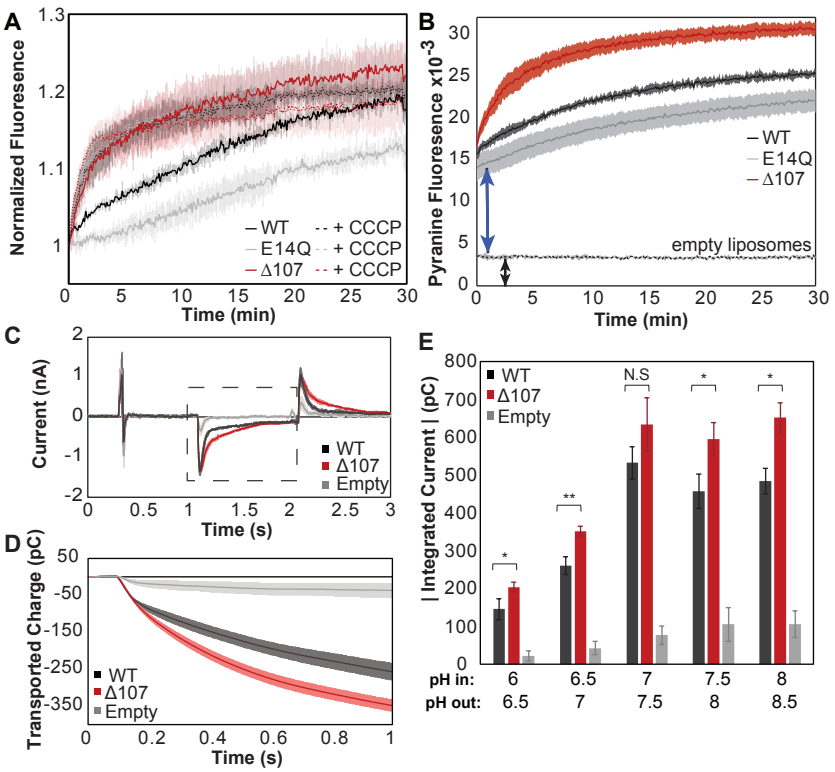


Figure A.2: **C-terminal tail truncation enhances proton leak.** (A-B) Pyranine fluorescence directly reports on proton leak through EmrE. (A) WT-, Δ107- or E14Q-EmrE proteoliposomes with 1 mM internal pyranine and internal pH 6.5 were diluted 100-fold into pH 7.5 buffer (solid lines) or pH 7.5 buffer with CCCP (dashed lines) and fluorescence was normalized to time zero. CCCP is a protonophore, providing a positive control for maximal proton leak under these conditions. (B) Pyranine fluorescence normalized by subtracting the fluorescence of proteolipsomes diluted into pH 6.5 (no gradient, baseline) from the fluorescence of proteoliposomes diluted into pH 7.5 (transport) shows intraliposomal pH change with proteoliposomes in the lag time prior to initial fluorescence read and increased intraliposomal pH change for Δ107-EmrE than WT-EmrE or E14Q-EmrE. (C-E) Solid supported membrane electrophysiology data shows measurable charge movement through WT- and Δ107-EmrE proteoliposomes in the presence of a pH gradient alone, as compared to empty liposomes, with increased charge transport through Δ107-EmrE. (C) Current is recorded in real time as a matching pH internal buffer (pH 6.5) is flowed over the liposomes to establish baseline, then a higher pH (pH 7) buffer is flowed over the liposomes to create an outwardly directed proton gradient (dashed box), and finally the initial buffer (pH 6.5) is flowed back over the liposomes to reverse the charge movement and return to baseline. (D) The recorded current during the period of the applied gradient (dashed box, C) is integrated to determine the transported charge during that time. (E) The integrated currents were recorded for the same ΔpH at different base pH values (labels at bottom). While both WT- and Δ107-EmrE leaked more significantly at higher pH, Δ107-EmrE shows increased proton leak for all pH conditions tested.

EmrE proteoliposomes all show the same rapid increase in internal pH (Figure A.2A, dotted lines), and the results match the timescale and amplitude of proton leak observed for the Δ107-EmrE proteoliposomes.

Initiation of the assay by dilution prevents a clean measurement of the fluorescence baseline, obscuring any rapid change in internal liposome pH in the few seconds between dilution and initial fluorescence read. We therefore repeated the assay with side-by-side dilution of proteoliposomes into pH 7.5 buffer to match the internal pH (baseline), or pH 6.5 buffer to initiate transport. Using this baseline to correct the transport fluorescence reveals a rapid internal pH changes missed in the initial assay (Figure A.2B). The empty liposome control is completely flat, showing that the liposomes are tight and do not leak in the absence of transporter. However, the signal is non-zero (Figure A.2B, black arrow), most likely due to a small amount of residual pyranine on the exterior of the liposomes after encapsulation. The time 0 fluorescence of WT-, E14Q- and Δ107-EmrE proteoliposomes are much larger than for empty liposomes (Figure A.2B, blue arrow), indicating that there is an additional very rapid change in the internal pH in the presence of protein. While the encapsulated pyranine is protected from the direct impact of external pH change, proton transport (leak) through EmrE would change the internal pH. The transmembrane pH gradient may also affect the structure and dynamics of EmrE, altering the pK_a of key residues (E14 in WT- and Δ107-EmrE, H110 in WT- and E14Q-EmrE) and leading to rapid release of protons in the proteoliposome interior, or cause a rapid burst-phase of leak as the transporter transitions from a symmetric-pH conformation to an asymmetric-pH conformation. The behavior over longer times is similar to that observed with time 0 normalization in Figure A.2A: E14Q-EmrE proteoliposomes show gradually increasing internal pH, WT-EmrE proteoliposomes have enhanced leak and Δ107-EmrE proteoliposomes have significantly faster initial rate of internal pH change and overall greater increase in intralioposomal pH over the time of the assay, consistent with enhanced proton leak through Δ107-EmrE.

To further assess the initial very rapid pH change during the dead time of the pyranine assay, we used solid supported membrane electrophysiology (SSME) to measure ΔpH-driven current in proteoliposomes since this technique provides a continuous readout over short times immediately after gradient creation. A large number of proteoliposomes are adsorbed on the surface of a gold-coated sensor, enabling highly sensitive detection of electrogenic processes, including proton binding or conformational change (gating current) or proton transport (leak). The liposomes are first equilibrated with external buffer identical to the interior, and then a different external buffer is rapidly washed over the liposomes to create a transmembrane pH gradient while recording is in progress. The capacitive current is measured (Figure A.2C) and integrated to yield the total transported charge in response to the applied gradients (Figure A.2D). In the absence of drug, Empty liposomes have minimal charge movement as expected for minimal proton leak. However, both WT and Δ107-EmrE have measurable current in the presence of a pH gradient, with increased net charge movement at higher absolute pH and Δ107-EmrE has consistently higher net charge movement under all conditions (Figure A.2E). In symmetric pH environments, the only amino acid side chains with pK_a values near neutral pH are E14 (p$K_a = 6.8 \pm 0.1$ and 8.5 ± 0.2 at 25 °C) and H110 (6.98 ± 0.01, 7.05 ± 0.02). Δ107-EmrE is lacking H110, so any proton binding/release from H110 that contribute to the capacitative current will be absent in Δ107-EmrE, but the net charge transport (Figure A.2C, D) is greater for Δ107-EmrE than for WT-EmrE in direct contrast with a very simple model where proton binding/release from the C-terminal H110 accounts for the very fast current and pH change. Any conformational change involving the the C-terminal tail that contributes to the capacitative current should also decrease as the average pH increases and H110 protonation and net charge decrease, however the opposite pH dependence is observed for WT-EmrE, indicating that this is unlikely to be a major contributor to the observed current in the first second after gradient creation. Thus, the observed current must be due to a more complex impact of C-terminal tail truncation on

EmrE gating and transport. A model where movement of the C-terminal tail regulates access to E14 and C-terminal truncation alters this gating process would be consistent with the observed currents and their pH dependence (Figure A.2C–E), as well as the longer-timescale change in intra-liposome pH (Figure A.2B). It is important to note that the 1 second length of the SSME traces does not fully reach steady state (net transported charge is still increasing at the end of the assay, Figure A.2D), but instead represents net charge movement within 1 second. These assays provide information on the processes happening during the dead time of the pyranine assay (Figure A.2B) and are consistent with the different magnitude burst phase observed for WT- and Δ107-EmrE in that experiment. All together, this data supports a role for the C-terminal tail as a secondary gate that minimizes proton leak through WT-EmrE in the absence of substrate.

A.2.3 Re-assessing protonation state by NMR

Prior NMR pH titrations of WT-EmrE in the absence of drug-substrate revealed that the two essential E14 residues in the asymmetric homodimer have distinct pK_a values,[4] reflecting their unique structural environments. Upon binding TPP^+, one of the E14 residues is protected from protonation and no longer titrates, while the other ($E14^A$) retains a pK_a of 6.8 ± 0.1, similar to the drug-free state.[4] The only other titratable residue previously identified in WT-EmrE is the C-terminal histidine,[5] which also has a pK_a near neutral pH in WT-EmrE. Since H110 was removed by truncation in Δ107-EmrE, we expected only one protonation event in NMR pH titrations of TPP^+-bound Δ107-EmrE (Figure A.3A), corresponding to the one E14 residue that remains titratable when substrate is bound. A single protonation event would normally result in linear change in peak position from the chemical shift of the protonated state to the chemical shift of the deprotonated state over the course of the titration. This is because proton on-/off- is almost always in the fast-exchange limit for NMR, resulting in observation of a single peak at the population-weighted average

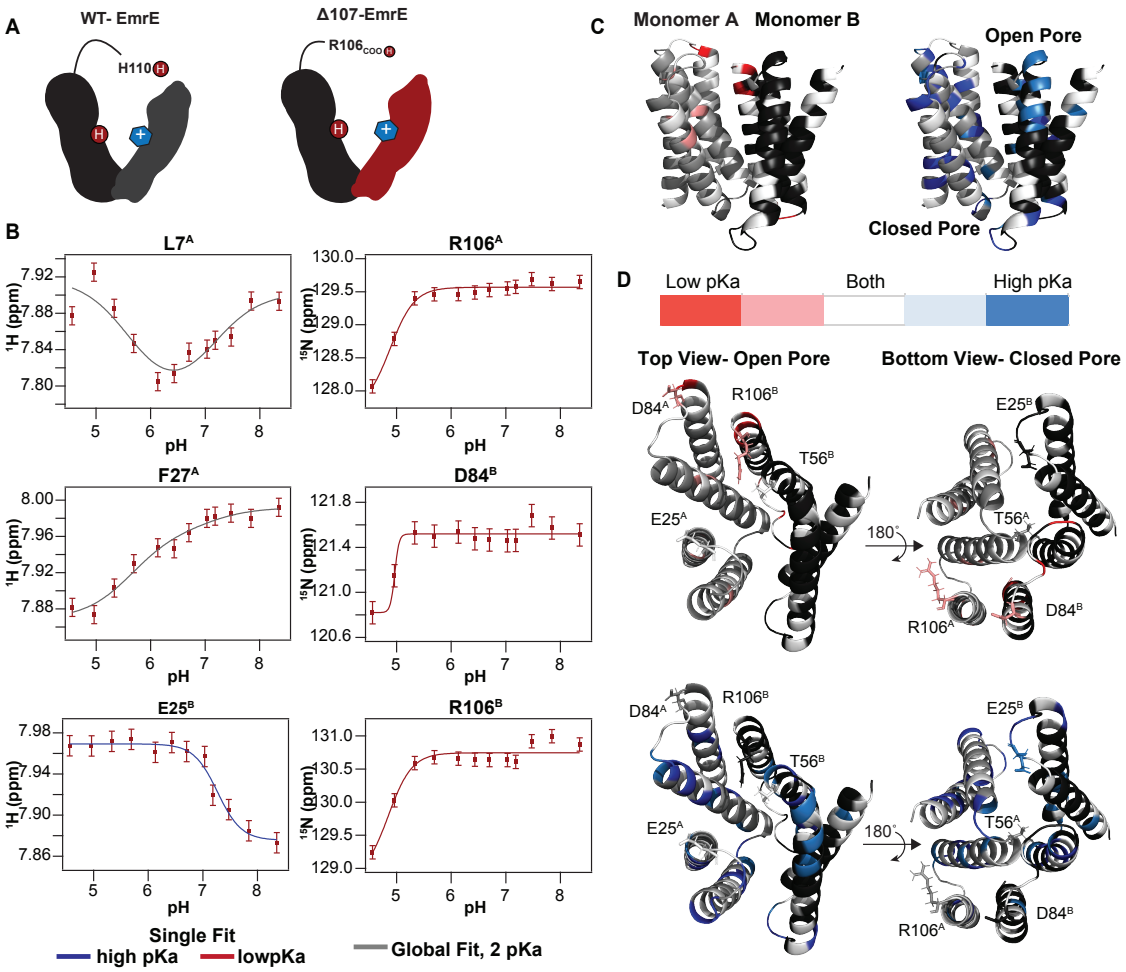


Figure A.3: **pH titration of TPP$^+$-bound Δ107-EmrE supports the possibility of secondary gating.** (A) In WT-EmrE bound to TPP$^+$, one E14 residue and one H110 residue, are the only titratable sites (dark red circles labeled H$^+$). In Δ107-EmrE, H110 is not present, suggesting that only one titratable group should remain (E14). (B) The proton and nitrogen chemical shifts for individual residues of TPP$^+$-bound Δ107-EmrE were recorded as a function of pH. The resulting titration profiles do not show the expected single-pKa pattern. Some are curved, consistent with multiple pKa values, and others are consistent with a single pK_a but at either high or low pH. All of the data can be globally fit to two pK_a values, using either a 2-pK_a fit (5.6 and 7.1, grey) or single pK_a fit at the relevant value (5.6, red; 7.1, blue). (C) Residues sensitive to each pK_a value are plotted on the faRM model (Ref. 3) using the indicated color scale. (D) Residues strongly sensing the lower pK_a value cluster around the C-terminus and 3–4 loop on both the open and closed face of the transporter, while the 1–2 loop and T56 on the open side of the pore sense both pKa values (Left).

chemical shift at each titration point. However, several peaks exhibit distinctly curved titration paths with transitions in two different pH ranges for TPP^+-bound Δ107-EmrE. Plotting the chemical shift of well-resolved peaks as a function of pH yields titration curves that can be fit using standard pK_a equations. In this case, the data is well fit with a global 2 pK_a model yielding apparent pK_a values of 5.6 ± 0.2 and 7.1 ± 0.2. The higher of these two pK_a values is close to the $E14^A$ pK_a in TPP^+-bound WT-EmrE and the residues most sensitive to this protonation event are found lining the transport pore near $E14^A$, supporting assignment of this pK_a to $E14^A$ (Figure A.3D).

This leaves the lower pK_a unaccounted for. Possibilities include the other acidic residues in the loops: E25 or D84, or the C-terminal carboxylate itself. The residues most sensitive to this lower pK_a are centered around the C-terminus, with the most prominent chemical shift changes at R106. Several other residues in the tail of subunit A and nearby regions of the protein (Figure A.3C) detect this lower pK_a value. This includes residues in the TM1–TM2 loop and TM3–TM4 of monomer B, which are on the same "open" face of EmrE. Residues E25 and D84 are located in these loops, have previously been suggested to be part of a secondary binding site for substrates like TPP^+.[6] There is no evidence that these residues titrate in this pH range in WT-EmrE or in other mutants for which we have carried out NMR pH titrations (9, 22, 35). Thus, there is no evidence for E25 or D84 titration in this pH range in the full-length transporter. However, truncation of the C-terminus in Δ107-EmrE could alter the structure, environment and pK_a of these residues. Indeed, the hydrogen bond between R106 and D84 observed in the WT-EmrE simulations (Figure 2.2B, D) is broken when the tail is truncated in Δ107-EmrE (Figure 2.2C), and D84 and R106 have some of the largest pH-dependent chemical shift changes. There are relatively few experimental reports of the pK_a of the terminal carboxylate in proteins, but it has been reported to have a pK_a as high as 5.9 for the partially-buried C-terminus of subunit c of F_1F_0-ATP Synthase.[7] The MD simulations show the C-terminal carboxylate in WT-EmrE hydrogen bonds with

T56 and K22 on the open face of the transporter (Figure 2.2B, D), but the C-terminus in Δ107-EmrE no longer interacts with these residues (Figure 2.2C). Examining the residues that sense this lower pK_a shows that the TM3–TM4 and TM1–TM2 loops on the open face have larger chemical shift changes associated with the low pK_a protonation event than those loops on the closed face of EmrE, while the C-terminal tail residues in both subunits detect the lower pK_a. Comparison with the MD simulations shows that the residues involved in the hydrogen bond networks anchoring the C-terminal tail over the pore align well with the full list of residues that are sensitive to the lower pK_a in the NMR titrations, including T56A (TM2–TM3 loop on the "open" face), V15–G17, I37, Y40, V69, and S72–L73. Thus, although this lower pK_a is likely an artifact of tail truncation, this data experimentally confirms the importance of the interactions between the tail and the rest of EmrE identified in the MD simulations as important for disrupting the water wire and occluding the E14 binding site.

A.2.4 Proton leak through Δ107-EmrE does not synergize with harmane

The well-established function of EmrE is proton-coupled antiport of toxic substrates, leading to toxin efflux and drug resistance. Recently, we discovered that some substrates, such as harmane, instead trigger uncoupled proton uniport, leading to ΔpH dissipation and defects in NADH production and growth in *E. coli*,[1] essentially causing susceptibility rather than resistance. We suspected the enhanced proton leak observed through Δ107-EmrE and this harmane-triggered proton leak might have common elements in their underlying mechanism. Using SSME, we first compared the inherent proton leak through WT- and Δ107-EmrE in the absence of substrate. For the same ΔpH driving force, Δ107-EmrE has ≈ 3-fold greater proton leak than WT-EmrE (Figure A.4A, D). However, upon addition of 16 μM harmane, there is a large increase in proton leak through WT-EmrE and a small increase in leak through Δ107-EmrE, such that this substrate triggers identical total leak through either transporter (Figure A.4B,D). This SSME-detected proton leak increases with harmane concentration and

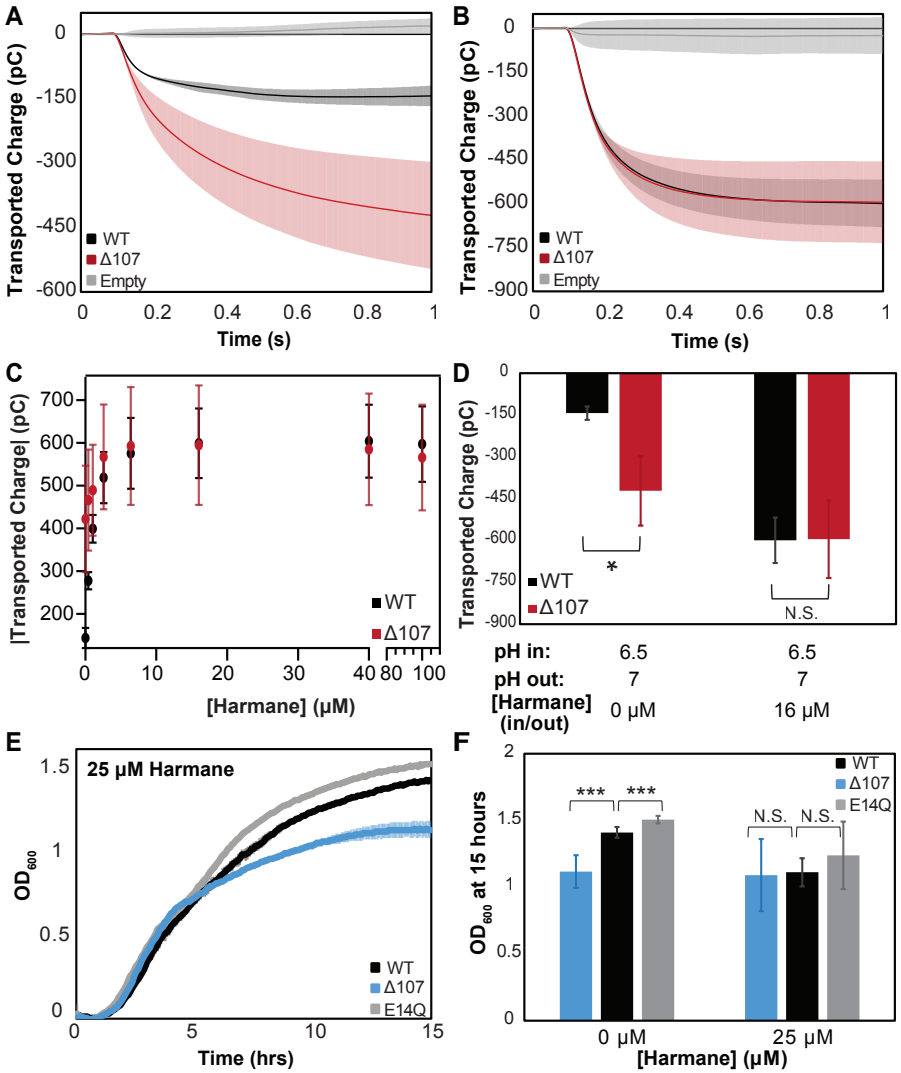


Figure A.4: **Intrinsic leak in Δ107-EmrE does not synergize with harmane-induced leak.** SSME traces of transported charge corresponding to proton leak in the absence (A) and presence (B) of harmane show that 16 μM harmane induces leak in WT-EmrE that is comparable to the leak observed through Δ107-EmrE in the absence of harmane. In the presence of increasing concentrations of harmane (C) the leak signal for WT-EmrE quickly converges to that of Δ107-EmrE. The leak observed for Δ107-EmrE is more variable, displaying larger standard deviations than WT-EmrE proteoliposomes (C). This could be due to greater variability in the unregulated transport activity of Δ107-EmrE compared to harmane-gated leak in WT-EmrE, and the impact of this unregulated behavior on the sensitivity of SSME to variation in the absolute number of proteoliposomes adsorbed on the surface sensor. Growth assays in the absence of substrate (Figure A.1A) show a clear growth defect for E. coli expressing Δ107-EmrE compared to WT-EmrE, which is eliminated when cells are grown in the presence of 25 μM (E-F).

is saturable in both WT- and Δ107-EmrE (Figure A.4C). If harmane acts as an allosteric regulator of the transporter that can unlock the secondary gate, then a saturating amount of harmane will result in the maximal signal for the WT transporter as observed. In Δ107-EmrE, C-terminal truncation removes this secondary gate and key residues in the allosteric site, rendering proton leak relatively independent of harmane (Figure A.4).

To test this theory in the native organism, we conducted *in vivo* growth assays with WT- and Δ107-EmrE in the presence of harmane. MG1655-Δ*emrE E. coli* cells constitutively expressing WT-, E14Q-, and Δ107-EmrE from a plasmid were grown in the presence of 25 μM harmane. The cells expressing E14Q grew equally well in the presence or absence of harmane as the mutation of the primary binding site prevents proton binding in the transport pore and abolishes any proton leak (Figure A.4E). In the presence of harmane the difference in growth between Δ107- and WT-EmrE is eliminated, with significant growth defect for both constructs relative to E14Q-EmrE (Figure A.3E, F). This *in vivo* data exactly matches the *in vitro* SSME results, demonstrating the importance of the C-terminal tail in gating and allosteric regulation of EmrE *in vivo*.

References

(1) Peyton J. Spreacker; Nathan E. Thomas; Will F. Beeninga; Merissa Brousseau; Colin J. Porter; Kylie M. Hibbs; Katherine A. Henzler-Wildman Activating Alternative Transport Modes in a Multidrug Resistance Efflux Pump to Confer Chemical Susceptibility | Nature Communications. *Nat. Commun.* **2022**, *13*, 7655.

(2) Baev, M. V.; Baev, D.; Radek, A. J.; Campbell, J. W. Growth of *Escherichia Coli* MG1655 on LB Medium: Monitoring Utilization of Sugars, Alcohols, and Organic Acids with Transcriptional Microarrays. *Appl. Microbiol. Biotechnol.* **2006**, *71*, 310–316.

(3) Vermaas, J. V.; Rempe, S. B.; Tajkhorshid, E. Electrostatic Lock in the Transport Cycle of the Multidrug Resistance Transporter EmrE. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, E7502–E7511.

(4) Morrison, E. A.; Robinson, A. E.; Liu, Y.; Henzler-Wildman, K. A. Asymmetric Protonation of EmrE. *Journal of General Physiology* **2015**, *146*, 445–461.

(5) Thomas, N. E.; Wu, C.; Morrison, E. A.; Robinson, A. E.; Werner, J. P.; Henzler-Wildman, K. A. The C Terminus of the Bacterial Multidrug Transporter EmrE Couples Drug Binding to Proton Release. *Journal of Biological Chemistry* **2018**, *293*, 19137–19147.

(6) Glaubitz, C.; Gröger, A.; Gottschalk, K.; Spooner, P.; Watts, A.; Schuldiner, S.; Kessler, H. ^{31}P-CP–MAS NMR Studies on TPP^+ Bound to the Ion-Coupled Multidrug Transport Protein EmrE. *FEBS Letters* **2000**, *480*, 127–131.

(7) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A Summary of the Measured pK Values of the Ionizable Groups in Folded Proteins. *Protein Sci.* **2009**, *18*, 247–251.