

QM/MM APPLICATIONS AND CORRECTIONS FOR CHEMICAL REACTIONS

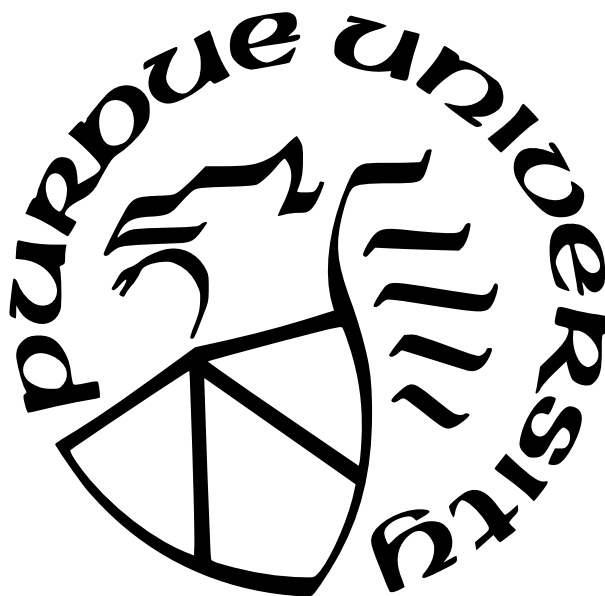
by
Bryant Kim

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



Department of Chemistry and Chemical Biology

Indianapolis, Indiana

May 2023

**THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL**

Dr. Jingzhi Pu, Ph.D., Chair

Department of Chemistry and Chemical Biology

Dr. Christoph Naumann, Ph.D.

Department of Chemistry and Chemical Biology

Dr. Jonah Vilseck, Ph.D.

IU School of Medicine, Department of Biochemistry and Molecular Biology

Dr. Ian Webb, Ph.D.

Department of Chemistry and Chemical Biology

Approved by:

Dr. Eric C. Long

I first dedicate this PhD thesis dissertation to my late aunt Connie Hong and my late grandfather Kyu Chong. Aunt Connie's passing from lung cancer during the early stages of my PhD candidacy was a difficult time, but her pride and encouragement in my pursuit of a career in healthcare was a driving force in my determination to succeed. She watched over me as a young child and always had high hopes for my future. Also, to my grandfather Kyu Chong, thank you for sharing with me your wartime experiences and your perspectives on life. Your belief in my success and your wise counsel will always stay with me. I am forever grateful for the unwavering love and support that both of you provided throughout my life, and it is an honor to dedicate my PhD to you both.

Also, to my family, I dedicate this dissertation with deep gratitude and profound admiration. You have been my unwavering source of strength and support, and I am beyond grateful for the love and encouragement you have given me. As I pursued my academic journey, you were my greatest cheerleaders, sharing in my triumphs and comforting me in my disappointments. You instilled in me a deep appreciation for learning and emphasized the importance of education. Through all of life's ups and downs, you have been my constants, my pillars of strength. I am immensely grateful for the love, support, and guidance you have given me throughout my life. I will always treasure the memories of our time together, as we navigated new experiences and bonded over family dinners and games. You have taught me the value of family, and I will carry that lesson with me always. As I complete this PhD, I do so with the knowledge that I would not have been able to accomplish it without you. Thank you for being my biggest supporters and for always believing in me. I love you all more than words can express.

ACKNOWLEDGMENTS

- I would first like to express my sincere appreciation to my committee members, Dr. Vilseck, Naumann, Webb, and Jingzhi Pu for their invaluable contributions to my doctoral studies and research. Each of them has played a major role in shaping my PhD candidacy, and their expertise and guidance have been invaluable throughout my doctoral journey. I enjoyed talking to Dr. Vilseck about his work in my aspirations for drug design, to Dr. Naumann who shared his understanding of quantum chemistry and Dr. Webb for useful career advice with his experience in the National Lab. Lastly, a lot of thanks goes to my advisor Dr. Jingzhi Pu, for believing in me every step of the way of my PhD career. Since meeting him in high school as a fellow soccer player, and meeting him later on as an undergraduate professor in physical chemistry, Dr. Pu has played a major role not only in my career but personal life. His dedication and care for my well-being and experience in the field is something that I'm so grateful for. Thank you for guiding me on the ride of my life and for helping me even afterwards in my career search. It goes without saying that none of this would have been possible without you, and I hope you understand my deepest and most thankful appreciation towards you.
- I would like to express my sincere gratitude to Jonathan, Jae, and Mike from EKPC for their invaluable contributions to my doctoral studies. When I first moved to Indiana, I didn't have any friends, and these guys helped introduce me to the area and support each other through high school. Without their encouragement and guidance, my experience would not have been the same. Their assistance and support also played a significant role in the completion of my thesis. Thank you for everything.
- I extend my heartfelt thanks to James, Andy, Cameron, Brandon, Destry, Chet, Sangwook, Sungsik, and Carrie from The Brotherhood for their encouragement and help during my doctoral studies. Jack, Andy, and Cameron were my roommates at Taylor University, and they really shaped my experience as an undergraduate student. Brandon, Destry, and Chet were good friends of mine that helped me enjoy my college

experience, while Sangwook and Sungsik were close friends of mine who taught me about Korean culture as exchange students. Additionally, Carrie has always been a source of joy both in college and now, supporting growth in life and in my career. Without their support, my doctoral studies would not have been the same. Thank you for everything.

- I would like to express my sincere appreciation to TJ, Ben, Sheryl, Garret, Greg, Paul, George, Kirsten, Chelsea, and Lauren from Eli Lilly for their invaluable contributions to my doctoral studies. This group of people helped shape my first experience in the industrial setting, teaching me about the joy of working in a chemistry lab and different instrumentation they could use to support pharmaceutical work. They also provided a great environment for growing professionally and personally, both in and out of the workplace. Their insights and advice were essential in the development of my research, and without their support, my doctoral studies would not have been the same. Thank you for everything.
- I extend my heartfelt gratitude to Patrick, Jacob, Zach, Eric, Shane, Ashley, Brionne, Meaghan, and Jamie from Mylan Technologies for their invaluable assistance and guidance during my doctoral studies. I worked alongside Patrick, Jacob, Zach, Eric, Shane, and Ashley on the night crew at Mylan, and they made me feel like home in Vermont. We did interesting work together in St Albans, Vermont, and their support was essential to my development as a researcher. Brionne, Meaghan, and Jamie also provided me with much-needed help when I suffered an accident to my knee. I'm so grateful for their assistance in my time of need. Without their support, my doctoral studies would not have been the same. Thank you for everything.
- I would like to express my sincere appreciation to Cameron, Rob, Tim, and Drew from AIT Bioscience for their invaluable contributions to my doctoral studies. All of these guys helped me and were supportive of my move to pursue my PhD candidacy. In the workplace, they were very professional and inspired me in the science and work that they were doing. They made me feel respected and valued in the workplace, and

I'm really grateful for that. Their involvement in my research was essential to my development as a researcher, and without their support, my doctoral studies would not have been the same. Thank you for everything.

- I extend my heartfelt thanks to Phil, Mert, Lauren, Sergey, Abhirup, Rae, Huafeng, Chuanjie, Yunrui, Ozan, and Joseph from Roivant Discovery for their invaluable contributions to my doctoral studies. Phil, Mert, Lauren, and Sergey were all a part of the same group, and we had great discussions about the science we were doing. Their expertise in computational chemistry and strong interest in the field were a huge motivation and inspiration for me. Huafeng and Chuanjie inspired the team and department to work towards a common goal of being innovative and creating new science. I also want to thank Yunrui, Ozan, and Joseph, who were the other interns in New York with me. It was amazing to experience different parts of the city with them and share our internship experiences with one another. A special thanks to Yunrui for philosophical discussions about science and life, and to Mert, who shared his academic experiences and demonstrated a lot of care in supporting me throughout the internship both in friendship and in experience. Without their support, my doctoral studies would not have been the same. Thank you for everything.
- I would like to express my sincere appreciation to Melissa, Kyle, Lauren, Matt, and Jasmine from the Graduate Students at IUPUI for their invaluable support and assistance during my doctoral studies. The graduate program can be a lonely game, and it goes without saying that without their support, my doctoral studies would not be the same. Whether it was watching Game of Thrones together, going out for trivia, or playing ultimate frisbee, it was an honor to share my PhD experience and time together with you all. Thank you for being there for me, and for making my doctoral journey a lot more enjoyable.
- I extend my sincere appreciation to Ryan Snyder and Ryan Young from My Lab for their invaluable contributions to my doctoral studies. To both of the Ryans, thank you for being in the same lab as me and discussing the science of our work. Specifically, for

Ryan Young, I appreciate some of the philosophical talks that we had and discussions on career growth. And to Ryan Snyder, in addition to working out technical details of our work and discussing career aspirations, our friendship both in and out of the lab, whether it be riding boats on a lake or playing ultimate frisbee, is greatly appreciated, and I thank you for sharing time together in this PhD program. Best of luck to both of you as you embark on journeys with your new family and kids.

- I would like to express my sincere appreciation to Dr. Basu, Deiss, and Long from the Faculty at IUPUI for their invaluable contributions to my academic journey and research. All of these professors were major contributors to my education at the School of Science in IUPUI. Each of them taught me different courses, and in addition to teaching me new aspects of chemistry, all of them provided insights into my career and what it means to be an academic scientist. Their guidance and mentorship were invaluable throughout my doctoral studies, and I am grateful for the opportunities to learn from them. Thank you for everything.
- I would like to express my gratitude to Joy Manhattan, the church community in New York, for welcoming me with open arms into a bright and vibrant city, and making my experience in New York a very special one filled with many joyous moments and laughter. A special thanks goes to Michelle who really encouraged and inspired me to see the light at the end of the tunnel and motivated me to embrace not only the academics of completing my PhD but also the meaning behind it all. I am grateful to Derek, Soyoung, Arah, Debbie, and Caleb for being a part of my life in New York, and for making my doctoral journey a lot more enjoyable. Thank you for being there for me, and for your support and kindness throughout my time in New York.
- I would like to extend my deepest gratitude to my family friend and mentor, Sung. Sung moved to Indiana at the same time as I did, and his unwavering support and guidance throughout my doctoral journey have been invaluable. His presence in my life has been a constant source of encouragement and inspiration. Thank you, Sung, for being there for me and for believing in me every step of the way.

- I would like to express my heartfelt gratitude to my grandparents, Kyu Young, Eun Ran, Hee Ok, and Ching Sung. Their care and love for me over the years have been unwavering, and their sacrifices and strength in immigrating to this country have shaped me into the person that I am today. Their hard work and dedication to our family have been a constant source of inspiration throughout my doctoral journey, and I would not be where I am today without them. Thank you for all that you have done for me, and for always believing in me.
- I would like to extend my deepest appreciation to my aunts and uncles, Yoo Jin, Marybeth, Jin, Tomoko, Connie, Susan, Kab, and Sooman. Throughout my life, they have always watched over me and provided me with their love and guidance, helping me to become the person that I am today. Their care and support have been a constant source of strength and inspiration throughout my doctoral journey, and I am forever grateful to have them in my life. Thank you for everything you have done for me, and for always being there for me.
- I want to express my deepest gratitude to my siblings Ellison, Courtney, and Harrison for being a significant part of my life and shaping me into the person I am today. Despite the ups and downs, the memories we share growing up will always be cherished. As they continue to embark on their careers and start their own families, I feel incredibly proud and grateful to have them as my siblings. To my sister Emica and brothers-in-law Benji, thank you for being such wonderful partners to my siblings and bringing so much joy and positivity to our family.
- I would like to express my gratitude to my niece and nephews, Lily and Benny, for being a source of joy and inspiration in our family. It has been a privilege to watch Lily grow and thrive, and I'm proud of her for her determination and hard work. To my soon to be niece or nephew, I'm grateful for the happiness and excitement you have already brought to our family, and I can't wait to meet you and watch you grow. Your presence in our lives is a true blessing.

- I would like to acknowledge my cousin Jessica Hong, who has been an important figure in my life since we were young. We have always had a strong relationship, and I am grateful for her constant support and encouragement. Her advice and guidance have been invaluable to me, particularly in my career growth, and I feel fortunate to have her in my life. Thank you, Jessica, for treating me like a younger brother and for being such a positive influence on me.
- I would like to express my gratitude to my cousin Carol and her husband Jason for the positive impact they have had on my life. From watching over me as a kid to spending time with me and their friends, they have been a guiding force in my life. I am especially thankful to Jason, who helped me when I was teaching SAT in Philadelphia. They are both great role models and friends, and I am so proud of their two amazing children, which is a testament to their skills as parents.
- I would like to thank my cousins Danny and Andrew for their support and friendship throughout my PhD journey. Danny, thank you for being a great sports buddy and a friend. Andrew, it was a pleasure spending time with you on both coasts, and I am grateful for the ways in which we have enriched each other's lives over the years.
- My cousins Emily, Wayne, Agnus, and Joan in San Diego have been an important part of my life, and I am grateful for the joy and love they have brought to our family. From family trips to playing games together, their contributions have shaped me into the person I am today. I am excited to see where their careers take them and am proud of all they have accomplished.
- I am grateful to my cousins Julian, Quincy, and Elliot for their unwavering love and support throughout my PhD journey. Spending time with them and sharing moments together, from family vacations to deep conversations about our aspirations, has been a source of inspiration and comfort for me. I am especially thankful for their encouragement and guidance during the challenging times of my studies. Their friendship and care have helped shape me into the person I am today, and I am blessed to have them in my life.

- To my beloved dog Gio, I would like to express my deepest appreciation and gratitude for your unwavering love, comfort, and support throughout my PhD journey. Your presence has been a constant source of joy and motivation, and I will always cherish the peaceful walks we took together where I could collect my thoughts and reflect on my work. The time we spent together learning tricks and making YouTube videos were some of the happiest moments of my life. It is no exaggeration to say that I could not have made it through this PhD without you. Thank you for always being there for me, Gio.
- Finally, my heartfelt gratitude goes out to my parents, Ik Joo and Hye Kyung, for their unwavering love, support, and guidance throughout my doctoral journey. Ever since I was a young child, they have always put my needs above their own, and have showered me with love and encouragement in everything I do. Their sacrifices and dedication to providing me with the best experiences in life, whether it be through boy scouts, travel, sports, education, or family camping trips, have shaped me into the person that I am today. I am forever grateful for their presence in my life, and for always being my biggest cheerleaders. Thank you for all that you have done for me, and for your unwavering love and support.

TABLE OF CONTENTS

LIST OF TABLES	15
LIST OF FIGURES	16
LIST OF SCHEMES	22
ABSTRACT	23
1 INTRODUCTION	24
2 REACTION PATH-FORCE MATCHING IN COLLECTIVE VARIABLES: DETERMINING AB INITIO QM/MM FREE ENERGY PROFILES BY FITTING MEAN FORCE	28
2.1 Introduction	29
2.2 Theory	32
2.2.1 RP-FM Is Equivalent to Fitting Free Energy Mean Force	32
2.2.2 RP-FM-CV Fits Mean Force on Collective Variables in Internal Coordinates	33
2.2.3 Determining Forces on CVs Using Redundant Internal Coordinate Transformation	35
2.2.4 Linearized Force Matching in RP-FM-CV Using Spline Functions	36
2.2.5 Force Modification for Iterative RP-FM-CV	37
2.3 Critical Test: Menshutkin Reaction $\text{NH}_3 + \text{CH}_3\text{Cl}$	38
2.4 Computational Details	41
2.4.1 Description of the Solute Model	42
2.4.2 Potential Energy Calculations	42
2.4.3 Definition of Collective Variables	42
2.4.4 Boundary Conditions and Treatment of Long-Range Electrostatics	43
2.4.5 Restraints and MD Simulations	43
2.4.6 String MFEP Simulations	44

2.4.7	Force Matching in Redundant Internal Coordinates	44
2.5	Results and Discussion	45
2.5.1	Free Energy Profiles	47
2.5.2	Force Correlations	49
2.5.3	Internal Force Corrections on CVs along MFEP	49
2.5.4	Tests of Different Sets of Redundant Internal Coordinates	50
2.5.5	Tests of Number of Configurations Included in FM	51
2.5.6	Tests of Basis-Set Convergence	52
2.5.7	RP-FM-CV Produces AI/MM-Quality Free Energy Paths	53
2.5.8	Convergence of the Overall Procedure	55
2.5.9	Radial Distribution Functions	56
2.6	Outlook	58
2.7	Concluding Remarks	61
2.8	Appendix	71
3	ACCURATE FREE ENERGY PROFILES IN CHEMICAL REACTIONS: A QM/MM STUDY OF THE ROLE OF PAIRWISE REPULSIVE CORRECTING POTEN- TIALS IN FORCE MATCHING	78
3.1	Introduction	78
3.2	Methods	85
3.3	Computational Details	87
3.4	Results and Discussion	88
3.4.1	Optimizing Radial Cutoff Distances for Pairwise Potential and Force Correction	88
3.4.2	Optimization of pairwise force matching using the Micro-genetic Al- gorithm (Micro-GA) and pairwise RP-FM in collective variables (CVs)	95
3.4.3	Force Matched Free Energy Correction for Assessing RP-FM Models	98
3.4.4	Coordinate Dependence and Benefits of Force Matched Free Energy Corrections in RP-FM-CV	101

3.4.5	Comparison of pairwise RP–FM and RP–FM–CV for reaction path corrections	103
3.5	Concluding Remarks	106
4	DOUBLY POLARIZED QM/MM WITH MACHINE LEARNING CHAPERONE POLARIZABILITY	114
4.1	Introduction	115
4.2	Methods	120
4.3	Computational Details	124
4.4	Results and Discussion	126
4.4.1	Molecular Polarizability	126
4.4.2	Free Energy Profile	128
4.4.3	Comparison of MFEPs and Shift of Transition State	132
4.4.4	Atomic Chaperone Polarizability	134
4.4.5	Local Environment and RDF	137
4.4.6	Convergence of Atomic Polarization Energy with Solvent Inclusion . .	138
4.5	Outlook	140
4.6	Concluding Remarks	144
4.7	Appendix	145
5	REACTION PATH–FORCE MATCHING IN COLLECTIVE VARIABLES AND DOUBLY POLARIZED QM/MM WITH MACHINE LEARNING CHAPERONE POLARIZABILITY	148
5.1	Introduction	148
5.2	Methods	151
5.3	Computational Details	152
5.4	Results	153
5.4.1	Molecular Polarizability	153
5.4.2	Atomic Chaperone Polarizability	155
5.4.3	Internal Force Correction in CVs	158
5.4.4	Comparison of MFEPs and Free Energy Profiles	159

5.4.5	Local Environment and RDF	162
5.5	Concluding Remarks	163
6	CONCLUDING REMARKS	166
	REFERENCES	167
A	PROOF OF COPYRIGHT PERMISSION(S)	182
B	SUPPORTING INFORMATION	184
B.1	Reaction Path-Force Matching in Collective Variables: Determining Ab Initio QM/MM Free Energy Profiles by Fitting Mean Force	184
B.2	Accurate Free Energy Profiles in Chemical Reactions: A QM/MM Study of the Role of Pairwise Repulsive Correcting Potentials in Force Matching	204
B.3	Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability	211
	PUBLICATION(S)	248
	ORAL PRESENTATION(S)	249
	POSTER PRESENTATION(S)	250
	VITA	252

LIST OF TABLES

2.1	Free energy barriers (ΔG^\ddagger), reaction free energies (ΔG_r), and transition-state geometries (Å) for the Menshutkin reaction between NH_3 and CH_3Cl in water. .	46
2.2	Computed free energy barriers (ΔG^\ddagger), reaction free energies (ΔG_r), and transition state geometries for the Menshutkin reaction between NH_3 and CH_3Cl in water over five cycles of RP FM-CV simulations at the MP2:AM1/MM level.	55
3.1	Reaction Barrier/Free Energy of $\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{CH}_3\text{NH}_3^+ + \text{Cl}$ in the Solution Phase, Geometrical Parameters/Force Correlation (%Fcorr) in CVs and Radial Cutoff Distances (r_c)	93
3.2	Cross-validation root mean-square errors (kcal/mol/Å)	94
3.3	Reaction Barrier/Free Energy of $\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{CH}_3\text{NH}_3^+ + \text{Cl}$ in the Solution Phase, for 1 st Iteration (itr=1)/Zeroth-Order (0 th -order) Predictions and Statistical Errors Estimated from Free Energy Corrections	100
3.4	Reaction Barrier of $\text{Cl} + \text{CH}_3\text{Cl} \rightarrow \text{CH}_3\text{Cl} + \text{Cl}$ in the Solution Phase, Geometrical Parameters/Force Correlation (%Fcorr) in CVs and Radial cutoff Distance (r_c)	105
4.1	Free energy barrier (ΔG^\ddagger) and reaction free energy (ΔG) of the Menshutkin reaction in solution.	130

LIST OF FIGURES

2.1	A schematic representation of the RP-FM-CV method.	37
2.2	Menshutkin reaction between ammonia and methyl chloride ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$).	38
2.3	Free energy profiles along the string MFEPs (with $\alpha = 0$ being reactant and 1 product) from the AM1/MM (dashed red) and RP-FM-CV simulations of the Menshutkin reaction in aqueous solution. Results from the RP-FM-CV AI:AM1/MM simulations were obtained by matching AM1/MM forces on the CVs to various target AI/MM levels using the default 6-31+G(d,p) basis set: B3LYP:AM1/MM (dotted black), BH&HLYP:AM1/MM (green with circles), and MP2:AM1/MM (solid blue).	47
2.4	Internal force correlations between the base AM1/MM and target AI/MM methods [at the B3LYP/MM and MP2/MM levels both using the 6-31+G(d,p) basis set]: before (red squares) and after (blue circles) applying the RP-FM-CV internal force corrections; the corresponding trend lines are shown as dashed and solid lines. Internal forces on the two bond CVs were computed based on 300 configurations sampled along the condensed-phase MFEP from the AM1/MM string simulations. The average internal force deviations (ΔF ; in kcal/mol/Å ²) between the base and target levels, before (red) and after (blue) force matching, are also shown for comparison.	62
2.5	Internal force corrections on CVs for the Menshutkin reaction in solution with AM1/MM being the base level and the target forces obtained at the B3LYP/MM and MP2/MM levels using the 6-31+G(d,p) basis set. Both the actual internal force differences (based on 300 solution-phase configurations along the string MFEP) between the base and target levels (labeled Target; red triangles) and the spline based force corrections (labeled Fit; blue lines) resulting from FM are shown for each CV (i.e., the N-C and C-Cl bonds).	63
2.6	Various redundant internal coordinate schemes tested for RP-FM-CV simulations of the Menshutkin reaction in solution. The two distance-based CVs for bond forming and breaking, i.e., N-C (1-5) and C-Cl (5-9) (shown in red in the Bonds section), are used consistently in both the string MFEP simulations and the FM calculations.	64
2.7	Free energy profiles for the Menshutkin reaction in solution obtained from the RP-FM-CV simulations at the MP2:AM1/MM level by using different redundant internal coordinate sets; the AM1/MM profile (dashed red) is shown for comparison. The RP-FM-CV results obtained using 28 (Int28; dotted black), 31 (Int31; solid green with circles), and 34 (Int34; solid blue) internal coordinates are shown (see Figure 2.6 for the definitions of the three schemes).	65

2.8	Free energy profiles for the Menshutkin reaction in solution obtained from the RP-FM-CV simulations at the MP2:AM1/MM level using different sample sizes for FM; the AM1/MM profile (dashed red) is shown for comparison. The RP-FM-CV results where FM in CVs were conducted using 300 (FM-300; dotted black), 1500 (FM-1500; green with circles), 3000 (FM-3000; solid blue), and 15000 (FM15000; pink with crosses) configurations are shown.	66
2.9	Free energy profiles for the Menshutkin reaction in solution obtained from the RP-FM-CV simulations by matching the AM1/MM forces on CVs to those determined at the MP2/MM level using various basis sets, including 6-31G(d), 6-31+G(d,p), 6-311++G(d,p), and 6-311+G(2df,2p).	67
2.10	Minimum free energy paths (MFEP) for the Menshutkin reaction in solution for the free energy profiles shown in Figure 2.3. The MFEPs from the RP-FM-CV AI:AM1/MM simulations were obtained by matching AM1/MM forces on CVs to various target AI/MM levels using the 6-31+G(d,p) basis set: B3LYP:AM1/MM (dotted black), BH&HLYP:AM1/MM (green with circles), and MP2:AM1/MM (solid blue), compared with AM1/MM (dashed red). The transition states (TS) located on the MFEPs are also marked: B3LYP:AM1/MM (square), BH&HLYP:AM1/MM (triangle), MP2:AM1/MM (diamond), and AM1/MM (circle).	68
2.11	Free energy profiles for the Menshutkin reaction in solution obtained from the RP-FM-CV simulations at the MP2:AM1/MM level using 6-31+G(d,p) basis set over five consecutive RP and FM cycles: the 1st (dotted black), 2nd (green with circles), 3rd (solid blue), 4th (pink with crosses), and 5th cycle (light blue with triangles), compared with AM1/MM (dashed red).	69
2.12	Solute-solvent radial distribution functions (RDFs) obtained from the RP-FM-CV simulations at the MP2:AM1/MM level using the 6-31+G(d,p) basis set, compared with the AM1/MM results. The RDFs for the solute (heavy atoms) and solvent (water oxygens: O_w) (i.e., N- O_w , C- O_w , and Cl- O_w) determined using an average of 3,600 configurations in each regions are shown: reactant (R; dotted red), transition state (TS; solid green), and product (P; dashed blue).	70
3.1	Potential of mean force for the Menshutkin reaction in aqueous solution for AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes.	81
3.2	Cross-validation root mean-square errors of atom type for AM1/MM (red), and B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes.	89
3.3	Cross-validation root mean-square errors along the reaction path for AM1/MM (red), and B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes.	90

3.4	Force correlation in CVs before (red) and after (blue) B3LYP force matched corrections for: (a) RP-FM-CV and (b) uniform radial cutoff schemes.	91
3.5	Distance-based force corrections for B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes in collective variables for: (a) N-C and (b) C-Cl.	92
3.6	Comparisons of correcting potential energy functions from the B3LYP/MM force matched generic radial cutoff scheme for atom type pairs on: (a) collective variables and (b) spectator degrees of freedom, where sampled pair distances are represented in dotted lines (300 samples within 0.252 Å windows).	96
3.7	Potential of mean force, force correction and free energy correction for the Menshutkin reaction in aqueous solution for B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes. (a) Solid lines are from 1 iteration of force matching and dashed lines are zeroth-order predictions for FEPs from free energy corrections. Distance-based force corrections for (b) N-C and (c) C-Cl. Total free energy correction (d) and decomposed free energy corrections in collective variables for (e) N-C and (f) C-Cl.	109
3.8	Potential of mean force for force matched collective variables for the Menshutkin reaction in aqueous solution for AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes.	110
3.9	Error bars on Potential of mean force calculations for the Menshutkin reaction in aqueous solution for: (a) B3LYP/MM force matched RP-FM-CV (black), and (2) B3LYP/MM force matched and micro-genetic algorithm in collective variables (blue) radial cutoff scheme.	110
3.10	Potential of mean force for the Finkelstein reaction in aqueous solution for AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes.	111
3.11	Figure 3.11. Potential of mean force, force correction and free energy correction for the Finkelstein reaction in aqueous solution for B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes. (a) Solid lines are from 1 iteration of force matching and dashed lines are zeroth-order predictions for FEPs from free energy corrections. Distance-based force corrections for (b) C-Cl and (c) C-Cl. Total free energy correction (d) and decomposed free energy corrections in collective variables for (e) C-Cl and (f) C-Cl.	112

3.12	Minimum free energy path and transition state geometries (circles) for the Menshutkin reaction based on AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes.	113
4.1	Solution-phase molecular polarizability as a function of the reaction coordinate for the Menshutkin reaction: AM1 (squares), B3LYP/aug-cc-pVTZ (circles), and their difference (triangles). The means (solid curves) and standard deviations (vertical bars) are computed based on samples within each string images (see SI.5 of B.3 for the tabulated statistical distributions).	127
4.2	Regressions of the molecular polarizabilities from AM1 (square with a dashed line) and from the chaperone-corrected AM1 (circles with a solid line) against the molecular polarizability from B3LYP/aug-cc-pVTZ; the corresponding root-mean-square-errors (RMSEs) in polarizability compared with the AI target values are also shown.	129
4.3	Free energy profiles as a function of the reaction coordinate for the Menshutkin reaction: AM1/MM (dashed line) and dp-QM/MM with polarizabilities corrected to B3LYP/aug-cc-pVTZ (solid line). The error bars relative to the free energy in the reactant state ($\alpha = 0$) along the string MFEP are estimated using a procedure developed by Zhu and Hummer (Ref. [22]), slightly modified for non-uniform collective-variable grids (see SI.1 of B.3 for details).	131
4.4	The minimum free energy path (MFEPs) as a function of the collective variables, i.e., the C-Cl and N-C bonds: AM1/MM (dashed line) and the dp-AM1/MM with chaperone polarizabilities corrected to the B3LYP/aug-cc-pVTZ level (solid line). The locations of free energy transition states are also marked: AM1/MM (open square) and dp-AM1/MM (open circle).	133
4.5	Atomic chaperone polarizabilities as a function of reaction coordinate for each solute atom in the Menshutkin reaction.	135
4.6	Convergence of atomic polarizability on the chlorine atom with respect to basis sets and AI methods.	136
4.7	Radial distribution functions between the chlorine atom and water oxygens (O_w) in the reactant (R), transition-state (TS), and product (P) region of the Menshutkin reaction: AM1/MM (dashed line) and dp-AM1/MM with polarizabilities corrected to the B3LYP/aug-cc-pVTZ level (solid line).	138
4.8	Decomposition of polarization energy to each water molecule around the chlorine atom: AM1/MM (squares), chaperone polarization energy (triangles), and B3LYP/aug cc-pVTZ/MM (circles).	140

5.1	The solution-phase molecular polarizabilities as a function of the reaction coordinate for the Menshutkin reaction are shown for AM1 (squares), B3LYP/MM (circles), and their difference (triangles). The means (solid curves) and standard deviations (vertical bars) are calculated based on the samples within each string image.	154
5.2	The regressions of molecular polarizabilities from AM1 (square with a dashed line) and from the chaperone-corrected AM1 ($\text{AM1} + \Delta\alpha^{\text{C}}$); circles with a solid line) against those from B3LYP/MM are shown, along with the corresponding root-mean-square errors (RMSEs) relative to the B3LYP reference values. . . .	156
5.3	The polarizabilities of atomic chaperones as a function of the reaction coordinate for each solute atom in the Menshutkin reaction.	157
5.4	Internal force corrections for the Menshutkin reaction in solution on CVs (i.e., the NC, C-Cl and NCl bonds) using the AM1/MM level as the base level and the B3LYP/MM level as the target level with the 6-31+G(d,p) basis set. . . .	158
5.5	The internal force correlations between the base AM1/MM and B3LYP/MM using the 6-31+G(d,p) basis set are shown before (red squares) and after (blue circles). The internal force corrections of RP-FM-CV are applied to polarization corrections; the corresponding trend lines are shown as dashed and solid lines. The internal force deviations (ΔF ; in kcal/mol/Å ²) between the base and target levels, before (red) and after (blue) force matching, are shown for comparison. .	160
5.6	The minimum free energy paths (MFEPs) as a function of collective variables, i.e., C-Cl and NC bonds, are shown by the AM1/MM (dashed line) and dp-AM1/MM with the chaperone polarizabilities fitted to the B3LYP/MM level (solid line). The locations of the free energy transition state are also marked: AM1/MM (open square) and dp-AM1/MM (open circle).	161
5.7	The Menshutkin reaction's free energy profiles as a function of the reaction coordinate: AM1/MM (dashed line) and Combined-AM1/MM with polarizabilities corrected to B3LYP/MM (solid line). The error bars associated with the free energy in the reactant state ($\alpha = 0$) along the string MFEP are estimated. . . .	162
5.8	The radial distribution functions between the chlorine atom and water oxygens (O_w) in the reactant (R), transition-state (TS), and product (P) region of the Menshutkin reaction are shown for AM1/MM (dashed line) and Combined-AM1/MM with polarizabilities corrected to the B3LYP/MM level (solid line). .	165
A.1	Kim, B., Snyder, R., Nagaraju, M., et al. Reaction Path-Force Matching in Collective Variables: Determining Ab Initio QM/MM Free Energy Profiles by Fitting Mean Force. <i>Journal of Chemical Theory and Computation</i> , 2021, 17(8), 4729-4737. Reprinted with permission from <i>Journal of Chemical Theory and Computation</i> . Copyright 2021 American Chemical Society.	182

A.2	Kim, B., Shao, Y., Pu, J. Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability. <i>Journal of Chemical Theory and Computation</i> , 2021, 17(12), 7816-7825. Reprinted with permission from <i>Journal of Chemical Theory and Computation</i> . Copyright 2021 American Chemical Society.	183
-----	--	-----

LIST OF SCHEMES

3.1	Schematic Representation of the Menshutkin Reaction ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$)	80
3.2	Schematic Representation of the Finkelstein Reaction ($\text{Cl}^- + \text{CH}_3\text{Cl} \rightarrow \text{ClCH}_3 + \text{Cl}^-$)	104
4.1	Schematic Representation of the Menshutkin Reaction from the Charge-Neutral Reactant State to the Charge-Separated Product State ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$)	119
4.2	Topology of Artificial Neural Network Containing Two CV inputs, Ten Hidden Neurons, and Nine Atomic Chaperone Polarizabilities in Output Layer	122

ABSTRACT

In this thesis, we present novel computational methods and frameworks to address the challenges associated with the determination of free energy profiles for condensed-phase chemical reactions using combined quantum mechanical and molecular mechanical (QM/MM) approaches. We focus on overcoming issues related to force matching, molecular polarizability, and convergence of free energy profiles. First, we introduce a method called Reaction Path-Force Matching in Collective Variables (RP-FM-CV) that efficiently carries out ab initio QM/MM free energy simulations through mean force fitting. This method provides accurate and robust simulations of solution-phase chemical reactions by significantly reducing deviations on the collective variables forces, thereby bringing simulated free energy profiles closer to experimental and benchmark AI/MM results. Second, we explore the role of pairwise repulsive correcting potentials in generating converged free energy profiles for chemical reactions using QM/MM simulations. We develop a free energy correcting model that sheds light on the behavior of repulsive pairwise potentials with large force deviations in collective variables. Our findings contribute to a deeper understanding of force matching models, paving the way for more accurate predictions of free energy profiles in chemical reactions. Next, we address the underpolarization problem in semiempirical (SE) molecular orbital methods by introducing a hybrid framework called doubly polarized QM/MM (dp-QM/MM). This framework improves the response property of SE/MM methods through high-level molecular polarizability fitting using machine learning (ML)-derived corrective polarizabilities, referred to as chaperone polarizabilities. We demonstrate the effectiveness of the dp-QM/MM method in simulating the Menshutkin reaction in water, showing that ML chaperones significantly reduce the error in solute molecular polarizability, bringing simulated free energy profiles closer to experimental results. In summary, this thesis presents a series of novel methods and frameworks that improve the accuracy and reliability of free energy profile estimations in condensed-phase chemical reactions using QM/MM simulations. By addressing the challenges of force matching, molecular polarizability, and convergence, these advancements have the potential to impact various fields, including computational chemistry, materials science, and drug design.

1. INTRODUCTION

The field of computational chemistry has grown rapidly in recent years, allowing scientists to use powerful computational tools such as combined quantum mechanical and molecular mechanical (QM/MM) simulations to study complex chemical systems. These simulations combine QM calculations, which accurately describe the electronic structures of molecules, with MM calculations, which describe the motion of atoms and molecules using classical potentials. In this approach, the QM and MM regions are chosen based on their chemical reactivity. The QM region includes the molecules and molecular fragments that are chemically active, whereas the MM region includes the rest of the system [1]. This combination allows for the prediction of the behavior of condensed-phase chemical processes, and has been used to study a wide range of systems, including reactions in solution and in enzymes [2]. Despite many successes, QM/MM simulations have their own limitations. One of the main limitations is their cost for accurately simulating chemical reactions in dynamic environments, for which entropic contributions to free energy can be important [3]. Such systems are ubiquitous in many areas of chemistry, including biochemistry, organic chemistry, and materials science [4]. Quantitatively reliable free energy simulation of these highly dynamical chemical processes requires adequate configurational sampling involving a large number of potential energy and force calculations, which can be computationally very demanding [5]–[8]. Another limitation of QM/MM simulations is their accuracy in describing the interactions between the QM and MM subsystems. These interactions can be complex and difficult to model accurately, which can lead to errors in the predictions made by the simulation [9]. This PhD thesis, titled “QM/MM Applications and Corrections for Chemical Reactions”, focuses on the development of QM/MM methods and their application to chemical reactions. The research presented in this thesis aims to address these limitations by developing new methods for QM/MM free energy simulations to improve their accuracy and efficiency. The thesis also presents a range of studies that demonstrate the applicability of these new methods to a variety of chemical systems. Through this research, the thesis is expected to contribute to the ongoing developments of QM/MM simulations and to promote their use in the study of chemical reactions in complex and dynamic environments.

Hydrolysis of adenosine triphosphate (ATP) is a fundamental biochemical process that provides energy for numerous cellular functions [10]. This research is motivated to develop computational methods for better understanding of ATP hydrolysis in the ATP-binding cassette (ABC) transporter Hemolysin B (HlyB), a crucial biological component in bacteria responsible for toxin secretion [11]. Understanding the common mechanisms in the ABC protein family is also of great importance due to their connections to diseases and medical conditions such as cystic fibrosis and multidrug resistance [12]. In HlyB, ATP hydrolysis is catalyzed by its nucleotide binding domains (NBD) located in the cytosolic portion of the protein. ATP molecules bind to the NBDs, undergo hydrolysis, and power the translocation of bacterial toxins out of the cell [11]. Understanding the free energy requirement for ATP hydrolysis in HlyB would provide crucial insights into the reaction mechanism, enzyme catalysis, and potential biochemical designs to target this process. To accurately model enzymatic ATP hydrolysis, QM calculations are required for a detailed description of the electronic structures of the reaction center. However, QM calculations can be computationally expensive, especially for large systems such as the solvated HlyB system [11]. To overcome this challenge, a QM/MM approach [1], [13]–[16] is employed, localizing the QM calculations to a reactive region while treating the rest of the system using classical mechanics [11]. To determine the free energy pathway and free energy profile of ATP hydrolysis, enhanced sampling techniques are utilized in molecular dynamics (MD) simulations [17]–[21]. In this study, the string method in collective variables [22] combined with the semiempirical QM/MM method is used to achieve efficient free energy pathway simulations [23]. This approach allows for modeling of large enzyme systems and helps obtain statistically reliable free energy profiles. Despite the advantages of the QM/MM approach, configurational sampling of the reactive subsystem at highly accurate ab initio QM levels remains computationally challenging. To address this issue, faster, lower-cost semiempirical QM methods such as AM1 [24] or PM3 [25] are often employed. However, these methods can be less reliable due to intrinsic deficiencies [26], [27]. This research aims to develop innovative computational approaches that strike a balance between accuracy and efficiency [28], with the ultimate goal of obtaining reliable free energy computations to provide mechanistic insights for complex chemical and biochemical systems.

The focus of this thesis is on developing a dual-level strategy that corrects semiempirical QM/MM methods to achieve the accuracy of ab initio QM/MM methods. This dual-level correction strategy is inspired by the ideas of the QM/MM approach, which also combines the accuracy of QM methods and the efficiency of MM methods [1]. In particular, we examine two types of dual-level corrections: a mean force correction on collective variables [26] and a polarizability correction on quantum mechanical atoms [27]. The mean force correction considers the effect of solvent on the reaction coordinate, while the polarizability correction factors in the solute polarization by the solvent and its impact on the electronic structure of the system. Our work in this direction has led to the developments of two new computational approaches: reaction path force matching in collective variables [26] and doubly polarized QM/MM [27]. The hypothesis of this thesis is that these dual-level correction methods will enable accurate chemical simulations to provide key insights into the energetics and mechanisms of chemical reactions. The new methods are expected to overcome some of the major limitations of semiempirical QM/MM methods and provide a more accurate and efficient way to simulate chemical reactions in condensed phases. The results of this research will also offer a great potential to answer many long-standing chemical questions and could be used to design new drugs, catalysts, and materials. The methods developed here will further expand the toolbox of QM/MM simulations [16], [29].

The current state of computer technology, including the availability of faster hardware and specialized chips, presents a unique opportunity for advancing research in the field of chemical simulations [30]. These advancements have enabled researchers to perform simulations with increasing accuracy and efficiency, thereby providing an avenue to address long-standing questions in chemistry that were previously hindered by computational limitations [31]. The research described in this thesis aims to leverage emerging technology to gain a deeper understanding of chemical reactions through more accurate predictions. This research is expected to have far-reaching implications for various fields, including chemistry, biology, and medicine [4]. The combination of computational chemistry and state-of-the-art computing technology offers an exciting opportunity to advance our understanding of complex chemical phenomena with unprecedented simulation accuracy.

The research methodology used in this thesis involves machine learning [32]–[40], genetic algorithms [41], and coordinate transformation techniques [42]–[44] to achieve successful dual-level corrections that result in cost-effective predictions of free energy profiles for chemical reactions. The string method in collective variables [22] is combined with QM/MM methods to achieve enhanced sampling for the mapping of free energy surfaces for chemical reactions. This combination has been shown to be effective in studying chemical reactions in complex environments [26], [27]. Dual-level corrections are applied to these surfaces to further improve the accuracy of the predictions. The results of the research are analyzed using statistical methods and are validated against experimental as well as high-level computational benchmarks for a prototype chemical reaction known as the Menshutkin reaction [45]. The Menshutkin reaction is a well-established reaction that has been widely used as a model system for studying chemical reactions in various environments [46]–[54].

The research outlined above will be organized into six chapters in this thesis: Introduction (Chapter 1), Reaction Path–Force Matching in Collective Variables: Determining Ab Initio QM/MM Free Energy Profiles by Fitting Mean Force (Chapter 2), Accurate Free Energy Profiles in Chemical Reactions: A QM/MM Study of the Role of Pairwise Repulsive Correcting Potentials in Force Matching (Chapter 3), Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability (Chapter 4), Reaction Path–Force Matching in Collective Variables and Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability (Chapter 5), and Concluding Remarks (Chapter 6). Chapters 2 and 4 will discuss state-of-the-art methods used to study chemical reactions in dynamic environments. Chapter 3 will provide a detailed analysis of force matching methods and Chapter 5 will combine the approaches together to achieve a holistic QM/MM dual-level correction method.

2. REACTION PATH–FORCE MATCHING IN COLLECTIVE VARIABLES: DETERMINING AB INITIO QM/MM FREE ENERGY PROFILES BY FITTING MEAN FORCE

First-principles determination of free energy profiles for condensed-phase chemical reactions is hampered by the daunting costs associated with configurational sampling on ab initio quantum mechanical/molecular mechanical (AI/MM) potential energy surfaces. Here, we report a new method that enables efficient AI/MM free energy simulations through mean force fitting. In this method, a free energy path in collective variables (CVs) is first determined on an efficient reactive aiding potential. Based on the configurations sampled along the free energy path, correcting forces to reproduce the AI/MM forces on the CVs are determined through force matching. The AI/MM free energy profile is then predicted from simulations on the aiding potential in conjunction with the correcting forces. Such cycles of correction–prediction are repeated until convergence is established. As the instantaneous forces on the CVs sampled in equilibrium ensembles along the free energy path are fitted, this procedure faithfully restores the target free energy profile by reproducing the free energy mean forces. Due to its close connection with the reaction path–force matching (RP–FM) framework recently introduced by us, we designate the new method as RP–FM in collective variables (RP–FM–CV). We demonstrate the effectiveness of this method on a type-II solution-phase S_N2 reaction, $\text{NH}_3 + \text{CH}_3\text{Cl}$ (the Menshutkin reaction), simulated with an explicit water solvent. To obtain the AI/MM free energy profiles, we employed the semiempirical AM1/MM Hamiltonian as the base level for determining the string minimum free energy pathway, along which the free energy mean forces are fitted to various target AI/MM levels using the Hartree–Fock (HF) theory, density functional theory (DFT), and the second-order Møller–Plesset perturbation (MP2) theory as the AI method. The forces on the bond-breaking and bond-forming CVs at both the base and target levels are obtained by force transformation from Cartesian to redundant internal coordinates under the Wilson **B**-matrix formalism, where the linearized FM is facilitated by the use of spline functions. For the Menshutkin reaction tested, our FM treatment greatly reduces the deviations on the CV forces, originally in the range of 12–33 to ~ 2 kcal/mol/Å². Comparisons with the

experimental and benchmark AI/MM results, tests of the new method under a variety of simulation protocols, and analyses of the solute–solvent radial distribution functions suggest that RP–FM–CV can be used as an efficient, accurate, and robust method for simulating solution–phase chemical reactions.

2.1 Introduction

The holy grail of simulating condensed–phase chemical/biochemical reactions is to obtain reliable free energy profiles through sampling highly accurate potential energy surfaces (PES) described by first–principles quantum mechanical (QM) methods such as ab initio molecular orbital [55] (AI–MO) and density functional theory [56], [57] (DFT) methods, which are collectively referred to as the AI methods here for convenience. Even with the aid of the combined quantum mechanical and molecular mechanical (QM/MM) [1], [13]–[16] technique, such simulations will likely remain impractical in the near future because of the daunting computational demands associated with free energy sampling on the already costly AI/MM PESs [8]. Alternatively, semiempirical (SE) MO [23] methods are often used in QM/MM for more efficient PES calculations; although they make adequate free energy sampling more affordable, their accuracy and reliability may not always be guaranteed. How to systematically improve an SE/MM method to reach AI/MM–level quality is a long–standing challenge in the fields of computational chemistry and computational enzymology [3]. Recently, we introduced a new computational framework called reaction path–force matching [28] (RP–FM) to address this challenge. The central idea of RP–FM is to bridge the SE/MM and AI/MM levels in the context of QM/MM free energy simulations through force [28], which encodes all dynamical information of the system, by making use of the force matching (FM) technique [58]–[66]. The second key element in RP–FM is that we conduct FM along free energy reaction pathways [11], [22], [67], which is a natural choice for reactive FM. Because FM is performed between two electronic–structure–based QM/MM potentials, RP–FM enables cost–effective fitting of highly accurate reactive potentials for studying chemical reactions. As the molecular configurations for FM and for determining free energy profiles are always sampled at an efficient SE/MM level, direct sampling of the expensive AI/MM surface is

avoided. The RP-FM method therefore offers a promising tool for accurate and efficient free energy simulations of condensed-phase reactions.

In our recently published work [28], we demonstrated the idea of RP-FM in the framework of optimizing specific reaction parameters (SRPs) for SE methods. Based on a set of condensed-phase configurations sampled along the free energy reaction pathway using SE/MM simulations, the selected SE-SRPs are adjusted until atomic forces match with those from the high-level AI/MM calculations. Then, the free energy pathway and the corresponding free energy profile are updated with the FM-optimized SE-SRP/MM simulations. Because of the self-consistent nature of the problem, RP-FM is conducted iteratively until the free energy results converge.

Although conceptually elegant, the RP-FM method can be difficult to apply to complex systems due to the nonlinear optimization involved when parametrizing a QM potential. In the framework of FM optimization of SE-SRPs, the forces from the SE QM calculations are not linearly dependent on the electronic-structure parameters to be adjusted, which makes the situation quite different from using FM to fit MM force-field parameters. One way to enable nonlinear FM, as we demonstrated previously, is to fit forces using nonlinear optimization algorithms such as the genetic algorithm (GA) [28]. For simple reactions such as a proton-transfer reaction in the gas phase and in solution, the GA-based nonlinear FM strategy works reasonably well. For example, for the RP-FM simulations of the proton-transfer reaction between ammonia and ammonium in the gas phase, the force deviation between PM3 and Hartree-Fock (HF) is reduced remarkably from an average of 12 kcal/mol/Å² per force component to less than 1 kcal/mol/Å², which brings the PM3 barrier height to agree with the HF/3-21G benchmark results after a change of 10 kcal/mol [28]. For RP-FM of the same reaction in solution, although we observed a similar convergence of the PM3-SRP/MM free energy profile toward its AI/MM benchmark, the solution-phase FM, based on explicit QM/MM configurations, seems to be more challenging for a GA optimizer to handle; the average force deviation plateaus at 3.5 kcal/mol/Å², a value significantly higher than seen in the gas-phase FM.

When size and complexity of the reactive system increase, nonlinear SE-SRP optimization can become a practical bottleneck in Cartesian-based FM. With large numerical errors

in force fitting, RP-FM may be insufficient on its own but can be complemented by the weighted thermodynamics perturbation (wTP) approach [68]; the powerful combination of the two outperforms the use of either method individually in reproducing the AI/MM free energy profiles [69]. Without the help of free energy perturbation, the problem of nonlinear FM, however, can be alleviated by parametrizing a classical energy component in (or on top of) the SE potential; when forces from such a classical energy term display a linear dependence on its parameters, the associated FM morphs into a classical one. One example of this type of approach is the FM-optimized density functional-tight binding (FM-DFTB) method developed by Goldman and co-workers [70], [71], who used FM to optimize the pairwise repulsive potential terms in DFTB to account for the force differences between DFTB and the target AI level. Another exciting direction is to introduce machine learning (ML)-optimized corrections on energy [72], [73], forces [34], or both [32], [33], [74] for SE methods. Although FM serves as an important component in these developments either for optimizing potentials [28], [32], [33], [70], [71], [74] or for reproducing high-level molecular dynamics (MD) trajectories on selected internal degrees of freedom, [34] a direct link between FM and determining the target-level free energy profiles is lacking. To overcome this hurdle, it is highly desirable to build a rigorous connection between FM and free energy, ideally through a linearized force-only-based framework.

In the process of forging this missing link and establishing the conceptual framework we desired, we noticed that collective variables (CVs) and the associated forces play important roles in free energy simulations such as the minimum free energy path (MFEP) simulations using the string method [11], [22], [67]. In the context of RP-FM, we found that instead of fitting all the atomic forces, matching the AI/MM target forces on the CVs offers a theoretically elegant way to reproduce the AI/MM free energy profiles. Following a similar line of reasoning by Voth and co-workers, who pointed out that mapping all-atom potentials to coarse-grained potentials by FM rigorously reproduces the many-body potential of mean force (PMF) [75], we show here that fitting the CV forces along the MFEP reproduces the free energy mean force at the target level, the integration of which directly leads to the high-level PMF coarse-grained to the consistent CV degrees of freedom. Under this strategy, because usually only a few selected CVs are subject to FM, the high-dimensional

nonlinear optimization problem in a complex parameter space will be reduced to a much lower dimension. In this paper, we report our development in this direction, which results in a new method we designate as reaction path–force matching in collective variables (RP–FM–CV). As we will demonstrate below, formulation of RP–FM in the CV space leads to a smooth connection between the target–level free energy profiles and mean force fitting. Because we directly operate on force, no explicit modifications of the potential energy function are needed in RP–FM–CV.

The rest of the chapter is organized as follows. The related theory is presented in Section 2.2. The benchmark system for testing the method is described and reviewed in Section 2.3. Section 2.4 provides the computational details. Results and discussion are given in Section 2.5. The relations of this work to others and its future are discussed in Section 2.6. Concluding remarks are presented in Section 2.7.

2.2 Theory

2.2.1 RP–FM Is Equivalent to Fitting Free Energy Mean Force

Although serving as a convenient vehicle for optimizing SE–SRP/MM potentials [28], RP–FM, from a free energy perspective, is equivalent to fitting the many–body PMF at the target AI/MM level. Such a free–energy–based understanding of the method can be shown by starting from the familiar expression of mean force $\langle \mathbf{F} \rangle$ of free energy on a reaction coordinate (RC) ξ , represented by a set of n collective variables, i.e., $\xi = (\xi_1, \dots, \xi_n)$

$$\langle \mathbf{F} \rangle_{\xi=\xi^*} = \frac{\int dx_1 \dots dx_N dp_1 \dots dp_N \delta(\xi_1 - \xi_1^*) \dots \delta(\xi_n - \xi_n^*) \exp\left(-\frac{H}{k_B T}\right) \mathbf{F}[\mathbf{f}(x_1, \dots, x_N)]}{\int dx_1 \dots dx_N dp_1 \dots dp_N \delta(\xi_1 - \xi_1^*) \dots \delta(\xi_n - \xi_n^*) \exp\left(-\frac{H}{k_B T}\right)} \quad (2.1)$$

where x_i denotes the i^{th} Cartesian coordinate out of N degrees of freedom, p_i is the conjugate momentum, H is the Hamiltonian, k_B is the Boltzmann constant, T is the temperature, and δ is the Dirac delta function; \mathbf{F} represents the instantaneous force on the reaction coor-

dinate and can be obtained from transformation of the Cartesian atomic force $f(x_1, \dots, x_N)$. Cast into the context of RP-FM, eq (2.1) indicates that matching the SE/MM atomic forces to the corresponding AI/MM atomic forces, both in Cartesian coordinates, would indirectly reproduce the free energy mean force $\langle \mathbf{F} \rangle$ at the target AI/MM level.

Here, we demonstrate an alternative idea, where instead of fitting all of the AI/MM Cartesian atomic forces in eq (2.1), we will conduct FM directly on the reaction coordinate ξ . Because the instantaneous AI/MM forces along the reaction coordinate are reproduced over an ensemble of configurations sampled on an efficient SE/MM potential, the resulting treatment is equivalent to directly fitting the free energy mean forces, the integration of which over the reaction coordinate would faithfully restore the AI/MM free energy profile (interchangeably referred to as PMF in this work, only for convenience of discussion when the distinction between them is small [76]).

2.2.2 RP-FM-CV Fits Mean Force on Collective Variables in Internal Coordinates

Free energy profiles for complex chemical/biochemical reactions can be conveniently obtained using the string method [67] through the determination of mean forces on multidimensional collective variables (CVs) along the minimum free energy pathway (MFEP) [11], [22]. Therefore, we choose to formulate FM of the instantaneous forces \mathbf{F} in the CV space using the ansatz in eq (2.1). After the change of variables, we write the free energy mean force in terms of a set of generalized coordinates $(\mathbf{q}_1, \mathbf{q}_s)$.

$$\langle \mathbf{F} \rangle_{\xi=\xi^*} = \frac{\int d\mathbf{q}_s d\mathbf{p}_q d\mathbf{p}_\xi \exp\left(-\frac{H}{k_B T}\right) \mathbf{F}(\mathbf{q}_1, \mathbf{q}_s)}{\int d\mathbf{q}_s d\mathbf{p}_q d\mathbf{p}_\xi \exp\left(-\frac{H}{k_B T}\right)} \quad (2.2)$$

where $\mathbf{q}_1 \equiv \xi$ now represents the reaction coordinate expressed in a set of CVs, which is also used consistently for defining the MFEP, and \mathbf{q}_s denotes its complementary set for completing the generalized coordinate system [77], [78]. The instantaneous forces \mathbf{F} on the

CVs for evaluating the free energy force (also known as the thermodynamics force) can be further expressed as [77]–[79]

$$\mathbf{F}(\mathbf{q}_1, \mathbf{q}_s) = \frac{1}{\beta} \frac{\partial \ln |\mathbf{J}(\mathbf{q}_1, \mathbf{q}_s)|}{\partial \mathbf{q}_1} - \frac{\partial U(\mathbf{q}_1, \mathbf{q}_s)}{\partial \mathbf{q}_1} \quad (2.3)$$

where \mathbf{J} is the Jacobian matrix that transforms the Cartesian to the generalized coordinate system, β is $(k_B T)^{-1}$, and $\frac{\partial U(\mathbf{q}_1, \mathbf{q}_s)}{\partial \mathbf{q}_1}$ denotes the corresponding mechanical force, which is evaluated as the partial derivative of the potential energy U with respect to the CVs. Although the partial derivative form of eq (2.3) appears to suggest that the calculation of the instantaneous forces on the CVs depends not only on the definition of the reaction coordinate \mathbf{q}_1 but also on the choice of the complementary generalized coordinate \mathbf{q}_s , the thermodynamics force $\langle \mathbf{F} \rangle$ integrated from eq (2.2, however, does not depend on the specific choice of the complementary generalized coordinate. Indeed, Ruiz–Montero et al. pointed out that as long as one can identify a set of complementary generalized coordinates that makes the union of \mathbf{q}_s and \mathbf{q}_1 an orthogonal set, the explicit dependence of the partial derivative term on \mathbf{q}_s can be removed and eq (2.3) can be conveniently written as [78]

$$\mathbf{F}(\mathbf{q}_1, \mathbf{q}_s) = -\frac{1}{\beta} \frac{(\nabla |\nabla \mathbf{q}_1|) \cdot \nabla \mathbf{q}_1}{|\nabla \mathbf{q}_1|^3} - \frac{\nabla U \cdot \nabla \mathbf{q}_1}{|\nabla \mathbf{q}_1|^2} \quad (2.4)$$

where ∇ denotes the first derivative operator with respect to the Cartesian coordinates. Note that for a one-dimensional case that uses a single bond as the CV, the mechanical force term in eq (2.4) leads to an expression equivalent to a simple projection of Cartesian forces along the bond vector.

As den Otter and Briels later pointed out [79], the existence of such orthogonal complementary generalized coordinate sets may not always be guaranteed, especially when the CVs are global in nature. We realized that for most QM/MM applications in reaction mechanism studies, it is usually sufficient to use internal coordinates such as bond distances, bond angles, and dihedral angles, which are all local variables, to describe the reaction progress. For these local CVs in internal coordinates, construction of the complementary generalized coordinates

that are orthogonal to the selected CVs is attainable: one can use the Cartesian coordinates of the atoms that are not involved in the CVs; these non-CV Cartesian coordinates by definition are orthogonal to the internal coordinates in the CVs. For a solute-solvent system, such a treatment readily justifies the omission of solvent coordinates from \mathbf{q}_s when evaluating the force on \mathbf{q}_1 , if the reaction coordinate only involves the solute atoms. For the solutes, since the CVs can couple to other solute degrees of freedom through shared atoms and chemical bonds, \mathbf{q}_s needs to be constructed explicitly with its various choices tested systematically for convergence. Our demonstration of the RP-FM-CV method in this paper will focus on the CVs (as well as the complementary generalized coordinate) that are defined by local internal coordinates. With this clarification of the coordinate system, now we identify eq (2.3 as the key equation for conducting FM in CVs.

In the context of FM in QM/MM, the Jacobian force term \mathbf{J} in eq (2.3, which arises from coordinate transformation (zero for a rectilinear transformation but nonzero for a curvilinear transformation), is purely geometrical, regardless of whether an SE/MM or AI/MM method is used for the potential energy calculations; therefore, \mathbf{J} does not contribute to the force differences between the two QM/MM levels involved in FM. By contrast, the $-\frac{\partial U}{\partial \mathbf{q}_1}$ term on the right-hand side of eq (2.3, which gives the mechanical forces on the CV internal coordinates, is PES dependent and will be subject to force matching.

At this point, we reiterate the rationale of conducting RP-FM in CVs, i.e., if one reproduces the potential-energy-dependent part of the instantaneous internal forces on the CVs in eq (2.3 at the AI/MM level, the ensemble-averaged free energy mean forces in eq (2.2 would be reproduced at the target level. Integration of the resulting AI/MM-quality mean force along the string MFEP expressed in the same set of CVs would faithfully restore the target free energy profile. Next, we present the practical procedure of obtaining the internal forces on the CVs.

2.2.3 Determining Forces on CVs Using Redundant Internal Coordinate Transformation

The internal forces on the CVs can be conveniently obtained through force transformation from Cartesian to internal coordinates using the Wilson \mathbf{B} -matrix formalism [80].

Because the number of possible internal coordinates that can be constructed for a polyatomic molecule quickly exceeds the degrees of freedom in the system, we choose to use redundant internal coordinates. To remove the linear dependency in the redundant internal coordinate set and transform the Cartesian forces to the internal forces on the CVs, we adopted a procedure developed by Pulay and co-workers [43], where the redundancy of the coordinate system is identified and removed when forming the generalized inverse of the \mathbf{G} -matrix (see Appendix B for details). Note that both the nonredundant and redundant forms of Cartesian-to-internal coordinate transformation have been widely used in geometry optimization [43], [44] and in generalized vibrational analysis along the reaction path [81], [82].

Unlike classical FM based on pairwise classical potentials, where force matching can be conveniently cast into a linearized least-square problem, our previous implementation of RP-FM between two QM/MM potentials employed a genetic algorithm to handle the nonlinear optimization of SE-SRPs for fitting atomic forces in Cartesian coordinates. To circumvent the need for nonlinear optimization in RP-FM-CV, we introduce a set of empirical force correction terms to directly fit the target forces on the CVs, which is described below. Alternatively, we have also formulated the RP-FM-CV method in a machine learning (ML) framework, which will be reported in a companion paper.

2.2.4 Linearized Force Matching in RP-FM-CV Using Spline Functions

After the internal forces on the CVs are obtained at both the base (SE/MM) and target (AI/MM) levels for configurations sampled along the free energy pathway, we conduct FM through a force correction term for each CV to minimize the force differences between the two levels. Specifically, we fit the internal force corrections using a set of grid-based cubic spline functions (see Appendix A of 2.8 for implementation details), which is a numerical treatment originally introduced by Voth and co-workers [59] for FM optimization of classical force fields. As shown by the original developers [59], FM under this framework can be cast into solving an overdetermined linear equation system. With this linearization treatment, our FM between the SE/MM and AI/MM levels on each CV is converted to a least-square

problem and then solved by QR decomposition [83] or singular value decomposition (SVD) [83], in a way similar to FM optimization of classical force fields [59].

2.2.5 Force Modification for Iterative RP–FM–CV

To obtain the updated free energy results through MD sampling, the spline-based correcting forces on the CVs resulting from FM are incorporated in the SE/MM forces by distributing the internal force correction on each CV to the associated Cartesian atomic force components using the chain rule. Note that the same Cartesian force modifications can also be obtained using the backward transformation of eq A22 in Appendix B of 2.8, which transforms the force corrections on the CVs from the redundant internal coordinates back to Cartesian coordinates [84]. The two procedures are equivalent, where the chain-rule procedure is used in the implementation for simplicity.

After incorporating the FM corrections, the modified SE/MM atomic forces are used to propagate the SE/MM MD trajectories for free energy sampling, with which the updated free energy pathways and free energy profiles are determined. The cycle of free energy path sampling and FM is repeated iteratively until convergence is established. A schematic representation of the RP–FM–CV procedure is shown in Figure 2.1.

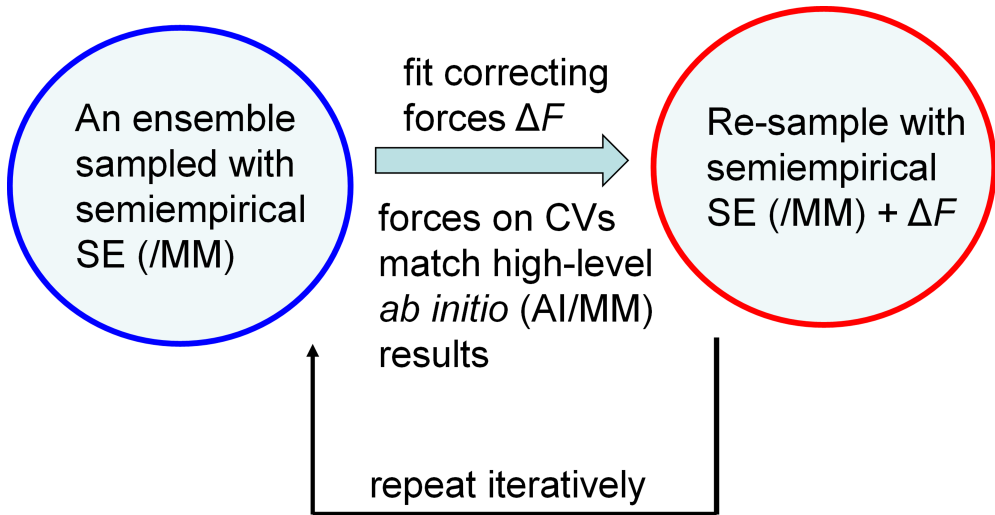


Figure 2.1. A schematic representation of the RP–FM–CV method.

2.3 Critical Test: Menshutkin Reaction $\text{NH}_3 + \text{CH}_3\text{Cl}$

To demonstrate its effectiveness, we applied the RP-FM-CV method to a type-II $\text{S}_{\text{N}}2$ reaction between ammonia and methyl chloride ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$) in an aqueous solution (Figure 2.2).

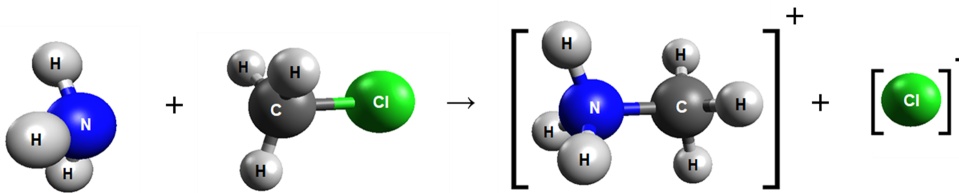


Figure 2.2. Menshutkin reaction between ammonia and methyl chloride ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$).

Following the literature convention, we refer to this reaction as the Menshutkin [45] reaction. Understanding kinetics and thermodynamics for the Menshutkin reaction through free energy simulation is challenging in that it requires quantitatively accurate descriptions of both the PES for the solute and the change in solvation effects when the solute system evolves from the charge-neutral reactant state to the charge-separated transition and product states; for a statistically reliable free energy description of this reaction, both these components need to be properly sampled over explicit solute/solvent configurations. Due to its fundamental importance, the Menshutkin reaction has been used as a workhorse for developing a host of computational methods for accurate and efficient treatments of PES and solvation over the past three decades [33], [34], [46]–[50], [53], [54], [69], [85]–[95].

Combined QM/MM methods offer a powerful tool for studying solution-phase chemical reactions and solvation effects due to solvent polarization [96]. The early QM/MM studies of the Menshutkin reaction were pioneered by Gao and co-workers [49], [85], [97]. The solution-phase free energy profiles in their work were obtained either by Monte-Carlo-based free energy of hydration calculations using an AI-level gas-phase minimum energy path under a static solvation assumption [49] or by PMF simulations using two-dimensional umbrella sampling at the semiempirical AM1/TIP3P level [85], where the latter allows polarization

of the solute to be treated at a consistent electronic-structure level through the QM/MM interaction Hamiltonian.

Although Gao’s QM/MM approach provided detailed molecular-level information, the high costs of sampling a large number of explicit solute/solvent configurations motivated the studies of the Menshutkin reaction by AI calculations at various MO and DFT levels coupled with implicit solvation treatments. These include the multipolar expansion model by Dillet et al. [86] at the HF level, the polarizable continuum model (PCM) calculations at the MP2/3-21G level by Fradera [48] and at the complete active space self-consistent field (CASSCF) level by Amovilli [47], and the generalized conductor-like screening model (GCOSMO) at the DFT and MP2 levels by Truong [53]. In particular, Truong et al. showed that by mixing a significant amount of HF exchange in hybrid DFT, the BH&HLYP method agrees well with the MP4 benchmark and experiments for the Menshutkin reaction in terms of its reaction energy and barrier height, both in the gas phase and in solution [53].

To strike a balance between the efficiency of using an implicit solvent and the microscopic/dynamic level of accuracy using an explicit solvent, several intermediate methods that bridge the continuum and QM/MM approaches have been developed. For example, Kato and co-workers [51], [87] explored the Menshutkin reaction using the reference interaction site model self-consistent field (RISM-SCF) method, which combines the RISM integral equation for solvent with the solute electronic structures for description of local solute-solvent interactions [98], [99]. Employing the free energy gradient (FEG) strategy of Okuyama-Yoshida et al. [100], Hirao et al. [50] optimized the solution-phase transition-state (TS) geometry for the Menshutkin reaction on a multidimensional free energy surface implicit of solvent coordinates based on the solute FEG derived from explicit QM/MM simulations. To reduce the computational costs of determining FEG explicitly from sampling the QM/MM potential energy surface, Galván et al. [90] developed a mean-field approach called the averaged solvent electrostatic potential QM/MM (ASEP-QM/MM), where they used a fixed solute geometry and charge distribution while sampling the solvent configurations to obtain the ASEP for subsequent implicit polarization of the solute at the quantum mechanical level; with this approach, the free energy of activation and free energy TS properties were characterized at the BH&HLYP/aug-cc-pVDZ level [90]. A related but different strategy was also

employed by Gordon and co-workers [54], who studied the Menshutkin reaction in the effective fragment potential (EFP) framework; in their approach, the electronic structure of the solute molecule is determined under the polarization of EFP generated from the explicitly represented surrounding solvent clusters [101].

As the mean-field QM/MM treatment helps eliminate the costs for explicit AI/MM MD simulations, the related approaches enable free energy to be computed with high-level AI methods. For example, Nakano et al. reported the free energy profiles for the Menshutkin reaction at the MP2/6-31+G(d,p)/MM level using their own mean-field QM/MM approach [102], which is similar in spirit to the QM/MM MFEP method developed earlier by Yang and co-workers [103]–[105], although the latter has not been applied to the Menshutkin reaction. To assess the performance of their mean-field method, Nakano et al. also obtained a benchmark QM/MM PMF at the MP2/6-31+G(d,p)/MM level using umbrella sampling [102]. One notable inconvenience in these mean-field QM/MM treatments (as well as in the implicit solvation calculations) is that, as the MFEP is optimized in terms of the QM solute coordinates, the dynamic sampling of the QM atoms is lacking; therefore, the missing vibrational entropy of the solute has to be estimated and added separately (e.g., using a harmonic approximation [104]). Unfortunately, for the Menshutkin reaction, these solute entropy corrections to the free energy profile seem to be substantial (ca., 7 and 9–13 kcal/mol [53], [90], [92] for the reaction free energy and free energy barrier, respectively), which makes a direct comparison between the mean-field QM/MM simulation results and experiments less straightforward.

The Menshutkin reaction has recently been revisited using explicit QM/MM simulations. On the one hand, alternative SE and solvent models have been tested. For example, Acevedo and Jorgensen [91] combined the semiempirical PDDG/PM3 method with the TIP4P water model and computed the free energy profiles for the Menshutkin reaction through Monte-Carlo-based free energy perturbation calculations. A similar approach has been employed by Vilseck et al. [94] to obtain the QM/MM free energy profiles for the same reaction based on the semiempirical RM1 method, where the more sophisticated CM1/3 charge model is used for treating QM/MM electrostatic interactions with the solvent. On the other hand, because of the inaccuracy in SE/MM methods and the daunting costs of AI/MM free en-

ergy simulations, a number of multiscale QM/MM methods utilizing AI information have been developed, aiming at simulating the Menshutkin reaction in a highly accurate but also affordable manner. For example, Tunon and co-workers [88], [89], [106] developed a dual-level QM/MM strategy, where a PES correction term is first obtained for an SE(/MM) method to fit the AI(/MM) energy results and then used in the subsequent PMF simulations. Technically, the energy correction needed for the two levels to match is treated as a spline interpolation function either along a one-dimensional reaction coordinate [88] or on a two-dimensional surface [106]. Depending on whether an unperturbed gas-phase Hamiltonian or an electrostatically perturbed QM/MM Hamiltonian is used, they developed two interpolated correction schemes, referred to as unperturbed interpolated correction (UIC) and perturbed interpolated correction (PIC), both applied to the Menshutkin reaction with AM1 being corrected to MP2 [88].

Most recently, the Menshutkin reaction has been revisited using a few newly emerging multiscale QM/MM techniques, including the machine learning approaches [33], [34], the force-matching-aided weighted thermodynamics perturbation (FM + wTP) method [69], and the RP-FM-CV method we present here. Despite the common theme that they all aim at reproducing the highly accurate AI/MM free energy profiles at a reduced cost, the ways these methods utilize the high-level information are considerably different. Therefore, assessing the RP-FM-CV method on the Menshutkin reaction would make it possible to cross-validate the related approaches against one another for consistent first-principles AI/MM free energy simulations.

2.4 Computational Details

This section contains the detailed descriptions of the RP-FM-CV free energy simulations outlined above. The general features of the simulations, including the solute/solvent models, computation of the potential energies, definition of the collective variables, boundary conditions, and electrostatic treatment, are described in Sections 2.4.1–4.4. The specific details associated with the restraints used in the simulations are given in Section 2.4.5. Ad-

ditional details for the string MFEP simulations and FM in redundant internal coordinates are provided in Sections 2.4.6 and 4.7.

2.4.1 Description of the Solute Model

The topologies for the solute molecules NH_3 and CH_3Cl were built based on similar residues available in the standard CHARMM topology files. Specifically, the atom types NH_3 , HC , CT3 , HA , and CLA are used for the nitrogen, hydrogens in NH_3 , carbon, hydrogens in CH_3 , and chlorine, respectively. With these atom types specified, the van der Waals (vdW) parameters were assigned based on the standard CHARMM22 force field [107] during the initial setup of the system. In the actual simulations, the values of these parameters, required for computing the nonbonded QM/MM interactions between the solute and solvent atoms, were replaced by their pair-specific version tailored for the Menshutkin reaction (see below).

2.4.2 Potential Energy Calculations

For the SE/MM simulations of the Menshutkin reaction in water, the solute molecules consisting of NH_3 and CH_3Cl are treated by the SE method AM1 [24], whereas the solvent molecules are treated by MM using the modified TIP3P model [108]. The MNDO97 package [109] incorporated into the CHARMM program [110] (version c42a2) was used for the AM1/MM calculations. The related AI/MM calculations were conducted using the Q-Chem package [111] (version 4.0.1) interfaced with CHARMM; the specific combinations of AI methods and basis sets used are discussed in Section 2.5. For the QM/MM vdW interactions, we adopted the pair-specific vdW parameters previously optimized by Gao et al. for the Menshutkin reaction [85], which were implemented in our simulations using the NBFIX facility in CHARMM (see Section SI.1 of B.1, Supporting Information).

2.4.3 Definition of Collective Variables

Let N and Cl represent the nitrogen and chlorine atoms in the NH_3 and CH_3Cl groups, respectively, and C represents the carbon atom in the transferred methyl group. To describe the free energy path for the Menshutkin reaction ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$), we use

the bond-breaking distance (C-Cl) and the bond-forming distance (N-C) as the two CVs in the string MFEP simulations; the same CVs are also used consistently in FM.

2.4.4 Boundary Conditions and Treatment of Long-Range Electrostatics

In the original and subsequent force-corrected SE/MM simulations, a $40 \times 40 \times 40$ Å³ cubic box of modified TIP3P water molecules is used to solvate the reactive solute system [108]. The SHAKE algorithm [112] is used to constrain the internal geometries of water during the MD simulations. In all cases, we adopted periodic boundary conditions in the simulations. In the SE/MM simulations, long-range electrostatics for MM/MM and QM/MM interactions are treated by the particle mesh Ewald (PME) [113] and QM/MM-PME [114], [115] methods, respectively. In both PME treatments, the parameter that represents the width of the Gaussian screening charge distributions is set to 0.34 Å⁻¹, and the reciprocal space summations are performed on a $40 \times 40 \times 40$ FFT grid, with maximally up to five k-vectors included in each Cartesian direction. For the real-space contribution of QM/MM-PME electrostatics, a switching function available in CHARMM is applied from 12 to 13 Å to smoothly attenuate the real-space QM/MM electrostatic interactions at a cutoff of 13 Å.

2.4.5 Restraints and MD Simulations

Colinear-like and C_{3v} geometry is imposed on the solute complex during all QM/MM MD simulations to reduce the systems distortion and prevent it from visiting irrelevant configurations that may slow down the MFEP convergence. A similar treatment was used by Truong and co-workers in their implicit solvation calculations [53], where they showed that the free energy barriers obtained for the Menshutkin reaction are not impacted significantly when applying these geometric constraints. Specifically, to impose the C_{3v} symmetry in CHARMM, we used three relative-distance (RESD) restraints to keep the three Hs in ammonia at the same distance from the central C. To prevent any potential drift of the solute toward the edge of the simulation box (commonly observed for small solutes), we also placed a one-sided quartic spherical repulsive potential at 8 Å away from the box center, using

the MMFP facility in CHARMM. The QM/MM MD simulations were carried out under constant-pressure and constant-temperature conditions at 1 atm and 298.15 K.

2.4.6 String MFEP Simulations

For the string MFEP simulations in CVs, we adopted a protocol we recently used for simulating the QM/MM free energy profiles of adenosine 5-triphosphate (ATP) hydrolysis in the ATP-binding cassette (ABC) transporter HlyB [11]. Here, we only provide a brief description of the simulation procedure; the detailed simulation parameters can be found in our published work [11]. The MFEP, represented by the two bond CVs described in Section 2.4.3, is discretized into 25 images of the system, whose initial coordinates were obtained from a QM/MM potential energy scan along the reaction coordinate. For each iteration of the string MFEP optimization, the free energy mean force on each CV was estimated from the CVs fluctuation during 20 ps QM/MM MD simulations in which the CVs are harmonically anchored at their previous path values using a uniform force constant of 1000 kcal/mol/Å². For the projection, reparametrization, and evolvement of the MFEP, as well as the integration of the free energy profiles, see our previous implementation [11] based on the original string method developed by Maragliano et al. [22].

2.4.7 Force Matching in Redundant Internal Coordinates

The generalized inverse of the **G**-matrix is formed by following the procedure outlined by Pulay and co-workers [43], where a threshold of 0.02 is used to identify the negligible eigenvalues that are related to redundancy in the coordinate system (see Appendix B of 2.8). The singular value decomposition (SVD) method, adapted from the Numerical Recipes in Fortran77 [83], is used as a default solver in the spline-based FM on CVs (see Appendix A of 2.8), where a maximum eigenvalue scaled by 10⁻⁶ is employed as a cutoff for removal of linear dependency when solving the overdetermined system under eq A15 (see Section SI.2 of B.1 for a concrete example of casting eq A15 in its matrix form). For each of the bond CVs, the internal force correction needed for matching the SE/MM and AI/MM forces is fitted into a spline function with a grid interval of 0.2 Å. The CHARMM code in its c42a2 version was

modified to incorporate the force correction terms on the selected CVs in internal coordinates by distributing the internal force corrections to the related Cartesian force components with the chain rule.

2.5 Results and Discussion

As Gao found in the early 1990s, due to charge separation in the Menshutkin reaction during its product formation, the presence of a polar solvent generates a tremendous amount of solvation free energy that stabilizes the products as well as the transition state, thereby lowering the free energy of activation compared with the gas phase [49]. As a result, although the reaction is endothermic in the gas phase, it becomes highly exergonic in aqueous solutions. However, it is known that the semiempirical AM1/MM method is unreliable in simulating this reaction, which gives a reaction free energy significantly higher than experiment [85]; this is possibly related to the inaccurate description of the chloride anion using a minimal basis set in the AM1 Hamiltonian.

To obtain the first-principles free energy profiles, we use RP-FM-CV to correct the AM1/MM forces to their AI/MM target values. Specifically, the AI levels for computing the target forces include two DFT methods, namely, B3LYP [116]–[118] and BH&HLYP [53], [117], [119], and the second-order Møller–Plesset perturbation (MP2) [120], [121] method, with 6-31+G(d,p) [122] as the default basis set. For brevity, we refer to the RP-FM-CV simulation methods using this default basis set as AI:SE/MM; for other basis sets or when multiple basis sets are compared explicitly, the more specific label AI/BasisSet:SE/MM is used instead for clarity. Force matching in each case was done based on 300 solution-phase configurations using a set of 28 redundant internal coordinates as a default. Unless stated otherwise, the results presented in this section are from the simulations using this default RP-FM-CV protocol. The convergence tests of the free energy results with respect to the redundant internal coordinate sets, sample sizes, and basis sets can be found in Sections 2.5.4–5.6. Due to the rapid convergence of the overall procedure when performing the method iteratively (see Section 2.5.8 for details), the results after a single cycle of RP-FM-CV are reported by default.

Next, we first present the ability of RP-FM-CV in converging a wide variety of target AI/MM results, including the free energy profiles, internal forces on the CVs, free energy pathways, and transition-state locations. Note that due to the daunting costs of obtaining the AI/MM benchmark free energy profiles simulated under the same sampling requirement, here we focus on cross-validating the RP-FM-CV method under various AI and FM conditions; a separate validation of the RP-FM-CV method against the directly obtained AI/MM benchmark results, both using a shorter sampling time, can be found in Section SI.8 of B.1. As an additional validation, the relevant AI/MM free energy results and transition-state geometries available in the literature are also compiled in Table 2.1 for comparison with our results.

Table 2.1. Free energy barriers (ΔG^\ddagger), reaction free energies (ΔG_r), and transition-state geometries (Å) for the Menshutkin reaction between NH_3 and CH_3Cl in water.

Method	ΔG^\ddagger (kcal/mol)	ΔG_r (kcal/mol)	N-C (Å)	C-Cl (Å)	Ref.
MP2(fc)/6-31+G(d,p)/AM1/TIP3P (UIC)	23.8	-31.1	2.24	1.99	[88]
MP2(fc)/6-31+G(d,p)/AM1/TIP3P (PIC)	19.1	-27.4	2.22	2.01	[88]
MP2/6-31+G(d,p)/MM benchmark	27.6	-16.9			[102]
HF/6-31G(d)/MM benchmark	21.2	-25.9			[34]
HF/6-31G(d):DFTB/MM ML int. force corr.	20.8	-23.8			[34]
B3LYP/6-31G(d)/MM benchmark	15.3	~ -28			[69]
B3LYP/6-31G(d):PM3/MM FM+wTP	15.7	~ -28			[69]
AM1/TIP3P (with QM/MM-cutoff)	29.3	-10.4	2.1	2.1	[88]
AM1/TIP3P (with QM/MM-cutoff)	26.3	-18	1.96	2.09	[85]
AM1/TIP3P (with QM/MM-Ewald; string)	30.9	-10.6	1.97	2.129	this work
HF/6-31G(d):AM1/MM	18.3	-33	2.27	2.248	this work
B3LYP:AM1/MM ^a	14.7	-27.2	2.213	2.194	this work
BH&HLYP:AM1/MM ^a	17.8	-28	2.187	2.222	this work
MP2/6-31G(d):AM1/MM	19.1	-28.7	2.171	2.196	this work
MP2:AM1/MM ^a	21.3	-26	2.17	2.193	this work
MP2/6-311++G(d,p):AM1/MM	22.2	-23.9	2.129	2.209	this work
MP2/6-311++G(2df,2p):AM1/MM	19.6	-24.6	2.145	2.168	this work
B3LYP/6-31G(d)/MM benchmark ^b	18.8	-29.4	2.295	2.133	this work
Experiment	23.5	-34 ± 10 -36 ± 6			[49] [94]

^a RP-FM-CV simulations with the target AI/MM forces evaluated at the default basis set 6-31+G(d,p)

^b DFT/MM-cutoff simulations with 30 iterations of string MFEP optimization; 1 ps sampling is used in each iteration for mean force evaluation (see SI.8).

2.5.1 Free Energy Profiles

For the present work, the base method we used in the RP-FM-CV simulations is AM1/MM, which by itself generates large errors in free energy. As shown in Figure 2.3, the AM1/MM simulations predict a free energy barrier of 30.9 kcal/mol for the Menshutkin reaction, which is 7.4 kcal/mol higher than the experimental value of 23.5 kcal/mol [49].

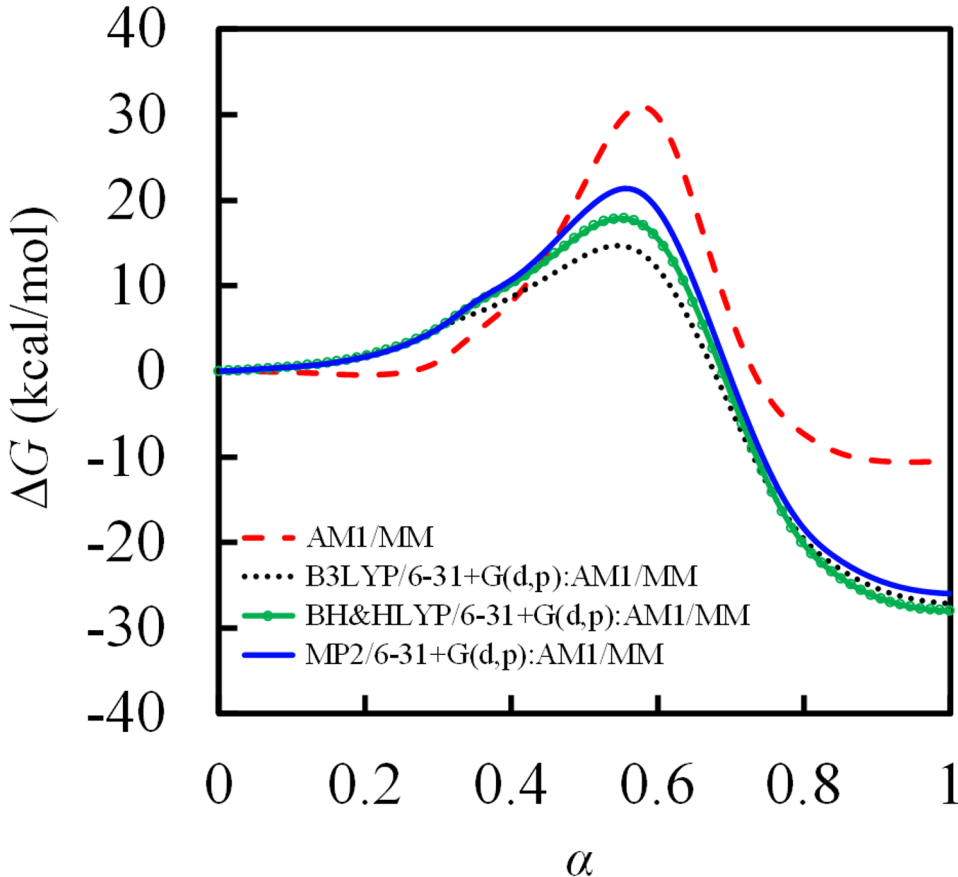


Figure 2.3. Free energy profiles along the string MFEPs (with $\alpha = 0$ being reactant and 1 product) from the AM1/MM (dashed red) and RP-FM-CV simulations of the Menshutkin reaction in aqueous solution. Results from the RP-FM-CV AI:AM1/MM simulations were obtained by matching AM1/MM forces on the CVs to various target AI/MM levels using the default 6-31+G(d,p) basis set: B3LYP:AM1/MM (dotted black), BH&HLYP:AM1/MM (green with circles), and MP2:AM1/MM (solid blue).

The reaction free energy for the Menshutkin reaction obtained from our AM1/MM simulations is -10.6 kcal/mol, which corresponds to an overestimate of 23.4 or 25.4 kcal/mol compared with the experimental value of -34 ± 10 [49] or -36 ± 6 kcal/mol [94], both established from the gas-phase thermodynamics and free energy of hydration data. Our free energy barrier is consistent with the AM1/MM simulations previously performed by Gao and Xia [85] and by Ruiz-Perna et al. [88] using electrostatic cutoff, who obtained slightly lower free energy barriers of 26.3 [85] and 29.3 [88] kcal/mol, respectively (see also Table 2.1). For the free energy of reaction, our result is also qualitatively comparable to the previous results of -18.0 [85] and -10.4 kcal/mol [88].

All three RP-FM-CV-based AI:AM1/MM methods improve the reaction free energy toward the experimental value. While improving the reaction free energy, the two DFT:AM1/MM methods lower the free energy barriers compared with the AM1/MM method. Specifically, the RP-FM-CV simulations that fit forces to the B3LYP/MM level yield a free energy barrier height of 14.7 kcal/mol [116]–[118], and the RP-FM-CV simulations at the BH&HLYP:AM1/MM level give a free energy barrier of 17.8 kcal/mol [53], [117], [119]. These results are in line with the literature observation that these DFT methods tend to underestimate the barrier height for this system. For example, based on the GCOSMO continuum solvation calculations, Truong et al. [53] showed that B3LYP underestimates the barrier height for the Menshutkin reaction, whereas BH&HLYP (with 50% HF exchange) produces much better results comparable to the data obtained at the highly correlated MP4(STDQ) level. This trend is successfully reproduced in all of our all-atom explicit-solvent RP-FM-CV simulations, where the highest-level AI:AM1/MM simulations that match forces to the MP2/6-31+G(d,p)/MM level strike a good balance between the free energy barrier and reaction free energy predictions. While improving the reaction free energy to -26.0 kcal/mol, compared with the experimental value of -34 ± 10 kcal/mol, our MP2:AM1/MM simulations maintain a free energy of activation at 21.3 kcal/mol, in close agreement with the experimental value of 23.5 kcal/mol established by Gao [49] on the basis of the $\text{NH}_3 + \text{CH}_3\text{I}$ reaction [52], [123].

Similarly, compared with the literature data shown in Table 2.1, our MP2:AM1/MM results are in great agreement with those of Ruiz-Perna et al. [88] using the perturbed interpo-

lated potential energy correction QM/MM method at a similar dual level of MP2(fc)/6-31+G(d,p)/AM1/TIP3P (PIC), which yields a free energy barrier and reaction free energy of 19.1 and -27.4 kcal/mol, respectively.

2.5.2 Force Correlations

The effectiveness of the RP-FM-CV method is also monitored by comparing the internal forces on the CVs between the base and target levels. Specifically, the internal forces obtained for the two bond-based CVs are compared between AM1/MM and the AI/MM levels that use B3LYP/6-31+G(d,p) and MP2/6-31+G(d,p). In Figure 2.4, we show the internal force correlations for the levels involved before and after force matching.

In the same figure, the average internal force deviations (ΔF) between the base and target levels are also given for comparison. As we can see from Figure 2.4, based on the average of 300 configurations, the internal forces on the CVs computed at the original AM1/MM level differ from their AI/MM targets by 12-32 kcal/mol/Å², depending on the bonds and the target levels [vs B3LYP/6-31+G(d,p)/MM: 16.5 (N-C) and 32.2 (C-Cl) kcal/mol/Å²; vs MP2/6-31+G(d,p)/MM: 12.4 (N-C) and 23.4 (C-Cl) kcal/mol/Å²]; after force matching, the average force differences between RP-FM-CV and the target levels are greatly reduced to only 2.0-2.1 kcal/mol/Å² for both bonds. These results demonstrate that the FM component in the RP-FM-CV method works effectively. As the AI/MM forces on the CVs are faithfully reproduced after FM, the force correlation results presented here also help us rationalize the improvements that we see in Section 2.5.1 for the free energy profiles.

2.5.3 Internal Force Corrections on CVs along MFEP

Although RP-FM-CV delivers great numerical agreement on the CV forces between the SE/MM and AI/MM levels based on the sampled configurations, one important question that remains is whether the spline-based correcting forces are well-behaved and smooth functions of the reaction coordinate along the MFEP. These properties of the force correction terms are highly desirable for numerically stable dynamics when the modified forces are plugged back into the MD simulations for obtaining the updated free energy profiles.

To demonstrate the smoothness of the spline-based internal force correction terms, in Figure 2.5 we plot the force deviations between the SE/MM and AI/MM levels as well as their FM-optimized spline fits for both CVs along their bond distances.

One can see that the spline functions nicely fit the averages of the individual force deviations sampled along the MFEP. We noticed that the distributions of the CV force deviations (i.e., the desired force corrections for accomplishing a perfect FM between the two levels) are indeed quite smooth, which justifies the use of spline functions in fitting these corrections. In spite of the deceptive smoothness of the fits, our spline-corrected internal forces successfully reproduce their instantaneous AI/MM internal force targets with small errors of 2.0–2.1 kcal/mol/Å² (see Section 2.5.2 and Figure 2.4), thereby capturing the detailed internal force fluctuations at the target levels. To this end, the spline-based force correction scheme serves the designed purpose of RP-FM-CV well, which, as we have discussed in Sections 2.2.1 and 2.2, is to fit free energy mean force through matching instantaneous forces for individual configurations in an ensemble.

Finally, the smoothness of the spline-based force corrections also indicates their numerical stability when incorporated in the SE/MM force calculations for FM-corrected MD trajectories, with which the free energy profiles and pathways can be updated in a robust way.

2.5.4 Tests of Different Sets of Redundant Internal Coordinates

In our formulation of the RP-FM-CV method, computation of the forces on the CVs is based on the force transformation from the Cartesian to a set of redundant internal coordinates, for which the definition is not unique. To test the robustness of the algorithm with respect to the choice of the internal coordinate system, we examined three different sets of redundant internal coordinates for the Menshutkin reaction. As shown in Figure 2.6, the first redundant internal coordinate set (also the default set), denoted Int28, includes 8 bonds, 15 angles, 1 doubly degenerate linear bend, and 3 torsions. The other two redundant sets, denoted Int31 and Int34, are constructed by adding three and six more torsions, respectively (see also Figure 2.6).

All three sets include the N–C and C–Cl bonds that define a common set of CVs in both the string MFEP and FM simulations.

In Figure 2.7, we compare the RP–FM–CV free energy profiles obtained at the MP2:AM1/MM level when the internal CV forces and FM are determined and conducted using the three different redundant internal coordinate systems described above.

Our results show that based on the redundant internal coordinate transformation, the internal forces on the CVs, in this case the bonds being broken (C–Cl) and formed (N–C), only vary marginally when using different redundant sets. On average (over 300 configurations), the internal forces using the Int31 and Int34 sets only differ from that using the Int28 set by less than $0.01 \text{ kcal/mol/\AA}^2$) at the MP2:AM1/MM level. Consequently, the free energy profiles resulting from the RP–FM–CV simulations using the three internal coordinate sets are almost identical. These results demonstrate the robustness of the RP–FM–CV method in converging the free energy results when conducting FM in different redundant internal coordinate systems. The invariance of the free energy profiles with respect to the three redundant internal coordinates tested also indicates that under the Int28 default set the coordinate system is already complete.

2.5.5 Tests of Number of Configurations Included in FM

One advantage of the RP–FM–CV approach is that once the configurations sampled at the efficient SE/MM level are collected, the computationally expensive single-point AI/MM force calculations can be conducted in an embarrassingly parallel manner. As long as one has access to enough central processing units (CPUs), the wall time for computing the target AI/MM forces does not grow with the number of configurations used. In practice, however, the free energy results from the RP–FM–CV simulations may vary with the number of configurations included in FM. In our default simulation scheme, we conducted FM based on 300 solution-phase configurations taken from 25 images each sampled along the string MFEP over a period of 60 ps. To test how sensitive the free energy results are to the sample sizes in fitting the internal CV forces, we repeated the RP–FM–CV simulations at the MP2:AM1/MM level with three additional FM schemes, in which 1500, 3000, and 15000

configurations are used respectively. The resulting free energy profiles using different sample sizes for FM are compared in Figure 2.8.

The results in Figure 2.8 show that the free energy profiles computed at the MP2:AM1/MM level converge well with respect to the FM sample sizes. The free energy profiles essentially overlap with one another even when the number of configurations for FM varies by 50-fold from 300 to 15000. These results once again demonstrate the robustness and statistical reliability of the RP-FM-CV method.

While our tests on the Menshutkin reaction suggest a good convergence of the free energy profile using a small to medium FM sample size, the free energy convergence for more complex systems with large dynamical fluctuations could be more challenging. For those systems, greater numbers of FM configurations drawn from long simulations may be required especially when slow non-CV degrees of freedom are present.

2.5.6 Tests of Basis-Set Convergence

Because of the computational costs associated with a great number of sequential potential energy calculations for configurational sampling, free energy simulations at AI/MM levels are often limited to single-determinant electronic-structure AI methods such as HF and hybrid DFT, whose N^4 scaling behavior (with N being the number of basis functions) allows them to be used in combination with relatively small double-basis sets. The use of larger basis sets at and beyond these levels would dramatically increase the computational costs and therefore is rarely seen in practical AI/MM free energy simulations. Therefore, having an affordable strategy that allows AI/MM free energy simulations to be used with large-sized basis sets would greatly ease some of the concerns regarding the otherwise unknown basis-size convergence behavior of the simulations.

In RP-FM-CV, because FM is decoupled from dynamical sampling and conducted separately in a parallel fashion, the AI/MM force calculations are no longer the computational bottleneck for the simulations and therefore can be done at post-HF correlated levels such as MP2 with large basis sets. This enables us to carry out FM at AI/MM levels using basis

sets in various sizes, including the very large ones, to systemically check convergence of the free energy results in a way routinely done for gas-phase quantum chemistry calculations.

In Figure 2.9, we compare the RP-FM-CV free energy files for the Menshutkin reaction obtained at the MP2/BasisSet:AM1/MM level using the 6-31G(d) [122], 6-31+G(d,p) [122], [124], 6-311++G(d,p) [122], [124], [125], and 6-311++G(2df,2p) [126] basis sets, which correspond to 61, 91, 116, and 170 basis functions, respectively.

For the smallest basis set we tested, i.e., 6-31G(d), the free energy profile deviates notably from the other results. When the medium- to large-sized basis sets are used with extra split-valence, polarization, and diffuse functions added, the reaction free energies converge to a similar value. Specifically, the MP2/6-31G(d):AM1/MM simulation gives a reaction free energy of -28.7 kcal/mol, compared with -26.0, -23.9, and -24.6 kcal/mol when the basis set is upgraded to 6-31+G(d,p), 6-311++G(d,p), and 6-311++G(2df,2p), respectively (see Table 2.1). On the other hand, for the RP-FM-CV simulations in which the internal forces are fitted to the MP2/MM level with 6-31G(d), 6-31+G(d,p), 6-311++G(d,p), and 6-311++G(2df,2p) basis sets, the free energy barriers are 19.1, 21.3, 22.2, and 19.6 kcal/mol, respectively. With these data, we conclude that our RP-FM-CV simulations show good convergence with the basis set.

It is worth noting that for the largest basis set we tested, direct QM/MM free energy simulations at the MP2/6-311++G(2df,2p)/MM level are out of reach but made possible by the RP-FM-CV method. Agreements among the results using the 6-31+G(d,p) basis set and beyond also suggest that AI/MM free energy simulations in condensed phases likely display a similar convergence behavior seen in gas-phase systems, as long as the basis sets used are sufficiently large. For obtaining reasonably converged results, we recommend inclusion of diffuse and polarization functions in any attempt of AI/MM free energy simulations.

2.5.7 RP-FM-CV Produces AI/MM-Quality Free Energy Paths

Above, we showed that RP-FM-CV generates the AI/MM-quality free energy profiles for the Menshutkin reaction. The next question we seek to answer is whether RP-FM-CV can improve the free energy path to the target-level quality. Note that although RP-FM-CV

is formulated to directly fit the free energy mean force (see Section 2.2), there is no a priori knowledge that the target-level free energy path would also be faithfully reproduced.

In Figure 2.10, we plot the MFEPs (in terms of the two bond CVs, i.e., the N–C and C–Cl bond distances) determined by the original AM1/MM simulations, as well as those obtained from the RP–FM–CV simulations at the B3LYP:AM1/MM, BH&HLYP:AM1/MM, and MP2:AM1/MM levels.

The MFEP obtained at the original AM1/MM level differs from the FM-optimized ones in predicting a more convex path as a result of a much tighter TS, i.e., the sum of the N–C and C–Cl bond distances along the MFEP are significantly shorter than that produced at the various AI:AM1/MM levels. After the RP–FM–CV force corrections, the MFEPs obtained at all three AI:AM1/MM levels essentially converge to one another, which indicates that the free energy paths at the target AI/MM levels are also successfully reproduced.

The corresponding CV bond distances that characterize the location of a free energy TS (defined as the highest free energy point along the MFEP) are given in Table 2.1. For the free energy TS located on the MFEP, the original AM1/MM level gives a N–C bond distance of 1.970 Å, which is significantly shorter than the C–Cl bond of 2.129 Å in the same TS; this trend is in great agreement with the values of 1.96 Å (N–C) and 2.09 Å (C–Cl) reported by Gao and Xia [85] from their earlier AM1/TIP3P simulations. The dual-level AI/MM free energy simulations reported by Ruiz–Perna et al. [88] suggest that the N–C bond is likely extended to 2.2 Å in the TS when the PES is corrected to the MP2(fc)/6–31+G(d,p)/TIP3P level; a similar trend has been observed from various AI calculations using an implicit solvent (e.g., see the data compiled by Vilseck et al. [94]). Our RP–FM–CV simulations at the B3LYP:AM1/MM and BH&HLYP:AM1/MM levels both successfully reproduce this feature, locating the free energy TS at 2.213 Å (N–C) and 2.194 Å (C–Cl) and at 2.187 Å (N–C) and 2.222 Å (C–Cl), respectively. Our MP2:AM1/MM simulations also converge the TS geometry toward the benchmark and literature results by giving distances of 2.170 and 2.193 Å for the N–C and C–Cl bonds, respectively.

2.5.8 Convergence of the Overall Procedure

Due to the self-consistent nature of RP-FM-CV, cycles consisting of the RP and FM steps ideally need to be conducted iteratively until convergence of the free energy profile is established. Using the MP2:AM1/MM and B3LYP:AM1/MM methods, we examined the convergence behavior of the overall procedure by performing multiple cycles of RP-FM-CV. In the first cycle of the simulations, we conducted 10 iterations of string MFEP optimization at the AM1/MM level followed by FM to fit the CV forces to the target AI/MM levels. In each of the subsequent cycles, we updated the MFEPs by repeating the string simulations under AM1/MM forces in conjunction with the CV force corrections obtained from the previous RP-FM-CV cycle. Such cycles of MFEP optimization and FM in CVs are repeated five times.

From Table 2.2, we can see that throughout the five cycles of RP-FM-CV simulations the free energy barriers and reaction free energies for the Menshutkin reaction obtained at the MP2:AM1/MM level display small fluctuations of 0.6 and 0.7 kcal/mol about the corresponding average values of 20.6 and -25.3 kcal/mol, respectively, whereas the first cycle produces 20.3 and -26.0 kcal/mol for these free energy results.

Table 2.2. Computed free energy barriers (ΔG^\ddagger), reaction free energies (ΔG_r), and transition state geometries for the Menshutkin reaction between NH_3 and CH_3Cl in water over five cycles of RP FM-CV simulations at the MP2:AM1/MM level.

Cycle	ΔG^\ddagger (kcal/mol)	ΔG_r (kcal/mol)	N-C (Å)	C-Cl (Å)
1	20.3	-26	2.17	2.193
2	20.8	-25.1	2.17	2.196
3	19.9	-25.8	2.174	2.195
4	20.1	-25.6	2.17	2.199
5	20.6	-24.2	2.135	2.226
Average	20.6 ± 0.6	-25.3 ± 0.7	2.164 ± 0.016	2.202 ± 0.014

In terms of geometry, the N-C and C-Cl bond distances found at the free energy TS throughout the five cycles fluctuate closely about their average values of 2.164 ± 0.016 and 2.202 ± 0.014 Å, respectively, compared with the values of 2.170 and 2.193 Å obtained after

the first cycle. For the Menshutkin reaction, we found that even one cycle of RP–FM–CV is sufficient enough to converge the free energy and TS geometry results reasonably well to the average values obtained after five cycles. The free energy profiles determined at the MP2:AM1/MM level from each of its five cycles are further compared in Figure 2.11, which shows that they overlap well with no systematic drift detected during the iterative applications of RP–FM–CV.

A similar convergence behavior is observed for the RP–FM–CV simulations at the B3LYP:AM1/MM level (see Section SI.6 of B.1). Altogether, these results strongly suggest a rapid self-consistent convergence of the RP–FM–CV procedure for the Menshutkin reaction studied here, which justifies our use of a single RP–FM–CV cycle as a default.

2.5.9 Radial Distribution Functions

To understand how the force correction terms applied in RP–FM–CV simulations would impact the solvent–solute interactions, we computed the radial distribution functions (RDFs) for the selected solute–solvent atom pairs in the reactant (R), transition state (TS), and product (P) regions along the MFEP. Specifically, the N–O_w, C–O_w, and Cl–O_w RDFs involving the water oxygen (O_w) atoms were obtained at the MP2:AM1/MM and AM1/MM levels and are compared in Figure 2.12; a similar comparison made for the B3LYP:AM1/MM level can be found in Section SI.7 of B.1.

The free energy barrier and reaction free energy we obtained from the AM1/MM simulations are 30.9 and –10.6 kcal/mol, which are lowered in the MP2:AM1/MM simulations to 21.3 and –26.0 kcal/mol, respectively (see also Table 2.1); this suggests that both the TS and the P state are more stabilized by FM than is the R state. If such a stabilization involves any changes in solvation, impacts on the solvent structures would be observed in the related RDFs. In the AM1/MM results (Figure 2.12), the first solvation peaks of all three solute–solvent RDFs are found higher and shifted toward shorter distances when the system evolves from the R, through the TS, to the P region, which is in line with enhanced solvation upon forming the ionic products in the Menshutkin reaction. This feature is largely preserved in the MP2:AM1/MM results (Figure 2.12) after the CV force corrections are applied, which

suggests that the physical description of solvation in the RP-FM-CV simulations is well retained.

On the other hand, FM seems to lead to a few quantitatively notable changes in RDFs. For example, the Cl-O_w RDF displays a lowered first peak in the TS region after the CV forces are corrected to the MP2:AM1/MM level, but no obvious changes in the peak height are found in the R and P regions; this observation suggests a less solvated TS and therefore a higher solvation barrier than without the FM corrections, which does not seem to directly contribute to the reduced free energy barrier seen in our MP2:AM1/MM simulations. Moreover, the first peak of the N-O_w RDF obtained from our AM1/MM simulations is found at 3.12 Å in the TS, which is 0.48 Å shorter than the corresponding location of 3.60 Å in the R state (Figure 2.12). After FM to the MP2/6-31+G(d,p)/MM level, the corresponding peak in the TS is moved to 3.36 Å, which becomes only 0.24 Å shorter than the peak location of 3.60 Å in the R state. This result also suggests that the enhanced solvation along the reaction coordinate that preferentially stabilizes the charge-separated TS over the charge-neutral R state is weakened after FM, which again would lead to a higher solvation barrier.

Based on the above data, we conclude that the lowered overall free energy barrier after FM does not correlate with a reduced solvation barrier; the improved free energy profile is predominantly a result of the modified intramolecular forces in the solute rather than changes in the solvent structures. This suggests that the free energy stabilization seen in the FM results is dominated by the force corrections on the CVs, as opposed to solvation itself. Note that in the present RP-FM-CV implementation, the CVs we used for correcting the internal forces only involve the solute coordinates; therefore, any changes in the solute-solvent interactions are likely caused indirectly by the solvents response to the modified solute charge distribution. Because RP-FM-CV does not modify SE-SRP parameters, any changes in solvation would be realized more through a modified solute geometry (e.g., a looser TS found after FM) than through an explicit alteration of the electronic-structure part of the SE/MM interaction Hamiltonian. To further improve the structural and dynamical descriptions of the solvent to the AI/MM levels, it is highly desirable to include solvent coordinates into the CVs for FM, which is a topic of our ongoing work.

2.6 Outlook

A common theme found in the recent developments of multilevel QM/MM free energy methods is to utilize high-level AI/MM information based on configurations efficiently sampled using low-level PES methods, such as SE/MM. Depending on how the high-level data are used, two related strategies, namely, energy matching and force matching, have emerged.

The most straightforward way of using the high-level information is to match the AI/MM total energy; this can often be done by fitting parameterized energy correction terms for the base level. Examples of the energy-matching-based methods include the interpolated PES correction approach using spline functions [88], [106] and the recent work of fitting AI/MM energy by machine learning [72], [73]. An obvious limitation of energy fitting is that even with the high-level energy data reproduced, there is little to no control on the improvement on atomic forces, which are essential for MD-based free energy simulations.

A different strategy of using the high-level information is to directly fit the AI/MM forces as the only target data [28], which can be viewed as the reactive version of the more generalized force matching strategy [127]. Connected to Voth and co-workers pioneering work on the multistate-empirical valence bond (MS-EVB) method [128], the multilevel QM(/MM) methods under the reactive FM umbrella include the SRP-fitting-based RP-FM [28], FM-DFTB [71], machine learning-based internal force correction [34], and RP-FM-CV reported in this paper. The FM strategy is especially appealing for QM/MM free energy simulations that use MD as the sampling tool. From a dynamics perspective, the multiple time step (MTS) integration approach developed by Nam [115], in which the SE/MM forces are directly corrected to their target AI/MM values at less frequent MD steps, can also be viewed as an FM QM/MM method with FM done on the fly.

As emphasized by us [28], because force serves as the central quantity that encodes all the dynamical information of the system, FM would restore the detailed dynamics at the target level. Therefore, in its purest form, the FM QM/MM strategy fits forces as the only objective quantities without an explicit use of any energy information. As forces are based on the first derivatives of the potential energy, the FM strategy can sometimes be used in a hybrid form in combination with either energy matching or a construction of the

potential energy function. For example, in our earlier implementation of RP-FM [28], we fit SRPs for an SE/MM method to restore the AI/MM atomic forces; as a byproduct, the FM-optimized SRPs also lead to an explicit SE potential energy function that gives the target forces although we never include the target-level energy in the objective function during the SRP-fitting process. In our recent work [69], we developed a hybrid strategy, where FM is first used to obtain SRPs to reproduce the target forces, on top of which weighted thermodynamics perturbation (wTP) utilizing the AI/MM energy data is further employed to restore the high-level free energy. In the FM-DFTB method developed by Kroonblawd et al. [71], parametrized pairwise energy terms are used to represent the repulsive potential part in the DFTB Hamiltonian; the linear dependence of the associated forces on the parameters makes these pairwise energy terms well suited for FM in a linear optimization framework. In a few recent machine learning (ML)-assisted QM/MM approaches developed by Riniker and co-workers [32], by York and co-workers [74], and by Shao and co-workers [33], both energy and force matching are accomplished; some of these works are enabled by the deep-learning tools developed by E and co-workers [129]–[131] or follow their strategy of folding both energy and the associated atomic forces into a combined loss function when optimizing the ML potentials. In all of these works, there are potential energy functions resulting from FM. When serving as a standalone objective, FM can be otherwise achieved without explicitly constructing the corresponding potential energy function. Examples of fitting forces without an explicit potential energy term include force corrections of Yang and co-workers [34] and our RP-FM-CV.

A particular advantage of the RP-FM-CV method is dimension reduction in terms of fitting the CV forces along a one-dimensional free energy path. This choice makes our method more convenient than fitting a multidimensional potential energy correction term (e.g., the work of Ruiz-Perna et al. [106]), which requires knowledge on the couplings among multiple reaction coordinates to maintain the global correctness of the PES and therefore would quickly become unmanageable beyond two dimensions. We note that fitting AI(/MM) data in high dimensions can be handled by alternative strategies such as the pairwise energy correction scheme [71] and the more generalized ML approaches [32], [33], [72]–[74], [129]–[131], by which multiple reaction coordinates can be incorporated explicitly or through atom–

centered local descriptors so that their couplings can be parametrically represented in the ML potentials.

Recently, Yang and co-workers also reported a force-based machine-learning QM/MM approach [34], where they obtained internal force corrections for DFTB/MM to match with the AI/MM results. Our work differs from theirs in the way the internal forces are defined. Yang and co-workers obtained their internal force expression with an aim to reproduce the MD trajectory integration step at the target AI/MM level. By contrast, our formalism directly aims at force matching. As a result, their trajectory matching formalism seems to involve additional mass factors compared with our force matching formalism (see Sections SI.3–SI.5 for details). For a special case of one-dimensional internal coordinate where a single bond is used as the only CV, the two formalisms conditionally converge to each other and to the projection operator formalism [42] (see Section SI.5 of B.1). For more complex reactions such as the Menshutkin reaction, where multidimensional nonorthogonal CVs are involved, the two strategies lead to internal forces that differ both in definition and in numerical values (see Sections SI.3 and SI.4 of B.1).

Besides the definition of internal forces, which is the major distinction between our method and Wu et al.s, RP-FM-CV is formulated in a different theoretical framework. Importantly, RP-FM-CV is framed in terms of fitting the free energy mean force, which builds a rigorous connection to fitting the high-level PMF. Interestingly, despite the very different definitions of internal forces, theoretical rationales, and technical details on how the force corrections are fitted (i.e., using spline functions vs machine learning), the two approaches both seem to satisfactorily reproduce their corresponding AI/MM free energy results for the Menshutkin reaction (see Table 2.1); this suggests that some of the numerical differences are perhaps averaged out when the internal force corrections are fitted over ensembles of configurations.

With the powerful deep-learning tools available now for molecular systems [129]–[131], combined energy and force matching is made possible to train ML models for AI/MM-quality free energy simulations [33], [74]. In these deep-learning works, the Cartesian atomic forces are fitted through differentiating the rotational- and translational-invariant ML energy. The internal force framework used by RP-FM-CV may provide an alternative way for learning

forces, as the internal forces by construction are invariant to rotation and translation. Moreover, due to the special role of reaction coordinate in chemical reactions, it is highly desirable for ML models to be able to selectively learn forces on the essential degrees of freedom; RP-FM-CV can complement ML to serve this purpose. To this end, our results suggest that it is very important to obtain the correct internal forces through proper coordinate transformation.

Other uses of the RP-FM-CV methods can also be envisioned. As we discussed above, because the internal forces obtained from RP-FM-CV can serve as a vehicle for fitting differentiable potential energy functions, the method can be used, for example, to optimize the parameters in the empirical energy correction term represented by a simple valence bond (SVB) potential [132]. The method can also be combined with the MTS [115] approach to directly correct the internal CV forces on the fly.

2.7 Concluding Remarks

In summary, we have developed RP-FM-CV, an FM-based multilevel QM/MM method, for determining first-principles free energy profiles for chemical reactions in condensed phases. At a conceptual level, our RP-FM-CV method reproduces the highly accurate AI/MM free energy profile by fitting the corresponding mean forces on a set of CVs based on which a free energy pathway is consistently defined. Mean force fitting in our method is accomplished by matching the target forces acting on the CVs, obtained properly from the redundant internal coordinate transformation, for the condensed-phase configurations sampled at an efficient SE/MM level. Application of the RP-FM-CV method to the Menshutkin reaction demonstrates its remarkable capability in reducing the errors on the CV forces, which greatly improves the quality of the free energy pathway and free energy profile to a level comparable to the AI/MM benchmarks and experimental results. This development therefore offers a systematic and practical strategy for first-principles free energy simulations; it is our expectation that this method will find more applications in AI/MM mechanistic studies of complex chemical and biochemical reactions, for which chemical accuracy and statistically adequate free energy sampling would otherwise be seemingly infeasible to achieve at the same time.

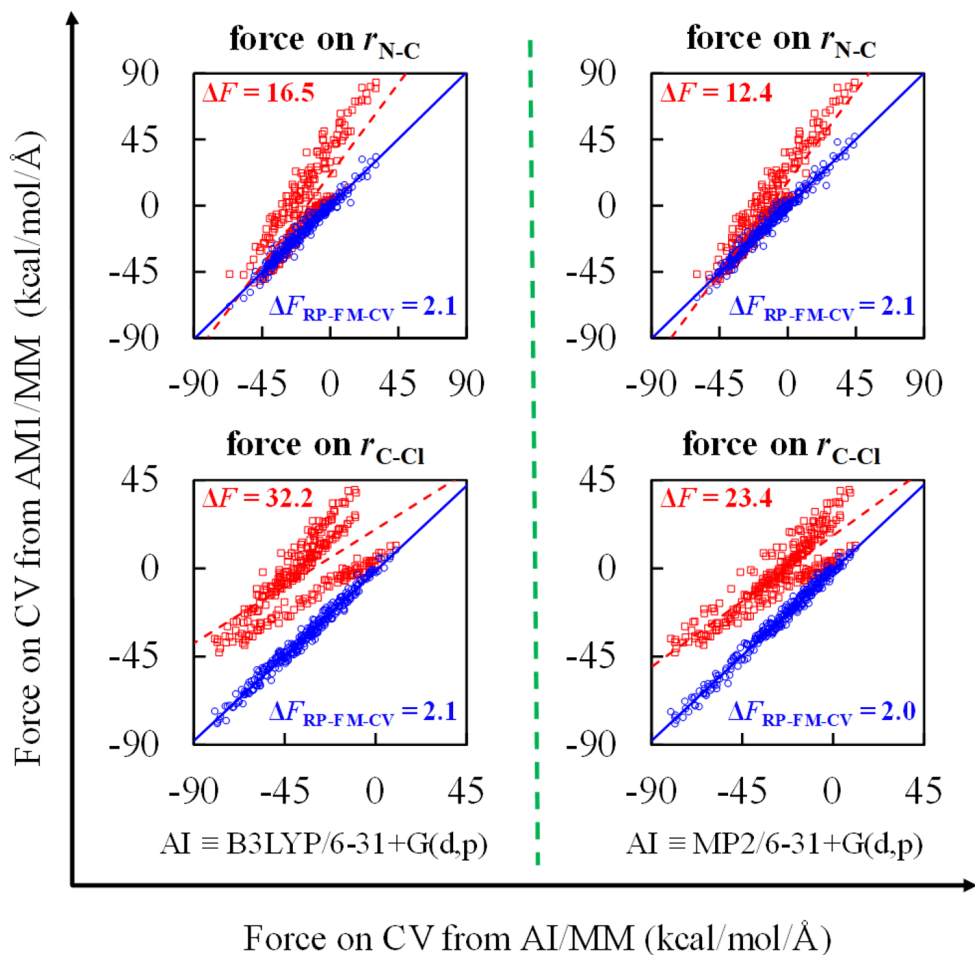


Figure 2.4. Internal force correlations between the base AM1/MM and target AI/MM methods [at the B3LYP/MM and MP2/MM levels both using the 6-31+G(d,p) basis set]: before (red squares) and after (blue circles) applying the RP-FM-CV internal force corrections; the corresponding trend lines are shown as dashed and solid lines. Internal forces on the two bond CVs were computed based on 300 configurations sampled along the condensed-phase MFEP from the AM1/MM string simulations. The average internal force deviations (ΔF ; in kcal/mol/Å²) between the base and target levels, before (red) and after (blue) force matching, are also shown for comparison.

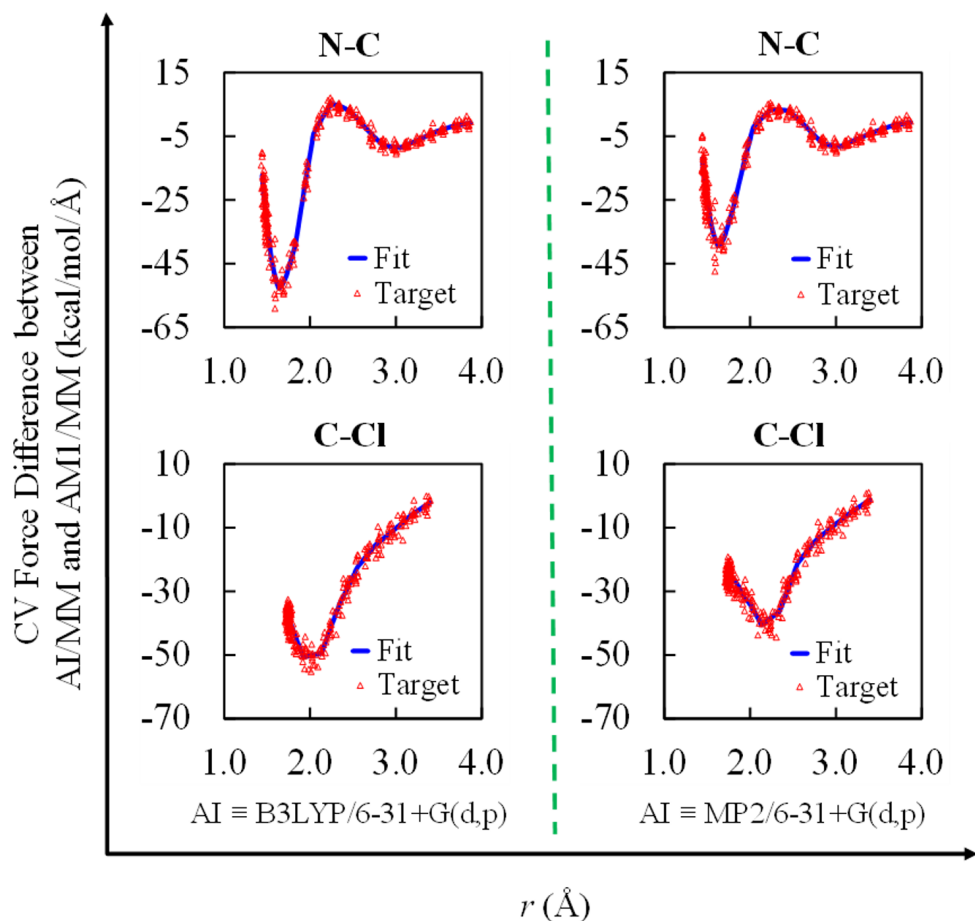
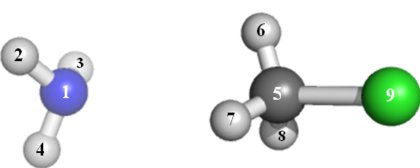


Figure 2.5. Internal force corrections on CVs for the Menshutkin reaction in solution with AM1/MM being the base level and the target forces obtained at the B3LYP/MM and MP2/MM levels using the 6-31+G(d,p) basis set. Both the actual internal force differences (based on 300 solution-phase configurations along the string MFEP) between the base and target levels (labeled Target; red triangles) and the spline based force corrections (labeled Fit; blue lines) resulting from FM are shown for each CV (i.e., the N-C and C-Cl bonds).



Bonds	1-5	5-9	1-2	1-3	1-4	5-6	5-7	5-8
Angles	1-2-3	1-2-4	1-3-4	2-1-5	3-1-5	4-1-5	1-5-6	1-5-7
Double Degenerate Bend	1=5=9							
Torsion	2-1-5-6	3-1-5-6	4-1-5-6	"Int28"				
	2-1-5-7	3-1-5-7	4-1-5-7	"Int31"				
	2-1-5-8	3-1-5-8	4-1-5-8	"Int34"				

Figure 2.6. Various redundant internal coordinate schemes tested for RP-FM-CV simulations of the Menshutkin reaction in solution. The two distance-based CVs for bond forming and breaking, i.e., N-C (1-5) and C-Cl (5-9) (shown in red in the Bonds section), are used consistently in both the string MFEP simulations and the FM calculations.

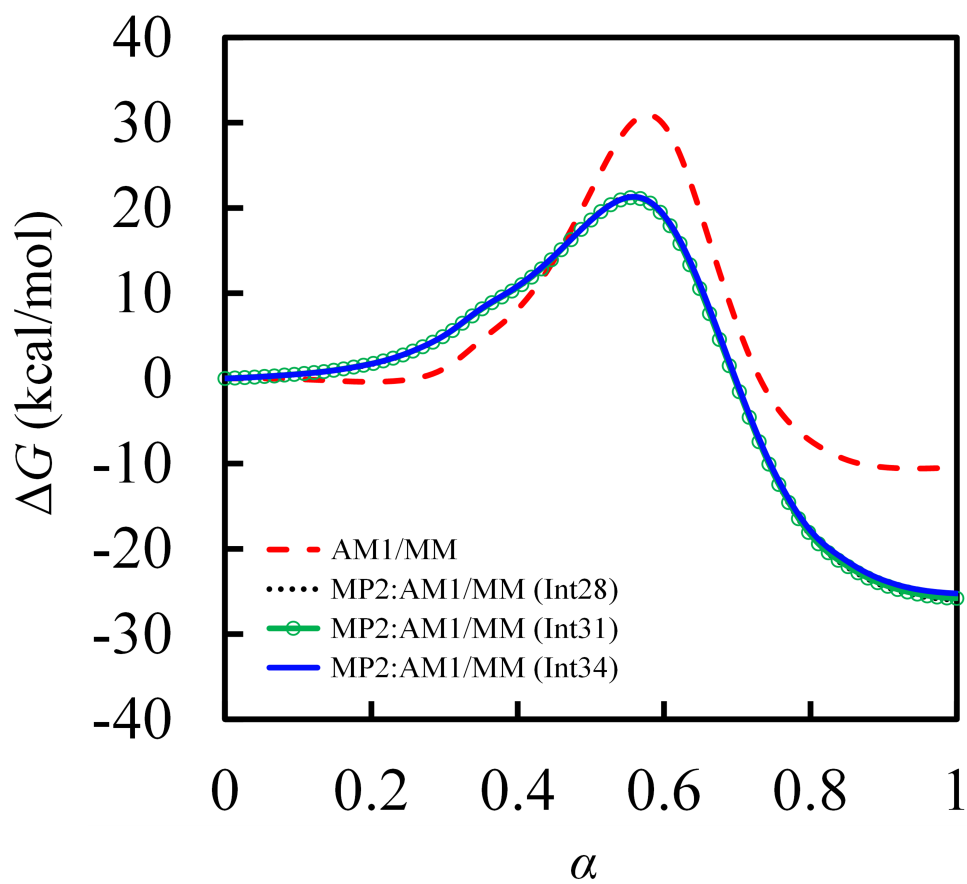


Figure 2.7. Free energy profiles for the Menshutkin reaction in solution obtained from the RP-FM-CV simulations at the MP2:AM1/MM level by using different redundant internal coordinate sets; the AM1/MM profile (dashed red) is shown for comparison. The RP-FM-CV results obtained using 28 (Int28; dotted black), 31 (Int31; solid green with circles), and 34 (Int34; solid blue) internal coordinates are shown (see Figure 2.6 for the definitions of the three schemes).

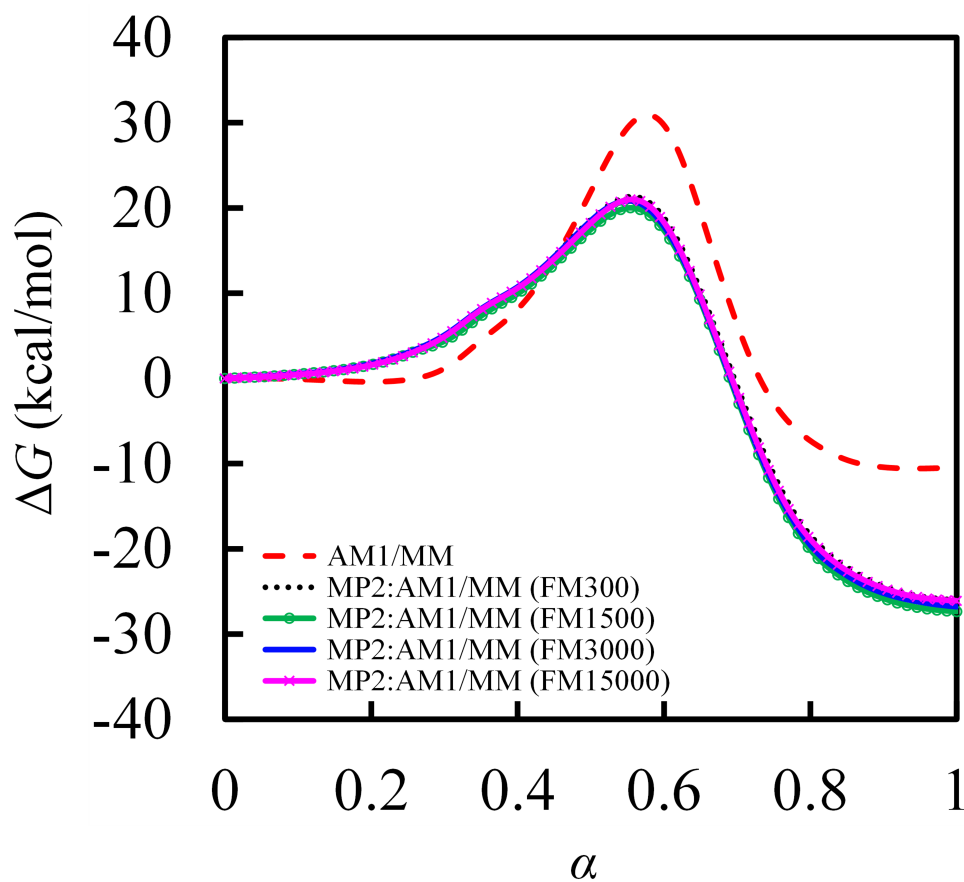


Figure 2.8. Free energy profiles for the Menshutkin reaction in solution obtained from the RP-FM-CV simulations at the MP2:AM1/MM level using different sample sizes for FM; the AM1/MM profile (dashed red) is shown for comparison. The RP-FM-CV results where FM in CVs were conducted using 300 (FM-300; dotted black), 1500 (FM-1500; green with circles), 3000 (FM-3000; solid blue), and 15000 (FM15000; pink with crosses) configurations are shown.

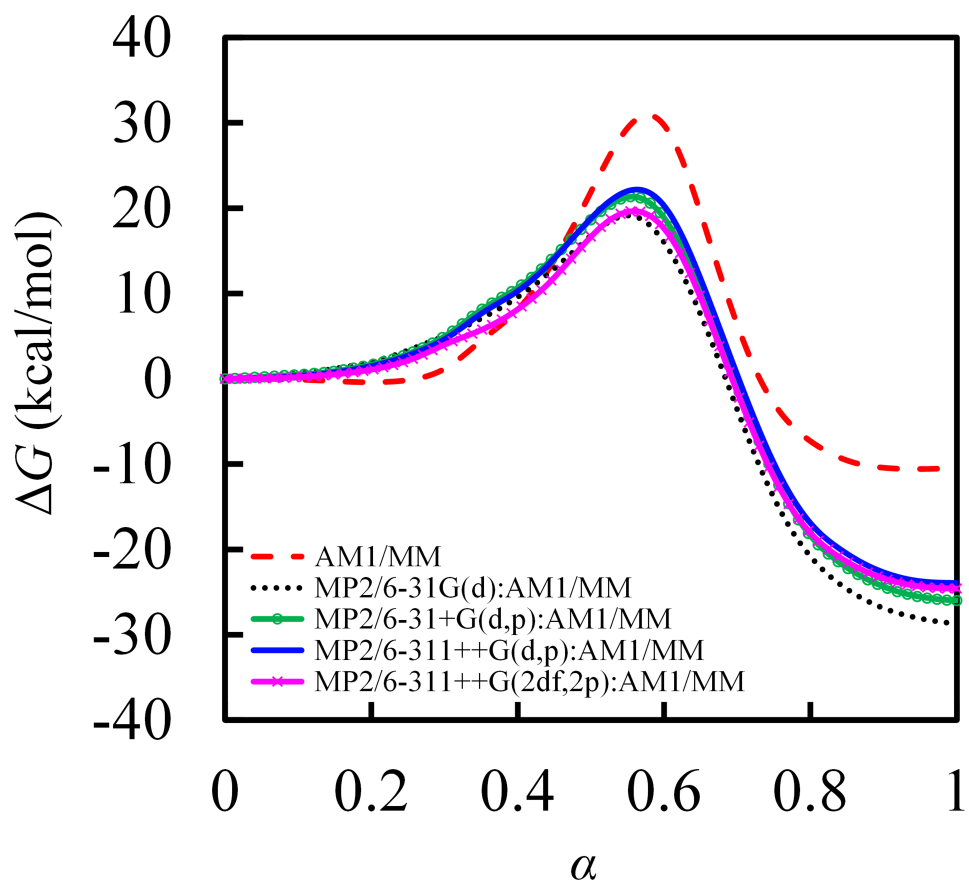


Figure 2.9. Free energy profiles for the Menshutkin reaction in solution obtained from the RP-FM-CV simulations by matching the AM1/MM forces on CVs to those determined at the MP2/MM level using various basis sets, including 6-31G(d), 6-31+G(d,p), 6-311++G(d,p), and 6-311++G(2df,2p).

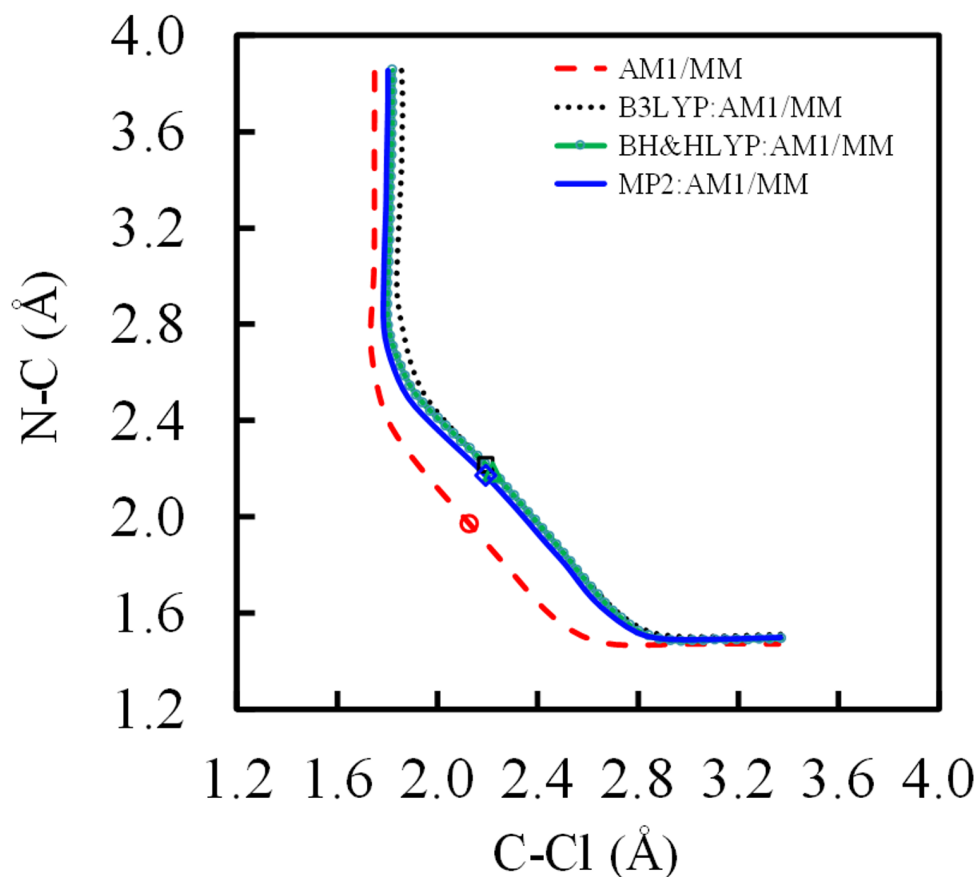


Figure 2.10. Minimum free energy paths (MFEP) for the Menshutkin reaction in solution for the free energy profiles shown in Figure 2.3. The MFEPs from the RP-FM-CV AI:AM1/MM simulations were obtained by matching AM1/MM forces on CVs to various target AI/MM levels using the 6-31+G(d,p) basis set: B3LYP:AM1/MM (dotted black), BH&HLYP:AM1/MM (green with circles), and MP2:AM1/MM (solid blue), compared with AM1/MM (dashed red). The transition states (TS) located on the MFEPs are also marked: B3LYP:AM1/MM (square), BH&HLYP:AM1/MM (triangle), MP2:AM1/MM (diamond), and AM1/MM (circle).

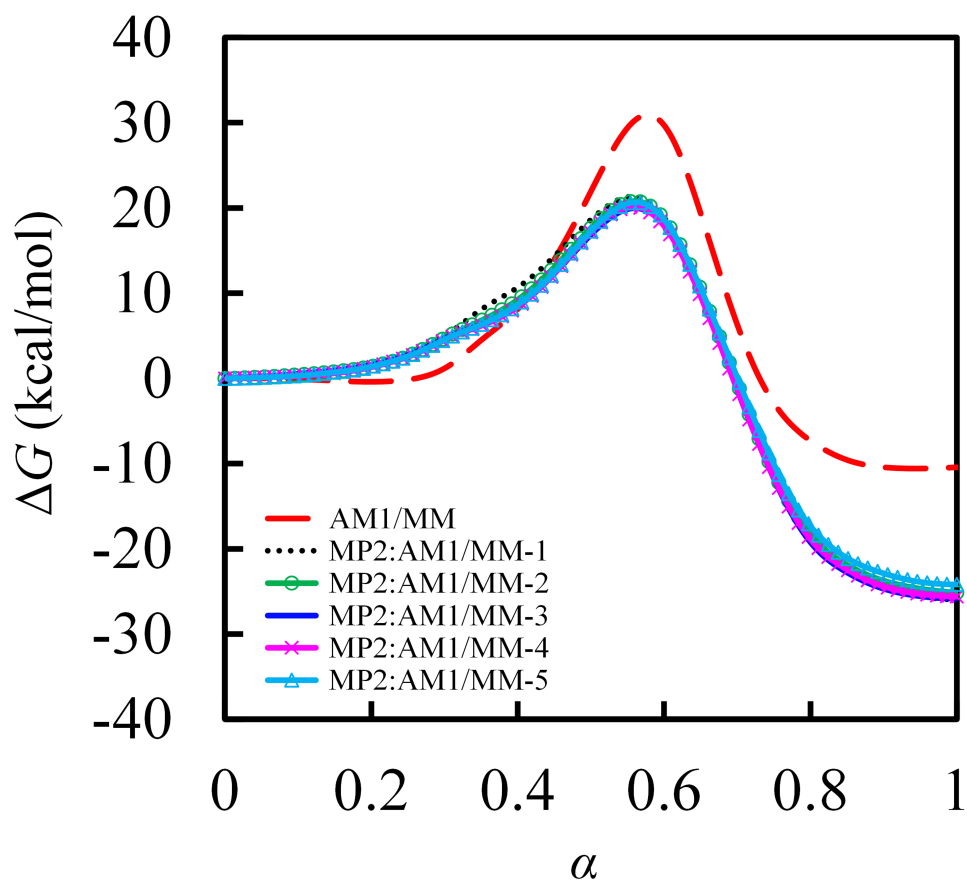


Figure 2.11. Free energy profiles for the Menshutkin reaction in solution obtained from the RP–FM–CV simulations at the MP2:AM1/MM level using 6–31+G(d,p) basis set over five consecutive RP and FM cycles: the 1st (dotted black), 2nd (green with circles), 3rd (solid blue), 4th (pink with crosses), and 5th cycle (light blue with triangles), compared with AM1/MM (dashed red).

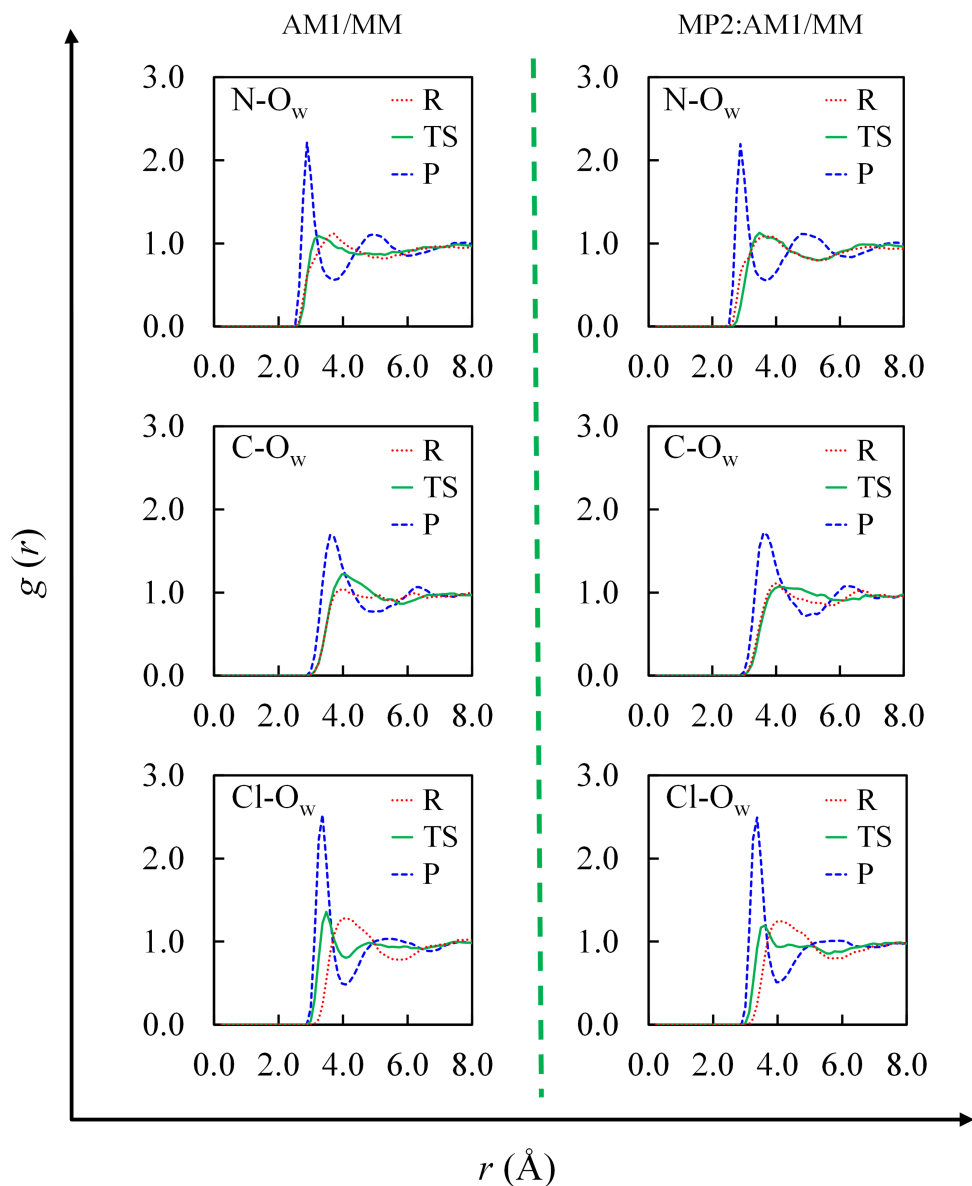


Figure 2.12. Solute-solvent radial distribution functions (RDFs) obtained from the RP-FM-CV simulations at the MP2:AM1/MM level using the 6-31+G(d,p) basis set, compared with the AM1/MM results. The RDFs for the solute (heavy atoms) and solvent (water oxygens: O_w) (i.e., N- O_w , C- O_w , and Cl- O_w) determined using an average of 3,600 configurations in each region are shown: reactant (R; dotted red), transition state (TS; solid green), and product (P; dashed blue).

2.8 Appendix

A. Force matching in CVs using spline functions

For FM in CVs, using a formalism that mimics the one used by Izvekov et al. [59] for Cartesian based FM, we define the objective function χ^2 as:

$$\chi^2 = \frac{1}{LN} \sum_{l=1}^L \sum_{i=1}^N \left| \Delta F_{il}^{\text{Ref}} - \Delta F_{il}^{\text{P}}(g_1^i, g_2^i, \dots, g_{m_i}^i) \right|^2 \quad (2.5)$$

where L denotes the number of sampled configurations for FM and N is the number of CVs for representing the MFEP; $\Delta F_{il}^{\text{Ref}}$ denotes the reference force correction needed for the internal force F on the i th CV in the l th configuration at the SE/MM level to match with the corresponding force at the target AI/MM level, i.e.,

$$\Delta F_{il}^{\text{Ref}} = F_{il}^{\text{AI/MM}} - F_{il}^{\text{SE/MM}} \quad (2.6)$$

Plugging Eq. (2.6) into Eq. (2.5) and then setting the objective function χ^2 to zero lead to the force matching condition:

$$\chi^2 = \frac{1}{LN} \sum_{l=1}^L \sum_{i=1}^N \left| F_{il}^{\text{AI/MM}} - F_{il}^{\text{SE/MM}} - \Delta F_{il}^{\text{P}}(g_1^i, g_2^i, \dots, g_{m_i}^i) \right|^2 = 0 \quad (2.7)$$

In Eqs. (2.5) and (2.7), $\Delta F_{il}^{\text{P}}(g_1^i, g_2^i, \dots, g_{m_i}^i)$ denotes the corresponding parametrized force correction term that is to be numerically determined for matching the internal forces between the SE/MM and AI/MM levels, where $(g_1^i, g_2^i, \dots, g_{m_i}^i)$ denotes a set of m_i parameters for fitting the force correction term for the i th CV. In the present work, we adopt a numerical treatment used by Voth and co-workers [59] in force-matching optimization of classical force fields, where the correcting force on each CV is expressed as a cubic spline function along evenly distributed grid points. Specifically, for the i th CV (of a bond-distance type) whose sampled values r^i fall in the interval of $[r_{\min}^i, r_{\max}^i]$, the corresponding spline function is defined as:

$$\begin{aligned}
\Delta F_{il}^P(g_1^i, g_2^i, \dots, g_{m_i}^i) &= f(r^i, \{r_k^i\}, \{f_k^i\}, \{f_k^{i''}\}) \\
&= A(r^i, \{r_k^i\})f_j^i + B(r^i, \{r_k^i\})f_{j+1}^i + C(r^i, \{r_k^i\})f_j^{i''} + D(r^i, \{r_k^i\})f_{j+1}^{i''} \\
&\quad (k=1, 2, \dots, n_{\text{grid}}^i; r^i \in [r_j^i, r_{j+1}^i])
\end{aligned} \tag{2.8}$$

where r_j^i denotes the position of the j th grid point over the radial mesh r_k^j consisting of n_{grid}^i grid points for the i th CV:

$$r_j^i = r_{\min}^i + (r_{\max}^i - r_{\min}^i) / (n_{\text{grid}}^i - 1) \times (j - 1) \quad (j=1, 2, \dots, n_{\text{grid}}^i) \tag{2.9}$$

and A , B , C , and D are derived quantities in cubic spline, [83] determined from the sampled CV value and its neighboring grid points, given that $r^i \in [r_j^i, r_{j+1}^i]$:

$$A = \frac{r_{j+1}^i - r^i}{r_{j+1}^i - r_j^i} \tag{2.10}$$

$$B = 1 - A = \frac{r^i - r_j^i}{r_{j+1}^i - r_j^i} \tag{2.11}$$

$$C = \frac{1}{6}(A^3 - A)(r_{j+1}^i - r_j^i)^2 \tag{2.12}$$

$$D = \frac{1}{6}(B^3 - B)(r_{j+1}^i - r_j^i)^2 \tag{2.13}$$

In Eq. (2.8), f_j^i and $f_j^{i''}$ denotes the parametrized force correction and its second derivative parameter with respect to the i th CV at the j th grid point of the spine function, respectively. Note that here we label these spline functions as the italic f to follow the literature convention, and they should not be confused with the Cartesian atomic force vectors which are labeled as bold non-italic in the text. As a result, the spline function described in

Eq. (2.8) contains altogether $m_i = 2n_{\text{grid}}^i$ adjustable parameters that need to be solved for the i th CVs FM condition, i.e.,

$$\Delta F_{il}^{\text{P}}(g_1^i, g_2^i, \dots, g_{m_i}^i) \equiv f\left(r^i, \{r_k^i\}, f_1^i, f_1^{i''}, f_2^i, f_2^{i''}, \dots, f_{n_{\text{grid}}^i}^i, f_{n_{\text{grid}}^i}^{i''}\right) \quad (2.14)$$

where

$$(g_1^i, g_2^i, \dots, g_{m_i=2n_{\text{grid}}^i}^i) \equiv \left(f_1^i, f_1^{i''}, f_2^i, f_2^{i''}, \dots, f_{n_{\text{grid}}^i}^i, f_{n_{\text{grid}}^i}^{i''}\right) \quad (2.15)$$

In FM, the numerical solution of Eq. (2.7) is obtained at a stationary condition that minimizes the objective function χ^2 with respect to the parameters $g_j^i (j = 1, 2, \dots, m_i)$:

$$\frac{d\chi^2}{dg_j^i} = \frac{1}{LN} \sum_{l=1}^L \sum_{i=1}^N 2 \left[\Delta F_{il}^{\text{Ref}} - \Delta F_{il}^{\text{P}}(g_1^i, g_2^i, \dots, g_{m_i}^i) \right] \left[-\frac{\partial \Delta F_{il}^{\text{P}}(g_1^i, g_2^i, \dots, g_{m_i}^i)}{\partial g_j^i} \right] = 0 \quad (2.16)$$

Using a short-hand notation:

$$\left(\Delta F_{il}^{\text{P}} \right)_{g_j^i}' = \frac{\partial \Delta F_{il}^{\text{P}}(g_1^i, g_2^i, \dots, g_{m_i}^i)}{\partial g_j^i} \quad (j = 1, 2, \dots, m_i) \quad (2.17)$$

as well as the specific functional form of ΔF_{il}^{P} defined in Eq. (2.8), we have:

$$\left(\Delta F_{il}^{\text{P}} \right)_{f_j^i}' = \left(\Delta F_{il}^{\text{P}} \right)_{g_{2j-1}^i}' = A \quad (2.18)$$

$$\left(\Delta F_{il}^{\text{P}} \right)_{f_{j+1}^i}' = \left(\Delta F_{il}^{\text{P}} \right)_{g_{2j+1}^i}' = B \quad (2.19)$$

$$\left(\Delta F_{il}^{\text{P}} \right)_{f_j^{i''}}' = \left(\Delta F_{il}^{\text{P}} \right)_{g_{2j}^i}' = C \quad (2.20)$$

$$\left(\Delta F_{il}^{\text{P}}\right)_{f_{j+1}^i}^{\cdot} = \left(\Delta F_{il}^{\text{P}}\right)_{g_{2j+2}^i}^{\cdot} = D \quad (2.21)$$

After rearrangement, Eq. (2.16) can be written as:

$$\sum_{l=1}^L \sum_{i=1}^N \left(\Delta F_{il}^{\text{P}}\right)_{g_j^i}^{\cdot} \Delta F_{il}^{\text{P}} = \sum_{l=1}^L \sum_{i=1}^N \left(\Delta F_{il}^{\text{P}}\right)_{g_j^i}^{\cdot} \Delta F_{il}^{\text{Ref}} \quad (j=1, 2, \dots, m_i) \quad (2.22)$$

Now consider a column vector \mathbf{g} as the union of all parameters for N collective variables,

$$(j=1, 2, \dots, m_i; i=1, 2, \dots, N) \quad (2.23)$$

where the superscript "T" denotes a transpose; the dimension M of the unionized parameter vector \mathbf{g} is:

$$M = \sum_i^N m_i \quad (2.24)$$

Then Eq. (2.16) can be written in a more compact matrix form:

$$\left(\Delta \mathbf{F}^{\text{P}}\right)_{\mathbf{g}}^{\cdot \text{T}} \Delta \mathbf{F}^{\text{P}} = \left(\Delta \mathbf{F}^{\text{P}}\right)_{\mathbf{g}}^{\cdot \text{T}} \Delta \mathbf{F}^{\text{Ref}} \quad (2.25)$$

With the following identity:

$$\Delta \mathbf{F}^{\text{P}} = \left(\Delta \mathbf{F}^{\text{P}}\right)_{\mathbf{g}}^{\cdot} \mathbf{g} \quad (2.26)$$

Eq. (2.25) can be written as:

$$\left(\Delta \mathbf{F}^{\text{P}}\right)_{\mathbf{g}}^{\cdot \text{T}} \left[\left(\Delta \mathbf{F}^{\text{P}}\right)_{\mathbf{g}}^{\cdot} \mathbf{g} \right] = \left(\Delta \mathbf{F}^{\text{P}}\right)_{\mathbf{g}}^{\cdot \text{T}} \Delta \mathbf{F}^{\text{Ref}} \quad (2.27)$$

This is equivalent to solving the parameters set \mathbf{g} for a linear equation system:

$$\left[\left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}} \left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}} \right] \mathbf{g} = \left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}} \Delta \mathbf{F}^{\text{Ref}} \quad (2.28)$$

Note that due to the overdetermined nature of Eq. (2.28) in FM, its numerical solution can be obtained by QR decomposition [83] or singular value decomposition (SVD), [83] with which Eq. (2.7) would be satisfied in a least square manner; for a perfect FM, the parametrized force correction in Eq. (2.28) would restore the reference force correction exactly:

$$\left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}} \mathbf{g} = \Delta \mathbf{F}^{\text{Ref}} \quad (2.29)$$

Therefore we identify Eq. (2.28) as the generalized key working equation for conducting FM in multidimensional CVs. Note that in Eqs. (2.25, 2.28, 2.29), $\Delta \mathbf{F}^{\text{P}}$ and $\Delta \mathbf{F}^{\text{Ref}}$ are both NL -dimensional column vectors, $\left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}}$ is an $NL \times M$ matrix with the leading dimension (i.e., number of rows) being NL , and \mathbf{g} is an M -dimensional column vector to be solved. For implementation purpose, consistency of Eq. (2.28) in its matrix form can be verified by analyzing the dimensionalities of the matrix operations involved:

$$(M \times NL)_{\left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}}} \times (NL \times M)_{\left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}}} \times (M \times 1)_{\text{g}} = (M \times NL)_{\left(\Delta \mathbf{F}^{\text{P}} \right)'_{\text{g}}} \times (NL \times 1)_{\Delta \mathbf{F}^{\text{Ref}}} = M \times 1 \quad (2.30)$$

For readers who are interested in more details about the implementation, a concrete example can be found in Section 2 of B.1, where we illustrate the matrix form of Eq. (2.28) for a two bond CV case based on a specific set of sample and grid distributions.

B. Determination of internal forces on CVs using redundant internal coordinate transformation

The transformation of forces from Cartesian to the selected redundant internal coordinates is conducted by using the procedure developed by Pulay and co-workers for geometry optimization. [43] Based on the Wilsons \mathbf{B} -matrix formalism, this procedure uses an eigenvalue decomposition technique to remove the linear dependence among the redundant

internal coordinates. The redundant internal coordinates and the Wilsons \mathbf{B} -matrix are connected through:

$$\mathbf{q} = \mathbf{B}\mathbf{X} \quad (2.31)$$

where \mathbf{q} is a set of N_R redundant internal coordinates (e.g., bonds, angles, torsions, doubly degenerate linear bends, out-of-plane wags, etc.), containing the CVs used in FM and the string MFEP simulations, \mathbf{X} represents the corresponding Cartesian displacement coordinates (in a dimension of $3n$, with n being the number of atoms involved in the coordinate system), and \mathbf{B} is the aforementioned Wilsons \mathbf{B} -matrix, [80] an $N_R \times 3n$ matrix accounting for the derivatives of the internal coordinates with respect to Cartesian displacement coordinates. For construction of the \mathbf{B} -matrix elements for bonds, angles, torsions, and out-of-plane wags, we follow the equations given in Wilson et al., [80] whereas for doubly degenerate linear bends, we follow the equations given in Califano [133] and an implementation by Jackels et al. [81] For force transformation from Cartesian to redundant internal coordinates, the \mathbf{B} -matrix is then used to form the condensed \mathbf{G} -matrix, which is an $N_R \times N_R$ dimension matrix defined as:

$$\mathbf{G} = \mathbf{B}\mathbf{u}\mathbf{B}^T \quad (2.32)$$

where \mathbf{u} is an arbitrary diagonal matrix (a $3n \times 3n$ identity matrix is used in the present work). Taking on the form of an eigenvalue equation, the condensed \mathbf{G} -matrix can be diagonalized as:

$$\mathbf{G}(\mathbf{K} \ \mathbf{L}) = (\mathbf{K} \ \mathbf{L}) \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad (2.33)$$

where \mathbf{K} is formed by $3n - 6$ eigenvectors of the \mathbf{G} -matrix that give non-zero eigenvalues corresponding to the diagonal elements of $\mathbf{\Lambda}$, and \mathbf{L} is the remaining $N_R - (3n - 6)$ redundant eigenvectors. In practice, to remove redundancy of the internal coordinate system, the \mathbf{L} eigenvectors in Eq. (2.33) are identified as the ones whose eigenvalues are below a pre selected

threshold; these numerically small eigenvalues are then set to zeros such that approximately $3n - 6$ largest eigenvalues are kept across the training samples. With \mathbf{K} , \mathbf{L} , and $\mathbf{\Lambda}$ in Eq. (2.33) determined, the generalized inverse of the \mathbf{G} -matrix, denoted \mathbf{G}^- , is constructed as:

$$\mathbf{G}^- = (\mathbf{K} \quad \mathbf{L}) \begin{pmatrix} \mathbf{\Lambda}^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{K}^T \\ \mathbf{L}^T \end{pmatrix} \quad (2.34)$$

where $\mathbf{\Lambda}^-$, represents the inverse of the non-zero eigenvalues, and $\begin{pmatrix} \mathbf{K}^T \\ \mathbf{L}^T \end{pmatrix}$ is the transpose of $\begin{pmatrix} \mathbf{K} & \mathbf{L} \end{pmatrix}$. Once the \mathbf{G}^- -matrix becomes available, the internal forces on the CVs in the redundant internal coordinates can be conveniently determined by the following transformation:

$$\mathbf{F} = \mathbf{G}^- \mathbf{B} \mathbf{f} \quad (2.35)$$

where the lower case \mathbf{f} is the Cartesian atomic forces obtained from conventional QM/MM simulations and \mathbf{F} represents the internal forces determined in the user-defined redundant internal coordinates \mathbf{q} . In RP-FM-CV, because the CVs form a subset of the redundant internal coordinate \mathbf{q} , this transformation procedure is used to obtain the internal forces on the CVs at both the SE/MM and AI/MM levels, which are subsequently used to determine the force corrections needed to match the internal CV forces at the two levels.

3. ACCURATE FREE ENERGY PROFILES IN CHEMICAL REACTIONS: A QM/MM STUDY OF THE ROLE OF PAIRWISE REPULSIVE CORRECTING POTENTIALS IN FORCE MATCHING

In this study, we examine the role of pairwise repulsive correcting potentials in generating converged free energy profiles for chemical reactions. Using a quantum mechanics/molecular mechanics (QM/MM) approach, we explore the accuracy of force matching with the pairwise RP-FM model in chemical reactions. The results of our study provide insights into the underlying framework for force fitting and help to establish a criterion for determining the strength and quality of future studies of chemical and biochemical solution phase reactions. By decomposing the contribution of free energy on each collective variable for asymmetric and symmetric reactions, we develop a free energy correcting model that sheds light on the behavior of repulsive pairwise potentials with large force deviations in collective variables. Our findings contribute to a deeper understanding of force matching models and pave the way for more accurate predictions of free energy profiles in chemical reactions.

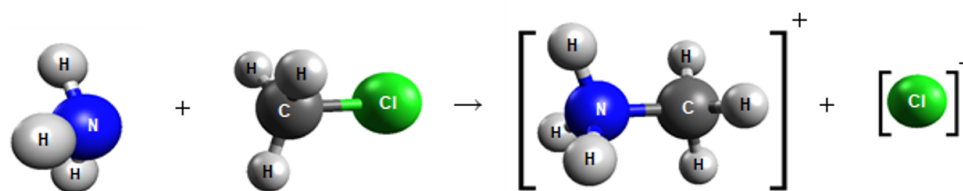
3.1 Introduction

A central problem within the fields of computational chemistry and computational enzymology is the daunting costs for sampling accurate ab initio quantum mechanical molecular mechanical (QM/MM) potential energy surfaces (AI/MM) for solution phase reactions. Alternatively, efficient semiempirical QM/MM (SE/MM) methods can be used to compute free energy profiles (FEPs), at a dramatic reduction in computational cost. However, efficiency in SE/MM methods is achieved at the expense of reliability and accuracy. Therefore, to address the challenge of achieving highly efficient and accurate methods for QM/MM free energy simulations, we study a recently developed reaction path force matching (RP-FM) model, reaction path-force matching in collective variables (RP-FM-CV) [26]. In the context of the original RP-FM approach [28], SE/MM forces are systematically fit to AI/MM forces based on specific reaction parameters along the free energy reaction path [22], [43],

[59], [78]–[80]. In the new RP–FM–CV model, SE/MM internal forces on select collective variables (CVs) are fit to target AI/MM methods. As a result of RP–FM, computations for FEPs, reaction path, transition state structures, etc. are improved. Contrary to internal force fitting in CVs, we study the recently developed RP–FM model based on pairwise atomic interactions [71]. Since this work builds on the previously developed RP–FM framework, we compare implementations of this method to RP–FM–CV. In general, the optimization of low to highly accurate method levels makes the RP–FM strategy ideal for studying chemical reactions where high-level AI/MM computations are infeasible. In terms of computational cost RP–FM reduces the brute-force sampling on AI/MM surfaces by several orders of magnitude, since expensive AI/MM sampling is avoided. Force fitting for complex molecular systems is challenging, therefore we investigate the pairwise approach on the classic Menshutkin reaction [45].

The Menshutkin reaction is a type II reaction between neutral ammonia and methyl chloride, which undergoes simultaneous bond breaking and formation of C–Cl and N–C respectively (Scheme 3.1). The neutral reactants in the reaction undergo charge separation which is destabilized in the gas phase, but in polar solvents, a favorable ion dipole interaction exists. The protection of partial charges by the solvent dipole generates a favorable solvation energy which lowers both the barrier and solvation free energies. The solution phase reaction has an experimental free energy barrier (ΔG^\ddagger) of 23.5 kcal/mol and an experimental free energy of reaction (ΔG) of –34 kcal/mol [49]. When SE/MM methods such as AM1/MM [24] are used to simulate this reaction in solution, an overestimated ΔG of –10.7 kcal/mol and an overestimated ΔG^\ddagger of 30.8 kcal/mol is obtained. As such, this reaction serves as a good prototype reaction for implementing the RP–FM strategy to address the central problem of obtaining reliable FEPs in an efficient manner.

Conceptually, the RP–FM strategy is defined as fitting a correction term for improving a low-level method to match the forces of a high-level target method. Regardless of method level, the strategy covers Cartesian to internal coordinate transformations, conformational changes, and electron transfer studies in CVs. The RP–FM strategy can thus be separated into two distinct approaches. The first RP–FM approach can be modeled by correcting the forces without explicitly defining the potential [34], [71]. In this approach, the potential of



Scheme 3.1. Schematic Representation of the Menshutkin Reaction ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$)

mean force (PMF) is constructed by integrating forces on a reaction coordinate (RC) for only select important degrees of freedom. A recent study with the RP-FM-CV approach was used to simulate the Menshutkin reaction by fitting the mean internal forces as a spline interpolation function [59] over a two-dimensional RC. At the B3LYP 6-31+G(d,p) method level [117], RP-FM (B3LYP/MM) yielded a reaction free energy and barrier of -26.9 and 14.5 kcal/mol, respectively (Figure 3.1).

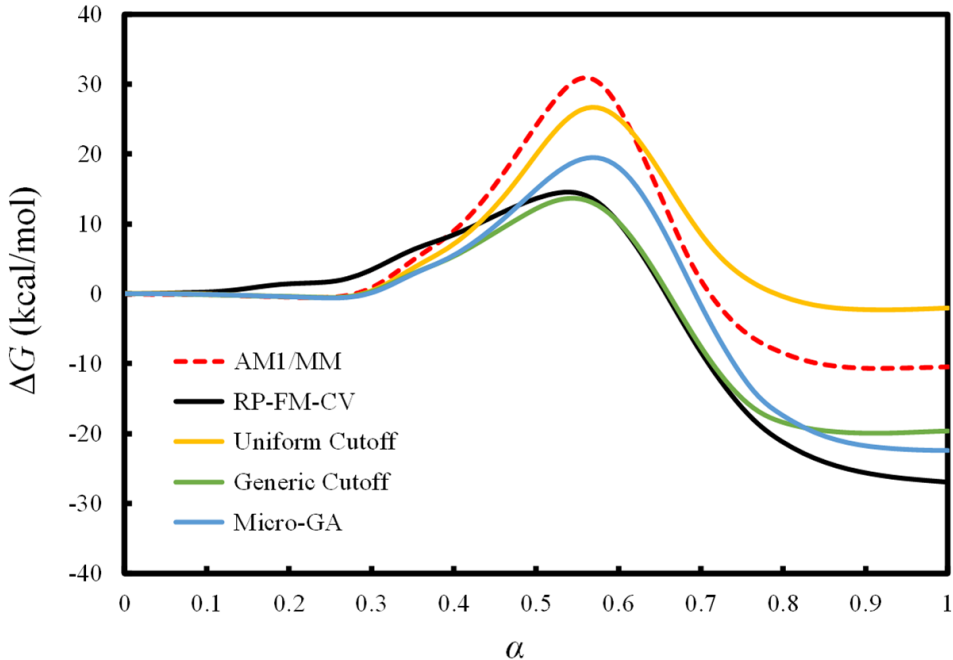


Figure 3.1. Potential of mean force for the Menshutkin reaction in aqueous solution for AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes.

A second RP-FM approach uses an explicitly defined correcting potential based on pairwise interatomic distances or machine learning (ML) methods. However, when using these methodologies, an error cancellation cannot be determined for the correcting potential since the correcting terms for the spectator degrees of freedom are collected. Recently, Wu et al. applied a ML QM/MM approach on the Menshutkin reaction to perform a potential based force correction along the RC. Here a one-dimensional RC was used to construct a distance-based ML potential by fitting the internal forces with a neural network. The ML potential

in this study was parameterized to fit the forces from the HF/G-31G(d) method to yield ML corrections with a reaction free energy and barrier of -23.8 and 20.8 kcal/mol, respectively. Currently, no benchmark criterion is available for variations of the original RP-FM strategy, therefore, we study a recently developed pairwise potential RP-FM method to compare to the explicit force sampling method to gain better insight for pairwise RP-FM and to develop a new tool for assessing the accuracy and robustness of RP-FM models.

Recently, a force matched pairwise potential computational model (pairwise RP-FM) was developed by Kroonblawd et. al, by using the RP-FM strategy. This model calculates a correcting force by differentiating a well-defined correcting potential. The functional form of this correcting potential is constructed by representing the pairwise atomic interactions as an approximated power series where parameters are linearly combined with correcting forces. As a result of linearity, the parameters are efficiently determined by minimizing the linear least-squares objective function by singular value decomposition. By computing the single point correcting force for target AI/MM and base SE/MM methods along the reaction path, a correcting potential can be applied to correct the atomic forces to a higher level of accuracy. For the pairwise study, a force matched semi-empirical SCC-DFTB model was developed to better understand polypeptide synthesis in prebiotic environments. As such, the pairwise RP-FM approach was applied to the glycine condensation reaction where reaction free energies of 15.3, 12.4, and 16.8 kcal/mol were computed for equal-weight, weight-by-type, and weight-by path models. Thus, the pairwise potential based model demonstrates the capability of providing qualitative differences for chemical reactions by using the RP-FM strategy.

In principle, fitting a pairwise correcting potential is reliable since an explicitly well-defined potential is constructed for all of the atoms in the system. This is advantageous since differentiation of the potential is not only linear with respect to the parameters, but this allows the Cartesian forces to directly be routed back to the potential. For the pairwise RP-FM study the sensitivity for the all atom RP-FM DFTB model was assessed by computing cross-validation root mean-square errors (RMSE) for total forces. For the glycine condensation reaction, solvent atoms are not differentiated from solute atoms. Therefore, the abundance of solvent molecules leads to a dramatic reduction in RMSE for oxygen and

hydrogen atoms due to overfitting of these atoms, having deviations of ~ 4 kcal/mol/Å. However, the RMSE for important solute atoms remains quite large, having deviations of ~ 5.5 and ~ 7 kcal/mol/Å for nitrogen and carbon, respectively. Therefore, we propose that the RMSE on all atoms is not sufficient for assessing the accuracy and robustness for a RP-FM model. The overfitting and indifference of atomic pairs within the system creates a certain degree of bias along the path since artifacts from spectator atoms lead to over correction of the forces. Therefore, the criteria for a RP-FM model based in CVs, is more meaningful since the forces are unbiased within the reaction region whereas all atomic force fitting classically corrects the forces as a force field. Secondly, special attention must be taken for the selection of radial cutoff distances for atomic pairs due to the fluctuating FEPs from varying combinations of radial cutoff distances. In the pairwise study, the radial cutoff distances from the mio-1-1 parameter sets were used without modification to avoid time-consuming nonlinear optimizations, but these radial cutoff distances are system specific where short distances truncate sampling to exclude important long-range reactant, transition and product state regions of the reaction. Such effects lead to increased deviation on solute atoms and ultimately inaccurate PMFs. In this study, the semi-empirical SCC-DFTB model was force matched to a benchmark DFT method, to yield barrier and reaction free energies deviating ~ 10 kcal/mol from the target AI/MM method. The argument made in this study is that an improvement is made for the PMF since the reported results matches experimental results. However, the goal of the RP-FM strategy is to efficiently obtain reliable FEPs by RP-FM to infeasible higher levels of theory as opposed to agreeing with experimental results. Finally, the pairwise study describes the glycine condensation reaction as a linear process where the proton from the amine acid of the reactant glycine is directly transferred to the hydroxyl base of the product glycine with water as a byproduct. However, an argument that is made is that the proton transfer in glycine condensation is a complex mechanism, since proton hopping between zwitterionic and neutral states is a collective process. In the current implementation, the RC is not continuous since the path omits a potential shuttling of the proton via water chain and other potential reaction pathways through zwitterionic and neutral intermediate species. The nonexistence of such sampling could lead to biased fitting which is another potential source of error. In the pairwise RP-FM study, the free energy for

the glycine condensation study is mapped based on the coordination number of the nearest neighbor atoms, however fitting only important degrees of freedom involved in the bond breaking and forming processes, is a system specific alternative for obtaining meaningful representations of the free energy.

To gain an intuitive understanding of the pairwise RP-FM method used in the glycine condensation study, we implement the potential fitting procedure to the more straightforward Menshutkin reaction, used in RP-FM-CV. At the core, RP-FM-CV differs from pairwise RP-FM since the forces are not differentiated from a well-defined potential. Instead, the free energy is obtained by integrating over the internal forces in CVs along the reaction path. One criticism of RP-FM-CV is the coordinate dependence in CVs since a coordinate transformation type utility is employed. This contrasts the pairwise RP-FM method where fitting in CVs is inconsequential since all atomic forces are fit. However, since the FEPs are based on the RC, only important degrees of freedom, the breaking and forming bonds, are needed for computing accurate PMFs [77]. Therefore, in the RP-FM-CV method, only internal forces in CVs are fit to directly correct base SE/MM to target AI/MM levels. Similarly, pairwise RP-FM should restore AI/MM accuracy by fitting forces in CVs along the reaction path. Another difference between RP-FM-CV and the pairwise study is that target forces are partially generated with expensive DFT sampling, in contrast to RP-FM-CV where sampling is explicitly obtained from the affordable AM1/MM level. In principle, a combination of both types of utilities may be required to further improve accuracy and efficiency of computed PMFs, but we see advantages in parameterizing both a well-defined potential, in addition to CV based RP-FM with spline corrections.

Since the glycine condensation reaction is complex, we are verifying the pairwise RP-FM method by condensing the system to a simple, two-dimensional, Type II SN2 Menshutkin reaction. We assess the quality of the newly developed RP-FM model by first comparing the PMFs obtained from each model. With this rationale we argue that all atomic force fitting is unnecessary for obtaining reliable and accurate free energies. Also, we compare the deviation of internal forces in CVs to verify whether proper force fitting is performed on important degrees of freedom. With this study, we aim to clarify whether correcting potentials as a polynomial function and mean force in CVs are procedures capable of obtaining converged

results by using the same base and target method. In terms of free energy and geometries [94], converged results and reduced force deviations in CVs is a strong argument that internal RP-FM in CVs is sufficient for studying reactive chemical systems. In our study, we use B3LYP/MM as the target AI/MM method and the AM1/MM as the base SE/MM method. Although hybrid DFT methods are shown to underestimate the barrier height [53], the choice in method level is arbitrary since the focus of this study is obtaining converged RP-FM results. In this study, we investigate the implementation of a newly developed RP-FM model in the Menshutkin reaction, and we assess the benefits and risks to develop a criterion for measuring the fitness and robustness of future RP-FM models. An overview of the methods is provided in section 3.2, with computational details provided in section 3.3 results provided in section 3.4, and a conclusion is given in section 3.5.

3.2 Methods

The pairwise RP-FM model fits forces from a polynomial based pairwise potential to form a restoring correcting potential [70]. In this instance, the polynomial of the correcting potential ($E_{CP}^{\alpha\beta}$) is expressed as follows:

$$E_{CP}^{\alpha\beta}(r_{ij}) = \left\{ \sum_{n=2}^9 c_n^{\alpha\beta} (r_c^{\alpha\beta} - r_{ij})^n \right\} \quad (3.1)$$

where r_{ij} is the separation distance between atom i of pair type α and atom j of pair type β and where $r_c^{\alpha\beta}$ and $c_n^{\alpha\beta}$ are the radial cutoff distances and coefficients to be determined later. Here, $r_{ij} \leq r_c^{\alpha\beta}$ otherwise the radial cutoff distance is set to 0. Correcting potentials were fit for N-H, N-C, N-Cl, H-H, C-H, C-Cl, and Cl-H as shown in Eq. S1 in B.2) of the supplementary materials. Since there are 8 parameters for 7 atom pair types, we fit a total of 56 parameters. In the previous implementation of pairwise RP-FM, solute atoms are not differentiated from solvent atoms, thus recurring pairs in both solvent and solute atoms, are pairs which contributes equal potential. In the modified model, solvent atoms are removed from the fitting procedure since free energy is directly computed from the reaction in CVs. To obtain the functional form for the correcting potential, a design matrix (**B** matrix) is

constructed from pairwise forces from molecular dynamics (MD) as shown in Eqs. (S2–S4) in B.2 of the supplementary materials. Using the chain rule to sum together the interatomic pairwise forces, AI/MM accuracy is restored from the force matched forces of the correcting potential. Since the gradient is linear with respect to the parameterized coefficients, $c_n^{\alpha\beta}$, a linear least squares expression for force correction can be minimized by using singular value decomposition, as shown in Eq. 3.2.

$$\chi^2 = |A \cdot c - b| \quad (3.2)$$

In essence, the A is constructed from the \mathbf{B} matrix which is a proper construction of the differentiated Cartesian forces evaluated for each individual element (Figure S2 and S3 of in B.2 in the supplementary materials). Furthermore, c represents the correction term, where b is the difference between the target and base level Cartesian forces.

The foundation of the RP–FM strategy is to fit efficient many–body PMFs to reproduce target AI/MM accuracy. By generating a one–dimensional cut of the CV based potential energy surface, forces are collected on an initial RC to obtain the minimum free energy path (MFEP). This increases the statistical integrity of the FEP, since averaged samples are collected over the reaction path. Although it is possible to strenuously fit all the atomic forces, it is not necessary for reproducing the AI/MM mean force in CVs. By using the finite string method in CVs, we express our free energy mean force $\langle F \rangle_\xi^*$ as a set of generalized coordinates q as shown in Eq. 3.3.

$$\langle F \rangle_\xi^* = \frac{\int dq dp_q dp_\xi F(q_1, q_s)}{\int dq dp_q \dots dp_\xi \exp\left(-\frac{H}{k_B T}\right)} \quad (3.3)$$

The instantaneous force $F(q_1, q_s)$ can further be expressed by Eq. 3.4.

$$F(q_1, q_s) = \frac{\partial \ln |J(q_1, q_s)|}{\partial q_1} - \beta \frac{\partial U(q_1, q_s)}{\partial q_1} \quad (3.4)$$

Where J is the Jacobian matrix that transforms the Cartesian to the generalized coordinate, and where the second term on the right-hand side $\frac{\partial U(q_1, q_s)}{\partial q_1}$ denotes the force on the select CVs. However, in this formalism there is an explicit dependence on the partial derivative q_s . In order to remove such dependency, only coordinates orthogonal to the CVs are selected, following the rationale of Ruiz-Montero et al. in Eq. 3.5.

$$F(q_1, q_s) = -\frac{(\nabla |\nabla q_1|) \cdot \nabla q_1}{|\nabla q_1|^3} - \beta \frac{\nabla U \cdot \nabla q_1}{|\nabla q_1|^2} \quad (3.5)$$

Here, ∇ denotes the first derivative operator with respect to Cartesian coordinates and $\frac{(\nabla |\nabla q_1|) \cdot \nabla q_1}{|\nabla q_1|^3}$ denotes the transformation term of the coordinates from the sampled configurations. Since configurations from the Jacobian term are shared in both SE/MM and AI/MM calculations, these terms cancel out. As such, RP-FM is applied to the Cartesian based mechanical force on the potential energy surface term, $-\nabla U$. Therefore, we say RP-FM restores the Cartesian AI/MM forces, since CV based high-level AI/MM forces are fit to restore the reaction FEP. From RP-FM, potential-based force correction terms are obtained for use in free energy computations. In the linear least squares fitting procedure for pairwise RP-FM, force fitting in CVs is linear, therefore force correction terms for each pair type between SE/MM and AI/MM levels in CVs, is determinable by singular value decomposition. These modified potential-based force corrections are then incorporated as additional force components on original SE/MM forces for the CVs in a repeat MD simulation. As a result, the incorporation of force matched terms, leads to an updated FEP which converges the base SE/MM to AI/MM level accuracy.

3.3 Computational Details

Using standard CHARMM force fields, van der Waals parameters are computed between the QM and MM atoms in a combined QM/MM treatment of the reaction in solution [49], [85]. In the solution reaction, the QM treatment of the solute, is calculated using AM1 and the MM treatment of the solvent is performed on the modified TIP3P model [108]. CHARMM program (version c42a2) was used for the AM1 calculations. In addition, the

Q-Chem package (version 4.0.1) interfaced with CHARMM was used to compute forces for B3LYP hybrid nonlocal density functional theory (DFT) with the 6-31+G(d,p) basis set. To describe the MFEP for the $\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$ reaction, we adopted two bond distances for the CVs, defined as the breaking bond (N-C) and the forming bond (C-Cl), as shown in Scheme 1. This RC was used in obtaining the solution-phase reaction paths within a cubic periodic box of $40 \times 40 \times 40^3$ of modified TIP3P water. The SHAKE algorithm [112] was then used to constrain the internal geometries of the waters during the MD simulations, and long-range electrostatics for MM/MM and QM/MM interactions were treated by Particle Mesh Ewald. For real-space contribution of QM/MM-Ewald electrostatics, a switching function is applied between 12 to 13 Å to smoothly attenuate the real-space QM/MM interaction at a radial cutoff distance of 13 Å. A collinear C_{3v} geometry was imposed in CHARMM on the solute molecules during the MFEP samplings to reduce system distortion to irrelevant configurations that could potentially slow down MFEP convergence [53]. Also, to avoid systematic solute drifting towards the edge of the simulation box, we added a SBOUND-like one-sided quartic sphere repulsive potential. Using the following conditions, a constant pressure of 1 atm and a constant temperature of 298.15 K, was simulated through the use of MD. The MFEP is discretized into 25 images, with initial coordinates obtained from QM/MM potential energy scans along the RC. The mean force in CVs were estimated by simulating for 20 ps sampling on restrained QM/MM MD simulations with CVs restrained at the path values with a harmonic force constant of 1,000 kcal/mol/Å. The projection, reparameterization, and evolvement of the paths, as well as integration of the FEPs follows the string method developed by Maragliano et al., where convergence of the MFEP is monitored by fluctuations on CVs through subsequent iterations.

3.4 Results and Discussion

3.4.1 Optimizing Radial Cutoff Distances for Pairwise Potential and Force Correction

The pairwise potential for each atomic pair type in the Menshutkin reaction was first constructed with a uniform radial cutoff distance of 2 Å, to avoid time-consuming nonlinear

optimizations. For this study, we use the pairwise potential on solute atoms instead of solvent coordinates since the PMF is implicitly affected by the solvent and since free energy is directly related to the QM coordinates. Furthermore, with a radial cutoff distance of 2 Å, the H–H pair was removed from sampling to avoid SCF failure from the large correcting potential (Figure S4 in B.2) of the supplementary materials). Likewise, hydrogen pair removal in the previous pairwise study is adopted to avoid poor sampling. As a side remark, careful construction for the pairwise potential is recommended to avoid collapsing the quantum mechanical system from large correcting potentials. Compared to the SE/MM method, pairwise RP–FM with a uniform radial cutoff scheme, increases the reaction free energy by 8.4 kcal/mol and reduces the barrier by 4.2 kcal/mol (Figure 3.1). In Table 3.2, the RMSE of atoms are listed, and in Figure 3.2 the atomic deviations are shown to be reduced by 9.0, 7.3, 14.0, and 6.9 for N, H, C, and Cl respectively, where overall RMSE for uniform radial cutoff distances is reduced from 14.5 to 6.4 kcal/mol/Å.

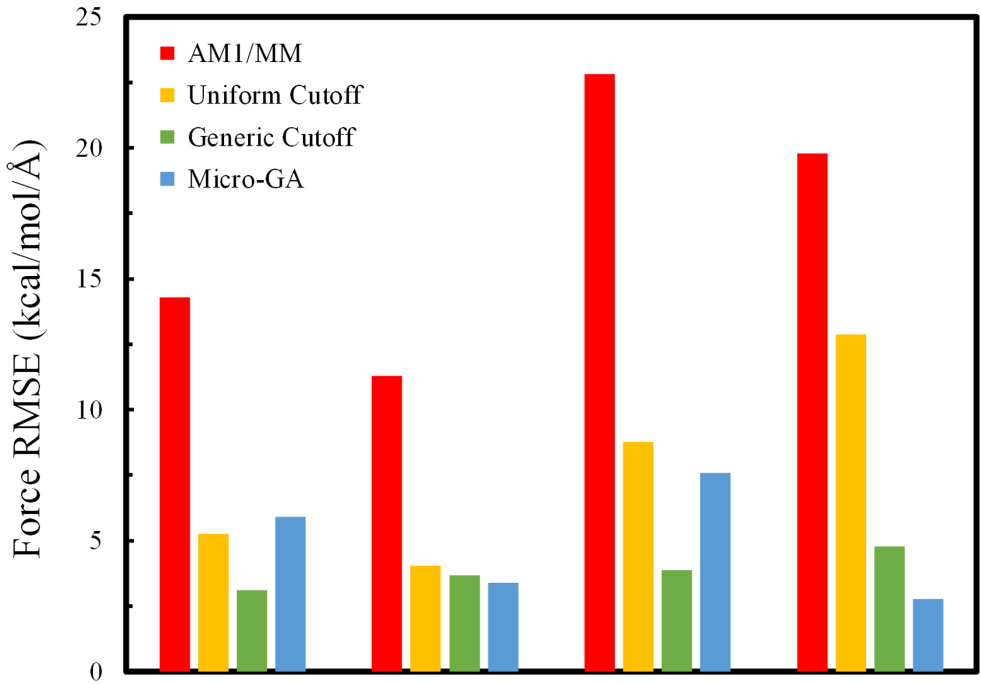


Figure 3.2. Cross-validation root mean-square errors of atom type for AM1/MM (red), and B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes.

The free energy barrier is thus similar to the original base method whereas the increase in free energy is markedly different from the expected reduction from solvation free energy. As shown in Figure 3.3, a uniform radial cutoff distance results in a biased force correction along the reaction path, where reactant and product state forces are sufficiently fit below 5.0 kcal/mol/Å in contrast to transition state samples, with deviations as large as 12.2 kcal/mol/Å (Figure 3.3).

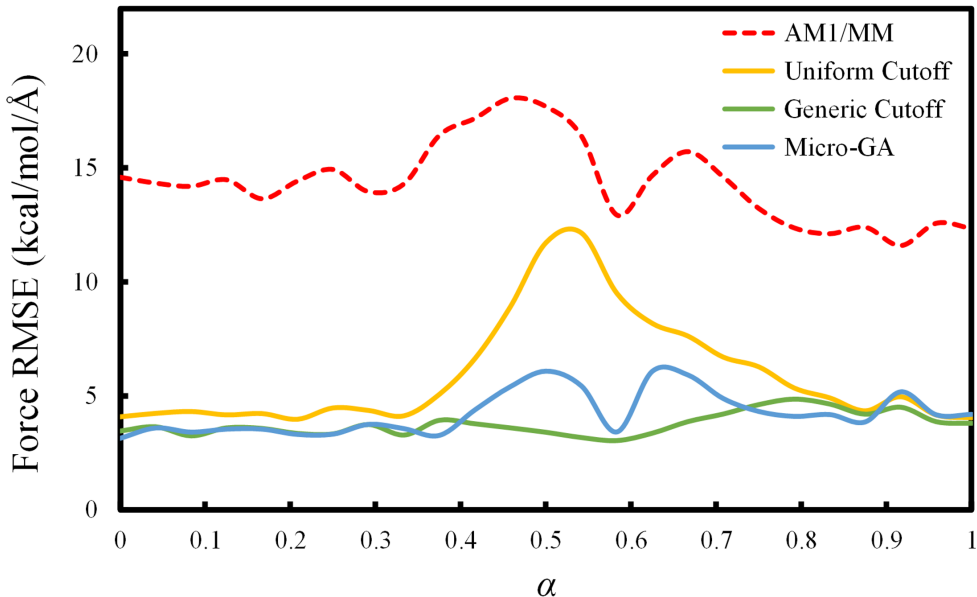


Figure 3.3. Cross-validation root mean-square errors along the reaction path for AM1/MM (red), and B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes.

Since the PMF is directly related to the forces on the CVs we computed the internal force correlation between SE/MM and AI/MM methods, with correlations of 6.1 and 17.2 for the N-C and C-Cl bonds respectively (Figure 3.4).

The large reduction in RMSE in Figure 3.2 for hydrogen explains the overall reduction in deviation since the hydrogen biases the error at the expense of the Cl atom. While RMSE on atoms is lower than force correlations in CVs, this is not reflective of the proper force correction since spectator degrees of freedom are overfit. Thus, the overall lower error for atoms with a uniform radial cutoff of 2 Å is inadequate for determining the quality of RP-

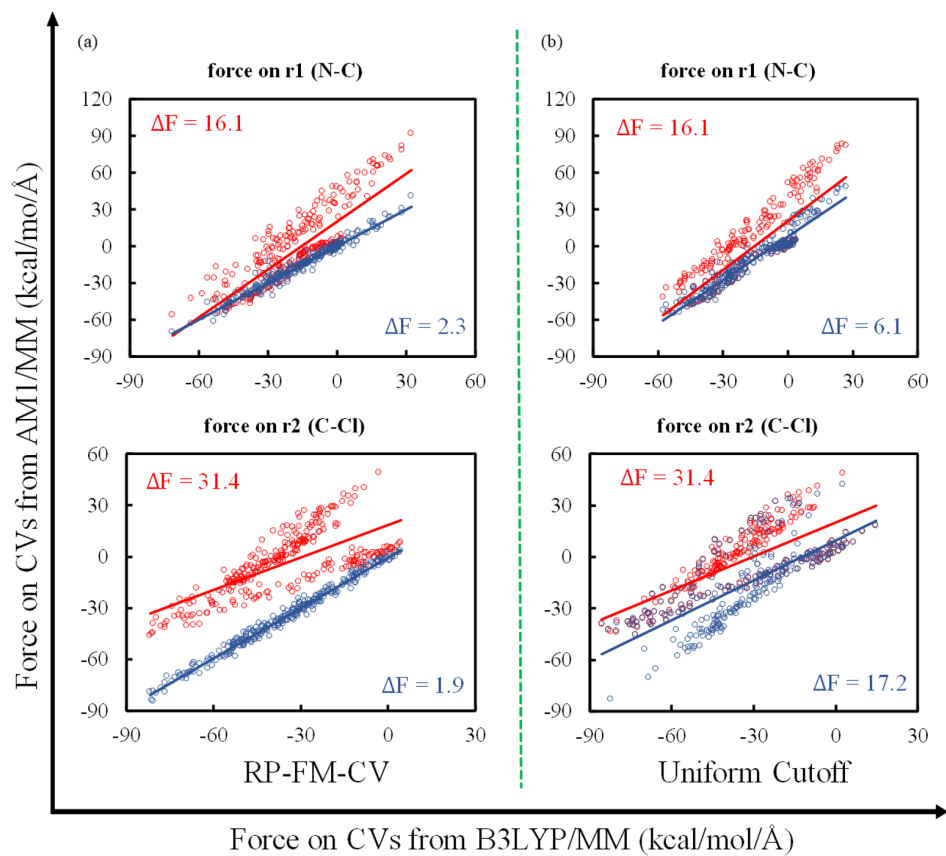


Figure 3.4. Force correlation in CVs before (red) and after (blue) B3LYP force matched corrections for: (a) RP-FM-CV and (b) uniform radial cutoff schemes.

FM since large deviations exist for the corrected forces in CVs. Furthermore, the quality of fitting is directly dependent on the length of the radial cutoff distance, where shorter radial cutoff distances results in truncated sampling (Figure 3.5).

For B3LYP/MM target forces, the N-C force correction is depicted as having a signature dip in force correction followed by a slight increase in barrier where C-Cl has an initial dip followed by a slow rise. Thus, limited sampling is obtained since important samples from reactant and product state structures from N-C and C-Cl are respectively omitted. Therefore, sufficient sampling is required for parameterizing the reaction path with statistical integrity. Conversely, increasing the radial cutoff to distances beyond the transition state introduces artificial behaviors which extrapolates the force correction in CVs (Figure S5 in

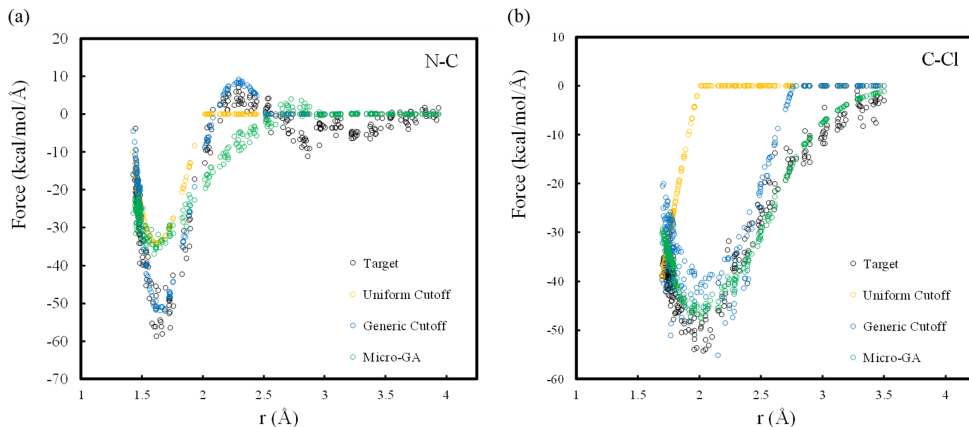


Figure 3.5. Distance-based force corrections for B3LYP/MM force matched uniform (yellow), generic (green) and micro-genetic algorithm (blue) radial cutoff schemes in collective variables for: (a) N-C and (b) C-Cl.

B.2) of the supplementary materials). In the previous pairwise study, radial cutoff distances approximately equal to 2 Å are predefined from the mio-1-1 parameter set to avoid nonlinear optimization. In contrast to using a predefined parameter set, a uniform set of radial cutoff distances are not sufficient, so the parameterized functions are not linear with respect to the gradient. As such, a more specific criterion is required for selecting the optimal range of radial cutoff distances for each atom pair type.

In the pairwise RP-FM study, radial cutoff distances were chosen from the mio-1-1 parameter set without modification. Using a uniform radial cutoff distance of 2 Å reduces the overall RMSE but the PMF is not improved. As an alternative, a generic radial cutoff scheme ($r_{generic}^c$), listed in Table 3.1, was used for each atom pair type based on the equilibrium distances between two atoms in a pair [134] plus 1.0 Å in Eq. 3.6.

$$r_{generic}^c = r_i + r_j + 1.0 \text{ Å} \quad (3.6)$$

Since the N-Cl bond distance is composed of both N-C and C-Cl pairs, pair distances were instead substituted for atoms in Eq. 3.6. For simplicity, an arbitrary distance of 1.0 Å is used to extend the radial cutoff length to sample non-equilibrium configurations beyond

Table 3.1. Reaction Barrier/Free Energy of $\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{CH}_3\text{NH}_3^+ + \text{Cl}$ in the Solution Phase, Geometrical Parameters/Force Correlation (%Fcorr) in CVs and Radial Cutoff Distances (r_c)

	(kcal/mol)		geometry (Å)		F_{corr} (kcal/mol/Å)		r_c (Å)						
	ΔG_s^\ddagger	ΔG_s	$CV_{\text{N-C}}$	$CV_{\text{C-Cl}}$	$CV_{\text{N-C}}$	$CV_{\text{C-Cl}}$	N-H	N-C	N-Cl	H-H	C-H	C-Cl	Cl-H
Uniform Cutoff	26.6	-2.3	1.988	2.138	6.1	17.2	2.000	2.000	2.000		2.000	2.000	2.000
Generic Cutoff	13.7	-20.0	2.152	2.178	3.8	7.1	2.120	2.520	4.280	1.740	2.140	2.760	2.360
Micro-GA	19.5	-22.4	2.105	2.278	6.7	3.5	3.770	4.254	5.676	3.198	2.391	4.870	0.999
Extended Cutoff	14.1	-23.1	2.106	2.227	7.6	7.1		3.945				3.510	
Micro-GA-CV	14.1	-18.5	2.148	2.197	4.2	6.2		2.028				3.068	
expt	23.5	-34 ± 10											
AM1/MM	30.8	-10.7	1.996	2.099	16.1	31.4							
RP-FM-CV	14.5	-26.9	2.170	2.228	2.3	1.9							

Table 3.2. Cross-validation root mean-square errors (kcal/mol/Å)

	AM1/MM	Uniform Cutoff	Generic Cutoff	Micro-GA
N	14.3	5.3	5.2	5.9
H	11.3	4.0	4.1	3.4
C	22.8	8.8	5.4	7.6
Cl	19.8	12.9	8.0	2.8
average	14.5	6.4	4.5	4.3

the van der Waals radius. In comparison to the original AM1/MM method the reaction free energy and barrier is lowered to -20.0 kcal/mol and 13.7 kcal/mol, respectively, as shown in Figure 3.1. Compared to the AM1/MM method, the reaction free energy and barrier is lowered by 9.3 kcal/mol and 17.1 kcal/mol respectively. Therefore, pairwise RP-FM is largely dependent on the choice of cut off distances, where elongation of radial cutoff distances allows for more inclusive sampling (Figure 3.5). However, inconsistencies in force corrections for the generic radial cutoff scheme at 2 Å in Figure 3.5b suggests potential interferences from spectator pairs on C-Cl. To further assess the quality of the force matched PMFs for the generic radial cutoff scheme we computed the overall RMSE, in addition to deviations for all atoms and for sampled configurations along the reactive path. The overall deviation is reduced to 3.8 kcal/mol/Å (Table 3.2), where deviations along the path is reduced to below 4.8 kcal/mol/Å (Figure 3.3).

In regard to distribution of deviation on atomic forces, the deviations for N, H, C, and Cl are reduced to below 4.9 kcal/mol/Å (Figure 3.2). However, for force correlations in CVs, the deviation remains large, having values of 3.8% and 7.1% for N-C and C-Cl respectively (Table 3.1). Therefore, improvement in PMF for the generic radial cutoff scheme is distinct from improving the force correction in CVs since the correlation for N-C and C-Cl remains large.

As shown in Figure 3.5, by extending the radial cutoff distances, more samples are utilized into the functional form of the potential to generate a more encompassing force correction,

which are otherwise prematurely truncated. However, elongation of the radial cutoff distance is not beneficial since the force correction is extrapolated throughout the target force corrections (Figure S5 in B.2) of the supplementary materials). Thus, an optimal range for the radial cutoff distance exists where short radial cutoff distances excludes samples and where large radial cutoff distances flattens the force correction. Furthermore, compared to the uniform radial cutoff scheme, the generic scheme agrees well with the expected reduction in barrier and reaction free energies from solvation free energy. The improvement in the PMF can be attributed to the inclusion of detailed features in the reaction, but more importantly, the correcting potential is seen as largely dependent on select radial cutoff distances. The periodicity of the force correction makes the optimization of radial cutoff distances challenging since radial cutoff is nonlinear. In this instance, the role of predefined parameters is unclear since sampling is system specific where careful attention is needed for the selection of radial cutoff distances.

3.4.2 Optimization of pairwise force matching using the Micro-genetic Algorithm (Micro-GA) and pairwise RP-FM in collective variables (CVs)

In this study, the PMFs are improved as a result of modifying the radial cutoffs but the correction should be dictated based on the deviation in CVs. As such, the extent to which force deviation in CVs is improved is examined in the pairwise fitting procedure to determine whether the correct forces in CVs is sufficient for obtaining reliable PMFs. Furthermore, since spectator atoms are orthogonal to the RC, the contribution to free energy from force corrections in CVs is independent of the complementary coordinates [78]. This is further evidenced in Figure 3.6 for the generic radial cutoff scheme, where maximum potential corrections on effective regions of CV pairs is large (N-C: -16.6 and C-Cl: -26.3 kcal/mol) in contrast to spectator pairs (N-H: 3.2, N-Cl: -0.58, H-H: 9.8, C-H: 4.4 and Cl-H: -1.1 kcal/mol).

Since pairwise RP-FM is dependent on select radial cutoff distances, we improve the fitness of RP-FM by employing a micro-genetic algorithm (Micro-GA) for nonlinear optimization [41]. To employ the Micro-GA, varying radial cutoff distances are inserted into the **B** matrix to compute optimized candidate parameters. Based on translationally and

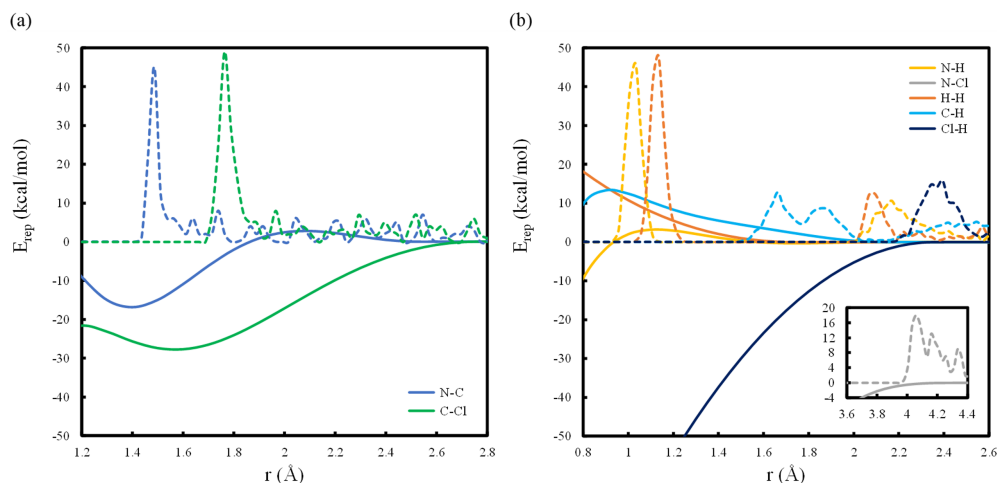


Figure 3.6. Comparisons of correcting potential energy functions from the B3LYP/MM force matched generic radial cutoff scheme for atom type pairs on: (a) collective variables and (b) spectator degrees of freedom, where sampled pair distances are represented in dotted lines (300 samples within 0.252 Å windows).

rotationally invariant internal forces, the objective function of the Micro-GA is minimized by reducing the force correlation in CVs from a combination of radial cutoff distances from all atom pair types. For this procedure, the initial population is randomly generated by mixing the computed candidate parameters into genetic pools for exchange of genetic information, based on a 10-digit binary code, where a population of 8 chromatids is propagated through 50 cycles to achieve convergence over 100 generations. By iteratively performing the SVD routine, the penalty function is minimized while maximum fitness is achieved for radial cutoff distances of all atom pair types. As such, the overall force correlation from both CVs in the Micro-GA radial cutoff scheme was reduced from 16.1% and 34.1% to 6.7% and 3.5% for N-C and C-Cl respectively (Table 3.1). Interestingly, the reduced radial cutoff for Cl-H in Micro-GA, suggests the need for suppression of the pairwise interaction, to avoid interferences in fitting forces on C-Cl. Furthermore, a reaction free energy and barrier of -22.4 kcal/mol and 19.5 kcal/mol was respectively obtained (Figure 3.1). Overall, the Micro-GA is time consuming, with the advantage of reducing the deviation in CVs to the

greatest extent. However, the optimized scheme remains insufficient for correcting the forces since correlations in CVs are not satisfactory.

Since the PMF is dependent on only important degrees of freedom, further optimization is implemented on only CVs to understand the direct relationship of pairwise RP-FM on free energy. A one-dimensional free energy scan for pairwise RP-FM on only N-C and C-Cl pairs, demonstrates the wide array of computed PMFs and deviations from target internal forces in CVs (Table S1 in B.2) of the supplementary materials). Furthermore, the impact of sample truncation in Figure 3.5 shows that identical radial cutoff distances for both CVs is not necessary. Since the selection of radial cutoff distances in CVs is case-by-case we employed a two-dimensional scan with 0.1 Å windows (Figure S6 in B.2) of the supplementary materials) and a Micro-GA to fit forces for only CV pairs. For Micro-GA on only CVs (Micro-GA-CV) the resulting radial cutoff distances converged to values of 2.028 and 3.068 Å, for N-C and C-Cl respectively (Table 3.1). Achieving a maximum fitness below the range of samples for both CVs is understandable since the magnitude of force corrections beyond the optimized radial cutoffs (2.6 Å) is small (Figure 3.7b-c).

However, the importance of sampling cannot be discredited since sampling directly influences the resulting fit from the decomposed design matrix. Therefore, an extended radial cutoff scheme is incorporated to account for long-range sampling (Figure 3.8).

A reaction free energy and barrier of -18.5 kcal/mol and 14.1 kcal/mol is respectively obtained for the Micro-GA-CV radial cutoff scheme, where the extended radial cutoff scheme, yields a reaction free energy and barrier of -23.1 kcal/mol and 14.1 kcal/mol respectively. In regard to force correlation in CVs, the Micro-GA-CV procedure outperforms the extended radial cutoff scheme, with reduced deviations of 44.7% and 12.7% for N-C and C-Cl respectively. Compared to the Micro-GA scheme, the Micro-GA-CV performs similarly with an average force correlation on both CVs differing by 1.9%. Therefore, the resulting deviations in the Micro-GA-CV scheme remains insufficient in regard to force fitting. Compared to the RP-FM-CV force correlation (N-C: 2.3% and C-Cl: 1.9%), the average deviation from both CVs for the Micro-GA-CV procedure, is greater by 147.6%.

3.4.3 Force Matched Free Energy Correction for Assessing RP-FM Models

The range in barrier energy from both pairwise RP-FM with Micro-GA-CV and RP-FM-CV is in excellent agreement (within 2.8%), while the free energy of reaction exceeds the generous experimental range of -44 kcal/mol to -24 kcal/mol for the pairwise RP-FM models (-23.1 kcal/mol to -18.5 kcal/mol). Interestingly, from a one-dimensional free energy scan, we see the possibility for obtaining comparable FEPs to RP-FM-CV results at the expense of large force deviations in CVs (Table S1 in B.2) of the supplementary materials). As such, no predicting power seemingly exists for force fitting in CVs since the FEP is an inadequate marker for determining the quality of RP-FM. Therefore, to assess the effects of force deviation on PMFs we developed a model for computing a zeroth-order (0th-order) approximations for force matched free energy corrections. In theory, the force matched free energy correction is equivalent to differentiating the free energy in CVs where subtle errors in the force correction could accumulate to influence the outcome of the integrated force corrections. To understand the relationship between RP-FM and free energy, we calculate the force matched free energy correction by integrating the force corrections for discretized images along the RC on each CV.

In essence, the Gibbs free energy, $\Delta G(\alpha)$, is defined as the integration of average force corrections ($\Delta \bar{f}$) along the reaction path (for a boundary condition from $0 \geq \alpha \leq 1$):

$$\Delta G(\alpha) = \int_0^1 \Delta \bar{f} \cdot d\alpha \tag{3.7}$$

Since the path evolution in CVs is not constant, in the framework of the free energy variable (z) the free energy correction is given by

$$\Delta G(z) = \int \frac{dG}{dz} \cdot \frac{dz}{d\alpha} \cdot d\alpha \tag{3.8}$$

As such, $\Delta G(z)$ is alternatively expressed as

$$\begin{aligned}
\Delta G(z) &= \sum_{i=1}^{N_{CV}} \sum_{j=1}^{J_{img}(\alpha)} \frac{(\Delta \bar{f}_{j-1}^i + \Delta \bar{f}_j^i)}{2} \times (z_j^i - z_{j-1}^i) \\
J_{img} &= \frac{\alpha - 0}{\Delta \alpha} \\
\Delta \alpha &= \frac{1}{(n_{img} - 1)}
\end{aligned} \tag{3.9}$$

where NCV is number of CVs, and nimg is the total number of images along the path.

To illustrate the effect of RP-FM on free energy, we computed the force matched free energy correction term for the SN2 reaction at the B3LYP/MM method level. In Figure 3.7a, the predicting PMF (dotted line) is calculated for RP-FM-CV, extended and Micro-GA-CV radial cutoff schemes using Eq. 3.9. Compared to the simulated results in Figure 3.7a, the 0th-order predictions from all schemes differs by less than 0.7 kcal/mol and 2.1 kcal/mol for reaction barrier and free energy respectively. In regard to distance-based force corrections in Figures 3.7b-c, the expected shapes for truncation and extrapolation of sampling in CVs is consistent with previous observations. In Figures 3.7d-f, the computed free energy corrections are provided with decomposed free energy contributions from each CV. Interestingly, a maximum correction unique to the activation barrier is captured at the transition state ($\alpha \approx 0.6$) for RP-FM-CV and for both radial cutoff schemes.

The force correlation is reduced to 4.2% and 6.2% for N-C and C-Cl respectively using the Micro-GA-CV (Table 3.1). In this case, the FEP agrees well with RP-FM-CV for the barrier height and somewhat well for the reaction free energy. Conversely, for a Micro-GA-CV scheme set to a radial cutoff of 4.6 Å on both CVs, a barrier and reaction free energy differing by 1.6 kcal/mol and 1.3 kcal/mol is respectively obtained (Table S1 in B.2) of the supplementary materials). Although the FEP results for radial cutoffs set to 4.6 Å are in better agreement to RP-FM-CV, the deviations in CVs remains substantial, having force correlations of 9.1% and 7.1% for N-C and C-Cl respectively. Consequently, the FEP results are not reliable for assessing the accuracy of a RP-FM model. Therefore, to better measure the quality of deviations on activation barrier and free energy in CVs, we decomposed the free energy correction along the RC for each CV for pairwise RP-FM and RP-FM-CV

Table 3.3. Reaction Barrier/Free Energy of $\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{CH}_3\text{NH}_3^+ + \text{Cl}$ in the Solution Phase, for 1st Iteration (itr=1)/Zeroth-Order (0th-order) Predictions and Statistical Errors Estimated from Free Energy Corrections

	(kcal/mol)		Statistical error (\pm kcal/mol)					
	ΔG_s^\ddagger	ΔG_s	$\Delta G_{\text{N-C}}^\ddagger$	$\Delta G_{\text{N-C}}$	$\Delta G_{\text{C-Cl}}^\ddagger$	$\Delta G_{\text{C-Cl}}$	$\Delta G_{\text{total}}^\ddagger$	ΔG_{total}
RP-FM-CV	18.6	-26.7	0.26	0.36	0.18	0.31	0.44	0.67
Extended Cutoff	17.0	-22.4	3.60	8.30	0.50	7.50	4.10	15.80
Micro-GA-CV	15.8	-18.4	2.74	4.68	0.39	7.30	3.13	11.98

models by using Eq. 3.9. In the combined free energy correction plot, we see a static free energy correction beyond $\alpha = 0.75$ for Micro-GA-CV and extended radial cutoff schemes, of approximately -8 kcal/mol and -12 kcal/mol respectively (Figure 3.7d). The static free energy correction contrasts the steady decline in free energy captured in RP-FM-CV and in the target method corrections. Therefore, the misrepresentation of free energy can be traced to the poor fit on C-Cl (Figure 3.7f). Moreover, compared to Micro-GA-CV, the reduction in free energy by 3.2 kcal/mol for the extended radial cutoff scheme in Figure 3.7a, can further be attributed to the lower quality fitting on N-C (Figure 3.7e). The reduction in free energy is further evidenced by the larger statistical error for free energy correction on N-C for the extended radial cutoff scheme (8.30 kcal/mol) compared to Micro-GA-CV (4.68 kcal/mol) in Table 3.3.

The large difference in free energy errors on N-C contrasts the C-Cl statistical errors, which are within 0.20 kcal/mol for both extended and Micro-GA-CV radial cutoff schemes (Table 3.3). As such, the resulting PMFs for various radial cutoff schemes is traceable to the collection of error cancellations in RP-FM in CVs. Additionally, we see that RP-FM is not uniquely defined to one-dimension since the free energy correction is dependent on the one-to-one mapping of integrated forces along the reaction path in two dimensions. As such, the large deviations in CVs in the force matched free energy correction explains the unpredictability in PMFs and erroneous agreements to RP-FM-CV. Therefore, the criteria for assessing whether a RP-FM model is robust should not be placed in the resulting PMF,

rather the quality of RP-FM should rest in the free energy correction for all CVs along the reaction path. Overall, we see reduction in force deviation as a qualitative measure for fitness in RP-FM, where RP-FM-CV minimizes the deviations in CVs to the greatest extent (below 3%) with combined statistical errors of 0.44 and 0.67 kcal/mol in Figure 3.9a for barrier and reaction free energies respectively.

In Table 3.3, the Micro-GA-CV outperforms the extended radial cutoff scheme, but in comparison to RP-FM-CV, the statistical error increases by 611.3 % and 1,688.1% for barrier and reaction free energies respectively due to error accumulation from reactant to product states (Figure 3.9). Therefore, the quality of RP-FM is sufficiently conveyed by the deviation of force corrections in CVs, however systematic correction of the forces along the entire path is emphasized for avoiding accumulated errors in large chemical systems with multiple CVs. In the RP-FM-CV method, the reduced deviation in CVs is achieved by correcting the internal forces along the discretized reaction path with the aid of spline functions [59]. Thus, in addition to correctly matching SE/MM forces to target AI/MM levels, the systematic correction of forces along the reaction path for all CVs is required for obtaining reliable FEPs, which is measurable by the developed model for computing 0th-order approximations for force matched free energy corrections.

3.4.4 Coordinate Dependence and Benefits of Force Matched Free Energy Corrections in RP-FM-CV

In this study, PMFs from initial iterations for extended radial cutoff, Micro-GA-CV and RP-FM-CV, are reproduced within 0.7 kcal/mol and 2.1 kcal/mol for reaction barrier and free energies respectively in 0th-order predictions. This approach is completely separate from free energy computations using the string simulation since the free energy forces are indirectly collected from the averaged force fluctuations from the reference value of the restrained internal coordinates. This provides supporting evidence that computing free energy based on redundant internal coordinate transformation and free energy forces from string simulations, are procedures that are both equivalent and intrinsically connected. Furthermore, the mapping procedure from coordinate transformation to internal degrees of freedom is correct since the RP-FM-CV correcting forces are identical to target forces and correct-

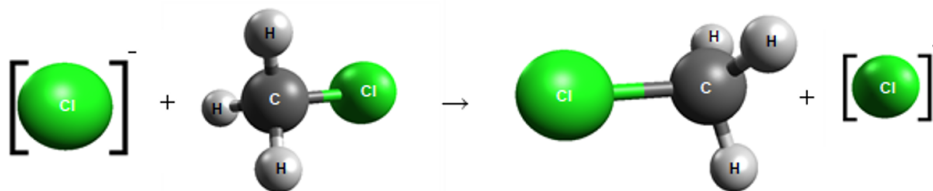
ing free energies (Figure 3.7), in comparison to simulated results (Table 3.3). Since the CV force is equivalent to the internal forces from coordinate transformation, we develop a new approach for predicting PMFs based on constraining the CV forces along the reaction path as opposed to collecting the fluctuated forces from restrained CVs.

One of the main criticisms of the RP-FM-CV method is the internal coordinate dependence. In contrast, forces from the string method are coordinate independent since they are externally determined from the restrained simulation. However, following the argument of Ruiz-Montero et al, the CV forces along the RC are described as coordinate independent since the CV forces are not dependent on orthogonal spectator degrees of freedom. With this rationale, the complimentary coordinates are not explicitly defined, so CVs can be regarded as being coordinate independent. This statement is further evidenced by the ability to reproduce the required free energy correction from SE/MM PMF with the integrated RP-FM free energy correction in CVs. One speculation is that averaging from restrained sampling is the cause for coordinate dependence removal. This argument can be made since the internal forces are based off of the optimized configurations from the restrained string simulation, where the optimized configurations are constructed from the averaged free energy force in CVs. Therefore, the free energy forces in CVs is embedded within the configurations for which redundant internal coordinate transformation is performed. Since the average fluctuations are captured in the environment of the spectator atoms and since nonredundant internal coordinate transformation accounts for the spectator atoms specific to certain RC, the uniform definition of internal coordinates along the RC combined with the grid-based force correction removes the coordinate dependence since the nonredundant forces are averaged along the RC, as shown in the integration procedure for the Gibbs free energy correction, following Eq. 3.9. Alternatively, since only CV internal forces are fit, the fluctuations from the spectator degrees of freedom can be regarded as equivalent to the coordinate mapping procedure in coordinate transformation. According to den Otter and Briels [79], the global nature of CVs discredits the removal of complementary coordinates, however, convergence in 0th-order predictions with restrained MD, supports the theoretic rigor of RP-FM in CVs to restore PMFs.

The integration of the force matched force corrections is not only beneficial for understanding the relationship between the force correlation and the PMF, but it is also a very powerful tool since it spares the need for repeating the MD procedure in RP-FM. As such, force matched free energy corrections is fast, but the assumption is that the reaction path is unchanged. As such, we offer an efficient tool for analyzing the impact of RP-FM on PMFs by identifying the most dynamic regions of force fitting along the reaction path. More specifically, this tool can be used to assess negative or positive correlations from RP-FM in CVs to pinpoint regions on bonds where errors are located. Finally, by assessing the free energy corrections in CVs, we present a potential tool for recommending a better way to fit forces for generating accurate high method level FEPs.

3.4.5 Comparison of pairwise RP-FM and RP-FM-CV for reaction path corrections

In the original pairwise RP-FM study, the asymmetric condensation reaction for glycine involves the formation and breakage of two N-C bonds. Although the reaction is asymmetric, all of the atom pair types exist on both sides of the reaction, therefore radial cutoff treatment for N-C pairs is the same on both sides of the RC. However, in the quantum mechanical sense the N-C pair from one glycine should be markedly different compared to the target glycine N-C pair, so the correction should not be the same due to environmental dependencies. In this respect, the proposed function of the pairwise RP-FM model is too classical to cover such distinct situations. Since the correcting potential is refit in the pairwise model, the RP-FM is expressed as a physical system correction as opposed to refitting the force as an explicit correction term. As such, the difference between high and low method levels is classically corrected, much like the force fields approach from the predefined pairwise potential. However, the potential based correction also contradicts the classical approach since the correction is based on the quantum force corrections between SE/MM and target AI/MM methods. Consequently, the argument whether quantum corrections are sufficient for representing the pairwise potential remains unclear. In contrast to ML and pairwise potentials, a method like RP-FM-CV does not require a classical/quantum correcting potential, since the integration of non-classical forces, achieves quantum behavior



Scheme 3.2. Schematic Representation of the Finkelstein Reaction ($\text{Cl}^- + \text{CH}_3\text{Cl} \rightarrow \text{ClCH}_3 + \text{Cl}^-$)

with the aid of spline functions. Despite the differences between the pairwise potential and explicit force correcting models, a consistent shared feature is the qualitative acquirement of quantum behavior from correcting forces in CVs along the reaction path.

Compared to the Menshutkin reaction, the glycine condensation reaction is balanced, which could lead to the equivalent combination of negative and positive correction parameters. To investigate the transferability of the pairwise RP-FM to non-trivial symmetric reactions, we computed the FEP for the simple symmetric $\text{Cl}^- + \text{CH}_3\text{Cl}$ Finkelstein reaction, where a methyl group is transferred between two chlorine atoms (Scheme 3.2). The rationale for testing this scheme is to see whether the dispersion of same pair types on both sides of the RC is capable for obtaining reliable PMFs. As such, extended and Micro-GA-CV schemes are simulated with C-Cl radial cutoff distances of 3.805 Å and 4.716 Å respectively. The force matched results are comparable, having reduced barrier energies within 2.8 kcal/mol of the original AM1/MM barrier (Figure 3.10).

However, compared to the RP-FM-CV results, the pairwise RP-FM methods are unable to raise the barrier energy to target levels due to smaller force corrections at distances away from the transition state (Figure 3.11b-c). In free energy correction plots in Figure 3.11e-f, the forming bond is shown to have a larger positive correction compared to the breaking bond as evidenced by the overall increase in free energy for the target method. However, as shown in Figure 3.11d, the magnitude of free energy correction is similar for all methods, but the accumulation of subtle errors leading towards the transition state results in the lowered activation energy.

Table 3.4. Reaction Barrier of $\text{Cl} + \text{CH}_3\text{Cl} \rightarrow \text{CH}_3\text{Cl} + \text{Cl}$ in the Solution Phase, Geometrical Parameters/Force Correlation (%Fcorr) in CVs and Radial cutoff Distance (r_c)

	(kcal/mol)	geometry (Å)	F_{corr} (kcal/mol/Å)	r_c (Å)
	ΔG_s^\ddagger	$CV_{\text{C-Cl}}$	$CV_{\text{C-Cl}}$	C-Cl
Extended Cutoff	16.7	2.324	2.1	3.805
Micro-GA-CV	18.5	2.323	2.0	4.716
expt	~26	2.375		
AM1/MM	22.7	2.169	25.0	
RP-FM-CV	19.5	2.327	1.7	

Compared to the assymetric Menshutkin reaction, the force deviation in CVs for all models in the force matched symmetric reaction, is dramatically reduced from 25.0% to below 2.1% (Table 3.4).

Therefore, an improvement in FEP, geometry and force correlation make pairwise RP-FM strategy better for symmetric reactions, but the procedure remains error prone since reduced deviations in CVs is connected to the accumulation of errors in free energy correction.

In principle, a flexible enough polynomial should be capable of describing the global behavior for correcting the reaction path. To better understand the RP-FM strategy, we study the pairwise model as an interesting conceptual take for the same force fitting problem, where forces from constructed pairwise potentials are matched to achieve free energies at higher levels of quantum theory. As such, independent of the reaction state, a classically derived correction term is proposed for tuning the force with enough flexibility to predict the reactive chemical behavior. In regard to parameterization, the pairwise routine uses the same number of parameters as the spline procedure in RP-FM-CV, but the force deviation remains large in CVs. In this study, we see that shorter radial cutoff distances achieve the lowest deviation in CVs at the expense of truncating the sampled forces. Conversely, extending the radial cutoff distance to regions beyond the sampling region, results in a generalized extrapolation of the force correction along the reaction path (Figure S5 in B.2)

of the supplementary materials). As such, free energy is radial cutoff dependent, and selection of radial cutoffs is nonlinear. Based on the MFEP, similar improvements to transition state geometries are made as a result of RP-FM (Figure 3.12).

The qualitative agreement in path evolution and geometry demonstrates the convergence of RP-FM to higher levels of theory. As such, the prediction of the reaction path from RP-FM to higher theory levels, results in a looser transition state with extended forming and breaking bonds, in contrast to tighter transition state structures from SE/MM. This agreement holds for both pairwise RP-FM and RP-FM-CV, so it suggests that RP-FM not only improves the computed FEPs, but it also converges to more optimized geometry state structures.

3.5 Concluding Remarks

In summary, pairwise reaction-path force-matching (RP-FM) offers an interesting conceptual take on RP-FM, but the procedure is inadequate for fitting the CV forces along the reaction path. In this study, we address the central problem of expensive AI/MM sampling by using the RP-FM strategy to achieve high level accuracy at the cost of SE/MM theories. The proposed model is carried out on the prototype Menshutkin reaction in solution, where SE/MM overestimates both reaction barrier and free energies. From the point of new methodological development, we compare two RP-FM approaches based on the correcting potentials and explicit force sampling models. In particular, we compare a recently developed potential based RP-FM approach, which computes the correcting force from a well-defined pairwise potential, with the recently developed reaction-path force-matching in CVs (RP-FM-CV) method. On the basis of the results, we conclude the following:

(1) In the pairwise RP-FM model, the overall atomic deviation is reduced between base SE/MM and target AI/MM level forces, but the deviations in CVs is large due to overfitting of spectator degrees of freedom. Furthermore, selection of radial cutoff distances is system specific, with distances directly related to the periodicity of force corrections in CVs. A generic cut off scheme is shown to improve the deviations compared to uniform radial cutoff distances, but the deviations in CVs remains insufficient. Furthermore, radial cutoff analysis

reveals FEPs in pairwise RP-FM models to be unstable, where the integration of internal force corrections in CVs is attributed to the accumulation of force correcting errors along the reaction path. The lower deviation in CVs in RP-FM-CV (less than 3%) demonstrates the importance for directly fitting the forces along the RC with grid-based force corrections. Moreover, the improvement in direct fitting is traceable to improvement of internal force corrections specific to each CVs. While deviations in CVs shows qualitative strength in fitting, the PMF is an arbitrary measure for the quality of fitting since carefully handled radial cutoffs is equivalent to generating converged FEPs.

(2) By reproducing the PMF from integration of internal force in CVs, we see that internal coordinate transformation in RP-FM-CV is equivalent to the free energy forces from restrained MD simulations. As such, the restoration of PMFs to target force corrections is achieved by combining the computed force corrections along the reaction path for CVs. Regardless of coordinate dependence, the agreement in PMFs for varying radial cutoffs and for RP-FMs demonstrates the intrinsic relationship between restrained MD and coordinate mapping. Therefore, a rapid analysis tool for measuring the robustness and reliability for generating converged FEPs has been developed in this study. Such analysis reveals pairwise RP-FM to better suit symmetric reactive processed, since the procedure adopts a more classical force fields type approach. Nonetheless it is clear that RP-FM is able to predict optimized geometries structures and reaction paths, in addition to reproducing high accuracy ab initio forces for computing the FEP.

In summary, computed FEPs from correcting forces in RP-FM are shown to be sensitive to deviations in CVs. In this study, a criterion was developed to identify the strength of RP-FM, where a suggested deviation below 3% in CVs is recommended for obtaining reliable FEPs. As such, our approach reveals an underlying framework for RP-FM models, and provides a deeper understanding to addressing force fitting problems. In addition to benchmarking the B3LYP/MM method for Menshutkin and Finkelstein reactions, we gain more confidence for assessing RP-FM strategies in fitting multi-dimensional internal coordinate transformations, conformational changes, or electron transfer studies in CVs. Therefore, by developing a standard for implementing RP-FM strategies to future studies, we plan to fur-

ther investigate more complex reactions to better understand the mechanisms and functions of chemical and biochemical reactions in solutions.

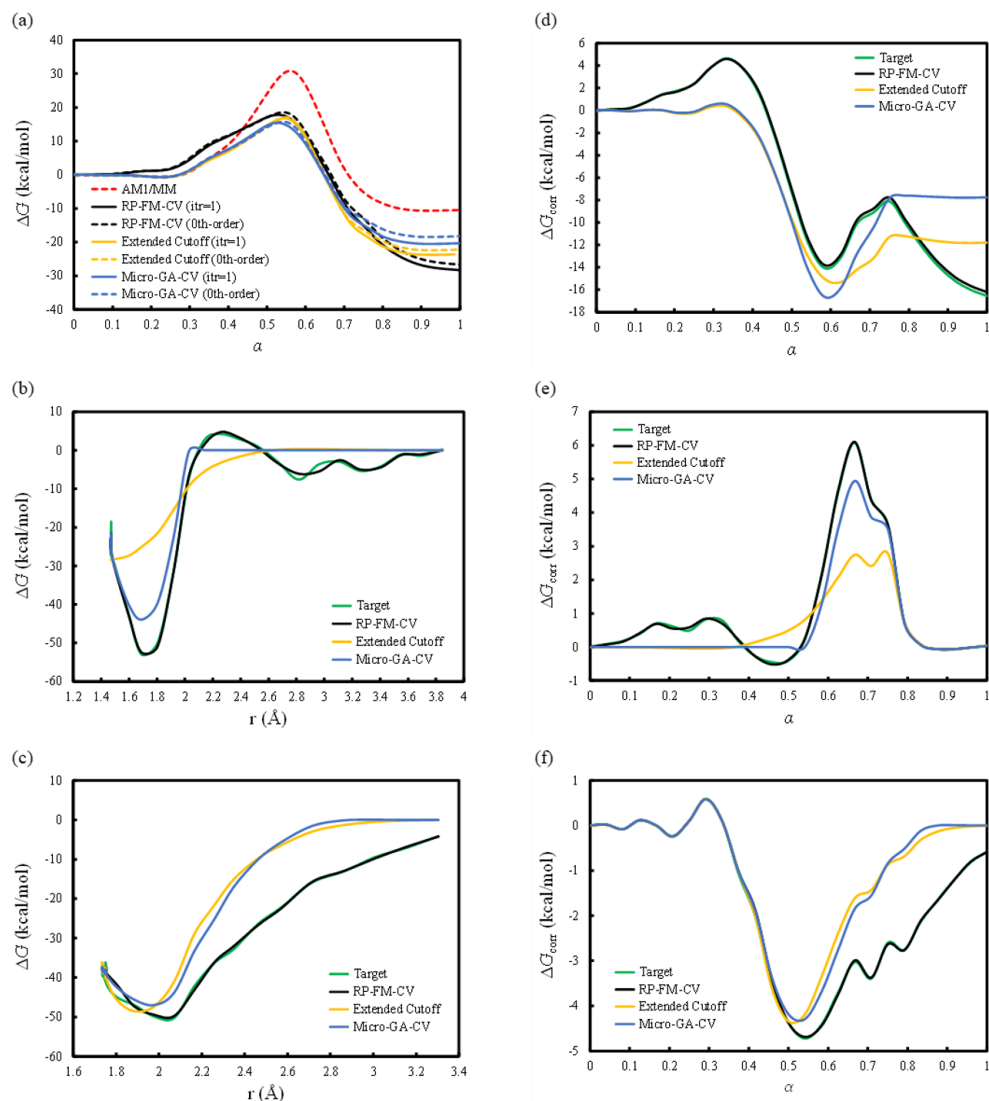


Figure 3.7. Potential of mean force, force correction and free energy correction for the Menshutkin reaction in aqueous solution for B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes. (a) Solid lines are from 1 iteration of force matching and dashed lines are zeroth-order predictions for FEPs from free energy corrections. Distance-based force corrections for (b) N-C and (c) C-Cl. Total free energy correction (d) and decomposed free energy corrections in collective variables for (e) N-C and (f) C-Cl.

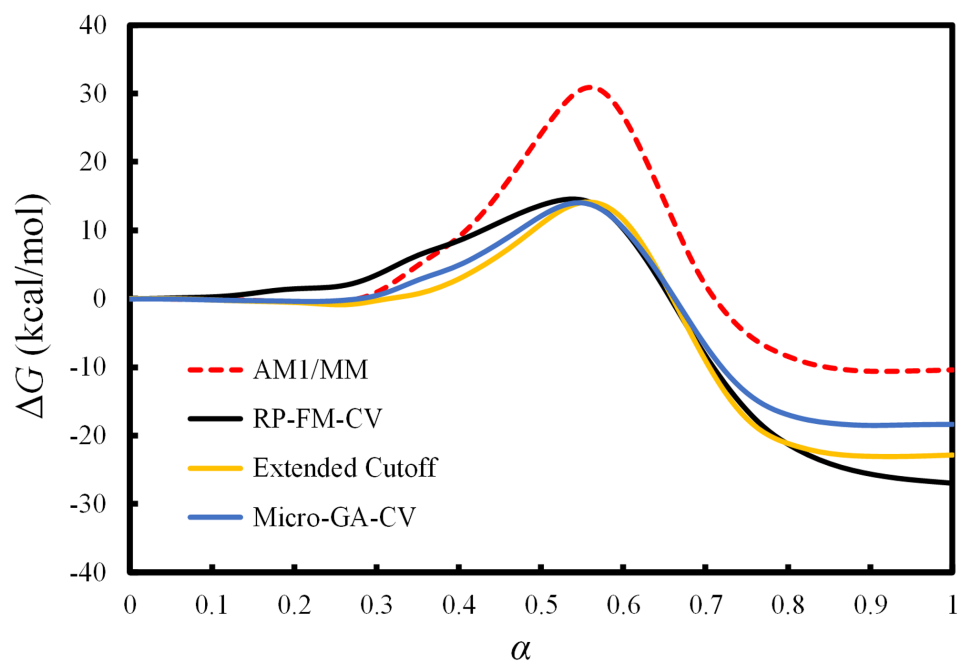


Figure 3.8. Potential of mean force for force matched collective variables for the Menshutkin reaction in aqueous solution for AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes.

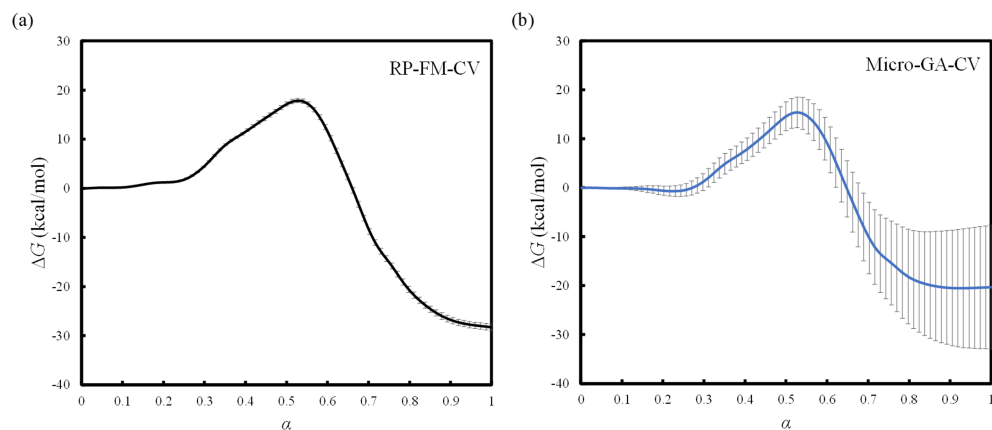


Figure 3.9. Error bars on Potential of mean force calculations for the Menshutkin reaction in aqueous solution for: (a) B3LYP/MM force matched RP-FM-CV (black), and (2) B3LYP/MM force matched and micro-genetic algorithm in collective variables (blue) radial cutoff scheme.

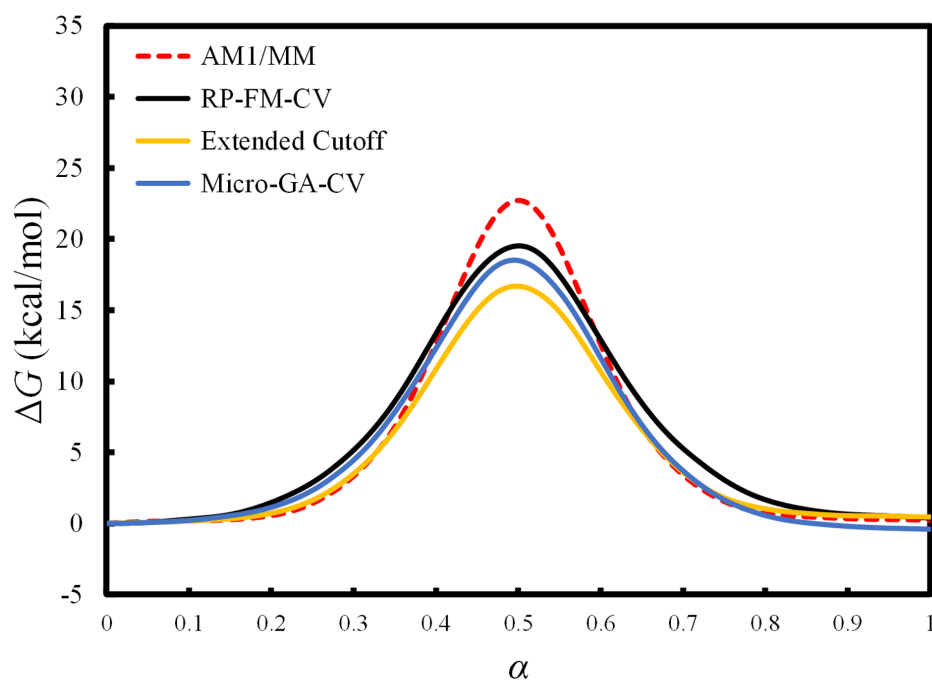


Figure 3.10. Potential of mean force for the Finkelstein reaction in aqueous solution for AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes.

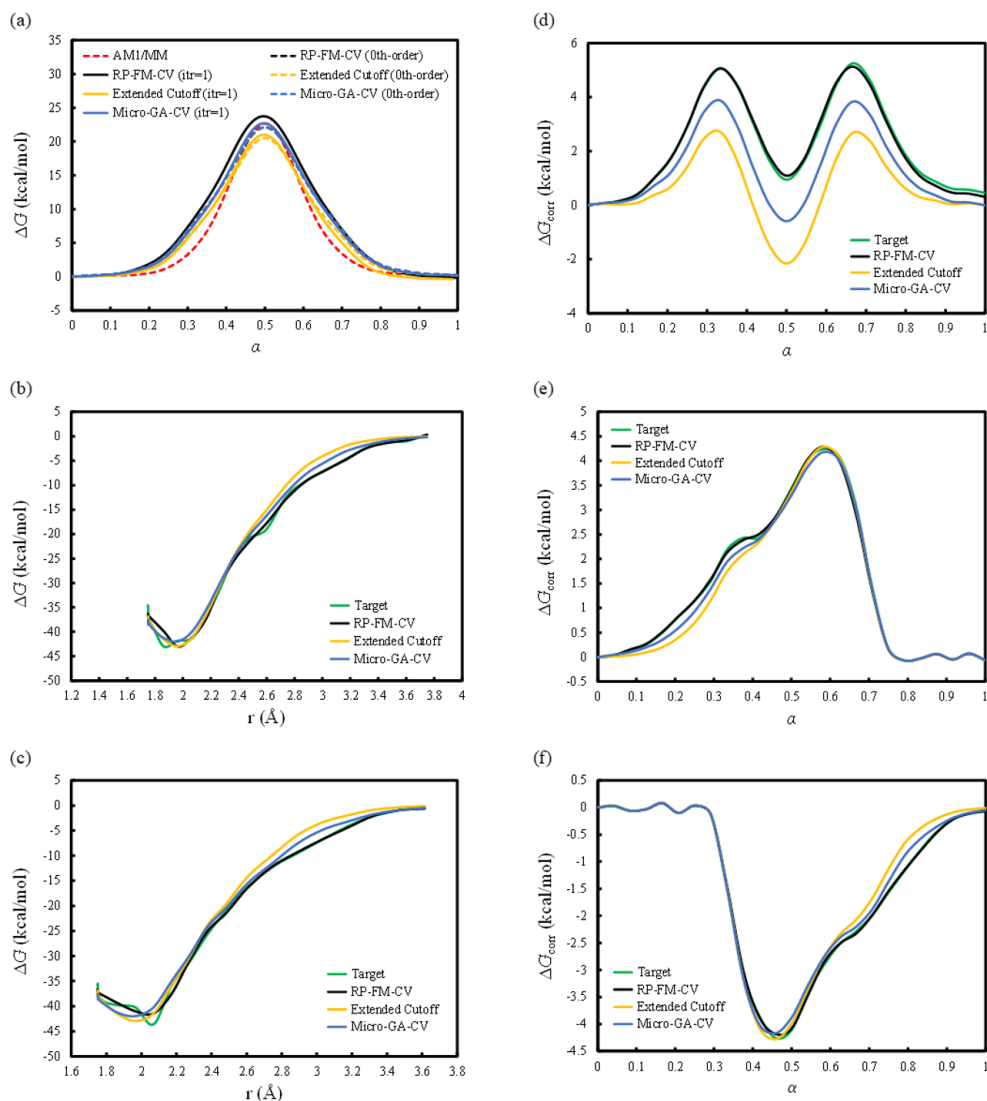


Figure 3.11. Figure 3.11. Potential of mean force, force correction and free energy correction for the Finkelstein reaction in aqueous solution for B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes. (a) Solid lines are from 1 iteration of force matching and dashed lines are zeroth-order predictions for FEPs from free energy corrections. Distance-based force corrections for (b) C-Cl and (c) C-Cl. Total free energy correction (d) and decomposed free energy corrections in collective variables for (e) C-Cl and (f) C-Cl.

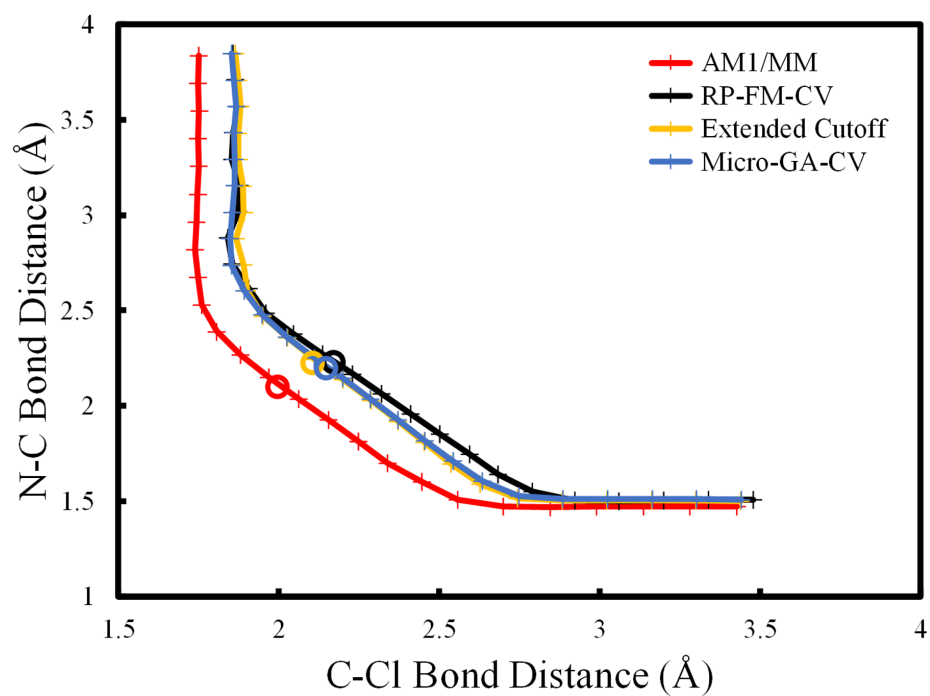


Figure 3.12. Minimum free energy path and transition state geometries (circles) for the Menshutkin reaction based on AM1/MM (red), B3LYP/MM force matched RP-FM-CV (black), and B3LYP/MM force matched extended (yellow) and micro-genetic algorithm in collective variables (blue) radial cutoff schemes.

4. DOUBLY POLARIZED QM/MM WITH MACHINE LEARNING CHAPERONE POLARIZABILITY

A major shortcoming of semiempirical (SE) molecular orbital methods is their severe underestimation of molecular polarizability compared with experimental and ab initio (AI) benchmark data. In a combined quantum mechanical and molecular mechanical (QM/MM) treatment of solution-phase reactions, solute described by SE methods therefore tends to generate inadequate electronic polarization response to solvent electric fields, which often leads to large errors in free energy profiles. To address this problem, here we present a hybrid framework that improves the response property of SE/MM methods through high-level molecular-polarizability fitting. Specifically, we place on QM atoms a set of corrective polarizabilities (referred to as chaperone polarizabilities), whose magnitudes are determined from machine learning (ML) to reproduce the condensed-phase AI molecular polarizability along the minimum free energy path. These chaperone polarizabilities are then used in a machinery similar to a polarizable force field calculation to compensate for the missing polarization energy in the conventional SE/MM simulations. Because QM atoms in this treatment host SE wave functions as well as classical polarizabilities, both polarized by MM electric fields, we name this method doubly polarized QM/MM (dp-QM/MM). We demonstrate the new method on the free energy simulations of the Menshutkin reaction in water. Using AM1/MM as a base method, we show that ML chaperones greatly reduce the error in the solute molecular polarizability from 6.78 to 0.03 Å³ with respect to the density functional theory benchmark. The chaperone correction leads to ~ 10 kcal/mol of additional polarization energy in the product region, bringing the simulated free energy profiles to closer agreement with the experimental results. Furthermore, the solutesolvent radial distribution functions show that the chaperone polarizabilities modify the free energy profiles through enhanced solvation corrections when the system evolves from the charge-neutral reactant state to the charge-separated transition and product states. These results suggest that the dp-QM/MM method, enabled by ML chaperone polarizabilities, provides a very physical remedy for the underpolarization problem in SE/MM-based free energy simulations.

4.1 Introduction

First-principles free energy simulation of condensed-phase reactions is computationally demanding, as the associated configuration-space sampling involves a great number of ab initio (AI) quantum mechanical potential energy calculations for large-scale systems. The problem can be alleviated using the combined quantum mechanical and molecular mechanical (QM/MM) approach [1], [13], [14], [16], [135], by which QM is only applied to a small-sized reactive subsystem with the description of the rest of the system left to efficient MM. In practice, semiempirical (SE) QM methods are much more widely used than AI methods in QM/MM free energy simulations to achieve the computational efficiency required for adequate sampling, especially when obtaining statistically robust free energy results is of a major concern [8].

Despite generally good cost effectiveness, one major shortcoming of SE methods that prevents them from being applied reliably to condensed-phase simulations is their systematic underestimation of molecular polarizability relative to experimental and AI benchmarks [136]–[143]. This causes the associated SE wave functions to respond too weakly to applied electric fields to generate strong enough induced dipoles for adequate electronic polarization. In QM/MM, a QM method that produces correct molecular polarizability is highly desirable, especially for modeling reactions in a polar solvent and in enzymes, as a large fraction of QM/MM electrostatic interaction for such systems can be caused by induction. Therefore, how to remedy SE/MM methods to offer reliable intermolecular polarization otherwise accessible with AI/MM methods is of great interest.

Underestimation of polarizability by SE methods is well documented in the literature [141], [142], [144]–[146]. The magnitude of the underestimation can range from ~ 25 –40% for neutral molecules [138], [142], [143], up to ~ 40 –55% for charged systems [142], or even more severe for highly polarizable systems. Gao and co-workers found that the SE AM1 method, when used in self-consistent reaction field (SCRF) and QM/MM calculations, underestimates the induced molecular dipoles for water [144]; a similar observation was also found on the basis of the AM1 molecular-orbital derived polarization potential [145]. In a linear-scaling SE study using the divide-and-conquer approach, Merz and co-workers [141] found that

the induced molecular dipoles computed for bulk water are 0.5 D (67%) lower than the experimental values, which is likely related to an underestimated molecular polarizability produced at the SE levels [146]. For water, both the AM1 and PM3 SE methods yield a molecular polarizability of 3.4 au (0.506 \AA^3) [146], which corresponds to an underestimation of $\sim 66\%$ when compared to the experimental value of 9.91 au (or 1.475 \AA^3) [147]. Given a strong dependence of polarizability on the size of the basis functions [143], [148], this underestimation is likely an intrinsic deficiency in the SE methods caused by the use of a minimum basis set. Although the prediction of bulk properties can be significantly improved through adjusting the SE parameters, such a strategy seems to be less effective for restoring the correct molecular polarizability for water [146], as reparametrization itself does not lift the basis-set limitation of the SE methods.

To address the underlying inadequacy of SE wave functions at a fundamental level, a few methods aimed at restoring the correct electronic polarization response have been developed [142], [149]. Using a chemical-potential equalization (CPE) formalism [150], York and co-workers have developed a charge-dependent response density treatment [142], in which energy is expanded in terms of electron density response using the SE density as a reference. As their method provides a theoretically elegant framework to rebuild the SE response property in a charge-dependent manner, it successfully reproduces high-level molecular polarizabilities for a large set of 1132 RNA-catalysis-relevant molecules and ions in various charge states from +2 to -2 [142]. In addition, the charge-dependent density-expansion idea has also been extended to the self-consistent-charge density-functional tight-binding (SCC-DFTB) method [151]–[153] for an improved description of intermolecular polarization. Gao and Truhlar have developed the polarized molecular orbital (PMO) model [143], [149], [154], in which SE methods are reformulated with p functions added on hydrogen atoms to increase the flexibility of wave function for improved molecular polarizability. Built upon an ab initio foundation [143], the PMO method has been parameterized and tested for H and O systems [149], including water clusters, and recently extended to C for organic chemistry [154] with a more balanced description of a broad range of properties, including atomic charge, polarizability, dipole moment, conformational energy, and chemical reactivity.

Both the CPE-based response density and PMO methods tackle the underpolarization problem of SE methods by modifying the Hamiltonians at a quantum mechanical level. Considering the central role of polarizability in the classical description of intermolecular forces [155] and its wide usage in developing polarizable force fields [156]–[160], here we offer an alternative solution for the problem from a different angle. In this paper, we present a hybrid framework that improves the electronic polarization response property of SE/MM methods through fitting high-level AI molecular polarizability in condensed phases. Specifically, we accomplish molecular-polarizability matching by placing a set of corrective classical atomic polarizabilities, named chaperone polarizabilities, on the QM atoms. Under the polarization of MM charges, the induced dipoles created on the chaperone polarizabilities would contribute additional polarization energy that compensates for the missing polarization in the conventional SE/MM simulations. Because both SE wave functions and classical chaperone polarizabilities hosted on QM atoms are polarized by MM electric fields, we designate this method doubly polarized QM/MM (dp-QM/MM).

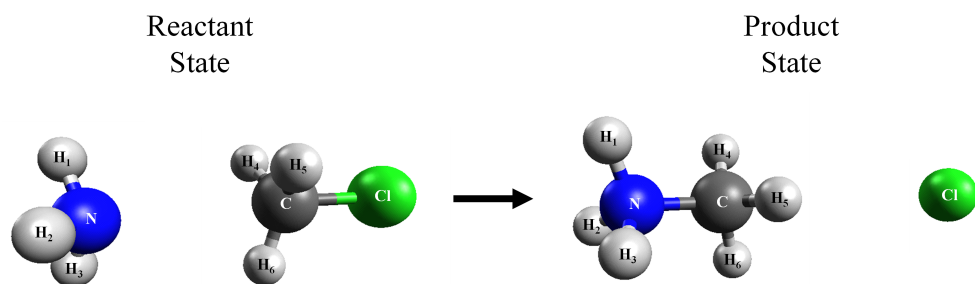
Our focal point here is to use the dp-QM/MM method to conduct free energy simulations of chemical reactions in condensed phases, for which polarizability corrections need to be determined as a function of the reaction coordinate on the fly. To predict chaperone polarizabilities along a reaction coordinate for dynamical configurations in general, we utilize advanced machine learning (ML) algorithms, which have increasingly been used in recent years to predict molecular properties and to supplement electronic structure methods [34]–[37], [72], [161]–[163]. While ML approaches have been used for predicting polarizabilities for stable molecules [38], ML prediction of polarizability for reactive chemical systems with bond dissociation and formation is less well established. As we will see in the Results and Discussion section, a reliable ML model for reactively predicting polarizability is essential in dp-QM/MM because the chaperone corrections that are needed for SE/MM polarizabilities to match with their AI/MM target values generally vary along the reaction coordinate in a nonuniform manner. In our method, this is addressed by learning polarizability along a minimum free energy path (MFEP) defined in a set of collective variables (CVs); these CVs are also included explicitly as input features for training our ML model.

For the output features of the model, although molecular polarizability of the QM subsystem is the central physical observable for us to predict, in practice, we distribute molecular polarizability and the related chaperone corrections to atomic centers, a choice made for a number of reasons. First, distributed atomic chaperone polarizabilities are advantageous in describing the heterogeneity of the polarization correction (with an atomic resolution) throughout the molecule. Second, although the atomic chaperone polarizabilities we use in this work are isotropic, anisotropic molecular chaperone polarizability can be reconstructed from them in a straightforward manner. Third, chaperone polarizability in its atomic form can be seamlessly plugged into the machinery of polarizable force field calculations. Last but not least, the repulsive part of van der Waals potential on each QM atom would effectively prevent the associated chaperone polarizability from being “catastrophically” polarized by MM charges during molecular dynamics.

A practical issue of using atomic chaperone polarizabilities is that atomic polarizabilities are not physical observables; therefore, the way of converting them from molecular polarizability, e.g., through nonlinear transformation [164]–[166] or partitioning [140], [167] schemes, or purely from ML decomposition [38], is not unique. Based on Applequists atom dipole interaction model [164], Thole showed that when short-range intramolecular polarization is properly damped, molecular polarizability can be decomposed and numerically fitted to a set of universal element-based atomic polarizabilities [165]. By contrast, the molecular-polarizability-partitioning study by Cramer, Truhlar, and co-workers suggests a strong dependence of atomic polarizability on local chemical environment [167]. For certain flexibility, atom-specific polarizabilities are used in our ML model; element-based polarizabilities will be examined in the future.

To validate the dp-QM/MM method, we applied it to the type-II S_N2 Menshutkin [45] reaction between ammonia and methyl chloride in water (see Scheme 4.1), for which significant discrepancies exist between SE/MM and AI/MM free energy profiles [26], [69].

Due to a drastic change of dipole moment during the charge-separating process, polarizability plays an important role for quantifying the solvation free energy contribution for this reaction. With the dp-QM/MM method, we demonstrate that a large fraction of the free



Scheme 4.1. Schematic Representation of the Menshutkin Reaction from the Charge-Neutral Reactant State to the Charge-Separated Product State ($\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{NH}_3\text{CH}_3^+ + \text{Cl}^-$)

energy error using SE/MM can be removed when proper polarization response is attained with chaperone polarizabilities.

The outline of the paper is as follows. Section 4.2 describes the key equations for computing energy and gradients associated with the chaperone-polarizability corrections. Section 4.3 discusses the computational details for the dp-QM/MM free-energy simulations. The results are presented and discussed in Section 4.4. The outlook of the method is discussed in Section 4.5, followed by the concluding remarks in Section 4.6.

4.2 Methods

To match molecular polarizability between the SE/MM and AI/MM levels, we introduce a set of corrective chaperone polarizabilities ($\Delta\alpha_i^C$) on the QM atoms:

$$\Delta\alpha_i^C = \alpha_i^{\text{AI/MM}}(\mathbf{R}, \mathbf{r}) - \alpha_i^{\text{SE/MM}}(\mathbf{R}, \mathbf{r}) \approx \alpha_i^{\text{AI-PCM}}(\mathbf{R}) - \alpha_i^{\text{SE-PCM}}(\mathbf{R}) \quad (4.1)$$

where $\alpha_i^{\text{AI/MM}}$ and $\alpha_i^{\text{SE/MM}}$ represent the distributed atomic polarizabilities on QM atom i , determined at the AI/MM and SE/MM levels, respectively, as a function of the QM solute geometry \mathbf{R} and MM solvent geometry \mathbf{r} . Note that although underestimations of polarizability in SE methods documented in the literature are mainly based on gas-phase systems, the polarizability corrections derived from the gas-phase SE and AI wave functions are not used here because they are unlikely applicable to the condensed-phase simulations, for which the chaperone polarizabilities should reflect the different response properties of solution-phase wave functions instead. To recover the correct response for the solvated wave function, we choose to define the chaperone polarizabilities in Eq. 4.1 using the polarizabilities obtained for the condensed-phase QM/MM systems. For technical simplicity, in Eq. 4.1, we further approximate the response correction for the solution-phase QM/MM wave functions with implicit solvation, where $\alpha_i^{\text{AI-PCM}}$ and $\alpha_i^{\text{SE-PCM}}$, both dependent only on the solute geometry \mathbf{R} , represent the polarizabilities computed at the AI and SE levels using the polarizable continuum model (PCM) [168]. A similar strategy has been adopted by Gao et al. [144] when developing atomic charges derived from solution-phase electrostatic potentials, where they found that explicit QM/MM and continuum solvation calculations generate comparable

induced dipole polarization. Because chaperone polarizability describes how differently the solution-phase SE and AI wave functions respond to electric fields, the difference itself is expected to be less sensitive to the explicit or implicit treatment of solvent.

For a QM/MM system consisting of N chaperone-hosting QM atoms, the polarization energy correction ΔE^{pol} due to the presence of chaperone polarizabilities is:

$$\Delta E^{\text{pol}} = -\frac{1}{2} \sum_{i=1}^N \Delta \alpha_i^{\text{C}} |\mathbf{E}_i^0|^2 \quad (4.2)$$

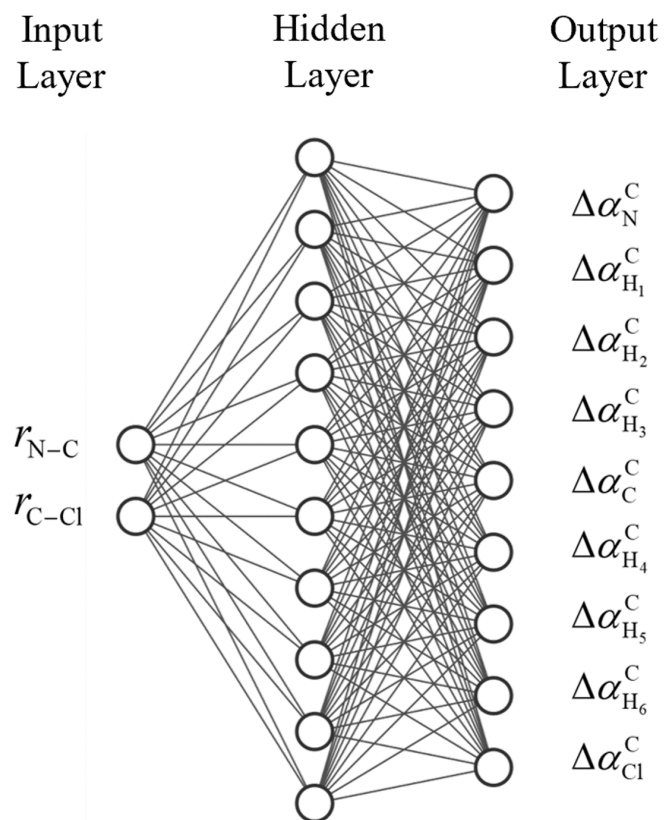
Where \mathbf{E}_i^0 denotes the permanent electric field generated by MM charges on QM atom i (see Appendix 4.7 for the definition of \mathbf{E}_i^0 and the derivation of Eq. 4.2). Once determined, ΔE^{pol} is added to the SE/MM Hamiltonian to define the total energy at the dp-SE/MM level.

To capture its reaction-coordinate dependence, we determine $\Delta \alpha_i^{\text{C}}$ on the fly during QM/MM molecular dynamics (MD) simulations through machine learning (ML). Specifically, an artificial neural network (ANN) is used, which takes molecular features as input and is optimized through a hidden layer to predict chaperone polarizabilities in the output layer (Scheme 4.2).

In the present work, we use a set of collective variables (CVs) as input features because they provide a direct link between the evolvement of polarizability and the string minimum free energy path (MFEP) that we use for free energy simulations. With a hyperbolic tangent activation function used in this three-layer ANN, $\Delta \alpha_i^{\text{C}}$ in the output layer can be written as:

$$\Delta \alpha_i^{\text{C}} = b_i^{\text{O}} + \sum_{j=1}^L W_{i,j}^{\text{O}} \tanh \left(b_j^{\text{H}} + \sum_{k=1}^M W_{j,k}^{\text{H}} p_k \right) \quad (4.3)$$

Where p_k denotes the k -th input feature among a total of M input features, $W_{j,k}^{\text{H}}$ and b_j^{H} are weights and biases in the hidden layer that consists of L nodes, and $W_{i,j}^{\text{O}}$ and b_i^{O} are weights and biases in the output layer.



Scheme 4.2. Topology of Artificial Neural Network Containing Two CV inputs, Ten Hidden Neurons, and Nine Atomic Chaperone Polarizabilities in Output Layer

To conduct dp-SE/MM MD simulations, forces associated with chaperone polarizabilities through the polarization energy correction term need to be incorporated consistently. With the product rule applied to Eq. 4.2, the nuclear gradients of the polarization energy correction ΔE^{pol} can be obtained by differentiating each of its components:

$$\frac{d\Delta E^{\text{pol}}}{d\mathbf{x}_a} = -\frac{1}{2} \sum_{i=1}^N \left(|\mathbf{E}_i^0|^2 \frac{d\Delta\alpha_i^{\text{C}}}{d\mathbf{x}_a} + 2\Delta\alpha_i^{\text{C}} \mathbf{E}_i^0 \cdot \frac{d\mathbf{E}_i^0}{d\mathbf{x}_a} \right) \quad (4.4)$$

Where \mathbf{x}_a denotes the Cartesian coordinates of QM center a . Note that in Eq. 4.4, there are two contributions to the gradients of the polarization energy correction. The first contribution in Eq. 4.4 is a novel term that arises from the reaction-coordinate dependence of chaperone polarizability introduced in the present work. Compared with the second contribution, which is more conventional, the chaperone gradient term (i.e., $\frac{d\Delta\alpha_i^{\text{C}}}{d\mathbf{x}_a}$) due to reactive fitting is absent in classical polarizable force field calculations where nonreactive fixed-valued atomic polarizabilities are used.

Under the current formulation of dp-QM/MM, because the permanent electric field \mathbf{E}_i^0 at a particular QM atom i does not depend on the position of any other QM atom (i.e., $\frac{d\mathbf{E}_i^0}{d\mathbf{x}_a} = 0$ for $i \neq a$; see also Appendix 4.7), Eq. 4.4 can be further simplified to a more operational form:

$$\frac{d\Delta E^{\text{pol}}}{d\mathbf{x}_a} = -\frac{1}{2} \left[\left(\sum_{i=1}^N |\mathbf{E}_i^0|^2 \frac{d\Delta\alpha_i^{\text{C}}}{d\mathbf{x}_a} \right) + 2\Delta\alpha_a^{\text{C}} \mathbf{E}_a^0 \cdot \frac{d\mathbf{E}_a^0}{d\mathbf{x}_a} \right] \quad (4.5)$$

In the context of ML, the Cartesian gradients of the chaperone polarizabilities $\Delta\alpha_i^{\text{C}}$ in Eq. 4.5 can be determined by first differentiating Eq. 4.3 with respect to the CV input features of the ANN, followed by a distribution of the gradient on each CV to its associated Cartesian components with the chain rule; the combination of the two steps can be cast into a compact form:

$$\frac{d\Delta\alpha_i^{\text{C}}}{d\mathbf{x}_a} = \sum_{j=1}^L \sum_{k=1}^M W_{j,k}^{\text{H}} W_{i,j}^{\text{O}} \left[1 - \tanh \left(b_j^{\text{H}} + \sum_{k=1}^M W_{j,k}^{\text{H}} p_k \right)^2 \right] \frac{dp_k}{d\mathbf{x}_a} \quad (4.6)$$

The field gradient term in Eq. 4.5 can be determined by differentiation of the permanent electric field \mathbf{E}_a^0 at QM atom a (see Appendix 4.7 Eq. A3), with respect to the position of the same center a , following the conventional point dipole formalism [158].

$$\frac{d\mathbf{E}_a^0}{d\mathbf{x}_a} = \Delta\alpha_a^c \mathbf{E}_a^0 \sum_{b=1}^S \mathbf{T}_{(2)}^{ab} Q_b \quad (4.7)$$

Where $T_{ab}^{(2)}$ is the second-order interaction tensor [169] between multipoles (see Eq. A9 of 2.8 for definition) and Q 's are the partial charges of the surrounding S solvent MM atoms. Because the electric field is collected based on QM and MM interacting pairs, there are field gradients also on each solvent MM center b , which can be obtained through a similar differentiation procedure, or simply by negating the related component in the sum over solvent in Eq. 4.7.

$$\frac{d\mathbf{E}_a^0}{d\mathbf{x}_b} = -\Delta\alpha_a^c \mathbf{E}_a^0 \mathbf{T}_{(2)}^{ab} Q_b \quad (4.8)$$

4.3 Computational Details

To validate the dp-QM/MM method, we tested it on the solution-phase S_N2 Menshutkin [45] reaction between ammonia and methyl chloride (Scheme 4.1), which has served as a paradigm system for developing QM and QM/MM free energy simulation methods [26], [34], [39], [46], [53], [69], [85], [88], [102]. For this reaction, we treat the solute molecules by QM and solvate them in a $40 \times 40 \times 40 \text{ \AA}^3$ cubic box of MM waters described by the modified TIP3P model [108]. The AM1 [24] method and the density functional theory (DFT) method B3LYP [116]–[118] with the aug-cc-pVTZ basis set [170]–[172] are used as the QM levels of theory in SE/MM and AI/MM, respectively. For the QM/MM van der Waals (vdW) interactions, we adopted the pair-specific parameters previously optimized by Gao and Xia [85] for the Menshutkin reaction, following our previous implementation [26] using the NBFix facility in CHARMM [110]. The cutoff distance for the nonbonded vdW interactions is set to 14 \AA , whereas the long-range-enabled MM/MM and QM/MM electrostatic interactions

are treated by the particle mesh Ewald (PME) [113] and QM/MM-PME [115] methods, respectively. The QM/MM system is modeled under periodic boundary conditions, where the internal geometries of water are constrained with the SHAKE [112] algorithm during all molecular dynamics (MD) simulations.

To obtain chaperone polarizabilities for the system, condensed-phase configurations are sampled along the minimum free energy path (MFEP) determined at the AM1/MM level using the string method in collective variables (CVs) as we implemented in a previous study [11] following the original procedure described by Maragliano et al. [22]. For the Menshutkin reaction (see Scheme 4.1 for the atom labels), we use the breaking and forming bond distances, i.e., the carbon-chlorine bond ($r_{\text{C-Cl}}$) and the nitrogen-carbon bond ($r_{\text{N-C}}$), as the CVs to describe the string MFEP. The MFEP represented in terms of the two CVs is discretized into 16 images of the system. For each iteration of MFEP optimization, we estimate the free energy mean force on each CV based on its fluctuation averaged over 20 ps QM/MM MD simulations in which the CVs are harmonically anchored at their previous path values with a uniform force constant of 1000 kcal/mol/Å². The detailed string simulation parameters for projection, reparametrization, and evolution of the path, as well as those for free energy profile integration, can be found in our previous work [11]. The error bars of free energy along the string MFEP are estimated based on a procedure developed by Zhu and Hummer [173], slightly modified for nonuniform-CV-grid cases [see Supporting Information Section 1 (SI.1 of B.3)].

To reduce unnecessary distortions of the system during the MD simulations, a collinear [53] restraint is placed on the N, C, and Cl atoms with the system also being maintained approximately at a C_{3v} symmetry [26]. To prevent the QM solute from drifting out of the simulation box, we also placed a spherical restraint at 8 Å away from the center using the miscellaneous mean field potential (MMFP) command in CHARMM. For the QM/MM MD simulations, we used an integration time step of 1 fs, and the system temperature and pressure are maintained at 298.15 K and 1 atm, respectively. All the SE/MM energy and force calculations were conducted using the semiempirical SQUANTM module in the CHARMM program (version c42a2) [110]. Solution-phase atomic chaperone polarizabilities are obtained based on the Hirshfeld partitioning method [167], [174], [175] using the PCM

implicit solvation model [168] in the Gaussian16 program [176] (see SI.2 of B.3 for details). Further modifications were also made to CHARMM for the implementation of chaperone polarization energy and the associated forces.

For ANN training, we used Tensorflow [40], an end-to-end open source platform for machine learning, to train our model to correlate the CV input features with the target chaperone polarizabilities. Regarding the hidden layers of the ANN, ten nodes in a single hidden layer were chosen for simplicity. Additionally, the Adam algorithm [177], a stochastic gradient descent method, is used to optimize the weights and biases in the ANN, where the learning rate hyperparameter is set to 0.001 and a mean square error metric is used for the loss function. To avoid overfitting and to maintain a certain level of flexibility for chaperone-polarizability predictions during molecular dynamics, we randomly divided the samples into the training and testing data sets, where 80% of the samples are used for ANN training, and the remaining 20% are used to verify the quality of the resulting ANN. The selected architecture and hyperparameters seem sufficient, as the chaperone-polarizability predictions perform consistently well across the testing and training data sets (see SI.3 of B.3). In addition, the use of the CV-only input features is justified because the ANNs fitting quality is improved only marginally when more internal coordinates are included (see also SI.3 of B.3). We also modified CHARMM so that the optimized ANN is reconstructed internally for efficient polarizability predictions and gradient calculations during MD simulations (see SI.4 of B.3 for an example calculation).

4.4 Results and Discussion

4.4.1 Molecular Polarizability

In Figure 4.1, we plot the molecular polarizabilities [27] for the Menshutkin reaction as a function of a one-dimensional reaction coordinate condensed from the two bond CVs (i.e., $r_{\text{C-Cl}} - r_{\text{N-C}}$); the molecular polarizabilities were obtained based on the solute geometries in the SE/MM configurations sampled in solution along the MFEP.

Specifically, the solution-phase (PCM) molecular polarizabilities computed at the AI (B3LYP/aug-cc-pVTZ) and SE (AM1) levels, as well as their difference, are compared in

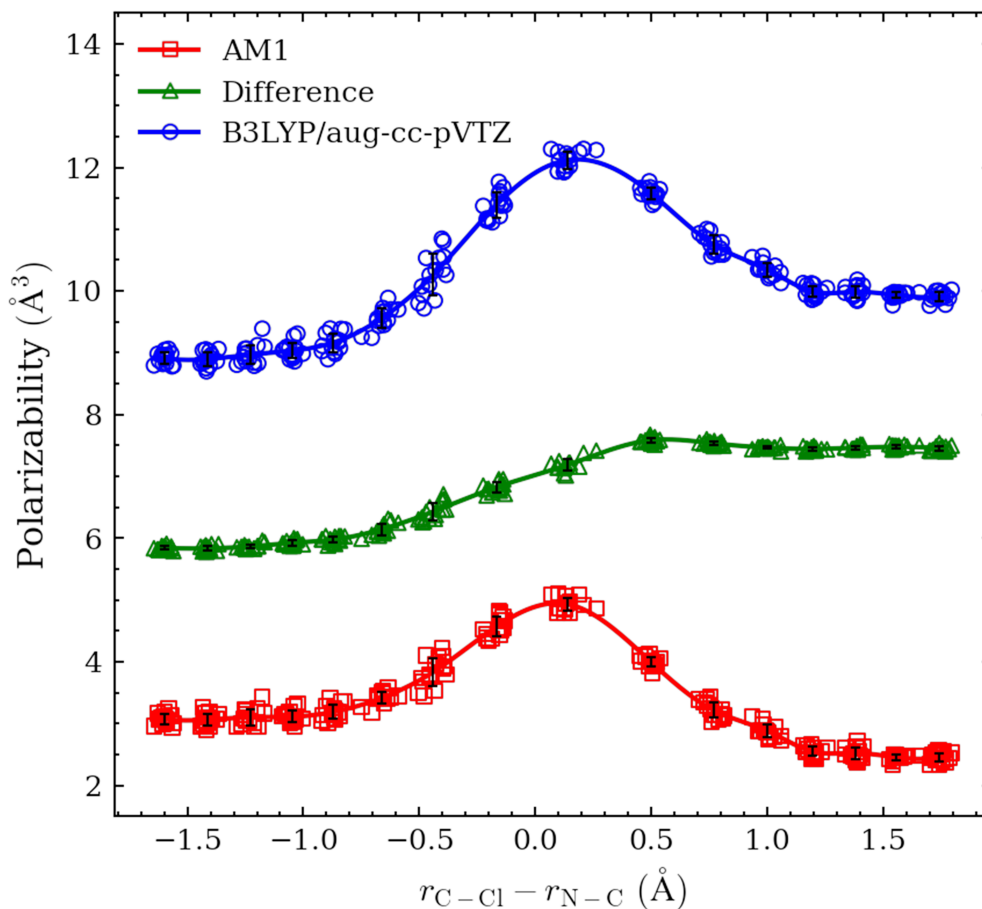


Figure 4.1. Solution-phase molecular polarizability as a function of the reaction coordinate for the Menshutkin reaction: AM1 (squares), B3LYP/aug-cc-pVTZ (circles), and their difference (triangles). The means (solid curves) and standard deviations (vertical bars) are computed based on samples within each string images (see SI.5 of B.3 for the tabulated statistical distributions).

Figure 4.1; the statistical distributions of the data are also displayed (see SI.5 of B.3 for the distribution data in a tabulated form). Using the B3LYP/aug-cc-pVTZ results as a benchmark (see SI. of B.36 for a convergence test of the benchmark with respect to the basis sets and AI levels of theory), we found that AM1 substantially underestimates the molecular polarizability for the Menshutkin reaction (Figure 4.1), which yields an overall percentage error of 67.4% along the entire MFEP. Compared with the product (P) region where the largest deviation (7.4 \AA^3 or 74.9%) occurs, the reactant (R) region displays a smaller deviation (5.8

\AA^3 or 65.5%); for the transition-state (TS) region, although a deviation similar to that in P is found (7.2 \AA^3), it gives the smallest percentage error (59.0%). This nonuniform distribution of underestimation in molecular polarizability causes significant errors in the free energy profile simulated at the AM1/MM level. More specifically, concerning the solute response to an external electric field, a greater molecular chaperone-polarizability correction of 1.6 \AA^3 is needed at the P state than at the R state for AM1 to match with the AI benchmark. Since induced dipoles in dp-QM/MM are proportional to chaperone polarizabilities (see Eq. A1 of 2.8), greater corrective dipoles are expected in the P region than in R. According to Eqs. 4.2 and A2 of 2.8, this would lead to a greater polarization energy correction that favors the free-energy stabilization of the product state.

Figure 4.2 shows the correlations between the solution-phase SE and AI molecular polarizabilities before and after the chaperone corrections predicted by our ANN model are applied. Compared with B3LYP/aug-cc-pVTZ, AM1 generates an average error of 6.78 \AA^3 (or a percentage error of 67.4%) for the molecular polarizability (data labeled as AM1).

The distribution of the AM1 data suggests the existence of two populations that contribute differently to the overall deviation. Further analysis shows a greater error in the product-formation branch of the MFEP than in the TS-formation branch (see SI.7 of B.3). With the ANN-based chaperone corrections (data labeled as $\text{AM1} + \Delta\alpha^c$), the B3LYP benchmark molecular polarizabilities are faithfully reproduced with only a small error of 0.03 \AA^3 (or 0.3%) along the entire MFEP; therefore, the bimodal error distribution is also eliminated. These results confirm the effectiveness of our ANN machine learning model in fitting molecular polarizability at the target AI level.

4.4.2 Free Energy Profile

Compared to experiments, AM1/MM is known to overestimate both the free energy barrier and the reaction free energy for the Menshutkin reaction [26], [85], [88]. As shown in Figure 4.3 and Table 4.1, the AM1/MM simulations we conducted here yield a free energy barrier of 28.3 kcal/mol , which is 4.8 kcal/mol higher than the experimental value of 23.5 kcal/mol [49].

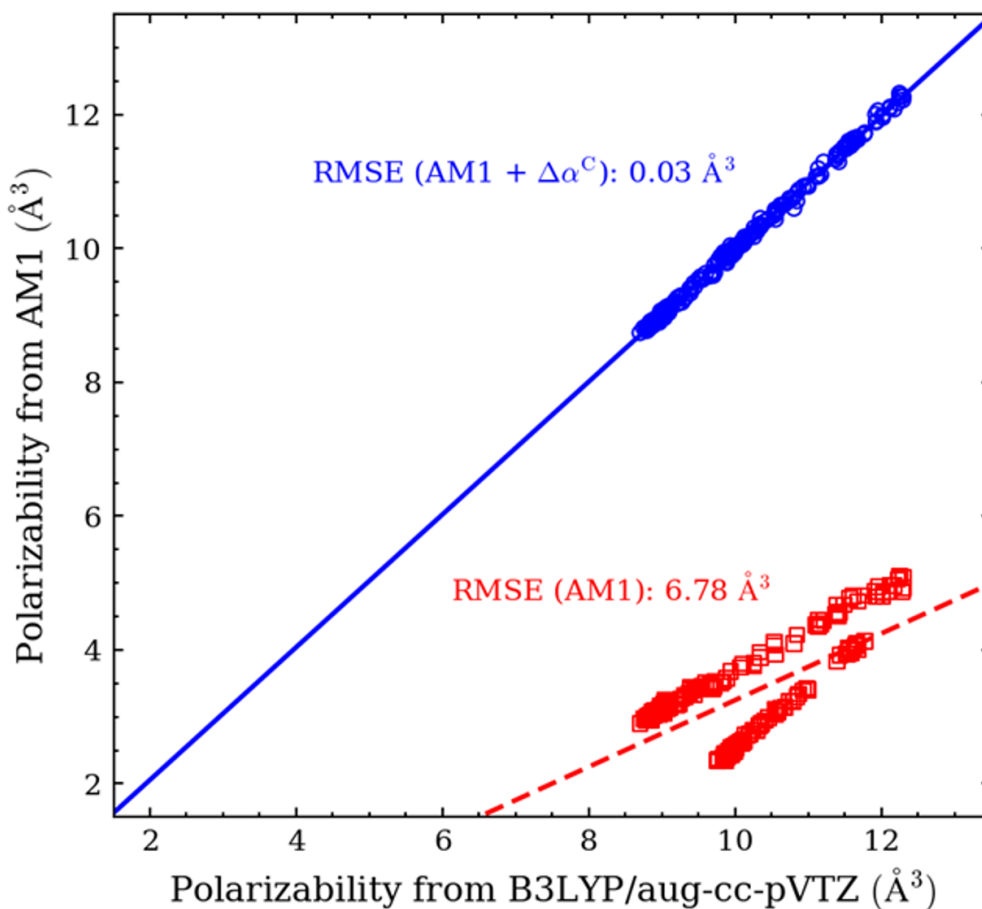


Figure 4.2. Regressions of the molecular polarizabilities from AM1 (square with a dashed line) and from the chaperon-corrected AM1 (circles with a solid line) against the molecular polarizability from B3LYP/aug-cc-pVTZ; the corresponding root-mean-square-errors (RMSEs) in polarizability compared with the AI target values are also shown.

The free energy of reaction obtained for the Menshutkin reaction from our dp-AM1/MM simulations is 13.3 kcal/mol, which corresponds to an overestimation of 20.7 kcal/mol compared with the experimental value of 34 ± 10 kcal/mol [49]. With polarizabilities corrected to the B3LYP/aug-cc-pVTZ level, the dp-AM1/MM simulation gives a free energy barrier and reaction free energy of 22.5 and 32.8 kcal/mol, respectively, with the corresponding errors relative to the experiment reduced to 1.0 and 1.2 kcal/mol.

Table 4.1. Free energy barrier (ΔG^\ddagger) and reaction free energy (ΔG) of the Menshutkin reaction in solution.

	(kcal/mol)	
	ΔG^\ddagger	ΔG
AM1/MM ^a	28.3	-13.3
dp-AM1/MM ^{a,b}	22.5	-32.8
expt. ^c	23.5	-34 \pm 10

^a this work

^b with polarizabilities corrected to B3LYP/aug-cc-pVTZ

^c Experimental values taken from Gao *et al.* (Ref. **dp77**).

Note that for the original AM1/MM simulations the error in the reaction free energy (20.7 kcal/mol) is more than fourfold of that in the free energy barrier (4.8 kcal/mol); therefore, the reduction of both free-energy errors to around 1.0 kcal/mol indicates a significant variation of the chaperone polarization correction in different regions of the reaction. Because the chaperone polarizabilities needed for AM1 to match with B3LYP/aug-cc-pVTZ are positive, the associated polarization energy correction is always negative (see Eq. 4.2), which leads to free-energy stabilization. Qualitatively, the observed changes in free energy are in accord with the molecular polarizability correction in Figure 4.1. The correction in molecular polarizability increases from around 6 Å³ in reactant to around 7 Å³ in the TS (located at $r_{\text{C-Cl}}-r_{\text{N-C}} = 0.14$ Å on the AM1/MM free energy profile), which continues its growth toward the product-forming side with a flattened peak of 7.4 Å³ found at $r_{\text{C-Cl}}-r_{\text{N-C}} = +0.5$ Å, followed by a slightly diminished plateau of 7.2 Å³ in the product region (see Figure 4.1). Based on the evolvement of the molecular polarizability correction itself, a greater amount of chaperone polarization energy is indeed expected in the product state than in the transition state, which explains a stronger stabilization correction to the reaction free energy than to the free energy barrier. Moreover, when the shift of the TS location is taken into account, the polarization correction for the free energy barrier will be further reduced (see Section 4.4.3).

As far as the product state is concerned, a greater chaperone polarizability results in a stronger corrective induced dipole on the solute (see eq A1 in Appendix 4.7), which leads to a greater amount of solvent-induced intermolecular polarization energy in the product than

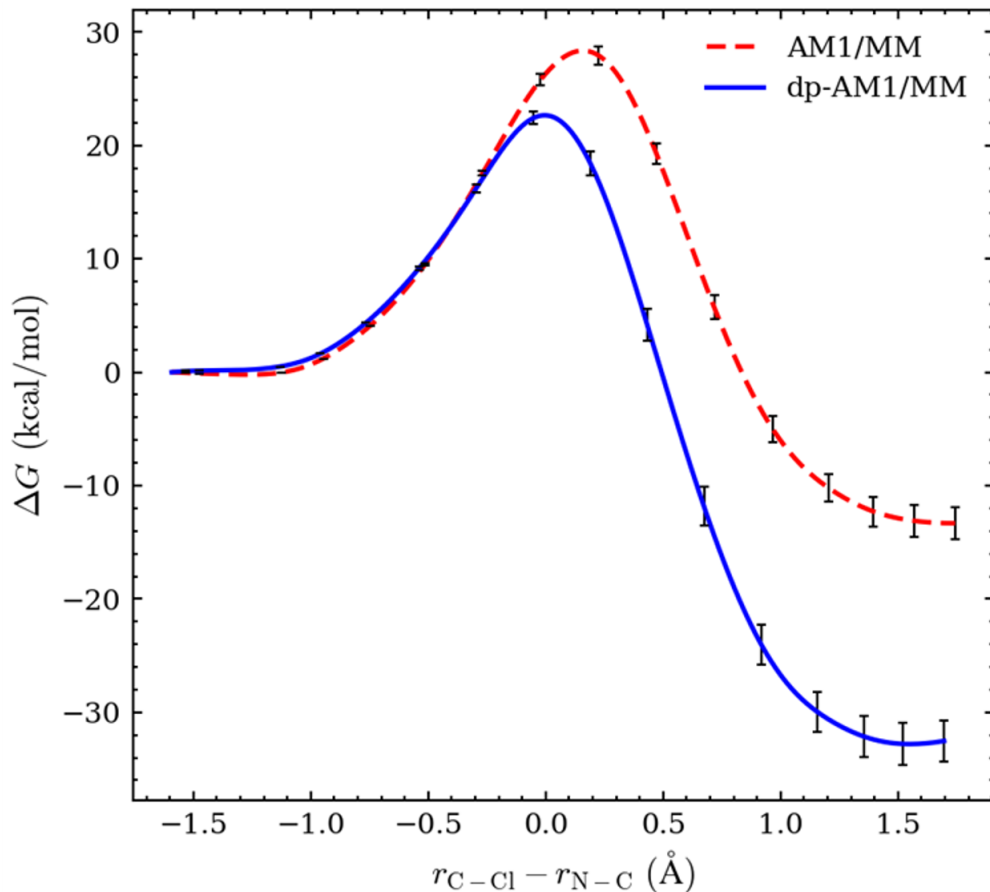


Figure 4.3. Free energy profiles as a function of the reaction coordinate for the Menshutkin reaction: AM1/MM (dashed line) and dp-QM/MM with polarizabilities corrected to B3LYP/aug-cc-pVTZ (solid line). The error bars relative to the free energy in the reactant state ($\alpha = 0$) along the string MFEP are estimated using a procedure developed by Zhu and Hummer (Ref. [22]), slightly modified for non-uniform collective-variable grids (see SI.1 of B.3 for details).

in the reactant. The lowered free energy of the reaction observed here, therefore, is in line with a greater induction contribution enhanced by the chaperone-polarizability correction upon product formation.

From a free energy perturbation perspective, the changes we see in the free energy profile when chaperone polarization is incorporated can be connected to the ensemble-averaged polarization energy correction $\langle \Delta E^{\text{pol}} \rangle$, for which both its polarizability and electric field

components (see eq 4.2) may vary along the reaction coordinate (RC). To see if the use of RC-dependent polarizability corrections is essential, we also tested a set of equally divided RC-independent atomic polarizabilities that yield an average molecular polarizability correction of 6.78 \AA^3 (i.e., the RMSE between AM1 and B3LYP in Figure 2). Our additional test shows substantial differences in $\langle \Delta E^{\text{pol}} \rangle$ when the uses of RC-dependent and RC-independent chaperone polarizabilities are compared (see SI.8 of B.3); this further suggests that reactive fitting of polarizability corrections is important.

The free energy profile for the Menshutkin reaction has been determined at the B3LYP/def2-TZVPPD/MM level using the weighted thermodynamics perturbation (wTP) method based on the force-matching recalibrated PM3*/MM potential, [69] at the B3LYP/6-31+G(d,p)/MM level using the internal CV force-matching corrections based on AM1/MM [26], and at the PM3*:B3LYP/ 6-31G(d)/MM level with the assistance of machine-learning potentials [39]; these AI/MM-quality simulations give free energy barriers of 18.5 [69], 14.7 [26], and 15.1 [39] kcal/mol, respectively. The observed trend is that although B3LYP underestimates the free energy barrier for the Menshutkin reaction, the barrier height tends to increase with the size of the basis set. The free energy barrier of 22.5 kcal/mol obtained here when we fit polarizability to the B3LYP/aug-cc-pVTZ level is in line with this trend, although neither energy nor force is fitted in dp-QM/MM. The quantitative differences in these benchmark results may also be related to a few simulation details including the use of a collinear vs. bent transition-state structure, which is known to introduce a small but noticeable difference (0.51 kcal/mol) in the barrier height [69].

4.4.3 Comparison of MFEPs and Shift of Transition State

Because chaperone polarizabilities also change the dynamics of the system through modified atomic forces, we compared the MFEPs obtained from the original AM1/MM and chaperone-corrected dp-AM1/MM simulations to see if the changes in the free energy profile are coupled to any significant perturbation of the free energy path (Figure 4.4).

We found no significant changes in the MFEP after adding the polarizability correction. Therefore, the improved free energy profile is attributed more to the corrected response

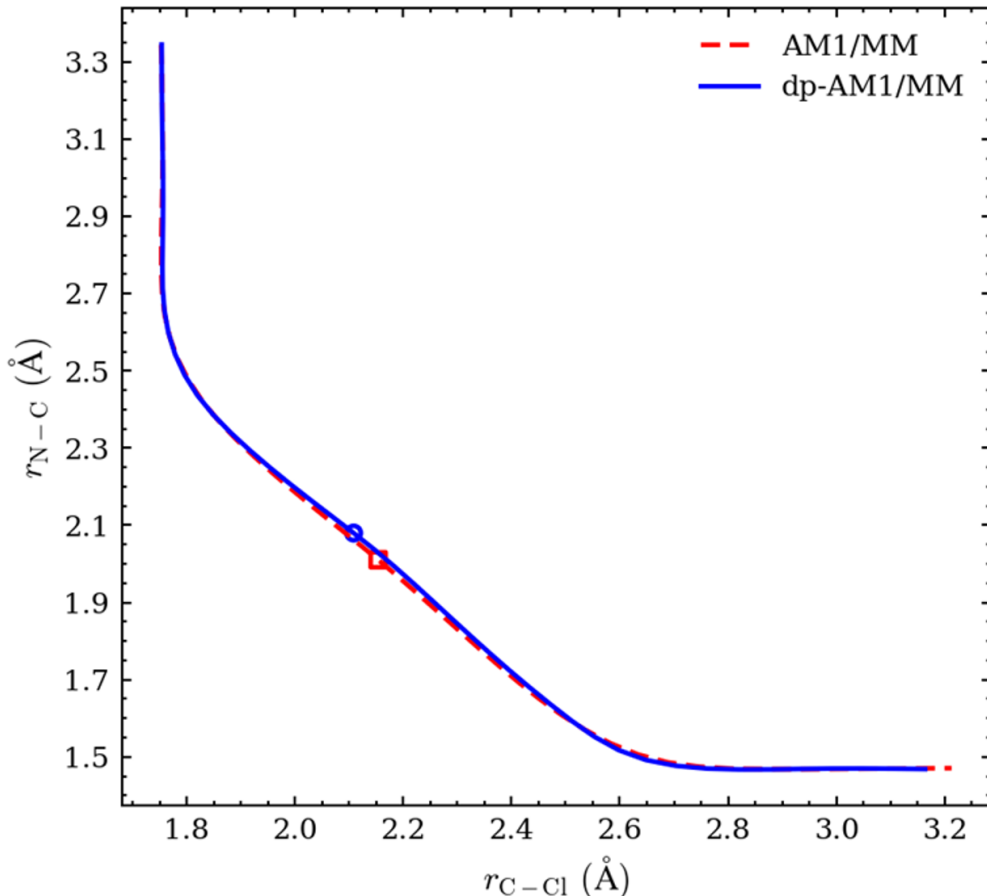


Figure 4.4. The minimum free energy path (MFEPs) as a function of the collective variables, i.e., the C–Cl and N–C bonds: AM1/MM (dashed line) and the dp-AM1/MM with chaperone polarizabilities corrected to the B3LYP/aug-cc-pVTZ level (solid line). The locations of free energy transition states are also marked: AM1/MM (open square) and dp-AM1/MM (open circle).

of the QM subsystem to solvation than to the changes of intramolecular interactions in the solute. This observation is consistent with our definition of the chaperone polarization energy in eq 4.2, in which the solute polarizabilities only interact with the surrounding solvent electric fields; because the mutual polarization of solute polarizabilities is neglected in this work, intramolecular forces in the solute are not directly perturbed. Our previous work shows that the AM1/MM MFEP for the Menshutkin reaction can be improved to an AI/MM-level of quality through intramolecular force matching [26]; therefore, a combination

of the chaperone-polarizability treatment and force matching is expected to improve both solutesolvent and solutesolute interactions in a balanced manner, which is our work underway.

Although the chaperone-polarizability correction does not move the MFEP, it does generate a notable shift in the location of the free energy transition state (TS). As shown in Figure 4, at the AM1/MM level, the free energy TS is located at $r_{\text{C-Cl}} = 2.15 \text{ \AA}$ and $r_{\text{N-C}} = 2.01 \text{ \AA}$, which is moved to $r_{\text{C-Cl}} = 2.11 \text{ \AA}$ and $r_{\text{N-C}} = 2.08 \text{ \AA}$ after the chaperone correction is turned on at the dp-AM1/MM level; this shift corresponds to an earlier TS on the chaperone-corrected MFEP, as the breaking bond $r_{\text{C-Cl}}$ is found at a shorter distance, whereas the forming bond $r_{\text{N-C}}$ becomes longer. Since intramolecular interactions in the solute are not directly perturbed here, the shift of the overall free energy TS can be traced back to a variational solvation effect associated with the chaperone polarization. In particular, for the TS located on the dp-AM1/MM MFEP, although the solute structure has not yet advanced to the energetic bottleneck in terms of intramolecular bond breaking/forming if the original AM1/MM description is followed, the build-up of its charge-transfer character has led to a chaperone polarization strong enough to make solvation stabilization a dominant free energy factor at and beyond this point. The shift of the TS to an earlier RC location can also be seen from the free energy profiles in Figure 3, where the $r_{\text{C-Cl}}-r_{\text{N-C}}$ value for the TS is moved from 0.14 \AA (AM1/MM) to 0.03 \AA (dp-AM1/MM). As this shift leads to a reduction in chaperone polarizability along the RC (see Figure 1), the variation of the TS location further contributes to the smaller change in the free energy barrier than in the reaction free energy after the chaperone polarization is turned on for AM1/MM.

4.4.4 Atomic Chaperone Polarizability

In Sections 4.1–4.3, we have shown the overall behaviors of molecular polarizability, free energy profile, and free energy path after the chaperone corrections are introduced. To understand the physical origin of these changes, we further examined the atomic chaperone polarizabilities as a function of the reaction coordinate (Figure 4.5).

Based on a comparison of the atomic polarizabilities in the solute, it is evident that the chlorine atom contributes most to the polarization correction in the free energy profile. When

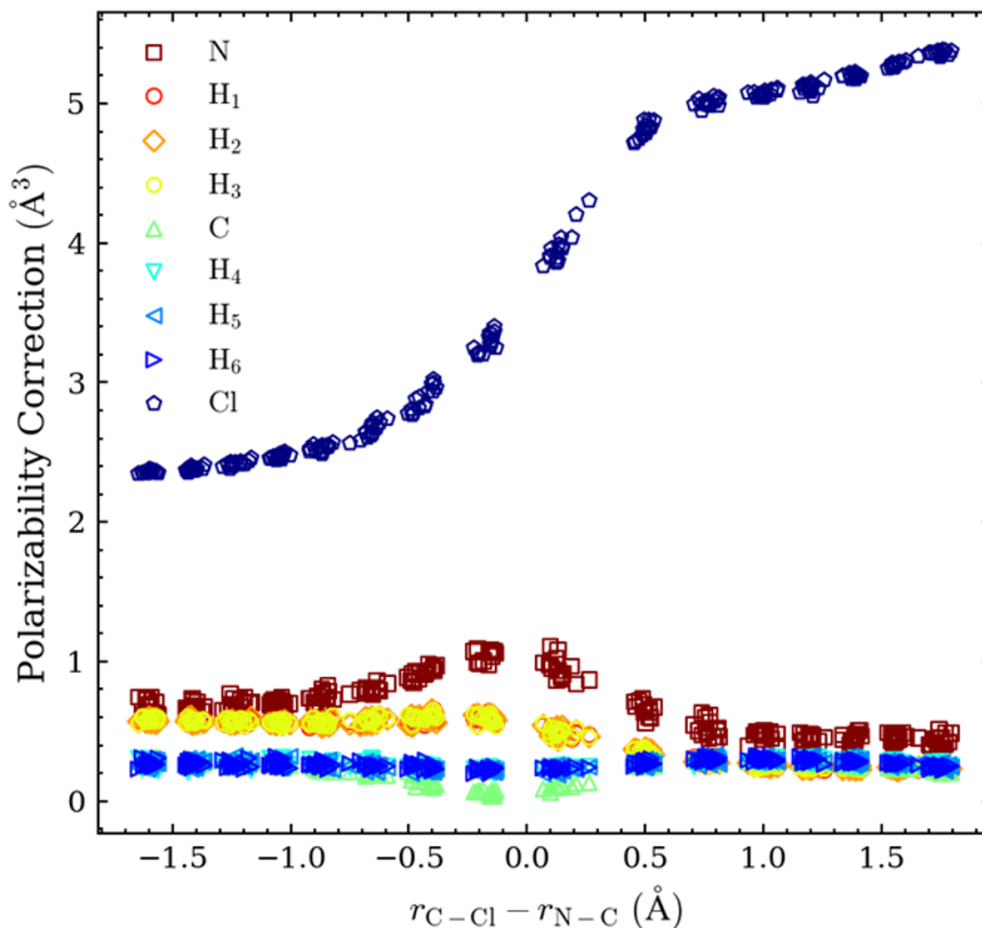


Figure 4.5. Atomic chaperone polarizabilities as a function of reaction coordinate for each solute atom in the Menshutkin reaction.

dissociated from the C–Cl covalent bond in the reactant to form the fully charged anion in the product, Cl increases its chaperone polarizability from 2.4 to 5.5 Å³, which corresponds to a change of more than twofold. Because chaperone polarizability reflects the extent to which AM1 underestimates the electronic polarization response for a given species, we can try to rationalize why the polarizability on Cl in the product is severely underestimated. Under the minimum basis convention used in AM1, higher-order angular momentum basis functions such as d-orbitals and additional diffuse functions are lacking; both of these are essential for a proper description of polarizability for large anions such as Cl[−].

The sensitivity of Cl’s atomic polarizability with respect to the AI levels of theory in the benchmark is further tested using an expanded list of basis sets, including Pople’s Gaussian basis sets [122], [124]–[126], [178], [179], Dunning’s correlation-consistent basis sets [170]–[172] (aug-cc-pVXZ, where X = D, T, Q), and Jensen’s polarization-consistent basis sets [180]–[183] (aug-pc-n, where n = 2, 3, 4), combined with an additional DFT method M06-2X [184] and the second-order Møller-Plesset perturbation [120] (MP2) theory (Figure 4.6).

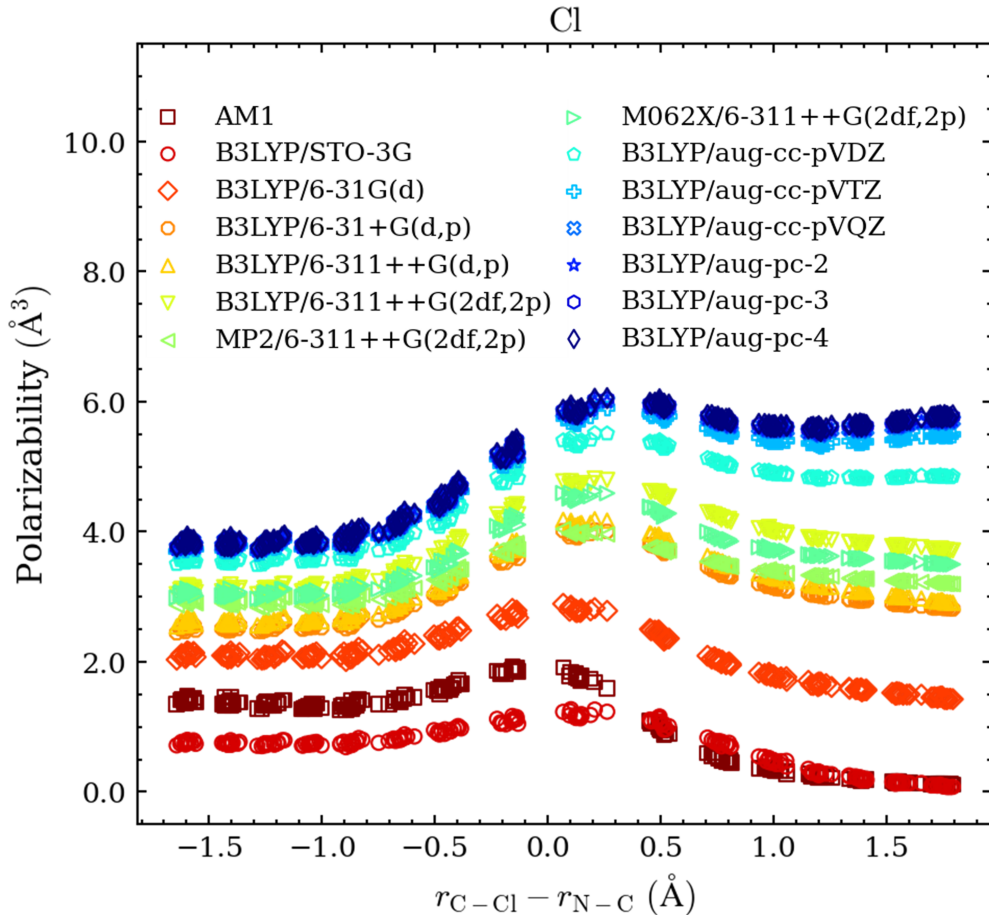


Figure 4.6. Convergence of atomic polarizability on the chlorine atom with respect to basis sets and AI methods.

The results confirm the general correlation between underestimation of polarizability and the use of limited basis sets. Specifically, when AI methods are used with an insufficient basis set (e.g., STO-3G), they also produce too low polarizabilities comparable to the AM1 result. With increases in the basis set size, greater AI polarizabilities are obtained for the chloride

anion, which converge to the results of using aug-cc-pVQZ and aug-pc-4, the two largest basis sets tested in this work.

In addition to the significant polarizability correction on Cl, which is expected to contribute predominantly to the overall change in the free energy result, the chaperone polarizabilities on other solute atoms also contribute to the change, although to a lesser degree (see SI.9 of B.3). Generally, we found that greater corrections are needed for the solvent-accessible terminal atoms than for the atoms buried inside; this is in line with quenched polarizabilities for interior atoms in molecules reported by Marenich et al. [167]. For example, the chaperone polarizabilities for the terminal hydrogens in NH_3 ($>0.5 \text{ \AA}^3$) are greater than those for the interior hydrogens in CH_3 ($<0.5 \text{ \AA}^3$) in the reactant state. When the atomic chaperone polarizabilities are summed into groups, greater corrections are found on the terminal groups (NH_3 and Cl) than on the inner group (CH_3). Altogether, these results suggest that the atomic chaperone polarizabilities are dependent on the local environment and that the atoms exposed to the solvent are more influenced by the polarizability corrections.

4.4.5 Local Environment and RDF

Polarization enhanced by chaperone polarizabilities on the solute would also generate significant impact on the local solvent environment. To characterize the changes in solvent structure, we compared the pair radial distribution functions (RDFs) between the solute atoms and water oxygens (O_w) obtained from the AM1/MM and dp-AM1/MM simulations. Because the chlorine atom hosts the largest polarizability correction, we highlight the RDFs for Cl- O_w in Figure 4.7, whereas the RDFs for other solute atoms can be found in SI.9 of B.3.

For Cl, a terminal atom exposed to the solvent, after the chaperone-polarizability correction is invoked, the peak height for its first solvation shell in the RDF of the charge-separated product (P) state is increased and the corresponding peak position is shifted from 3.3 \AA (AM1/MM) to a shorter distance of 3.0 \AA (dp-AM1/MM). Additionally, the short-range solvent structure also becomes more ordered as indicated by a second solvation shell

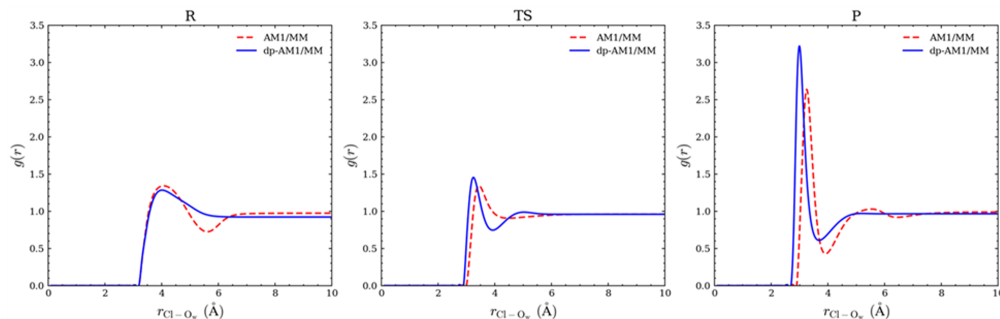


Figure 4.7. Radial distribution functions between the chlorine atom and water oxygens (O_w) in the reactant (R), transition-state (TS), and product (P) region of the Menshutkin reaction: AM1/MM (dashed line) and dp-AM1/MM with polarizabilities corrected to the B3LYP/aug-cc-pVTZ level (solid line).

that moves closer to the solute atom under the influence of chaperone polarization. Similar chaperone-enhanced solvation, although less pronounced, is also observed in the TS region, where charges are only partially separated. On the contrary, in the charge-neutral reactant (R) state, the chaperone-polarizability correction only plays a minor role in perturbing the solvent structure and no significant changes are found for the first solvation shell in the RDF. Thus, induced local solvation enhancement for Cl in the TS and P regions, reflected from the RDFs, provides another qualitative explanation for the lowered free energy barrier and reaction free energy observed in the chaperone-corrected dp-AM1/MM simulations.

4.4.6 Convergence of Atomic Polarization Energy with Solvent Inclusion

In the present study, we use a spherical cutoff distance when collecting the solvent electric fields acted on the solute chaperone polarizabilities (see Appendix 4.7), and therefore correction due to long-range solvent polarization is neglected. The adoption of this treatment assumes that the difference between SE/MM and AI/MM polarizations is dominated by short-range interactions. To further justify this treatment, we examined the convergence of the chaperone polarization energy with respect to the size of the solvent sphere included. Quantitatively, we first decomposed the chaperone polarization energy in eq 4.2 into individual QM solute atomic contributions. Then the polarization energy for a given QM center was

further decomposed into its pair-wise interactions with individual surrounding MM atoms. To examine the convergence of the chaperone polarization energy with respect to solvent molecules included in the calculation, we collected the polarization energy on each solute atom based on its interactions with individual water molecules; this was done by summing the pair-wise atomic polarization energies over the constituent hydrogen and oxygen (O_w) atoms in each water molecule. In Figure 4.8, we plot the decomposed chaperone polarization energies on the Cl atom, generated from 240 QM/MM configurations sampled in three string MFEP trajectories (960 ps in total), both as a function of solutesolvent distances (here, the Cl- O_w distance is used) and as a function of the reaction coordinate ($r_{C-Cl} - r_{N-C}$); the corresponding polarization energies at the AM1/MM and B3LYP/aug-cc-pVTZ/MM levels, approximated classically with the same decomposition using the Hirshfeld partitioned atomic polarizabilities, are also shown for comparison. Similar plots for other solute atoms can be found in SI.9 of B.3.

The “waterfall”-shaped distribution of the polarization energies in Figure 4.8 shows that the chaperone polarization energy that corrects the difference between AM1/MM and B3LYP/aug-cc-pVTZ/MM quickly decays and becomes negligible for solvent molecules beyond 10 Å away from Cl, which further justifies the use of a 12 Å cutoff distance for chaperone polarization in this study (see Appendix 4.7). A similar distance dependence of polarization energy has been reported for QM/MM calculations using polarizable force fields [185]. This decay behavior is understandable since the permanent electric field for determining chaperone polarization energy is inversely proportional to the cube of the solute-solvent distance (see also Appendix 4.7). The comparison between the AM1/MM and B3LYP/aug-cc-pVTZ/MM polarization energies as well as a much more negative chaperone polarization contribution in the product region than in the reactant region clearly shows a severe underestimation of polarization for the Cl^- anion in AM1/MM. Interestingly, the most significant polarization energy difference between the two levels is found at a solute-solvent distance of about 3.0 Å, which is in line with the peak position of the first solvation shell identified for Cl- O_w on its RDF in the product region as shown in Figure 4.7. Thus, in accord with a tighter solvation-shell structure, the polarization energy is also consistently corrected.

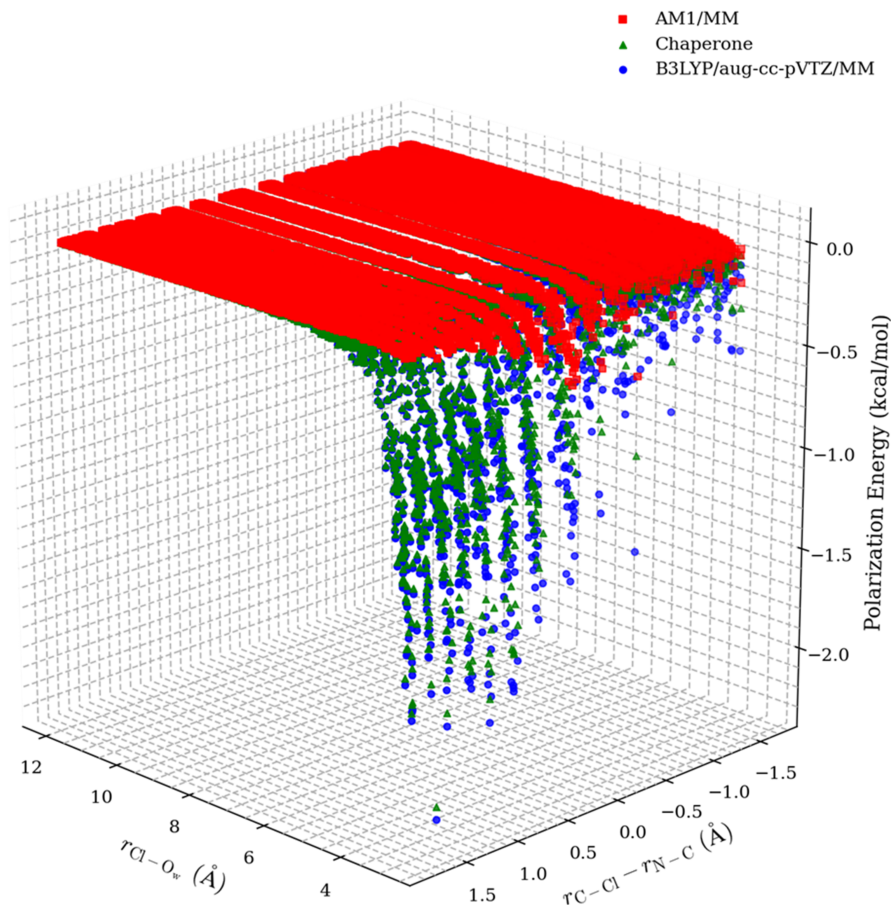


Figure 4.8. Decomposition of polarization energy to each water molecule around the chlorine atom: AM1/MM (squares), chaperone polarization energy (triangles), and B3LYP/aug cc-pVTZ/MM (circles).

4.5 Outlook

Due to the important roles played by intermolecular polarization in simulations of condensed-phase reactions, the ability to reliably predict molecular polarizability has been identified as one of the key elements for designing the next-generation SE methods [142]. Despite an impressively diverse collection of ingenious efforts in method development in this area, the related issues have not been fully resolved in a satisfactory manner. Within a QM framework, Clark and co-workers [139] adopted a post-self-consistent-field reparametrization strategy,

where the expectation value of polarizability is fitted to experiments by reparametrizing the related dipole and quadrupole types of integrals reusing the density matrix variationally converged from the standard SE calculations. Although offering accurate molecular polarizabilities for modeling quantitative structure–property relationship (QSPR) [139], their reparametrized variational treatment lacks a consistent incorporation of the improved polarizabilities into SE Hamiltonians for conducting condensed–phase calculations. The CPE [142], [186] and PMO [149] methods, on the other hand, improve polarizability by directly modifying SE Hamiltonians and wave functions; because both methods build chemical–environment information into polarizability, either through charge–dependent response density [142], [186] or through more polarizable wave functions enabled by additional atomic basis functions [149], they offer promising cures for the underpolarization problem at a more fundamental level. As the parametrized versions of CPE and PMO methods rely on fitting of gas–phase polarizabilities for stationary species in molecular databases, their applicability to free energy simulations of condensed–phase reactions has yet to be determined; often the quality of polarizability fitting has to be offset by the competing objective of obtaining desirable condensed–phase properties [149]. In comparison to these QM–based strategies, ML has recently been used as a pure classical fitting tool to generate highly accurate gas–phase molecular polarizabilities for druglike stable molecules [38]. Because of its nonreactive nature, there is a similar question whether the resulting AlphaML model can be directly used for simulating chemical reactions in condensed phases.

Compared with these existing approaches, the dp–QM/MM method we present here offers a few advantages. First, our method is a condensed–phase QM/MM technique, where we focus on recovering the proper polarization response for the solvated SE wave functions. In terms of fitting condensed–phase properties, our method is connected in spirit to the strategy employed by Gao and co–workers when developing dipole preserving and polarization–consistent charges (DPPC) [166]. In the DPPC method, polarizability is introduced as an additional constraint to serve as a vehicle for fitting condensed–phase atomic charges that are needed to preserve the induced dipole effect [166], whereas in dp–QM/MM, we directly fit chaperone polarizabilities and use them explicitly in polarization calculations. Second, the use of classical polarizability in dp–QM/MM allows us to construct a chaper-

one polarization energy term in the Hamiltonian with consistent nuclear gradients available, with which MD can be conducted for polarization-corrected free energy simulations. Third, our polarizability fitting is reactive, as we use ML to fit condensed-phase polarizabilities as a function of CV-based reaction coordinates along the MFEP. The free energy simulation results for the Menshutkin reaction and the related analyses indicate that the improvements made to the free energy profile are associated more with the newly incorporated physics than with the parametrization itself. As chaperone polarizabilities are applied in a corrective manner to recover the missing polarization that is mainly caused by the use of limited basis sets, the base level in the dp-QM/MM method is not limited to SE/MM; one can also apply the method between two AI/MM levels to produce the large-basis-set polarization results at the cost of a small-basis-set simulation.

There are still a few limitations in our current implementation of the dp-QM/MM method. First, in this work, only QM atoms are polarizable by MM charges, but not vice versa. One future direction is to extend the method to make it work with polarizable MM particles, with which the chaperone polarizabilities would become visible to the MM polarizable centers as well for mutual polarization. Second, in the present formalism, although QM atoms host classical polarizabilities, we do not allow them to interact with each other and therefore their mutual presence does not perturb the SE Hamiltonian to directly make the wave function more polarizable; this omission is made in part to avoid double counting for the solute-solute (QM-QM) intramolecular polarization, which is already covered by the SE wave function calculation. The magnitude of the error for neglecting this intramolecular chaperone polarization contribution is yet to be quantified. With short-range polarization properly damped, intramolecular chaperone polarization energy can be otherwise incorporated in induced dipole calculations based on the electric fields created using QM atomic charges, which would eventually lead to a perturbed SE wave function through the familiar Fock matrix modification procedure. Third, in this work, the target chaperone polarizabilities are determined approximately using implicit solvation calculations (e.g., with PCM). This choice, compared with explicit QM/MM, is technically more convenient for solution-phase systems with regard to a consistent electrostatics treatment when computing polarizability at both the SE and AI levels. For the method to be applied to heterogeneous

systems, such as enzymes, determination of chaperone polarizabilities using an explicit MM representation of the environment needs to be tested. Fourth, in our ML model, we place quantum mechanically prepartitioned atomic polarizabilities in the output layer of the ANN as targets for the ML fit. An alternative strategy is to leave decomposition of molecular polarizability to ML as a pattern-recognition task. For example, the AlphaML approach developed by Wilkins et al. [38] uses a symmetry-adapted Gaussian process regression (SA-GPR) model, which contains atomic-centered kernels that naturally yield additive atomic polarizabilities as a byproduct when fitting molecular polarizability as the direct observable for nonreactive charge-neutral systems. Decomposition of chaperone polarizabilities with ML can be tested for dp-QM/MM in the future; the requirement of reactive fitting, however, could make our case more challenging for the ML-based partition than with the physics-based prepartition scheme. Fifth, in our ANN model, only a few essential internal coordinates of the solute are used as input features, with which the symmetry of permutational invariance is not strictly imposed. Although this simple input scheme seems to work well for the Menshutkin reaction, it is highly desirable to extend input features using a generalized chemical environment representing scheme (e.g., SOAP [187]) so that solvent and other local coordinates can be treated in a permutationally invariant manner when needed. Due to a strong correlation between polarizability of a species and its charge state, which forms the foundation of the CPE-based response density method [142], it would be valuable to develop a charge-dependent ML-polarizability model that allows atomic charges to be explicitly incorporated in the input features; as population charges are considered a general property of QM wave function, the use of QM charges as input would also make the model more transferrable across different chemical situations. For the Menshutkin reaction, the charge-dependent scheme seems to be promising because the evolvments of Mulliken charges also synchronize well with the chemical-bond-based reaction coordinate (see SI.10 of B.3). Sixth, as shown in the MFEP and RDF results for the Menshutkin reaction, the chaperone-polarizability treatment mainly improves the description of solute-solvent intermolecular interactions, but its impact on improving intramolecular interactions in the solute is rather limited. A combination of our recently developed intramolecular reaction path force matching (RP-FM) methods [26], [28] and the dp-QM/MM method is expected to improve

both intra- and intermolecular interactions at SE/MM levels in a more balanced way. Finally, the idea of chaperone polarizability could also be applicable to pure SE simulations, for example, through the linear-scaling framework [141], for improving their predictions of condensed-phase bulk properties.

4.6 Concluding Remarks

In this paper, we have presented the dp-QM/MM method, a dual-level method that utilizes polarization-adequate AI information to guide an underpolarized SE method to respond properly to MM electric fields in SE/MM calculations. Formulated in a hybrid framework, the dp-QM/MM method corrects the underestimated polarizabilities at the SE/MM level to their condensed-phase AI targets along the MFEP by a set of ML classical chaperone polarizabilities. For the Menshutkin reaction we tested, the dp-QM/MM method greatly improves the quality of the free energy results to a level comparable to the AI/MM and experimental benchmarks. Overall, this method offers a new strategy for an improved description of intermolecular polarization in SE/MM free energy simulations. We expect to see further applications of this method to complex chemical and biochemical reactions, where the role of polarizability can never be underestimated.

4.7 Appendix

The system describing electrostatic interactions is defined as a collection of point charges and polarizable sites on which non-permanent dipoles are induced in response to the instantaneous electric field experienced at each site. Here, a chaperone polarizability, denoted by $\Delta\alpha_i^C$, can be assigned to each polarizable site i , to describe the correction needed at the indicated center in response to electric polarization. Assuming a linear response, a corrective induced dipole will be created at each QM atom center, as a result of polarization of the chaperone polarizability by the electric field acting on it,

$$\Delta\boldsymbol{\mu}_i^{\text{ind}} = \Delta\alpha_i^C \mathbf{E}_i^{\text{tot}} = \Delta\alpha_i^C (\mathbf{E}_i^{\text{ind}} + \mathbf{E}_i^0) = \Delta\alpha_i^C \mathbf{E}_i^0 \quad (4.9)$$

where μ_i^{ind} is the corrective induced dipole generated at the QM atom i . In Eq. (4.9), $\mathbf{E}_i^{\text{tot}}$, $\mathbf{E}_i^{\text{ind}}$, and \mathbf{E}_i^0 represent the total, induced, and permanent electric fields at QM atom i , respectively; note that in our current treatment, the chaperone polarizabilities are only added on the QM atoms to compensate for the missing QM/MM polarization, therefore neither polarization of MM atoms nor induced dipole-induced dipole interactions among QM atoms are considered, which makes the induced electric field contribution vanish (i.e., $\mathbf{E}_i^{\text{ind}} = 0$), and consequently the total electric field $\mathbf{E}_i^{\text{tot}}$ is the same as the permanent electric field \mathbf{E}_i^0 caused by the MM charges. For a QM subsystem of N chaperone-polarizability hosting atoms, the polarization energy correction is:

$$\Delta E^{\text{pol}} = -\frac{1}{2} \sum_{i=1}^N \Delta\boldsymbol{\mu}_i^{\text{ind}} \cdot \mathbf{E}_i^0 = -\frac{1}{2} \sum_{i=1}^N \Delta\alpha_i^C |\mathbf{E}_i^0|^2 \quad (4.10)$$

Note that the total polarization energy correction in Eq. (4.10) is derived from the corrective induced dipoles and is equivalent to the expression in Eq. (4.2) that directly uses polarizabilities. The permanent electric field \mathbf{E}_i^0 in Eq. (4.10) collects the electric fields (coulomb force per unit charge) generated at QM atom i by the point charges on the MM atoms (denoted \mathbf{Q}) within a sufficient cutoff of 12 Å, with each of its Cartesian component (in the x , y , and z directions) expressed as:

$$E_{i,\alpha}^0 = - \sum_{j \in \text{MM}} T_{\alpha}^{ij} Q_j \quad (\alpha = x, y, z) \quad (4.11)$$

Following the conventional notation of the interaction tensor between a pair of point charges at centers i and j , the many-body interaction tensor for interacting multipoles can be obtained by applying a derivative operation repeatedly: [169]

$$T_{\alpha\beta\gamma\dots\omega}^{ij} = \nabla_{\alpha}^i \nabla_{\beta}^i \nabla_{\gamma}^i \dots \nabla_{\omega}^i T^{ij} \quad (\alpha, \beta, \gamma, \dots, \omega \in x, y, z) \quad (4.12)$$

where the zeroth-order tensor T^{ij} is given as the reciprocal of the distance between centers i and j (denoted r_{ij}):

$$T^{ij} = \frac{1}{r_{ij}} \quad (4.13)$$

The first-order tensor describing the permanent electric field in Eq. (4.11) is then given as:

$$T_{\alpha}^{ij} = - \frac{\alpha_{ij}}{r_{ij}^3} \quad (\alpha = x, y, z) \quad (4.14)$$

Conveniently, one can write the first-order many-body interaction tensor in a vector form:

$$\mathbf{T}_{(1)}^{ij} = -\mathbf{T}_{(1)}^{ji} = \begin{pmatrix} -\frac{x}{r^3} \\ -\frac{y}{r^3} \\ -\frac{z}{r^3} \end{pmatrix}_{ij} = - \frac{\mathbf{r}_{ij}}{|\mathbf{r}_{ij}|^3} \quad (4.15)$$

where \mathbf{r}_{ij} denotes the vector pointing from center j to center i :

$$\mathbf{r}_{ij} = \begin{pmatrix} x_i - x_j \\ y_i - y_j \\ z_i - z_j \end{pmatrix} \quad (4.16)$$

Based on Eqs. (4.12, 4.13, 4.14, 4.15, 4.16), the second-order many-body interaction tensor $\mathbf{T}_{(2)}$ used in Eq. (4.7) and Eq. (4.8) for computing the chaperone gradients can be written explicitly as:

$$\mathbf{T}_{(2)}^{ij} = \mathbf{T}_{(2)}^{ji} = \begin{pmatrix} \frac{3x^2}{r^5} - \frac{1}{r^3} & \frac{3xy}{r^5} & \frac{3xz}{r^5} \\ \frac{3xy}{r^5} & \frac{3y^2}{r^5} - \frac{1}{r^3} & \frac{3yz}{r^5} \\ \frac{3xz}{r^5} & \frac{3yz}{r^5} & \frac{3z^2}{r^5} - \frac{1}{r^3} \end{pmatrix}_{ij} \quad (4.17)$$

5. REACTION PATH–FORCE MATCHING IN COLLECTIVE VARIABLES AND DOUBLY POLARIZED QM/MM WITH MACHINE LEARNING CHAPERONE POLARIZABILITY

Accurate calculations of free energy are essential for deeper mechanistic chemical understanding. In this work, we cycle between two strategies: Reaction Path–Force Matching in Collective Variables and Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability to directly correct the potential of mean force generated in collective variables, while correcting the intermolecular reaction of QM atoms to the MM electric field. Enabled with machine learning, a holistic correction based on intra and intermolecular corrections is developed to be used as an efficient, accurate and robust method for simulating solution phase chemical reactions. We apply this method to the Menshutkin reaction in aqueous solution and show that ML can improve the accuracy and physics of SE–QM/MM simulations, while maintaining a significant reduction in computational costs.

5.1 Introduction

The free energy of a chemical reaction is a central quantity that governs the thermodynamics of a reaction related to the barrier height and kinetics of a chemical reaction. Regarding chemical reactions in solution, free energy calculations can be enabled with the combined quantum mechanical and molecular mechanical (QM/MM) approach [1], [13]–[16], in which computationally intensive QM calculations are applied to the reactive region, in which efficient MM calculations are applied to the remainder of the system. To quantify the free energy of the reaction, the potential of mean force (PMF) in collective variables (CVs) along a reaction path can be determined with statistical sampling during molecular dynamics. In particular, one such sampling method is the string method [22], in which CVs are accelerated to the chemically relevant minimum free energy path (MFEP), where integration of the mean forces in CVs along the resulting MFEP [11], [67] is used to obtain the PMF. Despite the advantages of the string method in avoiding sampling on a full free energy surface, sampling with high accuracy ab initio QM/MM (AI/MM) sampling are computa-

tionally prohibitive due to the daunting computational costs. Therefore, as an alternative to using expensive AI/MM methods, semiempirical QM/MM (SE/MM) methods [8], [23] can be used to efficiently approximate the reaction free energy at the expense of accuracy.

Alternatively, higher accuracy PMFs can be achieved with the efficiency of an SE/MM method by using Reaction Path–Force Matching in Collective Variables (RP–FM–CV) [26], [28], a multi-level approach that restores the highly accurate AI/MM–free energy forces by using explicit SE/MM forces on CVs as a steppingstone. This involves a translational and rotational invariant fit of the free energy force, which can be determined by transforming the Cartesian force vector to internal forces on CVs expressed in redundant coordinates [43]. Once linearized, the resulting force correction can be supplemented onto CVs with a grid-based cubic spline function [75]. The application of RP–FM–CV for condensed phase reactions shows that errors in CV forces are significantly reduced, which leads to an improved PMF and MFEP which is comparable to experimental results.

Machine learning (ML) is a rapidly growing tool in the field of multi-level QM/MM molecular dynamics where many input features are considered for fitting, which is otherwise nontrivial for traditional interpolation/extrapolation schemes [34], [69]. Therefore, as an alternative to force matching each CV with a single dimension grid-based cubic spline function, we extend the RP–FM–CV strategy to ML to develop a multidimensional force correction that explicitly accounts for potential coupling between CVs. To do this, we first reconstruct the internal force correction in CVs with the formulation of the RP–FM–CV method. Second, an artificial neural network (ANN) is trained to predict the internal force corrections from a combination of pairwise distances of the molecule as input features. Finally, explicit SE/MM forces in CVs are updated on-the-fly to AI/MM quality in string simulations with the aid of a trained ANN. In summary, a highly accurate PMF is achieved with ML corrections, with a multidimensionally defined internal force correction in CVs.

Previous studies have shown that force matching (FM) in collective variables (CV’s) significantly corrects the structure of CVs without correcting the solvent structure, where a doubly polarized correction to polarizability improves the intermolecular interactions without correcting the reaction path. More precisely, while the correction of the total force on the CV’s significantly corrects the semiempirical (SE) structure from a single point force

calculation, the response of the solvent to the structural change in molecular dynamics is not necessarily considered. This behavior is represented by an additional stabilization of free energy from the polarization correction, in which structural changes in the CV’s on the reaction path are only marginally altered.

To overcome the omitted portions of both the RP–FM–CV and dp–QM/MM approaches, we combine both strategies to achieve a holistic correction within the framework of a consistent intramolecular and intermolecular correction. Against this background, the combination of polarizability and internal force corrections is not straightforward since the attraction and repulsion between CV and MM atoms are also indirectly included in the polarizability correction. Therefore, the contribution of the polarizability correction in CV’s must be removed to achieve a purified internal force correction that exclusively restores the total force in CV’s. Overall, the combined correction is performed iteratively, as a FM correction is required on a polarizability corrected reaction path. This is achieved by first correcting the molecular polarizability with chaperone atomic polarizabilities and equilibrating the system, which is then followed with an update to low-level Cartesian SE forces for CV atoms. This allows force contributions from the polarizability and the permanent electric field to be included into the internal force correction, which is determined by computing the difference between the resulting polarizability corrected force and high-level *ab initio* (AI) force. Specifically, the forces from the polarizability and the permanent electric field are determined by differentiating the weights and biases of the artificial neural network (ANN). Together, the polarizability correction on QM atoms and the internal force corrections in CVs are retrained in the ANN to simultaneously consider the intramolecular and intermolecular corrections in molecular dynamics. Such an approach is essential for mechanistic studies in enzyme reactions, as free energy in the substrate is influenced by a dynamic local environment that is subject to conformational changes in the presence of charged residues. As such, in addition to speeding up the calculation the main aim of this study is to consistently correct the CVs with the appropriate response to the environment to achieve a robust calculation that is consistent with pure AI calculations.

The outline of the work is as follows. First, we describe the key equation needed to obtain the internal force correction from an ANN and then we discuss the key equations for

the chaperone polarizabilities. Then we discuss the computational tools used to implement the combined/MM approach, followed by a discussion of the results and conclusions.

5.2 Methods

In the present study, we analyze the internal force correction in CVs ($\Delta\mathbf{F}_i^c$) from ML defined as

$$\Delta\mathbf{F}_i^c = \mathbf{F}_i^{\text{AI/MM}} - \mathbf{F}_i^{\text{SE/MM}} \quad (5.1)$$

where $\mathbf{F}_i^{\text{AI/MM}}$ and $\mathbf{F}_i^{\text{SE/MM}}$ represent the internal force of CV i , determined at the AI/MM and SE/MM levels, respectively. To determine the AI/MM and SE/MM internal forces on CVs, the Cartesian forces of the QM system are converted to internal forces expressed in redundant coordinates, as outlined in the RP-FM-CV strategy. This conversion requires a coordinate transformation procedure which relies on Wilsons \mathbf{B} -matrix [80], where the QM molecule is defined with all pairwise distances in the molecule.

Next, chaperone polarizability corrections are defined as

$$\Delta\alpha_i^c = \alpha_i^{\text{AI/MM}} - \alpha_i^{\text{SE/MM}} \quad (5.2)$$

where $\alpha_i^{\text{AI/MM}}$ and $\alpha_i^{\text{SE/MM}}$ represent the chaperon polarizability of QM atom i , determined at the AI/MM and SE/MM levels, respectively. To determine the chaperone corrections, atomic polarizabilities at the AI/MM and SE/MM level are computed using the Hirshfeld partitioning scheme as outlined in the dp-QM/MM strategy. This computation requires the application of an external electric field on the QM molecule, where changes in dipoles are used to compute the atomic polarizabilities.

One advantage of ML is the ability to learn internal force corrections in CVs and chaperone polarizabilities with a combination of input features. Furthermore, force/polarizability specific models can easily be combined together with separate models, after the corrections are determined. Next, we discuss the implementation of ML for predicting both the internal force corrections in CVs and chaperone polarizabilities during molecular dynamics.

Moreover, to capture reaction coordinate dependence of the corrections, we determine $\Delta\mathbf{F}_i^c$ and $\Delta\alpha_i^c$ on-the-fly during QM/MM simulations through ML. Specifically, an ANN is used, which takes molecular features as input and is optimized through a hidden layer to predict the internal force corrections in the output layer. When a hyperbolic tangent activation function is used in each of these two-layer ANNs, $\Delta\mathbf{F}_i^c$ and $\Delta\alpha_i^c$ in the output layer can be written in a general form as:

$$\begin{aligned}\Delta\mathbf{F}_i^c &= b_i^O + \sum_{j=1}^L W_{i,j}^O \tanh\left(b_j^H + \sum_{k=1}^M W_{j,k}^H p_k\right) \\ \Delta\alpha_i^c &= b_i^O + \sum_{j=1}^L W_{i,j}^O \tanh\left(b_j^H + \sum_{k=1}^M W_{j,k}^H p_k\right)\end{aligned}\tag{5.3}$$

where p_k denotes the k^{th} input feature among a total of M input features, $W_{j,k}^H$ and b_j^H are the weights and biases in the hidden layer that consists of L nodes, and $W_{i,j}^O$ and b_j^O are the weights and biases in the output layer. Once determined, a multidimensionally defined internal force correction and chaperone polarizabilities can be determined with the ANNs to correct the PMF and MFEP.

5.3 Computational Details

To validate the ML approach, the Menshutkin reaction [45] in solution between methyl chloride and ammonia is used as a prototype system. In this simulation, the QM system is solvated in a cubic periodic box of To validate the ML approach, the Menshutkin reaction [45] in solution between methyl chloride and ammonia is used as a prototype system. In this simulation, the QM system is solvated in a cubic periodic box of $50 \times 50 \times 50 \text{ \AA}^3$ with modified TIP3P [108] waters, in which van der Waals interactions between QM and MM atoms are adjusted to Lennard–Jones equations [49], [85] for the Menshutkin reaction using standard CHARMM force fields [110]. In addition, the SHAKE algorithm [112] is used to constrain the internal geometries of water. Furthermore, long-range electrostatics for MM/MM and QM/MM interactions are treated by Particle Mesh Ewald (PME) [113] and QM/MM–PME [114], [115] where interactions are smoothly attenuated from 12 to 13 \AA with

a switching function. In order to reduce system distortions, SBOUND, a harmonic spherical restraint is used to prevent solute drifting. For free energy calculations with the string method, the CVs composed of forming (r_{N-C}) and breaking (r_{C-Cl}) bonds are restrained in addition to an angle stabilizing CV (r_{N-Cl}) with a force constant of 1,000 kcal/mol/Å and are evenly discretized into 16 images along the reaction coordinate. Under constant temperature and pressure at 298.15 K and 1 atm, with a time step of 0.25 fs, each image is sampled for 20 ps, with configurations saved every 4 ps. The projection, reparameterization, and evolution of the paths, as well as integration of the PMF follows the string method developed by Maragliano et al. [22]. For all simulations AM1 [24] is chosen as the base SE/MM method where the B3LYP method with the 6-31+G(d,p) basis set [122], [124] is selected as the AI/MM method. Based on sampled configurations from molecular dynamics with the CHARMM (version c42a2) program [110], the Gaussian16 program [188] is used to calculate the SE/MM and AI/MM forces and polarizabilities, where atomic polarizabilities are computed with Hirshfeld partitioning. For ANN training, Tensorflow [40] is used to train CV input features to the target internal force corrections, where 19 and 22 neurons are used in a 2-layer ANN for RP-FM-CV and dp-QM/MM respectively. To linearly minimize the weights and biases of the ANN, the hyperbolic tangent function is used. ANN verification for over-fitting is evaluated by randomly dividing the samples into training, validation, and testing data sets, where 80% of samples are used for ANN training, where 20% of samples are used for the validation and testing data sets.

5.4 Results

5.4.1 Molecular Polarizability

In this work, the molecular polarizability is calculated with explicit solvent for the Menshutkin reaction along a force matched reaction path at both the B3LYP/MM and AM1/MM levels of theory (Figure 5.1).

The results of the force matched AM1/MM polarizabilities with explicit solvent confirms the underestimation of molecular polarizability, with an underestimation of $\sim 51\%$ compared to B3LYP/MM. According to AM1/MM theory, the molecular polarizability ranges between

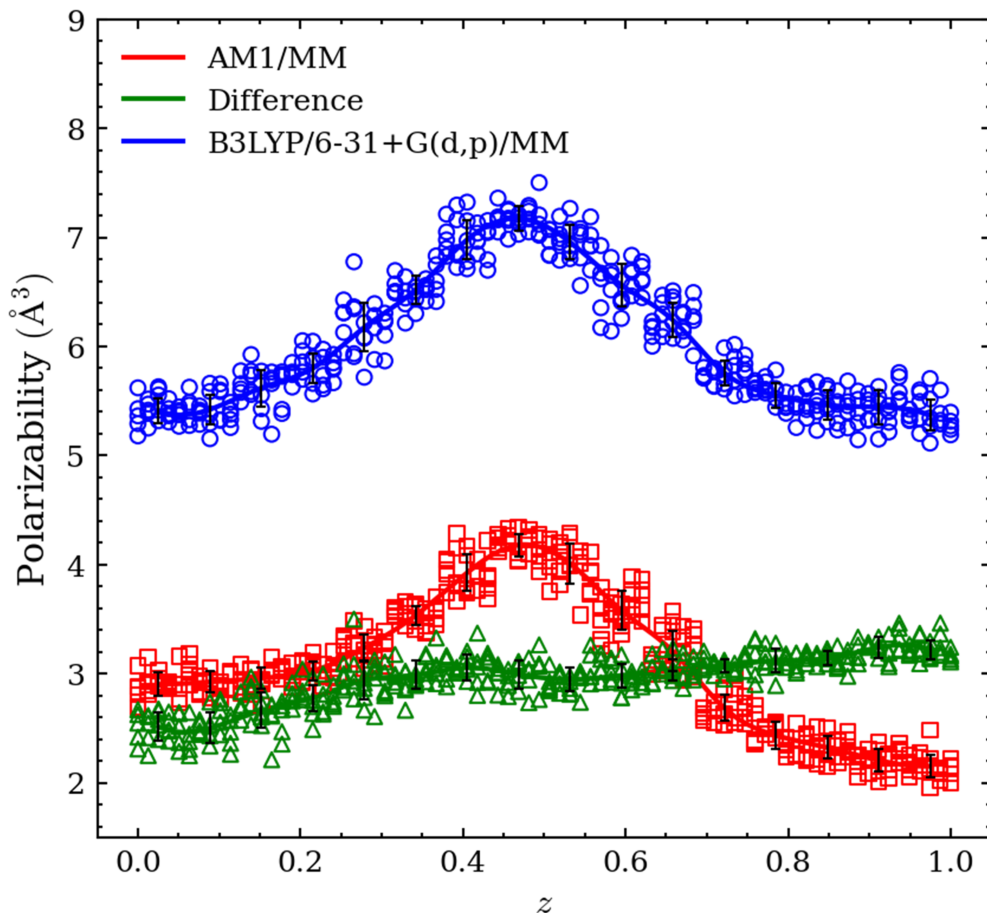


Figure 5.1. The solution-phase molecular polarizabilities as a function of the reaction coordinate for the Menshutkin reaction are shown for AM1 (squares), B3LYP/MM (circles), and their difference (triangles). The means (solid curves) and standard deviations (vertical bars) are calculated based on the samples within each string image.

2.0–4.6 \AA^3 along the reaction path, while the B3LYP/MM molecular polarizability ranges between 5.1–7.3 \AA^3 . From semiempirical configurations, the largest molecular polarizability at both levels of theory is located in the half-formed NC and half broken C–Cl bonds ($z = 0.48$), since charge separation at the transition generates a large dipole. In contrast, the largest difference in molecular polarizability between B3LYP/MM and AM1/MM is observed in the product state when the chlorine anion is completely dissociated from the methyl group ($z = 1.0$). The non-uniform difference along the reaction path illustrates the dynamic

evolution of the molecular polarizability from charge neutral to charge separated species of the Menshutkin reaction. At the AM1/MM level of theory, the molecular polarizability in the reactant state differs on average by $\sim 2.5 \text{ \AA}^3$ compared to the benchmark B3LYP/MM level, which is 0.8 \AA^3 lower than the average product state polarizability of $\sim 3.3 \text{ \AA}^3$. In particular, the presence of a flattened polarizability difference between the AM1/MM and B3LYP/MM polarizability is noticed between the reactant and product along the force matched reaction path with explicit solvent ($z = 0.55$). The flattened difference between the two calculations is considered to be the result of a more explicit calculation, in which solvent molecules within 12 \AA are included, rather than using an approximate representation with the continuum solvation model. In addition, a decrease in the polarizability difference can be attributed to the combined effects of explicit solvents and the modified reaction path from the force matching procedure. In general, the aim of this study is to correct the polarizability difference of two levels for a more heterogeneous local environment in molecular dynamics. Therefore, a correction derived from an explicit representation of the local environment is sought after with the aid of an artificial neural network. In this study, chaperone corrections from the artificial neural network are able to reduce the original AM1/MM molecular polarizability error by a factor of more than 25 from 2.96 to 0.11 \AA^3 (Figure 5.2).

Therefore, in combination with the force matching procedure, where collective variables are directly fit, B3LYP/MM intermolecular interactions are successfully corrected with the artificial neural network during updated free energy string sampling.

5.4.2 Atomic Chaperone Polarizability

The atomic chaperone polarizabilities are calculated along the force-matched reaction path with explicit solvents using Hirshfeld partitioning (Figure 5.3).

In particular, the correction of the chaperone chlorine anion shows the largest source of error in AM1/MM calculations, as it requires the largest correction of $\sim 2.5 \text{ \AA}^3$ in the product state. For other notable polarizability corrections on CVs, a positive peak is observed for nitrogen with the largest positive error in the transition state, in which a corresponding negative correction begins to develop for the carbon atom ($z = \sim 0.5$). The results of this

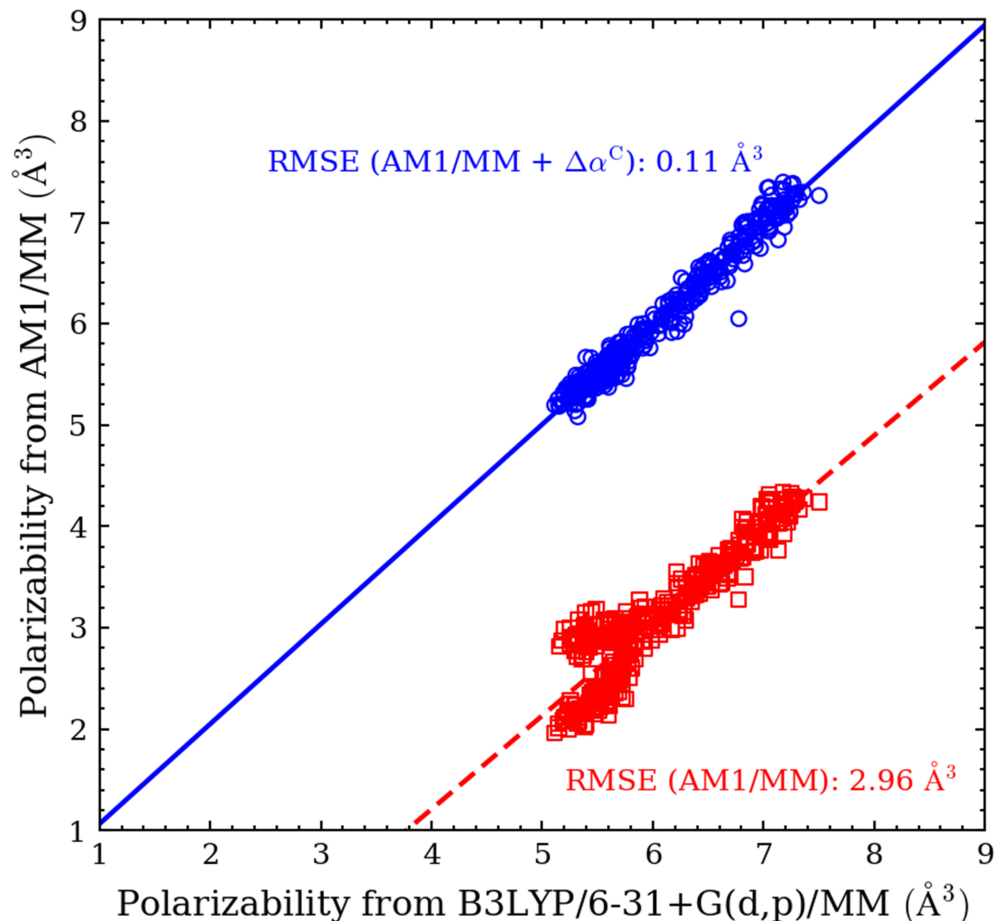


Figure 5.2. The regressions of molecular polarizabilities from AM1 (square with a dashed line) and from the chaperone-corrected AM1 (AM1+ $\Delta\alpha^C$; circles with a solid line) against those from B3LYP/MM are shown, along with the corresponding root-mean-square errors (RMSEs) relative to the B3LYP reference values.

study show the consistent underestimation of the AM1/MM polarizability for nitrogen along the force matched reaction path, while an overestimation of the polarizability is computed for carbon after the transition state. This observation is related to the flattened polarizability difference near the transition state ($z = 0.48$). In contrast, the chaperone corrections for hydrogen are not as significant, as the chaperone corrections are smaller in magnitude than the corrections to CVs. However, a larger polarizability correction is observed for the hydrogens on ammonia in the reactant before the ammonia is complexed with the methyl

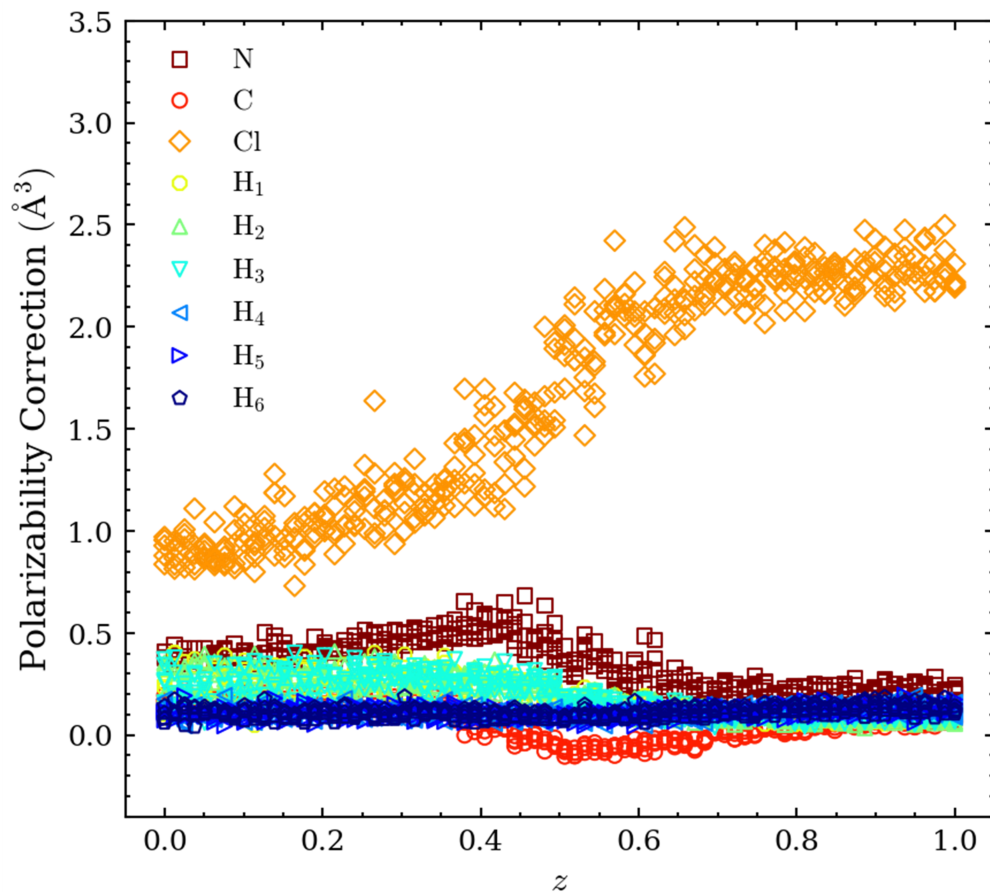


Figure 5.3. The polarizabilities of atomic chaperones as a function of the reaction coordinate for each solute atom in the Menshutkin reaction.

group of methyl chloride. Due to the greater exposure of these hydrogens to nearby solvent molecules, a greater correction is observed for hydrogens on ammonia, while the polarizability correction of the hydrogens on carbon are more quenched, as it is more buried in the inner atoms of the molecule. The chaperone corrections determined by explicit solvent for force-matched results are comparable to those determined by PCM, with the exception of increased variability in computed polarizabilities from using explicit solvent, in addition to molecular polarizability differences at the transition state for CV atoms, due to the loosening of the transition state in the force-matched reaction path.

5.4.3 Internal Force Correction in CVs

After equilibrating the simulation with the chaperone polarizabilities, a mean force correction in CVs is fit to B3LYP/MM forces with Reaction-Path Force-Matching (Figure 5.4).

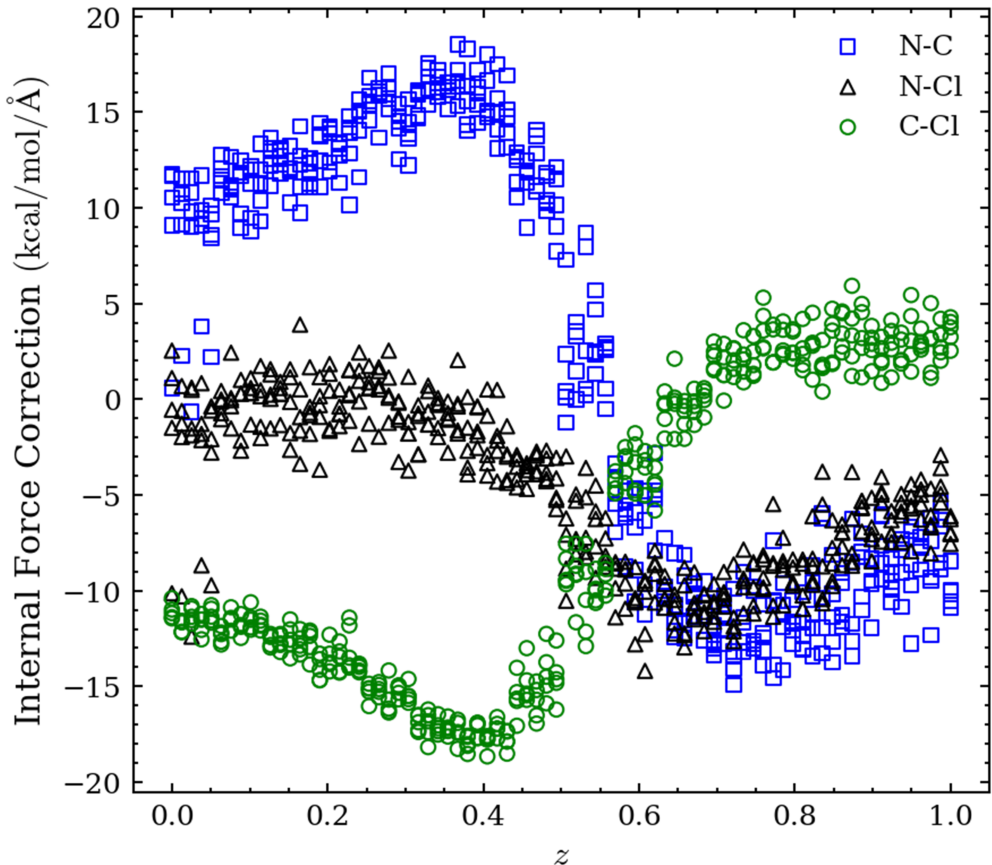


Figure 5.4. Internal force corrections for the Menshutkin reaction in solution on CVs (i.e., the NC, C-Cl and NCl bonds) using the AM1/MM level as the base level and the B3LYP/MM level as the target level with the 6-31+G(d,p) basis set.

In particular, the difference between B3LYP/MM and AM1/MM internal forces on restrained CVs is determined by decomposing the Cartesian forces of the QM system to 36 pairwise distances with Pulay’s procedure. The internal force correction is then determined after the chaperone correction to achieve a mean force correction in configurations containing enhanced intermolecular interactions at the B3LYP/MM level. Compared to previous work, the magnitude of the internal force corrections in N-C and C-Cl is diminished , with the

addition of chaperone corrections and partitioning of the internal force corrections to N–Cl. The observed inverse relationship between N–C and C–Cl along the reaction path is consistent with previous results, even with the addition of the negative internal force correction for N–Cl in the product. The large correction in the reactant leading to the transition state shows that a large internal force correction is required, i.e., for the dissociation of C–Cl and complexation of N–C, where the opposite effect is seen to a lesser extent for N–C formation and C–Cl dissociation after the transition state. As for the AM1/MM error to B3LYP/MM, the N–C and C–Cl CVs have a correlation greater than 10 kcal/mol/Å, where the N–Cl error for stabilizing the reactant/product contains a much smaller error of 6.75 kcal/mol/Å. The smaller error for N–Cl suggests that stabilizing the charge-separated product state in solution with AM1/MM is more problematic than stabilizing starting reactants which are neutral, a similar problem with the molecular polarizability error. To fit these forces, internal force corrections are trained to an artificial neural network instead of cubic spline functions. This application of internal force correction is advantageous, as it allows restrained contraction/expansion of CVs along the reaction path, such as N–Cl, to be fit to input features, as opposed to unidirectional bond shortening/lengthening, i.e. for N–C and C–Cl. Overall, internal force corrections on restrained collective variables are fit to within 1 kcal/mol/Å for chaperone corrected samples, with the use of a force matching artificial neural network (Figure 5.5).

5.4.4 Comparison of MFEPs and Free Energy Profiles

The string optimized minimum free energy paths are compared for both the AM1/MM and the combined/MM approaches (Figure 5.6).

In particular, the combined/MM reaction path for all images along the string is loosened, starting with a loosened N–C and C–Cl bond distance in the reactant. Regarding the product, a loosening of the C–Cl bond is observed, where N–C is shown to shorten only slightly. This is in contrast to the dp-QM/MM approach, where no significant changes are observed to the MFEP after the polarizability correction. For changes along the reaction path, the loosening of the MFEP from the combined/MM approach are more comparable to

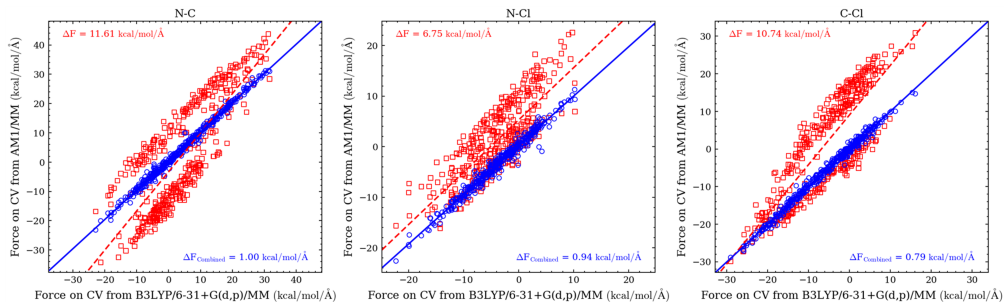


Figure 5.5. The internal force correlations between the base AM1/MM and B3LYP/MM using the 6-31+G(d,p) basis set are shown before (red squares) and after (blue circles). The internal force corrections of RP-FM-CV are applied to polarization corrections; the corresponding trend lines are shown as dashed and solid lines. The internal force deviations (ΔF ; in kcal/mol/ \AA^2) between the base and target levels, before (red) and after (blue) force matching, are shown for comparison.

the results of RP-FM-CV. Therefore, changes in the MFEP can be attributed towards the correction from intramolecular force matching rather than the correction to polarizability in the combined/MM approach. Nevertheless, while the path is loosened, the location of the transition state is changed compared to RP-FM-CV, since only N-C is loosened, where the length of the C-Cl bond is unchanged. This notable shift in the location of the transition state is directly related to enhanced solvation through the polarizability correction. Overall, the combined/MM effects are the direct result of intramolecular corrections on intermolecular corrections, all of which are related to improved geometry, solvation, and the resulting free energy profile from updated string sampling (Figure 5.7).

The free energy profile that results from the combination of chaperone polarizabilities with the internal force correction on CVs shows that the barrier height is significantly reduced, from about 30 kcal/mol to about 13 kcal/mol. Compared to our previous work, this dramatic change in barrier height can be attributed to the direct fit of the mean force in collective variables. Furthermore, the transition state peak is shifted from $z = 0.5$ to $z = 0.4$ towards the reactant, similar to the dp-QM/MM results, where only polarizability is corrected. Regarding the reaction free energy, the combined/MM correction generates -20

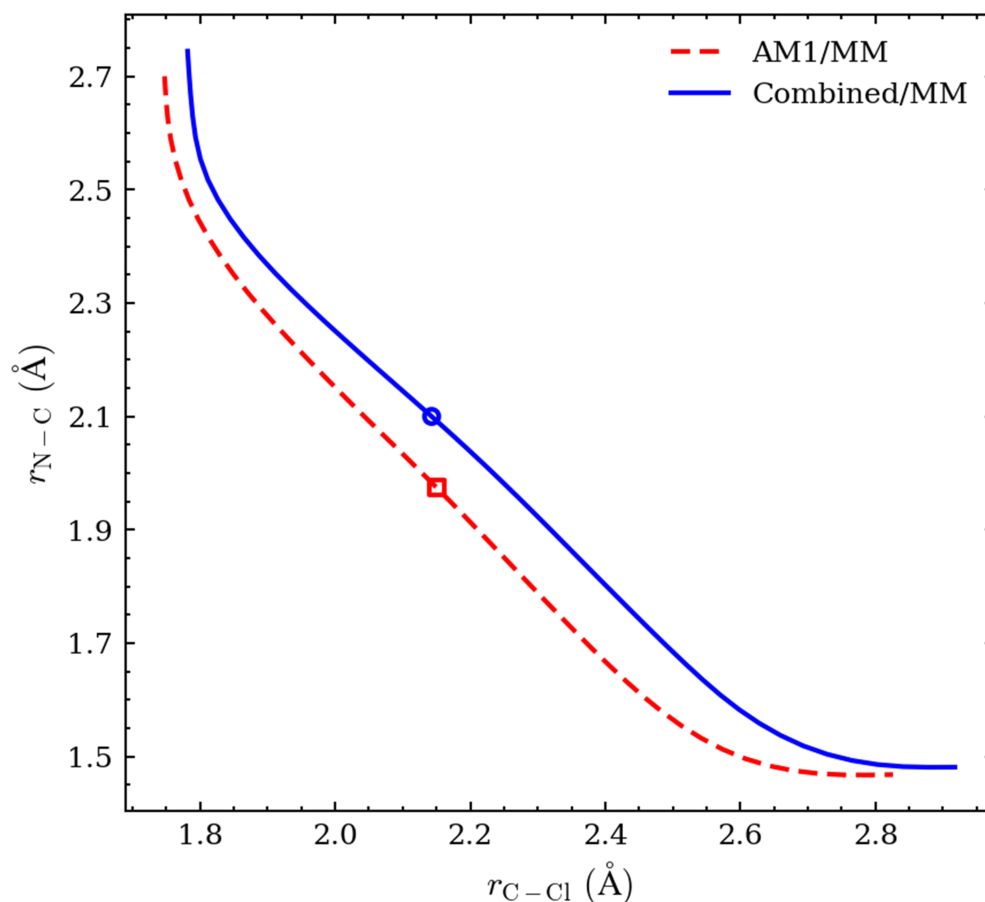


Figure 5.6. The minimum free energy paths (MFEPs) as a function of collective variables, i.e., C–Cl and NC bonds, are shown by the AM1/MM (dashed line) and dp-AM1/MM with the chaperone polarizabilities fitted to the B3LYP/MM level (solid line). The locations of the free energy transition state are also marked: AM1/MM (open square) and dp-AM1/MM (open circle).

kcal/mol of correction. Overall, the reaction-free energy correction is not additive, as the internal force is applied to the samples corrected by the chaperone polarizability. Therefore, the shifts and energies of the resulting free energy profile from the combined/MM approach reflect the balanced mean force correction in the presence of enhanced solvation in the final reaction path.

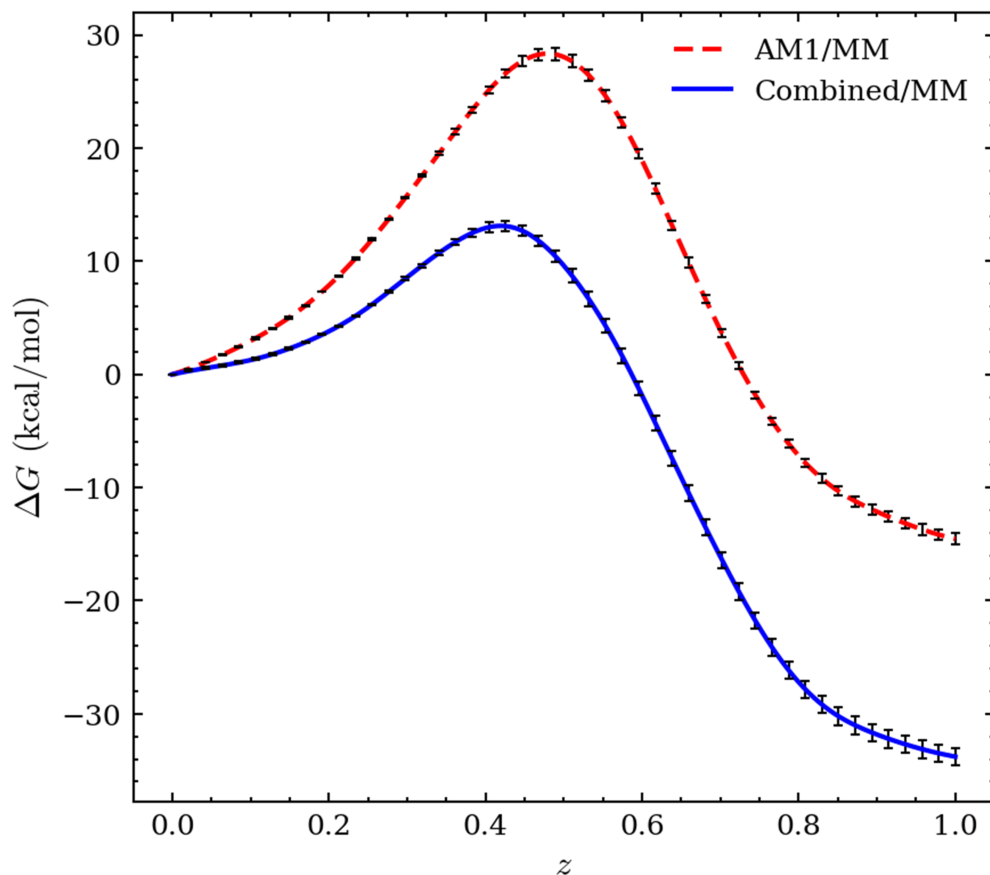


Figure 5.7. The Menshutkin reaction’s free energy profiles as a function of the reaction coordinate: AM1/MM (dashed line) and Combined-AM1/MM with polarizabilities corrected to B3LYP/MM (solid line). The error bars associated with the free energy in the reactant state ($\alpha = 0$) along the string MFEP are estimated.

5.4.5 Local Environment and RDF

The radial distribution functions of CV atoms for the combined/MM approach confirm enhanced solvation around nitrogen and chlorine upon product formation (Figure 5.8).

This is similar to previous results for chaperone corrections without force matching. However, due to the loosening of the transition state from force matching, solvation around the transition state for chlorine decreases significantly from 1.7 to 1.4 water molecules at 3.5 Å, while solvation around the transition state for nitrogen increases from 1.0 to 1.5 water molecules at 3.2 Å. This feature coincides with the resulting transition state from the

combined/MM approach, in which an unchanged C–Cl distance accompanies a loosened NC distance. More specifically, a slightly tighter C–Cl distance is more crowded than a loosened distance, thus attributing to lesser solvation. These observations coincide with the resulting transition state and the shift from the combined/MM approach. A final observation to note with the combined/MM approach is enhanced solvation for carbon in the reactant. This contrasts previous dp-QM/MM and RP-FM-CV approaches, since no significant changes are seen in solvation for carbon. However, the removal of C_{3v} and linear restraints is conducive to a more polarizable environment, as hydrogens on ammonia are no longer uniformly aligned with carbon in the reactant. Overall, the changes in free energy and string path optimization are observed from direct fits of mean forces in CVs, and as seen in RDFs, where improved solvation and corresponding changes in the location of the transition from chaperone polarization correction are also included in the combined/MM approach.

5.5 Concluding Remarks

In combination with modern high computing resources, a dual-level correction is developed with ML to assist in computational free energy studies for solution phase reactions where accurate free energy computations are essential for deeper mechanistic chemical understanding. As an alternative to force fitting to CVs with a grid-based cubic spline function, we extend the capabilities of the RP-FM-CV method with a chaperone polarizability correction using ML. With multiple MLs we are able to generate a coupled correction from CVs which further improves the deviation of forces between AI/MM and SE/MM forces in CVs along the MFEP and enhances QM solvation to surrounding solvent molecules. Moreover, a consistent correction to solvent is coupled with the correction to CVs, by correcting polarizability corrected samples with RP-FM-CV. With the combined/MM approach, the intermolecular corrections are consistently corrected in a physically reasonable manner along with the intramolecular correction as reflected in the lowering of the overestimated barrier and free energy in addition to enhanced solvation in the RDFs. Overall, we now have a validated force matching approach with ML that is more suitable to larger, more complex

reactions in which many CVs are coupled to the free energy process for heterogeneous local environments.

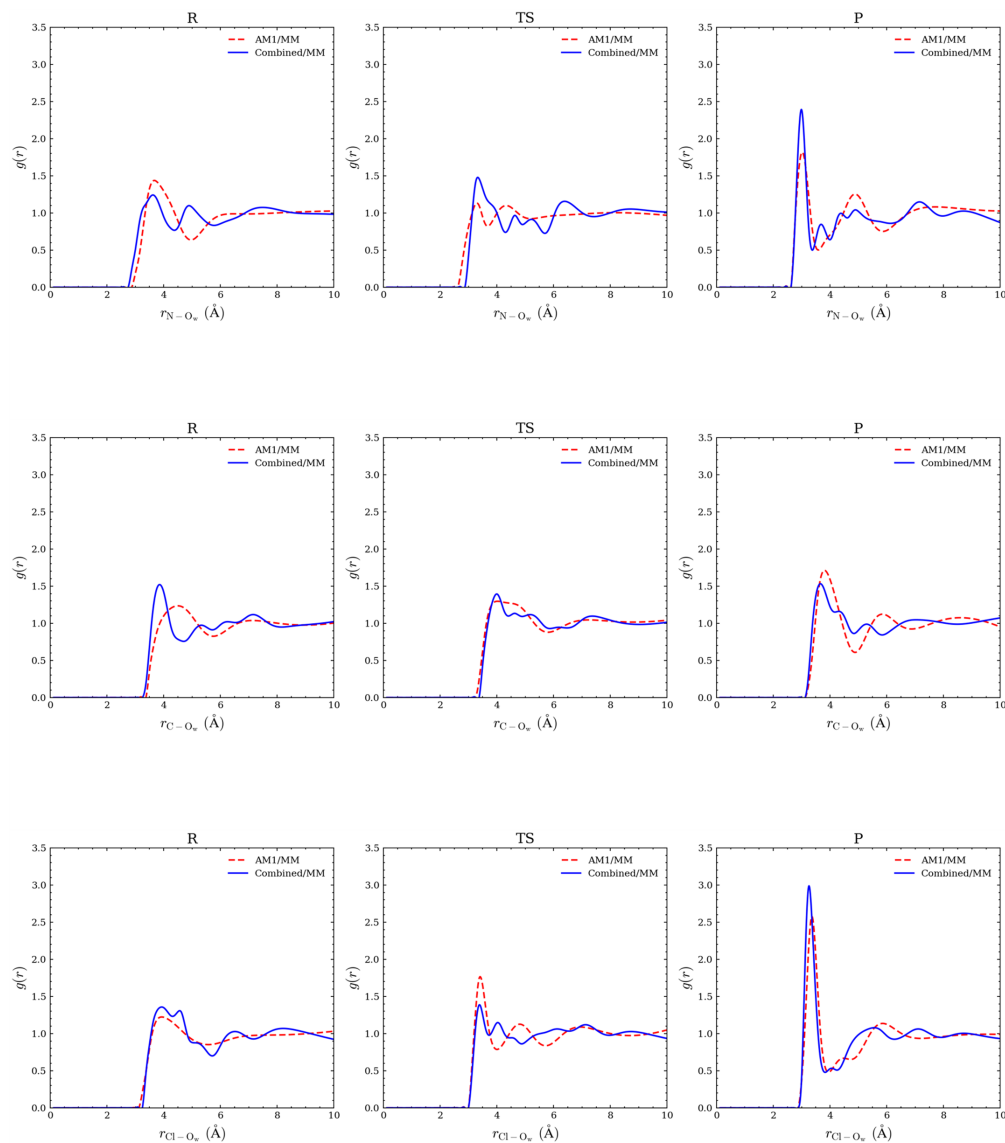


Figure 5.8. The radial distribution functions between the chlorine atom and water oxygens (O_w) in the reactant (R), transition-state (TS), and product (P) region of the Menshutkin reaction are shown for AM1/MM (dashed line) and Combined-AM1/MM with polarizabilities corrected to the B3LYP/MM level (solid line).

6. CONCLUDING REMARKS

In summary, this dissertation presents the development and application of various methods to address key challenges in determining first-principles QM/MM free energy profiles for chemical reactions in condensed phases. The RP-FM-CV method, an FM-based multilevel QM/MM approach, is developed to reproduce highly accurate AI/MM free energy profiles by fitting mean forces on a set of CVs. Furthermore, we compare different RP-FM approaches to assess their strengths and weaknesses, leading to the establishment of a criterion for obtaining reliable FEPs. Our findings contribute to a deeper understanding of force fitting models and offer insights into the implementation of RP-FM strategies in future studies. To address the underpolarization problem in SE/MM-based free energy simulations, we present the dp-QM/MM method, which utilizes AI information to guide an underpolarized SE method to respond properly to MM electric fields. This hybrid framework, employing ML chaperone polarizabilities, improves the quality of free energy results to a level comparable to AI/MM and experimental benchmarks. The dp-QM/MM method offers a new strategy for an improved description of intermolecular polarization in SE/MM free energy simulations, paving the way for further applications in complex chemical and biochemical reactions. Finally, we extend the capabilities of the RP-FM-CV method by incorporating a chaperone polarizability correction using ML. This dual-level correction enhances QM solvation and generates a coupled correction from CVs, providing a physically reasonable and consistent correction to both solute and solvent. The combined approach demonstrates its potential for application to larger, more complex reactions where multiple CVs are coupled to the free energy process in heterogeneous local environments. In conclusion, this dissertation presents a systematic and practical strategy for first-principles free energy simulations, opening new avenues for AI/MM mechanistic studies of complex chemical and biochemical reactions, where chemical accuracy and statistically adequate free energy sampling are essential for a deeper mechanistic understanding.

REFERENCES

- [1] A. Warshel and M. Levitt, "Theoretical studies of enzymatic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme," *J. Mol. Biol.*, vol. 103, pp. 227–249, 1976.
- [2] G. Groenhof, "Introduction to qm/mm simulations," in *Biomolecular Simulations: Methods and Protocols*, L. Monticelli and E. Salonen, Eds. Totowa, NJ: Humana Press, 2013, pp. 43–66, ISBN: 978-1-62703-017-5. DOI: [10.1007/978-1-62703-017-5_3](https://doi.org/10.1007/978-1-62703-017-5_3). [Online]. Available: https://doi.org/10.1007/978-1-62703-017-5_3.
- [3] Q. Cui, T. Pal, and L. Xie, "Biomolecular qm/mm simulations: What are some of the "burning iusses"?" *J. Phys. Chem. B*, vol. 125, pp. 689–702, 2021.
- [4] K. N. Houk and F. Liu, "Holy grails for computational organic chemistry and biochemistry," *Accounts of Chemical Research*, vol. 50, no. 3, pp. 539–543, 2017, PMID: 28945400. DOI: [10.1021/acs.accounts.6b00532](https://doi.org/10.1021/acs.accounts.6b00532). eprint: <https://doi.org/10.1021/acs.accounts.6b00532>. [Online]. Available: <https://doi.org/10.1021/acs.accounts.6b00532>.
- [5] L. Shen and W. Yang, "Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks," *Journal of Chemical Theory and Computation*, vol. 14, no. 3, pp. 1442–1455, 2018, PMID: 29438614. DOI: [10.1021/acs.jctc.7b01195](https://doi.org/10.1021/acs.jctc.7b01195). eprint: <https://doi.org/10.1021/acs.jctc.7b01195>. [Online]. Available: <https://doi.org/10.1021/acs.jctc.7b01195>.
- [6] Y.-J. Zhang, A. Khorshidi, G. Kastlunger, and A. A. Peterson, "The potential for machine learning in hybrid qm/mm calculations," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241 740, 2018. DOI: [10.1063/1.5029879](https://doi.org/10.1063/1.5029879). eprint: <https://doi.org/10.1063/1.5029879>. [Online]. Available: <https://doi.org/10.1063/1.5029879>.
- [7] K. E. Ranaghan and R. Lonsdale, "Qm/mm methods," in *Encyclopedia of Biophysics*, G. C. K. Roberts, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 2154–2156, ISBN: 978-3-642-16712-6. DOI: [10.1007/978-3-642-16712-6_275](https://doi.org/10.1007/978-3-642-16712-6_275). [Online]. Available: https://doi.org/10.1007/978-3-642-16712-6_275.
- [8] X. Lu, D. Fang, S. Ito, Y. Okamoto, V. Ovchinnikov, and Q. Cui, "Qm/mm free energy simulations: Recent progress and challenges," *Mol. Simul.*, vol. 42, p. 1056, 2016.
- [9] Q. Cui, T. Pal, and L. Xie, "Biomolecular qm/mm simulations: What are some of the burning issues?" *The Journal of Physical Chemistry B*, vol. 125, no. 3, pp. 689–702, 2021, PMID: 33401903. DOI: [10.1021/acs.jpcc.0c09898](https://doi.org/10.1021/acs.jpcc.0c09898). eprint: <https://doi.org/10.1021/acs.jpcc.0c09898>. [Online]. Available: <https://doi.org/10.1021/acs.jpcc.0c09898>.

- [10] B. L. Grigorenko, A. V. Rogov, I. A. Topol, S. K. Burt, H. M. Martinez, and A. V. Nemukhin, "Mechanism of the myosin catalyzed hydrolysis of atp as rationalized by molecular modeling," *Proceedings of the National Academy of Sciences*, vol. 104, no. 17, pp. 7057–7061, 2007. DOI: [10.1073/pnas.0701727104](https://doi.org/10.1073/pnas.0701727104). eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.0701727104>. [Online]. Available: <https://www.pnas.org/doi/abs/10.1073/pnas.0701727104>.
- [11] Y. Zhou, P. Ojeda-May, M. Nagaraju, B. Kim, and J. Pu, "Mapping free energy pathways for atp hydrolysis in the e. coli abc transporter hlyb by the string method," *Molecules*, vol. 23, p. 2652, 2018.
- [12] R. W. Robey, K. M. Pluchino, M. D. Hall, A. T. Fojo, S. E. Bates, and M. M. Gottesman, "Revisiting the role of abc transporters in multidrug-resistant cancer," *Nature Reviews Cancer*, vol. 18, no. 7, pp. 452–464, 2018, ISSN: 1474-1768. DOI: [10.1038/s41568-018-0005-8](https://doi.org/10.1038/s41568-018-0005-8). [Online]. Available: <https://doi.org/10.1038/s41568-018-0005-8>.
- [13] M. J. Field, P. A. Bash, and M. Karplus, "A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations," *J. Comput. Chem.*, vol. 11, p. 700, 1990.
- [14] U. C. Singh and P. A. Kollman, "A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the ch₃cl + cl⁻ exchange reaction and gas phase protonation of polyethers," *J. Comput. Chem.*, vol. 7, pp. 718–730, 1986.
- [15] J. Gao and M. A. Thompson, *Combined Quantum Mechanical and Molecular Mechanical Methods*. Washington, DC: ACS Symposium Series 712; American Chemical Society, 1998.
- [16] H. M. Senn and W. Thiel, "Qm/mm methods for biomolecular systems," *Angew. Chem. Int. Ed.*, vol. 48, pp. 1198–1229, 2009.
- [17] P.-Y. Chen and M. E. Tuckerman, "Molecular dynamics based enhanced sampling of collective variables with very large time steps," *The Journal of Chemical Physics*, vol. 148, no. 2, p. 024 106, 2018. DOI: [10.1063/1.4999447](https://doi.org/10.1063/1.4999447). eprint: <https://doi.org/10.1063/1.4999447>. [Online]. Available: <https://doi.org/10.1063/1.4999447>.
- [18] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, "Enhanced sampling in molecular dynamics," *The Journal of Chemical Physics*, vol. 151, no. 7, p. 070 902, 2019. DOI: [10.1063/1.5109531](https://doi.org/10.1063/1.5109531). eprint: <https://doi.org/10.1063/1.5109531>. [Online]. Available: <https://doi.org/10.1063/1.5109531>.
- [19] J. Kästner, "Umbrella sampling," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, pp. 932–942, Nov. 2011. DOI: [10.1002/wcms.66](https://doi.org/10.1002/wcms.66).
- [20] G. Bussi and A. Laio, "Using metadynamics to explore complex free-energy landscapes," *Nature Reviews Physics*, vol. 2, no. 4, pp. 200–212, 2020, ISSN: 2522-5820. DOI: [10.1038/s42254-020-0153-0](https://doi.org/10.1038/s42254-020-0153-0). [Online]. Available: <https://doi.org/10.1038/s42254-020-0153-0>.

- [21] A. Pérez de Alba Ortíz, J. Vreede, and B. Ensing, “The adaptive path collective variable: A versatile biasing approach to compute the average transition path and free energy of molecular transitions,” *Methods Mol Biol*, vol. 2022, pp. 255–290, 2019, ISSN: 1064-3745. DOI: [10.1007/978-1-4939-9608-7_11](https://doi.org/10.1007/978-1-4939-9608-7_11).
- [22] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, “String method in collective variables: Minimum free energy paths and isocommittor surfaces,” *J. Chem. Phys.*, vol. 125, p. 024106, 2006.
- [23] W. Thiel, “Semiempirical quantum-chemical methods,” *WIREs Comput. Mol. Sci.*, vol. 4, pp. 145–147, 2014.
- [24] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, “Am1: A new general purpose quantum mechanical molecular model,” *J. Am. Chem. Soc.*, vol. 107, pp. 3902–3909, 1985.
- [25] J. J. P. Stewart, “Optimization of parameters for semiempirical methods i. method,” *Journal of Computational Chemistry*, vol. 10, no. 2, pp. 209–220, 1989. DOI: <https://doi.org/10.1002/jcc.540100208>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540100208>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540100208>.
- [26] B. Kim, R. Snyder, M. Nagaraju, *et al.*, “Reaction path-force matching in collective variables: Determining ab initio qm/mm free energy profiles by fitting mean force,” *J. Chem. Theory Comput.*, vol. 17, p. 4961, 2021.
- [27] B. Kim, Y. Shao, and J. Pu, “Doubly polarized qm/mm with machine learning chaperone polarizability,” *Journal of Chemical Theory and Computation*, vol. 17, no. 12, pp. 7682–7695, 2021.
- [28] Y. Zhou and J. Pu, “Reaction path force matching: A new strategy of fitting specific reaction parameters for semiempirical methods in combined qm/mm simulations,” *J. Chem. Theory Comput.*, vol. 10, pp. 3038–3054, 2014.
- [29] H. Lin and D. G. Truhlar, “Qm/mm: What have we learned, where are we, and where do we go from here?” *Theoretical Chemistry Accounts*, vol. 117, no. 2, pp. 185–199, 2007, ISSN: 1432-2234. DOI: [10.1007/s00214-006-0143-z](https://doi.org/10.1007/s00214-006-0143-z). [Online]. Available: <https://doi.org/10.1007/s00214-006-0143-z>.
- [30] M. F. Guest, P. Sherwood, and J. A. Nichols, “Massive parallelism: The hardware for computational chemistry?” In *High-Performance Computing*, R. J. Allan, M. F. Guest, A. D. Simpson, D. S. Henty, and D. A. Nicole, Eds. Boston, MA: Springer US, 1999, pp. 259–272, ISBN: 978-1-4615-4873-7. DOI: [10.1007/978-1-4615-4873-7_28](https://doi.org/10.1007/978-1-4615-4873-7_28). [Online]. Available: https://doi.org/10.1007/978-1-4615-4873-7_28.
- [31] J. A. Keith, V. Vassilev-Galindo, B. Cheng, *et al.*, “Combining machine learning and computational chemistry for predictive insights into chemical systems,” *Chemical Reviews*, vol. 121, no. 16, pp. 9816–9872, 2021, PMID: 34232033. DOI: [10.1021/acs.chemrev.1c00107](https://doi.org/10.1021/acs.chemrev.1c00107). eprint: <https://doi.org/10.1021/acs.chemrev.1c00107>. [Online]. Available: <https://doi.org/10.1021/acs.chemrev.1c00107>.

- [32] L. Boselt, M. Thurlemann, and S. Riniker, "Machine learning in qm/mm molecular dynamics simulations of condensed-phase systems," *J. Chem. Theory Comput.*, vol. 17, pp. 2641–2658, 2021.
- [33] X. Pan, J. Yang, R. Van, *et al.*, "Machine learning assisted free energy simulation of solutionphase and enzyme reactions," submitted, 2021.
- [34] J. Wu, L. Shen, and W. Yang, "Internal force corrections with machine learning for quantum mechanics/molecular mechanics simulations," *J. Chem. Phys.*, vol. 147, p. 161 732, 2017.
- [35] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, "Machine learning force fields: Construction, validation, and outlook," *J. Phys. Chem. C*, vol. 121, p. 511, 2017.
- [36] Z. Li, J. R. Kermode, and A. De Vita, "Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces," *Phys. Rev. Lett.*, vol. 114, p. 096 405, 2015.
- [37] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Phys. Rev. Lett.*, vol. 108, p. 058 301, 2012.
- [38] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, and M. Ceriotti, "Accurate molecular polarizabilities with coupled cluster theory and machine learning," *Proc. Natl. Acad. Sci. USA*, vol. 116, p. 3401, 2019.
- [39] X. Pan, J. Yang, R. Van, *et al.*, "Machine-learning-assisted free energy simulation of solutionphase and enzyme reactions," *J. Chem. Theory Comput.*, vol. 17, p. 5745, 2021.
- [40] M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning.," *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, p. 265, 2016.
- [41] D. Carroll, *Genetic Algorithm Driver 1.70; University of Illinois: Champaign, IL*. 1998.
- [42] D.-h. Lu, M. Zhao, and D. G. Truhlar, "Projection operator method for geometry optimization with constraints," *J. Comput. Chem.*, vol. 12, pp. 376–384, 1991.
- [43] P. Pulay and G. Fogarasi, "Geometry optimization in redundant internal coordinates," *J. Chem. Phys.*, vol. 96, pp. 2856–2860, 1992.
- [44] C. Peng, P. Y. Ayala, H. B. Schlegel, and M. J. Frisch, "Using redundant internal coordinates to optimize equilibrium geometries and transition states," *J. Comput. Chem.*, vol. 17, pp. 49–56, 1996.
- [45] N. Menshutkin, "Beiträgen zur kenntnis der affinitätskoeffizienten der alkylhaloide und der organischen amine," *Z. Physik. Chem.*, vol. 5, pp. 589–600, 1890.
- [46] O. Acevedo and W. L. Jorgensen, "Exploring solvent effects upon the menshutkin reaction using a polarizable force field," *J. Phys. Chem. B*, vol. 114, pp. 8425–8430, 2010.

- [47] C. Amovilli, B. Mennucci, and F. M. Floris, “Mscf study of the sn2 menshutkin reaction in aqueous solution within the polarizable continuum model,” *J. Phys. Chem. B*, vol. 102, pp. 3023–3028, 1998.
- [48] X. Fradera, L. Amat, M. Torrent, *et al.*, “Analysis of the changes on the potential energy surface of menshutkin reactions induced by external perturbations,” *J. Mol. Struct.: THEOCHEM*, vol. 371, pp. 171–183, 1996.
- [49] J. Gao, “A priori computation of a solvent-enhanced sn2 reaction profile in water: The menshutkin reaction,” *J. Am. Chem. Soc.*, vol. 113, pp. 7796–7797, 1991.
- [50] H. Hirao, Y. Nagae, and M. Nagaoka, “Transition-state optimization by the free energy gradient method: Application to aqueous-phase menshutkin reaction between ammonia and methyl chloride,” *Chem. Phys. Lett.*, vol. 348, pp. 350–356, 2001.
- [51] K. Ohmiya and S. Kato, “Solution reaction path hamiltonian based on reference interaction site model self-consistent field method: Application to menshutkin-type reactions,” *J. Chem. Phys.*, vol. 119, pp. 1601–1610, 2003.
- [52] K. Okamoto, S. Fukui, and H. Shingu, “Kinetic studies of bimolecular nucleophilic substitution. vi. rates of the menshutkin reaction of methyl iodide with methylamines and ammonia in aqueous solutions,” *Bull. Chem. Soc. Jpn.*, vol. 40, pp. 1920–1925, 1967.
- [53] T. N. Truong, T.-T. T. Truong, and E. V. Stefanovich, “A general methodology for quantum modeling of free-energy profile of reaction in solution: An application to the menshutkin $\text{nh}_3 + \text{ch}_3\text{cl}$ reaction in water,” *J. Chem. Phys.*, vol. 107, pp. 1881–1889, 1997.
- [54] S. P. Webb and M. S. Gordon, “Solvation of the menshutkin reaction: A rigorous test of the effective fragment method,” *J. Phys. Chem. A*, vol. 103, pp. 1265–1273, 1999.
- [55] W. J. Hehre, L. Radom, P. v. R. Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory*. New York: John Wiley, 1986.
- [56] W. Kohn and L. J. Sham, “Self-consistent equations including exchange and correlation effects,” *Phys. Rev. A*, vol. 140, pp. 1133–1138, 1965.
- [57] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, USA, 1994.
- [58] F. Ercolessi and J. B. Adams, “Interatomic potentials from first-principles calculations: The force-matching method,” *Europhys. Lett.*, vol. 26, pp. 583–588, 1994.
- [59] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, “Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching,” *J. Chem. Phys.*, vol. 120, pp. 10 896–10 913, 2004.
- [60] S. Izvekov and G. A. Voth, “A multiscale coarse-graining method for biomolecular systems,” *J. Phys. Chem. B*, vol. 109, pp. 2469–2473, 2005.
- [61] A. Laio, S. Bernard, G. L. Chiarotti, S. Scandolo, and E. Tosatti, “Physics of iron at earth’s core conditions,” *Science*, vol. 287, pp. 1027–1030, 2000.

- [62] G. Csanyi, T. Albaret, M. C. Payne, and A. De Vita, "learn on the fly": A hybrid classical and quantum-mechanical molecular dynamics simulation," *Phys. Rev. Lett.*, vol. 93, p. 175 503, 2004.
- [63] P. Maurer, A. Laio, H. W. Hugosson, M. C. Colombo, and U. Rothlisberger, "Automated parametrization of biomolecular force fields from quantum mechanics/molecular mechanics (qm/mm) simulations through force matching," *J. Chem. Theory Comput.*, vol. 3, pp. 628–639, 2007.
- [64] O. Arkin-Ojo, Y. Song, and F. Wang, "Developing ab initio quality force field from condensed phase quantum-mechanics/molecular-mechanics calculations through the adaptive force matching method," *J. Chem. Phys.*, vol. 129, p. 064 108, 2008.
- [65] P. S. Hudson, S. Boresch, D. M. Rogers, and H. L. Woodcock, "Accelerating qm/mm free energy computations via intramolecular force matching," *J. Chem. Theory Comput.*, vol. 14, pp. 6327–6335, 2018.
- [66] T. J. Giese and D. M. York, "Development of a robust indirect approach for mm qm free energy calculations that combines force-matched reference potential and bennetts acceptance ratio methods," *J. Chem. Theory Comput.*, vol. 15, pp. 5543–5562, 2019, ISSN: 1549-9618. DOI: [10.1021/acs.jctc.9b00401](https://doi.org/10.1021/acs.jctc.9b00401). [Online]. Available: <https://doi.org/10.1021/acs.jctc.9b00401>.
- [67] W. E, W. Ren, and E. Vanden-Eijnden, "Finite temperature string method for the study of rare events," *J. Phys. Chem. B*, vol. 109, pp. 6688–6693, 2005.
- [68] P. Li, X. Jia, X. Pan, Y. Shao, and Y. Mei, "Accelerated computation of free energy profile at ab initio qm/mm accuracy via a semi-empirical reference-potential. i. weighted thermodynamics perturbation," *J. Chem. Theory Comput.*, vol. 14, pp. 5583–5596, 2018.
- [69] X. Pan, P. Li, J. Ho, J. Pu, Y. Mei, and Y. Shao, "Accelerated computation of free energy profile at ab initio quantum mechanical/molecular mechanical accuracy via a semi-empirical reference potential. ii. recalibrating semi-empirical parameters with force matching," *Phys. Chem. Chem. Phys.*, vol. 21, pp. 20 595–20 605, 2019.
- [70] N. Goldman, L. E. Fried, and L. Koziol, "Using force-matched potentials to improve the accuracy of density functional tight binding for reactive conditions," *J. Chem. Theory Comput.*, vol. 11, pp. 4530–4535, 2015.
- [71] M. P. Kroonblawd, F. Pietrucci, A. M. Saitta, and N. Goldman, "Generating converged accurate free energy surfaces for chemical reactions with a force-matched semiempirical model," *J. Chem. Theory Comput.*, vol. 14, pp. 2207–2218, 2018.
- [72] L. Shen, J. Wu, and W. Yang, "Multiscale quantum mechanics/molecular mechanics simulations with neural networks," *J. Chem. Theory Comput.*, vol. 12, pp. 4934–4946, 2016.
- [73] L. Shen and W. Yang, "Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks," *J. Chem. Theory Comput.*, vol. 14, pp. 1442–1455, 2018.

- [74] J. Zeng, T. J. Giese, S. Ekesan, and D. M. York, "Development of range-corrected deep learning potentials for fast, accurate quantum mechanical/molecular mechanical simulations of chemical reactions in solution," *ChemRxiv*, 2021.
- [75] W. G. Noid, J.-W. Chu, G. S. Ayton, *et al.*, "The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models," *J. Chem. Phys.*, vol. 128, p. 244 114, 2008.
- [76] K. Zinovjev, J. J. Ruiz-Pernia, and I. Tunon, "Toward an automatic determination of enzymatic reaction mechanisms and their activation free energies," *J. Chem. Theory Comput.*, vol. 9, pp. 3740–3749, 2013.
- [77] E. Darve and A. Pohorille, "Calculating free energies using average force," *J. Chem. Phys.*, vol. 115, pp. 9169–9183, 2001.
- [78] M. J. Ruiz-Montero, D. Frenkel, and J. J. Brey, "Efficient schemes to compute diffusive barrier crossing rates," *Mol. Phys.*, vol. 90, pp. 925–941, 1997.
- [79] W. K. den Otter and W. J. Briels, "The calculation of free-energy differences by constrained molecular-dynamics simulations," *J. Chem. Phys.*, vol. 109, pp. 4139–4146, 1998.
- [80] E. B. Wilson Jr., J. C. Decius, and P. C. Cross, *Molecular Vibrations*. New York: McGraw-Hill, 1955.
- [81] C. F. Jackels, Z. Gu, and D. G. Truhlar, "Reaction-path potential and vibrational frequencies in terms of curvilinear internal coordinates," *J. Chem. Phys.*, vol. 102, pp. 3188–3201, 1995.
- [82] Y.-Y. Chuang and D. G. Truhlar, "Reaction-path dynamics in redundant internal coordinates," *J. Phys. Chem. A*, vol. 102, pp. 242–247, 1998.
- [83] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in FORTRAN 77: The Art of Scientific Computing*, 2nd. New York: Cambridge University Press, 1992.
- [84] "In the present work, we only correct the internal forces on the cvs to reproduce the high-level mean forces along the free energy path defined in the same set of cvs. however, the force corrections on the non-cv degrees of freedom in the redundant internal coordinate system provided are also available as a byproduct of eq. (a22) in appendix b. under the cv-only fm scheme, when the backward coordinate transformation procedure is used to obtain the corresponding cartesian force corrections, one needs to neglect the internal force corrections on the non-cv degrees of freedom by setting them to zeros. effects of including the additional non-cv internal force corrections in fm are being examined in our ongoing work and will be reported separately," Unpublished Work.
- [85] J. Gao and X. Xia, "A two-dimensional energy surface for a type ii sn2 reaction in aqueous solution," *J. Am. Chem. Soc.*, vol. 115, pp. 9667–9675, 1993.

- [86] V. Dillet, D. Rinaldi, J. Bertran, and J.-L. Rivail, "Analytical energy derivatives for a realistic continuum model of solvation: Application to the analysis of solvent effects on reaction paths," *J. Chem. Phys.*, vol. 104, pp. 9437–9444, 1996.
- [87] K. Naka, H. Sato, A. Morita, F. Hirata, and S. Kato, "Rism-scf study of the free-energy profile of the menshutkin-type reaction $\text{nh}_3 + \text{ch}_3\text{cl} \rightarrow \text{nh}_3\text{ch}_3^+ + \text{cl}^-$ in aqueous solution," *Theor. Chem. Acc.*, vol. 102, pp. 165–169, 1999.
- [88] J. J. Ruiz-Pernia, E. Silla, I. Tunon, S. Marti, and V. Moliner, "Hybrid qm/mm potentials of mean force with interpolated corrections," *J. Phys. Chem. B*, vol. 108, pp. 8427–8433, 2004.
- [89] S. Marti, V. Moliner, and I. Tunon, "Improving the qm/mm description of chemical processes: A dual level strategy to explore the potential energy surface in vary large system," *J. Chem. Theory Comput.*, vol. 1, pp. 1008–1016, 2005.
- [90] I. Fdez Galvan, M. E. Martin, and M. A. Aguilar, "A new method to locate saddle points for reactions in solution by using the free-energy gradient method and the mean field approximation," *J. Comput. Chem.*, vol. 25, pp. 1227–1233, 2004.
- [91] O. Acevedo and W. L. Jorgensen, "Solvent effects on organic reactions from qm/mm simulations," in *Annual Reports in Computational Chemistry*, D. Spellmeyer, Ed. Amsterdam, The Netherlands: Elsevier, 2006, vol. 2, p. 263.
- [92] T. Yamamoto, "Variational and perturbative formulations of quantum mechanical/-molecular mechanical free energy with mean-field embedding and its analytical gradients," *J. Chem. Phys.*, vol. 129, p. 244 104, 2008.
- [93] Y. Komeiji, T. Ishikawa, Y. Mochizuki, H. Yamataka, and T. Nakano, "Fragment molecular orbital method-based molecular dynamics (fmo-md) as a simulator for chemical reactions in explicit solvation," *J. Comput. Chem.*, vol. 30, pp. 40–50, 2009.
- [94] J. Z. Vilseck, S. V. Sambasivarao, and O. Acevedo, "Optimal scaling factors for cm1 and cm3 atomic charges in rm1-based aqueous simulations," *J. Comput. Chem.*, vol. 32, pp. 2836–2842, 2011.
- [95] H. Nakano and T. Yamamoto, "Variational calculation of quantum mechanical/molecular mechanical free energy with electronic polarization of solvent," *J. Chem. Phys.*, vol. 136, p. 134 107, 2012.
- [96] J. Gao and X. Xia, "A priori evaluation of aqueous polarization effects through monte carlo qm-mm simulations," *Science*, vol. 258, pp. 631–635, 1992.
- [97] J. Gao, "Hybrid quantum and molecular mechanical simulations: An alternative avenue to solvent effects in organic chemistry," *Acc. Chem. Res.*, vol. 29, pp. 298–305, 1996.
- [98] S. Ten-no, F. Hirata, and S. Kato, "A hybrid approach for the solvent effect on the electronic structure of a solute based on the rism and hartree-fock equations," *Chem. Phys. Lett.*, vol. 214, pp. 391–396, 1993.

- [99] S. Ten-no, F. Hirata, and S. Kato, "Reference interaction site model self-consistent field study for solvation effect on carbonyl compounds in aqueous solution," *J. Chem. Phys.*, vol. 100, pp. 7443–7453, 1994.
- [100] N. Okuyama-Yoshida, M. Nagaoka, and T. Yamabe, "Transition-state optimization on free energy surface: Toward solution chemical reaction ergodography," *Int. J. Quantum Chem.*, vol. 70, pp. 95–103, 1998.
- [101] P. N. Day, J. H. Jensen, M. S. Gordon, *et al.*, "An effective fragment method for modeling solvent effects in quantum mechanical calculations," *J. Chem. Phys.*, vol. 105, pp. 1968–1986, 1996.
- [102] H. Nakano and T. Yamamoto, "Accurate and efficient treatment of continuous solute charge density in the mean-field qm/mm free energy calculation," *J. Chem. Theory Comput.*, vol. 9, pp. 188–203, 2013.
- [103] H. Hu, Z. Lu, and W. Yang, "Qm/mm minimum free-energy path: Methodology and application to triosephosphate isomerase," *J. Chem. Theory Comput.*, vol. 3, pp. 390–406, 2007.
- [104] H. Hu, Z. Lu, J. M. Parks, S. K. Burger, and W. Yang, "Quantum mechanics/molecular mechanics minimum free-energy path for accurate reaction energies in solution and enzymes: Sequential sampling and optimization on the potential of mean force surface," *J. Chem. Phys.*, vol. 128, p. 034 105, 2008.
- [105] H. Hu and W. Yang, "Free energies of chemical reactions in solution and in enzymes with ab initio qm/mm methods," *Annu. Rev. Phys. Chem.*, vol. 59, pp. 573–601, 2008.
- [106] J. J. Ruiz-Pernia, E. Silla, I. Tunon, and S. Marti, "Hybrid quantum mechanics/molecular mechanics simulations with two-dimensional interpolated corrections: Application to enzymatic processes," *J. Phys. Chem. B*, vol. 110, pp. 17 663–17 670, 2006.
- [107] A. D. MacKerell Jr., D. Bashford, M. Bellott, *et al.*, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B*, vol. 102, pp. 3586–3616, 1998.
- [108] E. Neria, S. Fischer, and M. Karplus, "Simulation of activation free energies in molecular systems," *J. Chem. Phys.*, vol. 105, pp. 1902–1921, 1996.
- [109] Computer Program, 1998.
- [110] B. R. Brooks, C. L. Brooks III, A. D. MacKerell Jr., *et al.*, "Charmm: The molecular simulation program," *J. Comput. Chem.*, vol. 30, pp. 1545–1614, 2009.
- [111] Y. Shao, L. F. Molnar, Y. Jung, *et al.*, "Advances in methods and algorithms in a modern quantum chemistry program package," *Phys. Chem. Chem. Phys.*, vol. 8, pp. 3172–3191, 2006.
- [112] J.-P. Ryckaert, G. Ciccotti, and H. C. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, pp. 327–337, 1977.

- [113] T. Darden, D. York, and L. Pedersen, "Particle mesh ewald: An $n \log(n)$ method for ewald sums in large systems," *J. Chem. Phys.*, vol. 98, pp. 10089–10092, 1993.
- [114] K. Nam, J. Gao, and D. M. York, "An efficient linear-scaling ewald method for long-range electrostatics in combined qm/mm calculations," *J. Comp. Theory Comput.*, vol. 1, pp. 2–13, 2005.
- [115] K. Nam, "Acceleration of ab initio qm/mm calculations under periodic boundary conditions by multiscale and multiple time step approaches," *J. Chem. Theory Comput.*, vol. 10, pp. 4175–4183, 2014.
- [116] A. D. Becke, "Densityfunctional thermochemistry. iii. the role of exact exchange," *J. Chem. Phys.*, vol. 98, pp. 5648–5652, 1993.
- [117] C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of electron density," *Phys. Rev. B*, vol. 37, pp. 785–789, 1988.
- [118] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields," *J. Phys. Chem.*, vol. 98, pp. 11 623–11 627, 1994.
- [119] A. D. Becke, "A new mixing of hartreefock and local density-functional theories," *J. Chem. Phys.*, vol. 98, pp. 1372–1377, 1993.
- [120] C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems," *Phys. Rev.*, vol. 46, pp. 618–622, 1934.
- [121] J. A. Pople, J. S. Binkley, and R. Seeger, "Theoretical models incorporating electron correlation," *Int. J. Quantum Chem.*, vol. Supp. Y-10, pp. 1–19, 1976.
- [122] M. M. Francl, W. J. Pietro, W. J. Hehre, *et al.*, "Selfconsistent molecular orbital methods. xxiii. a polarizationtype basis set for secondrow elements," *J. Chem. Phys.*, vol. 77, pp. 3654–3665, 1982.
- [123] K. Okamoto, S. Fukui, I. Nitta, and H. Shingu, "Kinetic studies of bimolecular nucleophilic substitution. vii. effect of hydroxylic solvents on the nucleophilicity of aliphatic amines in the menshutkin reaction," *Bull. Chem. Soc. Jpn.*, vol. 40, pp. 2354–2357, 1967.
- [124] M. J. Frisch, J. A. Pople, and J. S. Binkley, "Self-consistent molecular orbital methods 25. supplementary functions for gaussian basis sets," *J. Chem. Phys.*, vol. 80, pp. 3265–3269, 1984.
- [125] A. D. McLean and G. S. Chandler, "Contracted gaussian-basis sets for molecular calculations. 1. 2nd row atoms, $z=11-18$," *J. Chem. Phys.*, vol. 72, pp. 5639–5648, 1980.
- [126] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, "Selfconsistent molecular orbital methods. xx. a basis set for correlated wave functions," *J. Chem. Phys.*, vol. 72, pp. 650–654, 1980.

- [127] C. Knight, G. E. Lindberg, and G. A. Voth, "Multiscale reactive molecular dynamics," *J. Chem. Phys.*, vol. 137, 22A525, 2012.
- [128] C. Knight, C. M. Maupin, S. Izvekov, and G. A. Voth, "Defining condensed phase reactive force fields from ab initio molecular dynamics simulations: The case of the hydrated excess proton," *J. Chem. Theory Comput.*, vol. 6, pp. 3223–3232, 2010.
- [129] L. Zhang, J. Han, H. Wang, W. A. Saidi, R. Car, and W. E, "End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems," *Advances in Neural Information Processing Systems*, 2018.
- [130] H. Wang, L. Zhang, J. Han, and W. E, "Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics," *Comput. Phys. Commun.*, vol. 228, pp. 178–184, 2018.
- [131] Y. Zhang, H. Wang, W. Chen, *et al.*, "Dp-gen: A concurrent learning platform for the generation of reliable deep learning based potential energy models," *Comput. Phys. Commun.*, vol. 253, p. 107 206, 2020.
- [132] L. S. Devi-Kesavan, M. Garcia-Viloca, and J. Gao, "Semiempirical qm/mm potential with simple valence bond (svb) for enzyme reactions. application to the nucleophilic addition reaction in haloalkane dehalogenase," *Theor. Chem. Acc.*, vol. 109, pp. 133–139, 2003.
- [133] S. Califano, *Vibrational States*. New York: Wiley, 1976.
- [134] C. Barnes, "Inorganic chemistry (housecroft, catherine e.; sharpe, alan g.)," *Journal of Chemical Education*, vol. 80, no. 7, Appendix 6, 1013–1014, 2003, ISSN: 0021-9584. DOI: [10.1021/ed080p747](https://doi.org/10.1021/ed080p747). [Online]. Available: <https://doi.org/10.1021/ed080p747>.
- [135] J. Gao, M. A. Thompson, *et al.*, *Combined quantum mechanical and molecular mechanical methods*. ACS Publications, 1998, vol. 712.
- [136] M. J. S. Dewar, Y. Yamaguchi, and S. H. Suck, "Mndo calculations of molecular electric polarizabilities, hyperpolarizabilities, and nonlinear optical coefficients," *Chem. Phys. Lett.*, vol. 59, p. 541, 1978.
- [137] W. A. Parkinson and M. C. Zerner, "Hyperpolarizability determined from the intermediate neglect of differential overlap model," *J. Chem. Phys.*, vol. 94, p. 478, 1991.
- [138] N. Matsuzawa and D. A. Dixon, "Semiempirical calculations of hyperpolarizabilities for donor-acceptor molecules: Comparison to experiment," *J. Phys. Chem. A.*, vol. 96, p. 6232, 1992.
- [139] G. Schurer, P. Gedeck, M. Gottschalk, and T. Clark, "Accurate parametrized variational calculations of the molecular electronic polarizability by nddo-based methods," *Int. J. Quantum Chem.*, vol. 75, p. 17, 1999.
- [140] B. Martin, P. Gedeck, and T. Clark, "Additive nddo-based atomic polarizability model," *Int. J. Quant. Chem.*, vol. 77, p. 473, 2000.

- [141] G. Monard, M. I. Bernal-Uruchurtu, A. van der Vaart, K. M. Merz, and M. F. Ruiz-Lopez, "Simualtion of liquid water using semiempirical hamiltonians and the divide and conquer approach," *J. Phys. Chem. A*, vol. 109, p. 3425, 2005.
- [142] T. J. Giese and D. M. York, "Improvement of semiempirical response properties with charge-dependent response density," *J. Chem. Phys.*, vol. 123, p. 164 108, 2005.
- [143] L. Fiedler, J. Gao, and D. G. Truhlar, "Polarized molecular orbital model chemistry. 1. ab initio foundations," *J. Chem. Theory Comput.*, vol. 7, p. 852, 2011.
- [144] J. Gao, F. J. Luque, and M. Orozco, "Induced dipole moment and atomic charges based on avarage electrostatic potentials in aqueous solution," *J. Chem. Phys.*, vol. 98, p. 2975, 1993.
- [145] J. Gao, "A molecular-orbital derived polarization potential for liquid water," *J. Chem. Phys.*, vol. 109, p. 2346, 1998.
- [146] M. Welborn, J. Chen, L.-P. Wang, and T. Van Voorhis, "Why many semiempirical molecular orbital theories fail for liquid water and how to fix them," *J. Comput. Chem.*, vol. 36, p. 934, 2015.
- [147] W. F. Murphy, "The rayleigh depolarization ratio and rotational raman spectrum of water vapor and the polarizability components for the water molecule," *J. Chem. Phys.*, vol. 67, p. 5877, 1977.
- [148] D. Hait and M. Head-Gordon, "How accurate are static polarizability predictions from density functional theory? an assessment over 132 species at equilibrium geometry," *Phys. Chem. Chem. Phys.*, vol. 20, p. 19 800, 2018.
- [149] P. Zhang, L. Fiedler, H. R. Leverentz, D. G. Truhlar, and J. Gao, "Polarized molecular orbital model chemistry. 2. the pmo method," *J. Chem. Theory Comput.*, vol. 7, p. 857, 2011.
- [150] D. M. York and W. Yang, "A chemical potential equalization method for molecular simulations," *J. Chem. Phys.*, vol. 104, p. 159, 1996.
- [151] T. J. Giese and D. M. York, "Density-functional expansion methods: Grand challenges," *Theor. Chem. Acc.*, vol. 131, p. 1145, 2012.
- [152] S. Kaminski, T. J. Giese, M. Gaus, D. M. York, and M. Elstner, "Extended polarization in third-order scc-dftb from chemical-potential equalization," *J. Phys. Chem. A*, vol. 116, p. 9131, 2012.
- [153] A. S. Christensen, M. Elstner, and Q. Cui, "Improving intermolecular interactions in dftb3 using extended polarization from chemical-potential equalization," *J. Chem. Phys.*, vol. 143, p. 084 123, 2015.
- [154] M. Isegawa, L. Fiedler, H. R. Leverentz, *et al.*, "Polarized molecular orbital model chemistry 3. the pmo method extended to organic chemistry," *J. Chem. Theory Comput.*, vol. 9, p. 33, 2013.
- [155] A. Stone, *The theory of intermolecular forces*. oUP oxford, 2013.

- [156] D. N. Bernardo, Y. Ding, K. Krogh-Jespersen, and R. M. Levy, "An anisotropic polarizable water model: Incorporation of all-atom polarizabilities into molecular mechanics force fields," *J. Phys. Chem. B*, vol. 98, p. 4180, 1994.
- [157] P. Ren and J. W. Ponder, "Polarizable atomic multipole water model for molecular mechanics simulation," *J. Phys. Chem. B*, vol. 107, p. 5933, 2003.
- [158] W. Xie, J. Pu, A. D. MacKerell, and J. Gao, "Development of a polarizable intermolecular potential function (pipf) for liquid amides and alkanes," *J. Chem. Theory Comput.*, vol. 3, p. 1878, 2007.
- [159] J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell, "An empirical polarizable force field based on the classical drude oscillator model: Development history and recent applications," *Chem. Rev.*, vol. 116, p. 4983, 2016.
- [160] Z. Jing, C. Liu, S. Y. Cheng, *et al.*, "Polarizable force fields for biomolecular simulations: Recent advances and applications," *Annu. Rev. Biophys.*, vol. 48, p. 371, 2019.
- [161] T. Stecher, N. Bernstein, and G. Csányi, "Free energy surface reconstruction from umbrella samples using gaussian process regression," *J. Chem. Theory Comput.*, vol. 10, p. 4079, 2014.
- [162] J. P. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.*, vol. 98, p. 146 401, 2007.
- [163] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Phys. Rev. Lett.*, vol. 104, p. 136 403, 2010.
- [164] J. Applequist, J. R. Carl, and K.-K. Fung, "An atom dipole interaction model for molecular polarizability. application to polyatomic molecules and determination of atom polarizabilities," *J. Am. Chem. Soc.*, vol. 94, p. 2952, 1972.
- [165] B. T. Thole, "Molecular polarizabilities calculated with a modified dipole interaction," *Chem. Phys.*, vol. 59, p. 341, 1981.
- [166] P. Zhang, P. Bao, and J. Gao, "Dipole preserving and polarization consistent charges," *J. Comput. Chem.*, vol. 32, p. 2127, 2011.
- [167] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, "Reduced and quenched polarizabilities of interior atoms in molecules," *Chem. Sci.*, vol. 4, p. 2349, 2013.
- [168] J. Tomasi, B. Mennucci, and R. Cammi, "Quantum mechanical continuum solvation models," *Chem. Rev.*, vol. 105, p. 2999, 2005.
- [169] J. G. Ángyán, F. Colonna-Cesari, and O. Tapia, "Analytical first and second energy derivatives in the polarization model," *Chem. Phys. Lett.*, vol. 166, p. 180, 1990.
- [170] T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen," *J. Chem. Phys.*, vol. 90, p. 1007, 1989.

- [171] R. A. Kendall, T. H. Dunning, and R. J. Harrison, "Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions," *J. Chem. Phys.*, vol. 96, p. 6796, 1992.
- [172] D. E. Woon and T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. iii. the atoms aluminum through argon," *J. Chem. Phys.*, vol. 98, p. 1358, 1993.
- [173] F. Zhu and G. Hummer, "Convergence and error estimation in free energy calculations using the weighted histogram analysis method," *J. Comput. Chem.*, vol. 33, p. 453, 2012.
- [174] F. L. Hirshfeld, "Bonded-atom fragments for describing molecular charge densities," *Theoret. Chim. Acta*, vol. 44, p. 129, 1977.
- [175] A. Krishtal, P. Senet, M. Yang, and C. Van Alsenoy, "A hirshfeld partitioning of polarizabilities of water clusters," *J. Chem. Phys.*, vol. 125, p. 034312, 2006.
- [176] M. e. Frisch, G. Trucks, H. Schlegel, *et al.*, *Gaussian 16*, 2016.
- [177] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [178] W. J. Hehre, R. F. Stewart, and J. A. Pople, "Self-consistent molecular-orbital methods. i. use of gaussian expansions of slater-type atomic orbitals," *J. Chem. Phys.*, vol. 51, p. 2657, 1969.
- [179] W. J. Hehre, R. Ditchfield, R. F. Stewart, and J. A. Pople, "Self-consistent molecular orbital methods. iv. use of gaussian expansions of slater-type orbitals. extension to second-row molecules," *J. Chem. Phys.*, vol. 52, p. 2769, 1970.
- [180] F. Jensen, "Polarization consistent basis sets: Principles," *J. Chem. Phys.*, vol. 115, p. 9113, 2001.
- [181] F. Jensen, "Polarization consistent basis sets. ii. estimating the kohn-sham basis set limit," *J. Chem. Phys.*, vol. 116, p. 7372, 2002.
- [182] F. Jensen, "Polarization consistent basis sets. iii. the importance of diffuse functions," *J. Chem. Phys.*, vol. 117, p. 9234, 2002.
- [183] F. Jensen and T. Helgaker, "Polarization consistent basis sets. v. the elements si-cl," *J. Chem. Phys.*, vol. 121, p. 3463, 2004.
- [184] Y. Zhao and D. G. Truhlar, "The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four m06-class functionals and 12 other functionals," *Theor. Chem. Acc.*, vol. 120, p. 215, 2008.
- [185] J. Huang, Y. Mei, G. Konig, *et al.*, "An estimation of hybrid quantum mechanical molecular mechanical polarization energies for small molecules using polarizable force-field approaches," *J. Chem. Theory Comput.*, vol. 13, p. 679, 2017.

- [186] T. J. Giese and D. M. York, “Charge-dependent model for many-body polarization, exchange, and dispersion interactions in hybrid quantum mechanical/molecular mechanical calculations,” *J. Chem. Phys.*, vol. 127, p. 194 101, 2007.
- [187] A. P. Bartók, R. Kondor, and G. Csányi, “On representing chemical environments,” *Phys. Rev. B*, vol. 87, p. 184 115, 2013.
- [188] Computer Program, 2016.

A. PROOF OF COPYRIGHT PERMISSION(S)

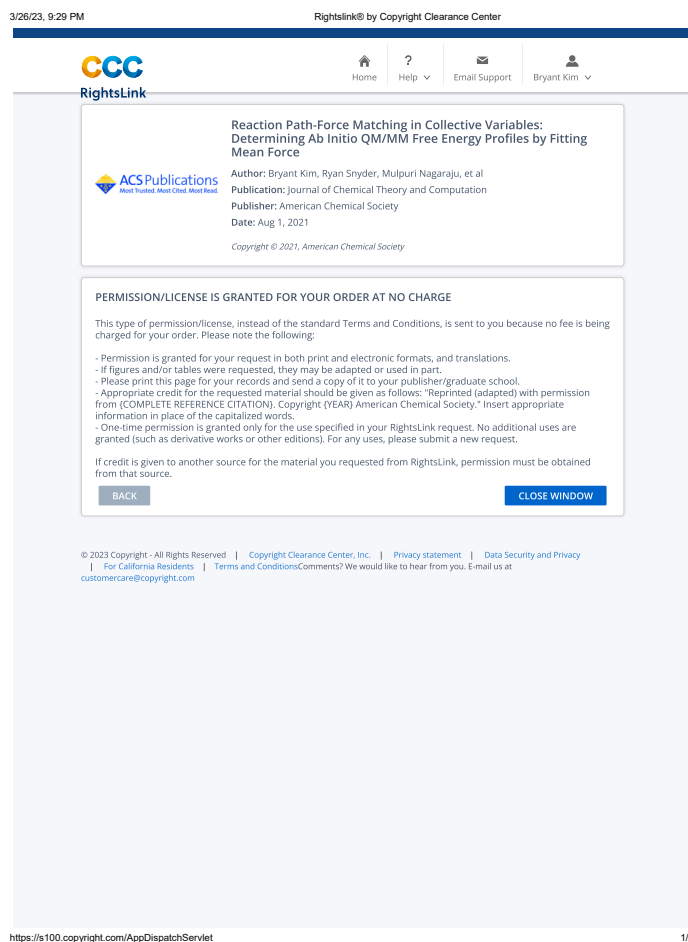




Figure A.1. Kim, B., Snyder, R., Nagaraju, M., et al. Reaction Path-Force Matching in Collective Variables: Determining Ab Initio QM/MM Free Energy Profiles by Fitting Mean Force. Journal of Chemical Theory and Computation, 2021, 17(8), 4729-4737. Reprinted with permission from Journal of Chemical Theory and Computation. Copyright 2021 American Chemical Society.

3/26/23, 9:31 PM Rightslink® by Copyright Clearance Center


[Home](#)
[?](#)
[Email Support](#)
[Bryant Kim](#)



Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability
 Author: Bryant Kim, Yihan Shao, Jingzhi Pu
 Publication: Journal of Chemical Theory and Computation
 Publisher: American Chemical Society
 Date: Dec 1, 2021
Copyright © 2021, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

[BACK](#)
[CLOSE WINDOW](#)

© 2023 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy
| For California Residents | Terms and Conditions

Comments? We would like to hear from you. E-mail us at customer-care@copyright.com

https://s100.copyright.com/AppDispatchServlet#formTop
1/1

Figure A.2. Kim, B., Shao, Y., Pu, J. Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability. Journal of Chemical Theory and Computation, 2021, 17(12), 7816-7825. Reprinted with permission from Journal of Chemical Theory and Computation. Copyright 2021 American Chemical Society.

B. SUPPORTING INFORMATION

B.1 Reaction Path-Force Matching in Collective Variables: Determining Ab Initio QM/MM Free Energy Profiles by Fitting Mean Force

The following Supporting Information section has been adapted from a previously published article (Kim, B., Snyder, R., Nagaraju, M., et al., "Reaction Path-Force Matching in Collective Variables: Determining Ab Initio QM/MM Free Energy Profiles by Fitting Mean Force," Journal of Chemical Theory and Computation, 2021, 17(8), 4729-4737) to provide further details and supporting evidence for the main findings presented in this thesis.

Supporting Information

Reaction Path-Force Matching in Collective Variables: Determining *Ab Initio* QM/MM Free Energy Profiles by Fitting Mean Force

Bryant Kim,[†] Ryan Snyder,[†] Mulpuri Nagaraju,[†] Yan Zhou,[†] Pedro Ojeda-May,[†] Seth Keeton,[†] Mellisa Hege,[†] Yihan Shao,^{*‡} and Jingzhi Pu^{*†}

[†]Department of Chemistry and Chemical Biology, Indiana University-Purdue University Indianapolis, 402 N. Blackford St., Indianapolis, IN 46202

[‡]Department of Chemistry and Biochemistry, University of Oklahoma, 101 Stephenson Pkwy, Norman, OK 73019

(*Correspondence: yihan.shao@ou.edu and jpu@iupui.edu)

1. Description of the modified solute/solvent van der Waals interactions by NBFIX in CHARMM

Following Gao and Xia,¹ we used the pair-wise Lennard-Jones (LJ) potentials adjusted for the Menshutkin reaction to treat the nonbonded van der Waals (vdW) interactions between the QM solute described by AM1 (or AI methods) and the MM solvent described by the TIP3P water model. The pair-specific vdW parameters of Gao and Xia¹ were implemented using the NBFIX facility in CHARMM. The LJ potential for computing the vdW interaction between a solute atom i and a solvent atom j is:

$$E_{\text{LJ}}(r_{ij}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] = -E_{\text{min},ij} \left[\left(\frac{r_{\text{min},ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{\text{min},ij}}{r_{ij}} \right)^6 \right] \quad (\text{S1})$$

where the pair-wise parameter σ_{ij} (which is the geometric mean of the associated atomic parameters σ_i and σ_j) and its CHARMM-compatible form $r_{\text{min},ij}$ are connected through:

$$r_{\text{min},ij} = \sqrt[6]{2} \sigma_{ij} = \sqrt[6]{2} \sqrt{\sigma_i \sigma_j} \quad (\text{S2})$$

The CHARMM-compatible parameter $E_{\text{min},ij}$, which corresponds to the minimum of LJ potential E_{LJ} obtained at $r = r_{\text{min},ij}$, is connected to the pair-wise interaction energy parameter ϵ_{ij} and its associated atomic parameters ϵ_i and ϵ_j through:

$$E_{\text{min},ij} = E_{\text{LJ}}(r_{\text{min},ij}) = -\epsilon_{ij} = -\sqrt{\epsilon_i \epsilon_j} \quad (\text{S3})$$

The σ_i and ε_i parameters used by Gao and Xia¹ in the original atomic form are given in Table S1. The equivalent parameters $E_{\min, ij}$ and $r_{\min, ij}$, made compatible with the CHARMM NBFIX convention through the conversions under Eqs. (S2-S3), are given in Table S2.

Table S1. The atomic vdW parameters used by Gao and Xia¹ in their AM1/TIP3P model for simulating the Menshutkin reaction^a

atom	σ (Å)	ε (kcal/mol)
H ₃ N-CH ₃ -Cl		
C	3.4000	0.1000
N	3.0875	0.1615
Cl	4.1964	0.1119
H _C	2.0000	0.0700
H _N	0.0000	0.0000
Water		
O	3.1506	0.1521
H	0.0000	0.0000

^aadopted from Table I of Gao and Xia *JACS* **1993**, 115, 9667.

Table S2. The pair-specific vdW parameters for QM/MM simulations of the Menshutkin reaction implemented in CHARMM for the present work^a

Pair	$r_{\min, ij}$ (Å) ^b	$E_{\min, ij}$ (kcal/mol)
solute-solvent		
C-O	3.67373	-0.12333
N-O	3.50084	-0.15673
Cl-O	4.08138	-0.13046
H _C -O	2.81763	-0.10318
H _N -O	0.00000	0.00000
C-H	0.00000	0.00000
N-H	0.00000	0.00000
Cl-H	0.00000	0.00000
H _C -H	0.00000	0.00000
H _N -H	0.00000	0.00000
solvent-solvent		
O-O	3.53643	-0.15210
H-H	0.00000	0.00000
H-O	0.00000	0.00000

^avdW parameters in the CHARMM NBFIX compatible form are converted from Table S1 using Eqs. (S2-S3)

^b $r_{\min, ij}$ is set to a dummy value of zero when $E_{\min, ij} = 0$

2. Illustration of Eq. (A15) using its matrix form for spline-based FM in a two-CV case

To help illustrate the implementation of the force matching procedure described in *Appendix A*, we give the explicit construction of Eq. (A15) for a specific example (see *Appendix A* for notations and symbols). For instance, in a case of two bond CVs denoted by r^1 and r^2 ($N=2$), we use a six-point spline grid $n_{\text{grid}}^1 = 6$ ($m_1 = 2n_{\text{grid}}^1 = 12$) and a four-point grid $n_{\text{grid}}^2 = 4$ ($m_2 = 2n_{\text{grid}}^2 = 8$) (therefore a total of $M=12+8=20$ spline parameters) for matching the CV internal forces, based on a collection of three configurations ($L=3$; and therefore $NL=2 \times 3=6$) that assumes the following sample distributions: $r^1 \in [r_5^1, r_6^1]$, $r^1 \in [r_3^1, r_4^1]$, and $r^1 \in [r_2^1, r_3^1]$, and $r^2 \in [r_3^2, r_4^2]$, and $r^2 \in [r_2^2, r_3^2]$. Starting from the linear system defined in Eq. (A15), by plugging in the specific functional form of the spline mesh defined in Eq. (A4) and Eqs. (A10a-d) for its first derivatives with respect to each of the M spline parameters, we have:

$$\begin{pmatrix}
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & A \\
 0 & 0 & C \\
 0 & A & B \\
 0 & C & D \\
 0 & B & 0 \\
 0 & D & 0 \\
 A & 0 & 0 \\
 C & 0 & 0 \\
 B & 0 & 0 \\
 D & 0 & 0
 \end{pmatrix}
 \begin{pmatrix}
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A & C & B & D \\
 0 & 0 & 0 & 0 & A & C & B & D & 0 & 0 & 0 & 0 \\
 0 & 0 & A & C & B & D & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{pmatrix}
 \begin{pmatrix}
 f_1^1 \\
 f_1^{1'} \\
 f_2^1 \\
 f_2^{1'} \\
 f_3^1 \\
 f_3^{1'} \\
 f_4^1 \\
 f_4^{1'} \\
 f_5^1 \\
 f_5^{1'} \\
 f_6^1 \\
 f_6^{1'} \\
 f_1^2 \\
 f_1^{2'} \\
 f_2^2 \\
 f_2^{2'} \\
 f_3^2 \\
 f_3^{2'} \\
 f_4^2 \\
 f_4^{2'}
 \end{pmatrix}
 =
 \begin{pmatrix}
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & A \\
 0 & 0 & C \\
 0 & A & B \\
 0 & C & D \\
 0 & B & 0 \\
 0 & D & 0 \\
 A & 0 & 0 \\
 B & 0 & 0 \\
 D & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & 0 \\
 0 & 0 & 0
 \end{pmatrix}
 \begin{pmatrix}
 F_1^{\text{LAJMM}} - F_1^{\text{LSEMM}} \\
 F_2^{\text{LAJMM}} - F_2^{\text{LSEMM}} \\
 F_3^{\text{LAJMM}} - F_3^{\text{LSEMM}} \\
 F_1^{\text{2AJMM}} - F_1^{\text{2SEMM}} \\
 F_2^{\text{2AJMM}} - F_2^{\text{2SEMM}} \\
 F_3^{\text{2AJMM}} - F_3^{\text{2SEMM}}
 \end{pmatrix}
 \quad (\text{S4})$$

where the 20-dimensional column vector \mathbf{f} on the left-hand-side of Eq. (S4) denotes the spline parameters to be determined and the six-dimensional column vector $\Delta\mathbf{F}$ on the right-hand-side of Eq. (S4) denotes the internal force corrections needed for each of the two CVs in the three sampled configurations. Note that in Eq. (S4), although the specific indices on the variables A , B , C , and D are omitted for brevity, their values vary with the row/column locations in the related matrices and should be computed using Eqs. (A6a-d) by plugging in the actual CV values sampled and the corresponding grid points. In a more generic and compact form, Eq. (S4) can be written in terms of the dimensions of the block matrices involved:

$$\begin{pmatrix} m_1 \times L & \mathbf{0} \\ \mathbf{0} & m_2 \times L \end{pmatrix} \begin{pmatrix} L \times m_1 & \mathbf{0} \\ \mathbf{0} & L \times m_2 \end{pmatrix} \begin{pmatrix} m_1 \times 1 \\ m_2 \times 1 \end{pmatrix} = \begin{pmatrix} m_1 \times L & \mathbf{0} \\ \mathbf{0} & m_2 \times L \end{pmatrix} \begin{pmatrix} L \times 1 \\ L \times 1 \end{pmatrix} \quad (\text{S5})$$

for which consistency in dimensionality for the related matrix operations, as presented in Eq. (A17), can be verified:

$$(M \times 2L)(2L \times M)(M \times 1) = (M \times 2L)(2L \times 1) = M \times 1 \quad (\text{S6})$$

3. Implementation of internal force correction defined by Yang and co-workers (Wu *et al.* JCP 2017, 147, 161732)

To compare the internal forces defined by us through coordinate transformation with those defined by Yang and co-workers,² we implemented the internal forces consistent with their definition. Because Wu *et al.*² only gave the expression of the internal force corrections instead of the internal forces themselves, we first derive the internal force expressions at both the base (SE/MM) and target (AI/MM) levels following their definition of the internal force corrections.

Adopting Eqs. (A6) & (A7) of Wu *et al.*² (JCP 2017, 147, 161732), we have:

$$\frac{[\mathbf{F}_C^H(t) - \mathbf{F}_C^L(t)] \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{C(Cl)}^{\text{corr}}(t) - F_{C(N)}^{\text{corr}}(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{[\mathbf{F}_{Cl}^H(t) - \mathbf{F}_{Cl}^L(t)] \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{Cl}^{\text{corr}}(t)}{m_{Cl}} = 0 \quad (\text{S7})$$

$$\frac{[\mathbf{F}_C^H(t) - \mathbf{F}_C^L(t)] \cdot \hat{\mathbf{r}}_{N-C}(t) - F_{C(N)}^{\text{corr}}(t) - F_{C(Cl)}^{\text{corr}}(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{[\mathbf{F}_N^H(t) - \mathbf{F}_N^L(t)] \cdot \hat{\mathbf{r}}_{N-C}(t) - F_N^{\text{corr}}(t)}{m_N} = 0 \quad (\text{S8})$$

where \mathbf{F} , F , $\hat{\mathbf{r}}$, and m represents the atomic Cartesian force vector, internal force, unit bond vector, and atomic mass, respectively; the detailed definitions of these variables follow the original reference of Wu *et al.* and therefore are not repeated here. Set the Cartesian forces at the low (L) level $\mathbf{F}_C^L(t)$, $\mathbf{F}_{Cl}^L(t)$, and $\mathbf{F}_N^L(t)$ in Eqs. (S7) & (S8) to zeros, then the related internal force corrections $F_{C(Cl)}^{\text{corr}}(t)$, $F_{C(N)}^{\text{corr}}(t)$, $F_{Cl}^{\text{corr}}(t)$, and $F_N^{\text{corr}}(t)$ become the corresponding internal forces at the high (H) level:

$$\frac{\mathbf{F}_C^H(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{C(Cl)}^H(t) - F_{C(N)}^H(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_{Cl}^H(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{Cl}^H(t)}{m_{Cl}} = 0 \quad (\text{S9})$$

$$\frac{\mathbf{F}_C^H(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - F_{C(N)}^H(t) - F_{C(Cl)}^H(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_N^H(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - F_N^H(t)}{m_N} = 0 \quad (\text{S10})$$

The low-level (L) internal forces can be obtained in the same way:

$$\frac{\mathbf{F}_C^L(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{C(Cl)}^L(t) - F_{C(N)}^L(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_{Cl}^L(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{Cl}^L(t)}{m_{Cl}} = 0 \quad (\text{S11})$$

$$\frac{\mathbf{F}_C^L(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - F_{C(N)}^L(t) - F_{C(Cl)}^L(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_N^L(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - F_N^L(t)}{m_N} = 0 \quad (\text{S12})$$

Or more generally, we can omit the level-specific labels from Eqs. (S9-S12) and obtain:

$$\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{C(Cl)}(t) - F_{C(N)}(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_{Cl}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{Cl}(t)}{m_{Cl}} = 0 \quad (S13)$$

$$\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - F_{C(N)}(t) - F_{C(Cl)}(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_N(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - F_N(t)}{m_N} = 0 \quad (S14)$$

Using the identities $F_{Cl}(t) = -F_{C(Cl)}(t)$ and $F_N(t) = -F_{C(N)}(t)$ resulting from Eqs. (A2) & (A3) in Wu *et al.*,² we re-write Eqs. (S13) & (S14) as,

$$\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - F_{C(Cl)}(t) - F_{C(N)}(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_{Cl}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) + F_{C(Cl)}(t)}{m_{Cl}} = 0 \quad (S15)$$

$$\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - F_{C(N)}(t) - F_{C(Cl)}(t) [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_N(t) \cdot \hat{\mathbf{r}}_{N-C}(t) + F_{C(N)}(t)}{m_N} = 0 \quad (S16)$$

We use Eqs. (S15) & (S16) as the working equations to solve the two unknown internal forces $F_{C(Cl)}(t)$ and $F_{C(N)}(t)$. If we set $F_{C(Cl)}(t) = x$ and $F_{C(N)}(t) = y$, Eqs. (S15) & (S16) become:

$$\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) - x - y [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_{Cl}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t) + x}{m_{Cl}} = 0 \quad (S17)$$

$$\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{N-C}(t) - y - x [\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} - \frac{\mathbf{F}_N(t) \cdot \hat{\mathbf{r}}_{N-C}(t) + y}{m_N} = 0 \quad (S18)$$

Solving Eqs. (S17) & (S18) by elimination and substitution, we have:

$$F_{C(Cl)}(t) = x = \frac{\left[\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)}{m_C} - \frac{\mathbf{F}_{Cl}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)}{m_{Cl}} \right] \left(\frac{1}{m_C} + \frac{1}{m_N} \right) - \left[\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{N-C}(t)}{m_C} - \frac{\mathbf{F}_N(t) \cdot \hat{\mathbf{r}}_{N-C}(t)}{m_N} \right] \left[\frac{\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)}{m_C} \right]}{\left(\frac{1}{m_C} + \frac{1}{m_{Cl}} \right) \left(\frac{1}{m_C} + \frac{1}{m_N} \right) - \left(\frac{[\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} \right)^2}} \quad (S19)$$

$$F_{C(N)}(t) = y = \frac{\left[\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)}{m_C} - \frac{\mathbf{F}_{Cl}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)}{m_{Cl}} \right] \left[\frac{\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)}{m_C} \right] - \left[\frac{\mathbf{F}_C(t) \cdot \hat{\mathbf{r}}_{N-C}(t)}{m_C} - \frac{\mathbf{F}_N(t) \cdot \hat{\mathbf{r}}_{N-C}(t)}{m_N} \right] \left(\frac{1}{m_C} + \frac{1}{m_{Cl}} \right)}{\left(\frac{[\hat{\mathbf{r}}_{N-C}(t) \cdot \hat{\mathbf{r}}_{Cl-C}(t)]}{m_C} \right)^2 - \left(\frac{1}{m_C} + \frac{1}{m_N} \right) \left(\frac{1}{m_C} + \frac{1}{m_{Cl}} \right)}} \quad (S20)$$

4. Comparison of the internal forces defined by Yang and co-workers (Wu *et al.* JCP 2017, 147, 161732) and those defined by us through redundant internal coordinate transformation

In this section, we compare the internal forces obtained by following Wu *et al.*'s definition [see Supporting Information Sec. 3 (SI.3) above] and our own definition through redundant internal coordinate transformation, which is used in the RP-FM-CV formalism. Note that based on Eqs. (S19) & (S20), which are derived from Wu *et al.*'s definition, the internal forces on the broken and formed bonds in the Menshutkin reaction depend only on the Cartesian forces of the three atoms involved in the two CV bonds. By contrast, our definition of the internal forces on the CVs is significantly different from theirs in this respect. In RP-FM-CV, we obtain the internal forces on the CVs by using redundant internal coordinate transformation, where the determination of internal forces on the two CV bonds involves not only the CVs but also other internal degrees of freedom in the redundant coordinate system. The use of this transformation makes the CV forces non-trivially dependent on the Cartesian forces from the non-CV atoms. In other words, our treatment distributes the Cartesian forces to a specific internal degree of freedom by explicitly considering its coupling to all other internal coordinates, which by construction jointly form a non-orthogonal coordinate set; the coupling between the CV and non-CV internal degrees of freedom would otherwise be ignored if the internal forces on the CVs only depend on the Cartesian forces on the CV atoms. For simple one-bond CV cases, the internal forces resulting from the two formalisms are similar, but differ by a few mass factors for a non-homonuclear bond, which can be traced back to the different philosophies of using a trajectory/position matching strategy by Wu *et al.*'s Eqs. (2-3) and our directly using the force matching strategy; for a single homonuclear diatomic bond, both converge to the projection operator formalism,³ which determines the internal force by projecting the total Cartesian force along the bond vector (see SI.5). For cases using multidimensional CVs, however, neglecting the couplings between the CVs and other internal degrees of freedom, which are needed for completing the coordinate system, could lead to qualitatively different internal forces.

To examine the consequences of internal force evaluation under these different definitions, we computed the internal forces on the two CV bonds in the Menshutkin reaction, i.e., the N-C and C-Cl bonds, by following Wu *et al.*'s definition [i.e., using Eqs. (S19) and (S20)], and compared them with our RP-FM-CV internal forces determined from the redundant internal coordinate transformation using our default set of 28 internal coordinates (referred to as "Int28"), whose definition can be found in Figure 6 in the text. To test our hypothesis that ignoring the couplings between the CV and non-CV internal degrees of freedom would lead to an incomplete coordinate system, we also computed the internal forces from the Cartesian-to-internal coordinate transformation by only including the two CV bonds in the Wilson **B**-matrix, with which the coupling between the CVs and the rest of the system is omitted; we refer to this incomplete coordinate system that uses only two internal coordinates as "Int2".

The internal forces on the two CV bonds using Int28, Int2, and Wu *et al.*'s definition based on Eqs. (S19) & (S20) (data referred to as "Yang'17"), are compared in Figure S1, where we show the internal forces at both the AM1/MM and MP2/6-31+G(d,p)/MM levels. From Figure S1, we can see that when only two internal coordinates are included in coordinate transformation, Int2 yields very similar internal forces to Yang'17, where both show very large fluctuations at a given bond distance. We suspect that these spurious fluctuations of the internal forces are perhaps caused by force contaminations from other non-CV degrees of freedom, which cannot be removed properly when their couplings to the CVs are neglected. By contrast, as long as the coordinate system is complete (e.g., using the Int28 redundant internal coordinate set), the resulting internal forces on the CVs are found to be smooth along each bond (and therefore they also evolve smoothly along the string MFEP) and display much smaller fluctuations, which seem to be more physical. In the bond-dissociation regions, the internal forces along the two CV bonds using Int28 gracefully attenuate to zeros at both the AM1/MM and MP2/6-31+G(d,p)/MM levels, but with Int2 and Yang'17, significant residual forces as well as large fluctuations are still observed even when these chemical bonds are fully broken.

The force correction terms for matching the internal forces between the AM1/MM and MP2/6-31+G(d,p)/MM levels using these three schemes (i.e., Int28, Int2, and Yang'17) are compared in Figure S2; the corresponding spline-function-based fits (as described in *Appendix A*) are also shown in the same figure for comparison. As the internal forces tend to display large fluctuations when the coordinate system is incomplete (e.g., Int2 and Yang'17), the internal force corrections also display much greater fluctuations, which not only makes the spline fits less accurate, but also generates quantitatively different internal force corrections. For example, the internal force correction for the N-C bond peaks at 0 kcal/mol/Å when Int28 is used, whereas Int2 and Yang'17 predict corrections as large as +30 kcal/mol/Å. A similar discrepancy is also observed for the C-Cl bond in its fully dissociated region (~ 3.5 Å and beyond). For Int28, the correction on the internal force in this region is close to zero, which indicates that both AM1/MM and MP2/6-31+G(d,p)/MM give a flat bond-stretching potential when the C-Cl bond is fully broken. By contrast, under the Int2 and Yang'17 schemes, the force deviations between the AM1/MM and MP2/6-31+G(d,p)/MM in the $r_{\text{C-Cl}} > 3.5$ Å region are still as large as 20 kcal/mol/Å; the large force deviations in this region imply that the two quantum mechanical methods, although both well-established, fail to consistently predict a flat potential at the dissociation limit, which seems rather unlikely.

Figure S1. Internal forces on the two CV bonds computed at the AM1/MM (squares) and MP2/6-31+G(d,p)/MM (triangles; denoted as AI/MM) levels using redundant internal coordinate transformation with 28 internal (Int28) and 2 internal (Int2) coordinates, compared with the internal forces following Wu *et al.*'s definition (Yang'17).

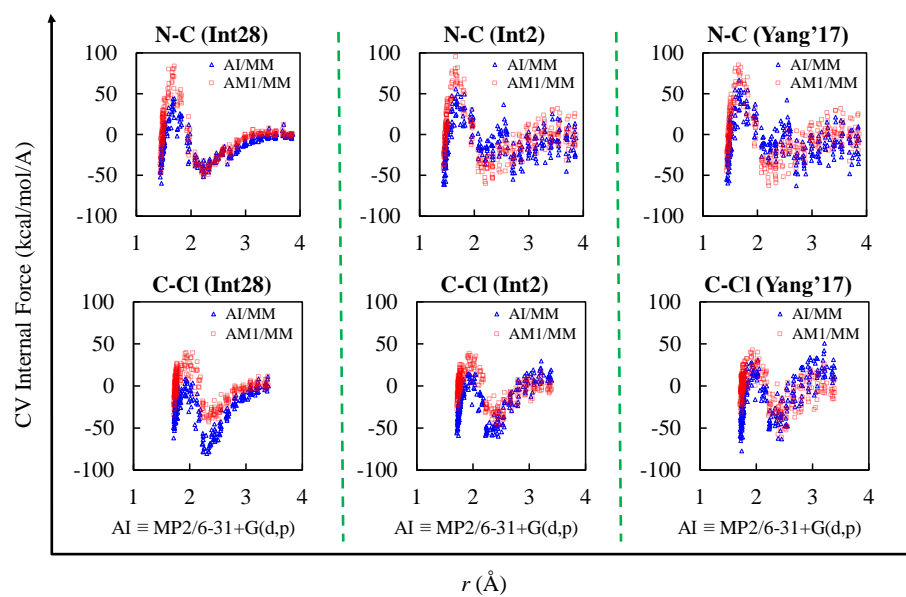
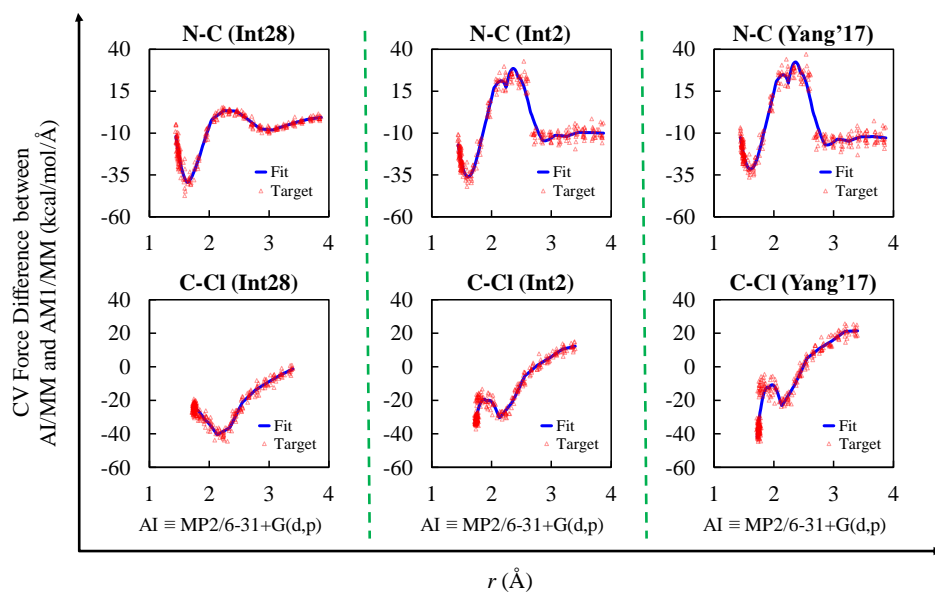


Figure S2. Internal force corrections on the two CV bonds for AM1/MM to match with MP2/6-31+G(d,p)/MM using redundant coordinate transformation with 28 internal coordinate (Int28) and 2 internal coordinates (Int2), compared with those using Wu *et al.*'s definition (Yang'17). The actual differences in internal forces between the target and AM1/MM levels are shown as open triangles (referred to as "Target"), whereas their spline-based fits (as described in *Appendix A*) (referred to as "Fit") are shown as solid lines.



5. Internal forces obtained for a one-dimensional system using Wu *et al.* (JCP 2017, 147, 161732) and our RP-FM-CV coordinate transformation, and their relations to the projection operator formalism for this case.

Although Wu *et al.*'s internal forces and ours obtained using redundant internal coordinate transformation do differ theoretically and numerically for cases that involve non-orthogonal multidimensional CVs (e.g. in the Menshutkin reaction; see our detailed discussion in SI.3-SI.4), the two internal force schemes become more connected to each other for simpler systems. In this section, we show the similarities between the two schemes when they are applied to a one-dimensional (1-d) internal coordinate case (i.e., only a single bond is included in the coordinate system). In particular, we found that for this 1-d case, Wu *et al.*'s scheme conditionally converges to our coordinate transformation scheme as well as to the projection operator formalism. We start our derivation from Eq. (7) of Wu *et al.* (JCP 2017, 147, 161732),² which is for the internal force correction of a single bond between atom i and j :

$$\frac{[\mathbf{F}_j^H(t) - \mathbf{F}_j^L(t)] \cdot \hat{\mathbf{r}}_{ij}(t) - F_j^{\text{corr}}(t)}{m_j} - \frac{[\mathbf{F}_i^H(t) - \mathbf{F}_i^L(t)] \cdot \hat{\mathbf{r}}_{ij}(t) - F_i^{\text{corr}}(t)}{m_i} = 0 \quad (\text{S21})$$

where \mathbf{F} , F , $\hat{\mathbf{r}}$, and m represents the atomic Cartesian force vector, internal force, unit bond vector from i to j [i.e., $\hat{\mathbf{r}}_{ij}(t) = \frac{\mathbf{r}_j - \mathbf{r}_i}{|\mathbf{r}_j - \mathbf{r}_i|}$], and atomic mass, respectively; again, the detailed

notations can be found in the original reference of Wu *et al.* and are not repeated here. Let the Cartesian forces at the low (L) level $\mathbf{F}_j^L(t)$ and $\mathbf{F}_i^L(t)$ be zeros in Eq. (S21), then the correcting (corr) forces between the two levels become the internal forces at the high (H) level:

$$\frac{\mathbf{F}_j^H(t) \cdot \hat{\mathbf{r}}_{ij}(t) - F_j^H(t)}{m_j} - \frac{\mathbf{F}_i^H(t) \cdot \hat{\mathbf{r}}_{ij}(t) - F_i^H(t)}{m_i} = 0 \quad (\text{S22})$$

Using the same strategy, we obtain the low (L) level internal forces:

$$\frac{\mathbf{F}_j^L(t) \cdot \hat{\mathbf{r}}_{ij}(t) - F_j^L(t)}{m_j} - \frac{\mathbf{F}_i^L(t) \cdot \hat{\mathbf{r}}_{ij}(t) - F_i^L(t)}{m_i} = 0 \quad (\text{S23})$$

After omitting the level-specific labels, we obtain a general expression:

$$\frac{\mathbf{F}_j(t) \cdot \hat{\mathbf{r}}_{ij}(t) - F_j(t)}{m_j} - \frac{\mathbf{F}_i(t) \cdot \hat{\mathbf{r}}_{ij}(t) - F_i(t)}{m_i} = 0 \quad (\text{S24})$$

Using the identity $F_i(t) = -F_j(t)$ resulting from Eq. (5) of Wu *et al.*, we have:

$$\frac{\mathbf{F}_j(t) \cdot \hat{\mathbf{r}}_{ij}(t) - F_j(t)}{m_j} - \frac{\mathbf{F}_i(t) \cdot \hat{\mathbf{r}}_{ij}(t) + F_j(t)}{m_i} = 0 \quad (\text{S25})$$

Solving Eq. (S25) for $F_j(t)$, we have:

$$F_j(t) = \frac{\frac{\mathbf{F}_j(t) \cdot \hat{\mathbf{r}}_{ij}(t)}{m_j} - \frac{\mathbf{F}_i(t) \cdot \hat{\mathbf{r}}_{ij}(t)}{m_i}}{\frac{1}{m_j} + \frac{1}{m_i}} = \frac{m_i \mathbf{F}_j(t) \cdot \hat{\mathbf{r}}_{ij}(t) - m_j \mathbf{F}_i(t) \cdot \hat{\mathbf{r}}_{ij}(t)}{m_i + m_j} \quad (\text{S26})$$

Note that the internal force F obtained from Eq. (S26) is mass dependent, which makes it resemble but not exact same as the projected force using the projection operator formalism (e.g., using Lu, Zhao & Truhlar³). By contrast, for a one-dimensional bond case, the **B**-matrix coordinate transformation scheme used in RP-FM-CV is equivalent to the projection operator formalism and both are mass independent. In terms of mass dependency of the internal force, Eq. (S26) can be rearranged as if one transforms the Cartesian to internal forces in the corresponding mass-scaled coordinate systems [note that the mass scaling factor here resembles, but differs from the mass-weighted treatment in Fukui's intrinsic reaction coordinate (IRC),⁴ where each Cartesian coordinate is weighted by a factor of \sqrt{m} instead of the atomic mass m itself]:

$$\frac{F_j(t)}{\mu_{ij}} = \frac{\mathbf{F}_j(t)}{m_j} \cdot \hat{\mathbf{r}}_{ij}(t) - \frac{\mathbf{F}_i(t)}{m_i} \cdot \hat{\mathbf{r}}_{ij}(t) \quad (\text{S27})$$

where μ_{ij} is the reduced mass for the bond ij ,

$$\mu_{ij} = \frac{m_i m_j}{m_i + m_j} \quad (\text{S28})$$

For a homonuclear diatomic bond, where $m_i = m_j$ and $\mu_{ij} = \frac{1}{2} m_i$, after mass cancellation in

Eq. (S28), Wu *et al.*'s formalism leads to an internal force expression,

$$F_j(t) = \frac{1}{2} [\mathbf{F}_j(t) \cdot \hat{\mathbf{r}}_{ij}(t) - \mathbf{F}_i(t) \cdot \hat{\mathbf{r}}_{ij}(t)] \quad (\text{S29})$$

which is *identical to* the results using the projection operator and the B-matrix based coordinate transformation (as shown below).

Using the notations compatible with Lu *et al.*,³ the internal force F using the projection operator formalism is:

$$F = |\mathbf{e}_{ij}| (\hat{\mathbf{e}}_{ij})^T \cdot \mathbf{f} = \frac{1}{\sqrt{2}} (\hat{\mathbf{e}}_{ij})^T \cdot \mathbf{f} \quad (\text{S30})$$

where \mathbf{f} is the Cartesian force vector for the system, and \mathbf{e}_{ij} is a column vector corresponding to the bond entry in the \mathbf{B} -matrix with $\hat{\mathbf{e}}_{ij}$ being its unit vector after normalization:

$$\mathbf{e}_{ij} = \left(0, \dots, \frac{x_i - x_j}{r_{ij}}, \frac{y_i - y_j}{r_{ij}}, \frac{z_i - z_j}{r_{ij}}, 0, \dots, \frac{x_j - x_i}{r_{ij}}, \frac{y_j - y_i}{r_{ij}}, \frac{z_j - z_i}{r_{ij}}, 0, \dots \right)^T \quad (\text{S31})$$

Using the notations consistent with *Appendix B* in the text, the internal force F obtained directly from the \mathbf{B} -matrix-based coordinate transformation [a non-redundant transformation in this case though; see Eqs. (12) & (14) in Jackels *et al.*,⁵ for example, for the related formula] is:

$$F = \left[\mathbf{u} \mathbf{B}^T (\mathbf{B} \mathbf{u} \mathbf{B}^T)^{-1} \right]^T \mathbf{f} = \left(\mathbf{B}^T \frac{1}{2} \right)^T \mathbf{f} = \frac{1}{2} \mathbf{B} \mathbf{f} \quad (\text{S32})$$

where \mathbf{u} and \mathbf{f} is a unit matrix and the Cartesian force vector for the system, respectively.

Plug Eq. (S31) into both Eq. (S30) and the \mathbf{B} -matrix in Eq. (S32), and equate \mathbf{f} in Eqs. (S30) & (S32) to \mathbf{F} in Eq. (S29) (for both represent the Cartesian forces, although using different symbols in the related literature), it is straightforward to show that the three formalisms, namely, Wu *et al.* [Eq. (S29), which holds for a homonuclear diatomic bond], the projection operator [Eq. (S30)], and the \mathbf{B} -matrix coordinate transformation [Eq. (S32)], generate the identical internal force. Note that for the 1-d bond CV case, the projection operator and the \mathbf{B} -matrix-based coordinate transformation are always equivalent, whereas the convergence of Wu *et al.*'s formalism to them is conditional when the related mass factors in Eq. (S27) cancel out.

Note that for a polyatomic molecular system, the conclusion of the above discussion is still valid as long as only a single bond is included in the 1-d internal coordinate system, for which cancellation of mass factors for Wu *et al.*'s scheme only applied to homonuclear diatomic bonds.

Another special condition for the mass factors in Eq. (S27) to completely cancel out is that if $\mathbf{F}_j = -\mathbf{F}_i$, e.g., for a system that contains *only two* atoms without a total net force, where the zero net force on the bond would make the Cartesian forces \mathbf{F} on the two atoms also cancel out. Note that this condition is not generally satisfied if one has a polyatomic system where the Cartesian forces on the two bonded atoms do not necessarily cancel out even if the total net force experienced by the whole molecular is zero. Under the two-atom and zero-net-force condition, we found that the mass factors cancel out regardless of the mass combinations.

For example, for a 1-d Harmonic oscillator consisting of two particles, the mass factors in Wu *et al.*'s Eq. (7) [or Eq. (S21) here] would eventually cancel out to make them irrelevant for solving the internal force. As we show below, for a Harmonic oscillator, the evaluation of internal force using Wu *et al.*'s Eq. (7) [or Eq. (S21) here] is equivalent to the projected force using a projection operator³ and to the transformed force using the \mathbf{B} -matrix formalism, both of which are mass independent.

For a two-atom molecular system, Eq. (S26) becomes:

$$F_2(t) = \frac{m_1 \mathbf{F}_2(t) \cdot \hat{\mathbf{r}}_{12}(t) - m_2 \mathbf{F}_1(t) \cdot \hat{\mathbf{r}}_{12}(t)}{m_1 + m_2} \quad (\text{S33})$$

If the total system experiences a zero net force, the Cartesian forces on the two atoms cancel out,

$$\mathbf{F}_1 = -\mathbf{F}_2 \quad (\text{S34})$$

Plugging Eq. (34) into Eq. (33) leads to an internal force independent of mass factors:

$$F_2(t) = \frac{m_1 \mathbf{F}_2(t) \cdot \hat{\mathbf{r}}_{12}(t) + m_2 \mathbf{F}_2(t) \cdot \hat{\mathbf{r}}_{12}(t)}{m_1 + m_2} = \mathbf{F}_2(t) \cdot \hat{\mathbf{r}}_{12}(t) \quad (\text{S35})$$

which turns out to be a special case of Eq. (S29) when Eq. (S34) holds for zero net force. Given the equivalence of Eq. (S29), Eq. (S30), and Eq. (S32), the internal force evaluated for a Harmonic oscillator using the three schemes would be identical and all mass independent.

6. Overall convergence of iterative RP-FM-CV at the B3LYP:AM1/MM level for the Menshutkin reaction

In addition to the test based on MP2:AM1/MM (see Sec. 5.8 in the text), the B3LYP:AM1/MM method was also used to examine the overall convergence behavior of the procedure when RP-FM-CV is performed iteratively for multiple cycles. We followed the same simulation protocols described in Sec. 5.8, by repeating five cycles of RP-FM-CV each consisting of ten iterations of string MFEP optimization and FM. The free energy results and key bond distances in TS from the B3LYP:AM1/MM simulations are given in Table S3. We can see that throughout the five cycles of RP-FM-CV simulations the free energy barriers and reaction free energies obtained at the B3LYP:AM1/MM level display small fluctuations (< 1.0 kcal/mol) about their average values at 13.5 ± 0.7 and -26.3 ± 0.5 kcal/mol, respectively. The N-C and C-Cl bond distances at the free energy TS throughout the five cycles also fluctuate closely about their average values of 2.235 ± 0.013 and 2.231 ± 0.023 Å, respectively. The overall convergence behavior is very similar to what we observed from the MP2:AM1/MM simulations.

Table S3. Computed free energy barriers (ΔG^\ddagger), reaction free energies (ΔG_r), and transition state geometries for the Menshutkin reaction between NH_3 and CH_3Cl in water over five cycles of RP-FM-CV simulations at the B3LYP^a:AM1/MM level.

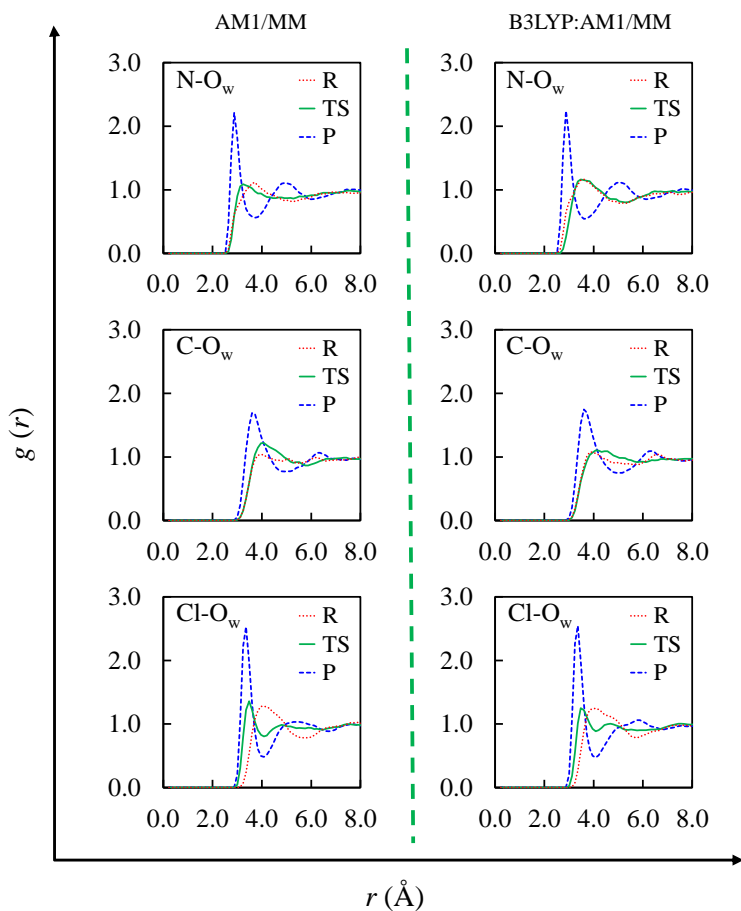
Cycle	ΔG^\ddagger (kcal/mol)	ΔG_r (kcal/mol)	N-C (Å)	C-Cl (Å)
1	14.7	-27.2	2.213	2.194
2	13.5	-26.6	2.242	2.224
3	13.1	-26.2	2.244	2.238
4	13.1	-25.9	2.239	2.247
5	12.9	-25.8	2.239	2.253
Average	13.5 ± 0.7	-26.3 ± 0.5	2.235 ± 0.013	2.231 ± 0.023

^ausing the default 6-31+G(d,p) basis set; see Sec. 5 in the text for the related notation.

7. Comparison of solute-solvent pair RDFs from the AM1/MM and B3LYP:AM1/MM RP-FM-CV simulations

The radial distribution functions (RDFs) for the selected solute-solvent atom pairs (see Sec. 5.9 in the text for description) were also computed at the B3LYP:AM1/MM level in the reactant (R), transition state (TS), and product (P) regions along the MFEP. Specifically, N-O_w, C-O_w, and Cl-O_w RDFs are plotted against the original AM1/MM results in Figure S3. Compared with the MP2:AM1/MM RDFs in Sec.5.9, a similar trend in peak height and position is found at the B3LYP:AM1/MM level, which confirms the corresponding observation that the physical description of solvent from the AM1/MM simulation is preserved in the RP-FM-CV simulations.

Figure S3. Solute-solvent radial distribution functions (RDFs) obtained from the RP-FM-CV simulations at the B3LYP:AM1/MM level using the 6-31+G(d,p) basis set, compared with the AM1/MM results.



8. Consistent AI/MM benchmark results

For the AM1/MM and AI:AM1/MM RP-FM-CV free energy path simulations, we conducted a moderate amount of sampling (20 ps sampling per image for the free energy force evaluations in each iteration of string MFEP optimization), which was combined with ten iterations of MFEP optimization and up to five cycles of RP updates and FM. Under RP-FM-CV, this combination has proved to give good convergence on various aspects of the free energy results (see Sec. 5.8 in the text). However, direct application of the same sampling and iterative protocols to the AI/MM MFEP simulations is computationally very demanding, which makes the corresponding benchmark impractical to obtain. Without an easy access to the benchmark AI/MM free energy results obtained in a consistent manner, we focused on validating the RP-FM-CV method against experiments and cross-validating it by varying the choice of AI method (B3LYP, BH&HLYP, and MP2) in the AI:AM1/MM simulations. Although a few AI/MM free energy simulation results are available in the literature (see Table 1 in the text), their uses of different AI levels, simulation protocols, and computer codes make it difficult to directly compare these literature results with our RP-FM-CV simulations. Therefore, we decided to obtain our own AI/MM free energy benchmark, which is made affordable through a relatively lighter computational/sampling requirement, so that we can validate our RP-FM-CV simulations against the consistent benchmark under the same simulation conditions.

For our AI/MM benchmark string simulations, we applied the B3LYP/6-31G(d)/MM method to the Menshutkin reaction. To save the computational cost, the relatively small 6-31G(d) basis set is used and free energy force sampling is shortened from 20 ps to 1 ps. To further alleviate the computational demand, a cutoff treatment is used to handle the non-bonded electrostatic interactions by using a cutoff distance of 14 Å. For the setup of the QM/MM system under periodic boundary conditions and other simulation parameters, we followed the simulation protocol detailed in Sec. 4 in the text.

To obtain a fair comparison with the AI/MM results, we performed the AM1/MM and RP-FM-CV simulations under the same setup and sampling duration (i.e., 1 ps). The benchmark B3LYP/6-31G(d)/MM MFEP, initiated from the path obtained by the 10th iteration of AM1/MM string simulations, is optimized for another 30 iterations of path optimization. Similarly, for the string MFEP simulations using RP-FM-CV at the B3LYP/6-31G(d):AM1/MM level, the same initial path is used, and we carried out 30 iterations of path optimization consistently for each cycle of FM. Including the 10th iteration of the AM1/MM string path, three cycles of RP updates are conducted for a total of 100 iterations, where a training set of 300 sampled configurations from the last three iterations of each FM cycle is used for force fitting. Since a shorter sampling time is accompanied with increased instability for string path optimization and FM, we employed a spline under tension technique,⁶ where the force correction is fitted along a scalar variable converted from the physical CV bond distances (Kim and Pu, unpublished). Along with the 30th iteration results from the 1 ps AM1/MM simulation, free energy profiles obtained from the B3LYP/6-31G(d):AM1/MM and benchmark B3LYP/6-31G(d)/MM string simulations are shown

in Figure S4, where the numerical values for the free energy barrier, reaction free energy, and CV bond distances in the free energy TS from these simulations are compared in Table S4.

Table S4. Free energy barriers (ΔG^\ddagger), reaction free energies (ΔG_r), and transition-state geometries for the Menshutkin reaction between NH_3 and CH_3Cl in water, computed by AM1/MM and RP-FM-CV at the B3LYP/6-31G(d)/MM level, compared with the B3LYP/6-31G(d)/MM benchmark results and with experiments

Method	ΔG^\ddagger (kcal/mol)	ΔG_r (kcal/mol)	N-C (Å)	C-Cl (Å)
AM1/MM ^a	40.2	-2.3	2.074	2.303
B3LYP/6-31G(d)/MM ^a	18.8	-29.4	2.295	2.133
B3LYP/6-31G(d):AM1/MM ^{a,b}	18.8	-29.4	2.246	2.249
Experiment	23.5 ^c	-34 ± 10^c -36 ± 6^d	-	-

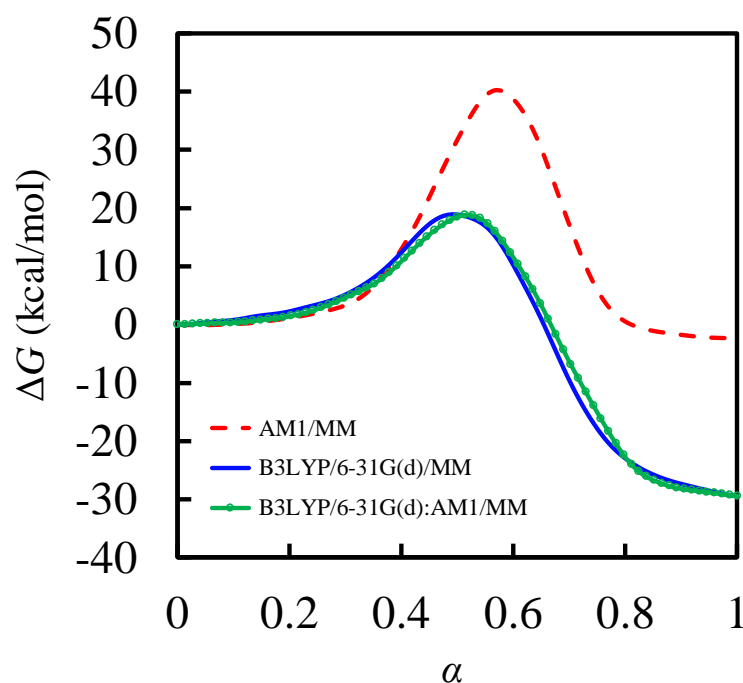
^afree energy forces are evaluated based on 1 ps sampling

^bRP-FM-CV with spline under tension

^cfrom Gao *JACS* **1991**, 113, 7796⁷

^dfrom Vilseck *et al. J. Comput. Chem.* **2011**, 32, 2836⁸

Figure S4. The benchmark B3LYP/6-31G(d)/MM free energy profile (solid line), compared with the AM1/MM free energy profile (dashed line) and the profile obtained from RP-FM-CV simulations at the B3LYP/6-31G(d):AM1/MM level (solid line with open circles). For comparison, all the string MFEP simulations at these levels were done consistently using 1 ps sampling for free energy force evaluation. Both the B3LYP/6-31G(d)/MM benchmark and semiempirical AM1/MM free energy profiles were obtained based on the corresponding string MFEPs optimized after 30 iterations, whereas the B3LYP/6-31G(d):AM1/MM free energy profile was obtained after three cycles of RP updates from the AM1/MM level in the presence of FM-derived CV-force corrections. In FM, 300 configurations from the last three iterations of AM1/MM (or CV-force corrected AM1/MM) string simulations were used to fit the B3LYP/6-31G(d)/MM single-point internal forces on CVs.



From Figure S4 and Table S4, we found that with a shorter sampling time (1 ps), the original AM1/MM simulations display even greater deviations from the experimental and AM1/MM benchmark results by giving a much higher free energy barrier of 40.2 kcal/mol and an overestimated reaction free energy of -2.3 kcal/mol (compared with 30.9 and -10.6 kcal/mol from the AM1/MM simulations using 20 ps free energy force sampling). By contrast, by fitting the internal CV forces to the benchmark level, our B3LYP/6-31G(d):AM1/MM RP-FM-CV

simulations generate a free energy profile that almost overlaps with the B3LYP/6-31G(d)/MM benchmark throughout the entire reaction coordinate range (Figure S4). As a result, the RP-FM-CV simulations at the B3LYP/6-31G(d):AM1/MM RP-FM-CV level agree perfectly with the B3LYP/6-31G(d)/MM benchmark simulations in producing a free energy barrier of 18.8 kcal/mol and a reaction free energy of -29.4 kcal/mol (Table S4). Similar agreements between RP-FM-CV and the B3LYP/6-31G(d)/MM benchmark simulations can also be found in Table S4 for the key bond distances in the free energy TS.

References

- (1) Gao, J.; Xia, X. A Two-Dimensional Energy Surface for a Type II SN2 Reaction in Aqueous Solution. *J. Am. Chem. Soc.* **1993**, *115*, 9667-9675.
- (2) Wu, J.; Shen, L.; Yang, W. Internal Force Corrections with Machine Learning for Quantum Mechanics/Molecular Mechanics Simulations. *J. Chem. Phys.* **2017**, *147*, 161732.
- (3) Lu, D.-h.; Zhao, M.; Truhlar, D. G. Projection Operator Method for Geometry Optimization with Constraints. *J. Comput. Chem.* **1991**, *12*, 376-384.
- (4) Fukui, K. The Path of Chemical-Reactions - The IRC Approach. *Acc. Chem. Res.* **1981**, *14*, 363-368.
- (5) Jackels, C. F.; Gu, Z.; Truhlar, D. G. Reaction-path potential and vibrational frequencies in terms of curvilinear internal coordinates. *J. Chem. Phys.* **1995**, *102*, 3188-3201.
- (6) Ruiz-Pernia, J. J.; Silla, E.; Tunon, I.; Marti, S.; Moliner, V. Hybrid QM/MM Potentials of Mean Force with Interpolated Corrections. *J. Phys. Chem. B* **2004**, *108*, 8427-8433.
- (7) Gao, J. A Priori Computation of a Solvent-Enhanced SN2 Reaction Profile in Water: The Menshutkin Reaction. *J. Am. Chem. Soc.* **1991**, *113*, 7796-7797.
- (8) Vilseck, J. Z.; Sambasivarao, S. V.; Acevedo, O. Optimal Scaling Factors for CM1 and CM3 Atomic Charges in RM1-Based Aqueous Simulations. *J. Comput. Chem.* **2011**, *32*, 2836-2842.

B.2 Accurate Free Energy Profiles in Chemical Reactions: A QM/MM Study of the Role of Pairwise Repulsive Correcting Potentials in Force Matching

The following Supporting Information section has been adapted from a previously published article (Kim et al., "Accurate Free Energy Profiles in Chemical Reactions: A QM/MM Study of the Role of Pairwise Repulsive Correcting Potentials in Force Matching") to provide further details and supporting evidence for the main findings presented in this thesis.

Supplementary Materials

Section A.

For the Menshutkin reaction study, we used 300 configurations from MD sampling at the AM1/MM method level. For each configuration, 9 solute atoms with 3 Cartesian coordinates were used to prepare each separate row of the design matrix (B matrix; 27 rows for 1 configuration). The possible pair types (n_{type}) for the reaction were first determined with Eq. (S1).

$$n_{type} = n_{elements} \times \frac{(n_{elements} - 1)}{2} + n_{diatomic} \quad (S1)$$

where $n_{elements}$ represents all solute atoms (N, H, C and Cl) and where $n_{diatomic}$ represents any diatomic pairs (H-H). Furthermore, the atomic pairs are constructed based on the separation distance between atom i of pair type α and atom j of pair type β . As such, the 7 pairs in the Menshutkin reaction are defined as follows: (1) N-H, (2) N-C, (3) N-Cl, (4) H-H, (5) C-H, (6) C-Cl, and (7) Cl-H. In order to construct the B matrix for SVD, a type table (Figure S1) is used to map solute atoms to the respective pair type.

	N	H	H	H	C	H	H	H	Cl
N		1	1	1	2	1	1	1	3
H	1		4	4	5	4	4	4	7
H	1	4		4	5	4	4	4	7
H	1	4	4		5	4	4	4	7
C	2	5	5	5		5	5	5	6
H	1	4	4	4	5		4	4	7
H	1	4	4	4	5	4		4	7
H	1	4	4	4	5	4	4		7
Cl	3	7	7	7	6	7	7	7	

Figure S1

To construct the B matrix the following computation in Eq. (S2) is implemented for each solute atom (i), where the sign of the 2nd atom (j) is reversed to account for the reciprocal force.

$$B_{\hat{e}_i}^{\alpha\beta} = \sum_{m=1}^{n_{config}} \sum_{n=2}^{n_{prm}+1} \sum_{i=1}^{n_{amm}-1} \sum_{j=i+1}^{n_{amm}} n \times (r_c^{\alpha\beta} - r_{ij})^{n-1} \times \frac{\hat{e}_i}{r_{ij}}$$

$$B_{\hat{e}_j}^{\alpha\beta} = \sum_{m=1}^{n_{config}} \sum_{n=2}^{n_{prm}+1} \sum_{i=1}^{n_{amm}-1} \sum_{j=i+1}^{n_{amm}} -n \times (r_c^{\alpha\beta} - r_{ij})^{n-1} \times \frac{\hat{e}_j}{r_{ij}} \quad (S2)$$

Here n_{config} is the number of sampled configurations (300 configurations), n_{prm} is the number of parameters (8 for $n=2$ to 9), n_{amm} is the number of solute atoms (9 solute atoms for the Menshutkin reaction), \hat{e} is the Cartesian direction (x, y, z), where $r_c^{\alpha\beta}$ and r_{ij} are reference radial

cutoff and actual radial distances, respectively (see Table 1 for reference radial cutoff distances). The Cartesian based elements are then summed into the B matrix rows using the following index system in Eq. (S3):

$$\begin{aligned}\hat{e}_i &= 3 \times n_{\text{atom}} \times (m-1) + 3 \times (i-1) + 1 \\ \hat{e}_j &= 3 \times n_{\text{atom}} \times (m-1) + 3 \times (j-1) + 1\end{aligned}\quad (\text{S3})$$

where parameters of the elements (prm_{column}) are assigned to the corresponding columns as follows in Eq. (S4):

$$prm_{\text{column}} = n_{\text{prm}} \times (n_{\text{type}} - 1) + (n - 1) \quad (\text{S4})$$

In Figure S2, the B matrix elements for the 1st configuration of the reactant state ($m=1$) in the Menshutkin reaction is displayed with N-involved pairs (blue), H- involved pairs (white), H-H pairs (yellow), C- involved pairs (gray) and Cl- involved pairs (green) in each respective element; where black elements correspond to non-interacting atoms and where purple elements correspond to predicted zero-interactions based on the radial cutoff distance of 2 Å.

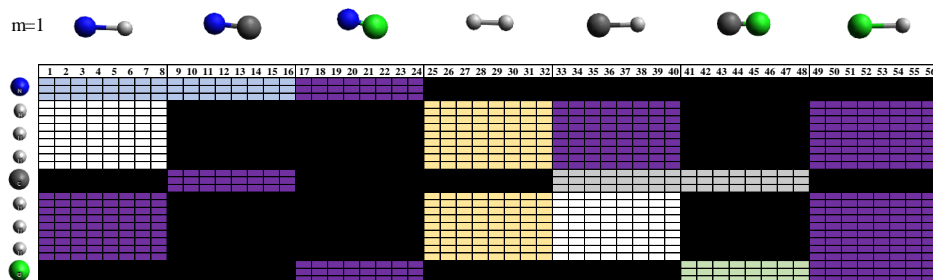


Figure S2

As such, the B matrix for 300 configurations in the Menshutkin reaction is 8100×56 where minimization of the objective function (singular value decomposition) is determined in Figure S3, using the Cartesian difference between target and base method (ΔF).

$$\chi^2 = |A \cdot c - b| = 0$$

$$A \cdot c = b$$

$$\begin{array}{c} A \\ \boxed{\begin{array}{c} B^T \\ 56 \times 8100 \end{array}} \times \boxed{\begin{array}{c} B \\ 8100 \times 56 \end{array}} \times \boxed{\begin{array}{c} c \\ 56 \times 1 \end{array}} = \boxed{\begin{array}{c} b \\ \begin{array}{c} B^T \\ 56 \times 8100 \end{array} \times \begin{array}{c} \Delta F \\ 8100 \times 1 \end{array} \end{array}}$$

Figure S3

Section B.

With a radial cutoff distance of 2 Å (Figure S4), a maximum correcting potential of 66.8 kcal/mol is applied in the MD simulation to the H-H pair at a bond distance 1.03 Å (Figure S4b). The large correcting potential for the H-H pair (orange) is obtained from overfitting, which causes the simulation to fail since the system is unable to achieve self-consistency within specified tolerances.

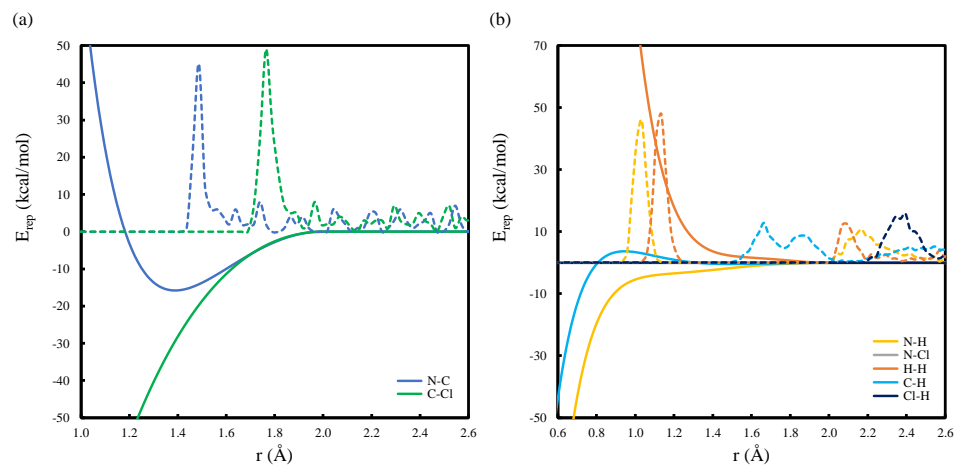


Figure S4

Section C.

In Figure S5, the radial cutoff distance is elongated to 10 Å for CVs to extend the range of sampling. The extension in radial cutoff distance results in an extrapolated force correction for (a) N-C and (b) C-Cl pairs.

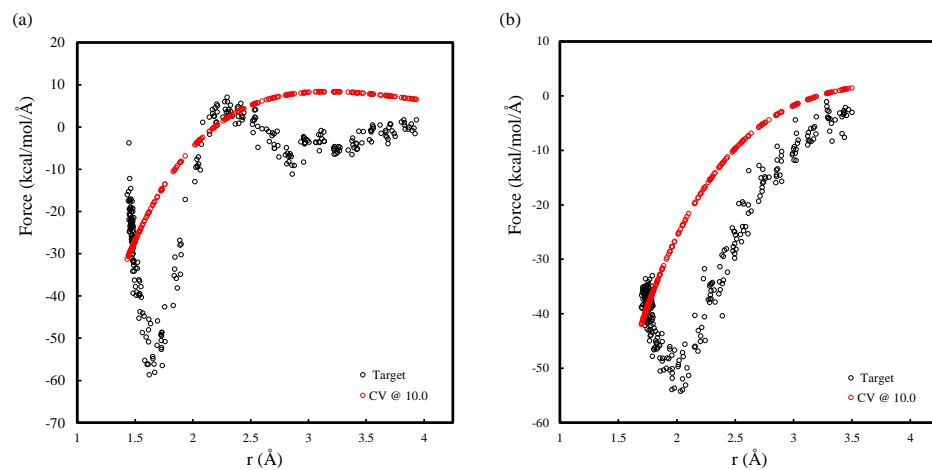


Figure S5

Section D.

A 2-D force correlation scan was performed with 0.1 Å windows as shown in Figure S6a (expanded) and Figure S6b (zoomed), to complement the genetic algorithm. In comparison to the predicted CV radial cutoff distances from the Micro-GA-CV, both scans predict the same region of optimized radial cutoff distances in CVs (Table S1 in Section E).

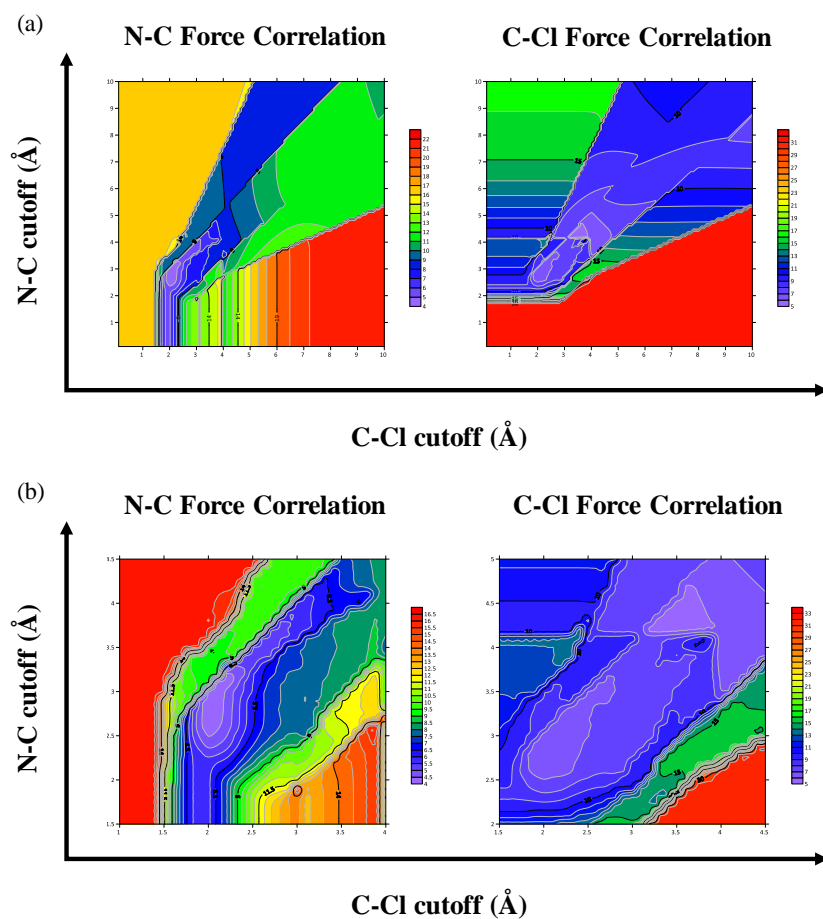


Figure S6

Section E.

A 1-D free energy scan (Table S1) was performed with varying radial cutoff schemes. Fluctuations in free energies/barriers reveals a large dependence on radial cutoff distances. Therefore, optimization of radial cutoff distances is required to obtain reliable FEPs, as shown by the varying degrees of deviation of force correlation in CVs and in free energy correction (Figure 7). Furthermore, a drastic change in free energy and barrier in Table S1 with CV radial cutoffs of ~ 4.0 Å reveals the importance of periodicity and the importance of including all CV force samples, where reactant state samples beyond 3.9 Å for N-C (up to ~ 3.945 Å) are omitted from the pairwise RP-FM procedure in contrast to C-Cl samples where all samples are included within a radial cutoff distance of ~ 3.510 Å (Figure 5).

Table S1: Reaction Barrier/Free Energy of $\text{NH}_3 + \text{CH}_3\text{Cl} \rightarrow \text{CH}_3\text{NH}_3^+ + \text{Cl}^-$ in the Solution Phase, Geometrical Parameters/Force Correlation ($\%F_{\text{corr}}$) in CVs and Radial cutoff Distances (r_c)

	(kcal/mol)		geometry (Å)		$\%F_{\text{corr}}$		r_c (Å)						
	ΔG_s^\ddagger	ΔG_s	$CV_{\text{N-C}}$	$CV_{\text{C-Cl}}$	$CV_{\text{N-C}}$	$CV_{\text{C-Cl}}$	N-H	N-C	N-Cl	H-H	C-H	C-Cl	Cl-H
Uniform Cutoff	26.6	-2.3	1.988	2.138	6.1	17.2	2.000	2.000	2.000		2.000	2.000	2.000
Generic Cutoff	13.7	-20.0	2.152	2.178	3.8	7.1	2.120	2.520	4.280	1.740	2.140	2.760	2.360
Micro-GA	19.5	-22.4	2.105	2.278	6.7	3.5	3.770	4.254	5.676	3.198	2.391	4.870	0.999
Micro-GA-No H-H	16.5	-20.0	2.110	2.280	9.0	8.6	3.183	3.425	1.908		0.683	4.408	0.500
Micro-GA-No N-Cl	20.5	-24.8	2.103	2.266	7.7	3.6	4.906	5.192		5.185	4.283	6.138	6.160
Micro-GA-No H-H/N-Cl	17.1	-21.2	2.117	2.286	6.1	4.9	3.770	4.254			2.370	5.353	5.192
Extended Cutoff	14.1	-23.1	2.106	2.227	7.6	7.1		3.945				3.510	
Micro-GA-CV	14.1	-18.5	2.148	2.197	4.2	6.2		2.028				3.068	
CV @ 2.0	22.7	-7.5	1.964	2.133	5.8	12.6		2.000				2.000	
CV @ 2.2	14.9	-17.2	2.082	2.143	6.2	9.7		2.200				2.200	
CV @ 2.4	11.7	-21.7	2.179	2.089	6.4	8.6		2.400				2.400	
CV @ 2.6	6.9	-28.1	2.212	2.046	6.9	7.8		2.600				2.600	
CV @ 2.8	4.5	-32.3	2.150	2.111	7.5	7.3		2.800				2.800	
CV @ 3.0	4.1	-34.2	2.146	2.112	7.7	7.1		3.000				3.000	
CV @ 3.2	5.5	-33.5	2.145	2.111	7.7	7.1		3.200				3.200	
CV @ 3.4	5.6	-32.4	2.148	2.111	8.0	7.2		3.400				3.400	
CV @ 3.6	2.5	-32.0	2.106	2.149	8.1	8.7		3.600				3.600	
CV @ 3.8	2.1	-32.9	2.112	2.155	8.3	8.3		3.800				3.800	
CV @ 3.9	2.3	-32.7	2.117	2.157	8.5	8.2		3.900				3.900	
CV @ 4.0	17.6	-22.3	2.109	2.225	8.1	6.2		4.000				4.000	
CV @ 4.1	16.5	-24.7	2.105	2.220	8.3	6.3		4.100				4.100	
CV @ 4.2	16.5	-25.7	2.101	2.217	8.5	6.4		4.200				4.200	
CV @ 4.3	17.9	-25.3	2.100	2.211	8.7	6.6		4.300				4.300	
CV @ 4.4	17.9	-25.7	2.098	2.207	8.8	6.7		4.400				4.400	
CV @ 4.6	16.1	-28.3	2.088	2.198	9.1	7.1		4.600				4.600	
CV @ 4.8	15.7	-29.6	2.082	2.192	9.4	7.4		4.800				4.800	
CV @ 5.0	16.5	-29.7	2.075	2.186	9.6	7.6		5.000				5.000	
CV @ 5.5	14.8	-31.2	2.064	2.176	10.4	8.1		5.500				5.500	
CV @ 6.0	12.8	-33.3	2.089	2.144	11.0	8.4		6.000				6.000	
CV @ 6.5	9.6	-35.5	2.085	2.143	11.3	8.6		6.500				6.500	
CV @ 7.0	10.3	-34.2	2.047	2.174	11.4	8.8		7.000				7.000	
CV @ 7.5	10.0	-33.8	2.046	2.180	11.4	9.0		7.500				7.500	
CV @ 8.0	9.3	-34.0	2.081	2.151	11.4	9.1		8.000				8.000	
CV @ 8.5	8.0	-34.4	2.046	2.182	11.2	9.2		8.500				8.500	
CV @ 9.0	12.4	-30.3	2.049	2.183	11.1	9.4		9.000				9.000	
CV @ 9.5	10.3	-30.9	2.049	2.185	11.0	9.5		9.500				9.500	
CV @ 10.0	9.1	-31.1	2.049	2.188	10.9	9.6		10.000				10.000	

B.3 Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability

The following Supporting Information section has been adapted from a previously published article (Kim, B., Shao, Y., Pu, J., "Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability," Journal of Chemical Theory and Computation, 2021, 17(12), 7816-7825.) to provide further details and supporting evidence for the main findings presented in this thesis.

Supporting Information

Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability

Bryant Kim,[†] Yihan Shao,^{*,‡} and Jingzhi Pu^{*,†}

[†]*Department of Chemistry and Chemical Biology, Indiana University-Purdue University Indianapolis, 402 N. Blackford St., Indianapolis, IN 46202*

[‡]*Department of Chemistry and Biochemistry, University of Oklahoma, 101 Stephenson Pkwy, Norman, OK 73019*

(*Correspondence: yihan.shao@ou.edu and jpu@iupui.edu)

S1. Estimate of string free energy error bars along nonuniform CV grids

To evaluate the statistical errors of the free energy profile along the minimum free energy path (MFEP) obtained from the string simulations, we followed a procedure developed by Zhu and Hummer (*J. Comput. Chem.* **2012**, 33, 453-465). In this procedure, the free energy variance is first expressed in terms of the variance of the free energy mean forces on the collective variables (CVs) through a quadrature relation as used in thermodynamic integration. For string simulations under strong harmonic restraints as we adopted in this work, the variance of mean forces are then estimated from the variance of mean values of the CVs and the harmonic force constants. In Zhu and Hummer's derivation, a uniform CV grid is used, which leads to their Eq. (32) for the evaluation of free energy variance. Because we have two bond-CVs, both distributed along non-uniform grids in the optimized MFEP, Zhu & Hummer's Eq. (32) needs to be slightly modified to handle these cases.

Next we derive the string free energy variance for nonuniform CV grids. We start from Zhu & Hummer's quadrature equation Eq. (28),

$$\begin{aligned} G_r(r_k) - G_r(r_j) &= \sum_{i=j}^{k-1} \frac{1}{2} (\bar{F}_i + \bar{F}_{i+1})(r_{i+1} - r_i) \\ &= \frac{1}{2} \bar{F}_j(r_{j+1} - r_j) + \frac{1}{2} \bar{F}_k(r_k - r_{k-1}) + \sum_{i=j+1}^{k-1} \frac{1}{2} \bar{F}_i(r_i - r_{i-1} + r_{i+1} - r_i) \\ &= \frac{1}{2} \bar{F}_j(r_{j+1} - r_j) + \frac{1}{2} \bar{F}_k(r_k - r_{k-1}) + \sum_{i=j+1}^{k-1} \frac{1}{2} \bar{F}_i(r_{i+1} - r_{i-1}) \end{aligned} \quad (\text{S1})$$

where $G_r(r_i)$ and \bar{F}_i represent the free energy associated with a particular bond-CV grid r and the associated mean force on that CV at its i th grid point/image ($r = r_i$) along the string MFEP; the detailed notations can be found in Zhu & Hummer's original reference and are not repeated here. Taking variance on both sides of Eq. (S1) and applying the identity of Zhu & Hummer's Eq. (29) [$\text{var}(\bar{F}_i) = K^2 \text{var}(\bar{x}_i)$], which expresses the variance of free energy mean force in terms of the variance of mean CV positions (\bar{x}_i) and the harmonic force constant (K) used in restraining the CV, we obtain a general form for the variance, applicable to both uniform and nonuniform CV-grid cases,

$$\begin{aligned} &\text{var}[G_r(r_k) - G_r(r_j)] \\ &= \frac{1}{4}(r_{j+1} - r_j)^2 \text{var}(\bar{F}_j) + \frac{1}{4}(r_k - r_{k-1})^2 \text{var}(\bar{F}_k) + \sum_{i=j+1}^{k-1} \frac{1}{4}(r_{i+1} - r_i)^2 [\text{var}(\bar{F}_i)] \\ &= \frac{1}{4} K^2 (r_{j+1} - r_j)^2 \text{var}(\bar{x}_j) + \frac{1}{4} K^2 (r_k - r_{k-1})^2 \text{var}(\bar{x}_k) + \frac{1}{4} K^2 \sum_{i=j+1}^{k-1} (r_{i+1} - r_{i-1})^2 [\text{var}(\bar{x}_i)] \end{aligned} \quad (\text{S2})$$

For a uniform CV grid, where $r_{i+1} - r_i = \Delta r$ [Zhu & Hummer's Eq. (30)] and $r_{i+1} - r_{i-1} = 2\Delta r$, one can show that Eq. (S2) becomes Eq. (S3), which recovers Zhu & Hummer's Eq. (32):

$$\begin{aligned}
& \text{var}[G_r(r_k) - G_r(r_j)] \\
&= \frac{1}{4} K^2 (r_{j+1} - r_j)^2 \text{var}(\bar{x}_j) + \frac{1}{4} K^2 (r_k - r_{k-1})^2 \text{var}(\bar{x}_k) + \frac{1}{4} K^2 \sum_{i=j+1}^{k-1} (r_{i+1} - r_{i-1})^2 [\text{var}(\bar{x}_i)] \\
&= \frac{1}{4} K^2 (\Delta r)^2 \text{var}(\bar{x}_j) + \frac{1}{4} K^2 (\Delta r)^2 \text{var}(\bar{x}_k) + \frac{1}{4} K^2 \sum_{i=j+1}^{k-1} (2\Delta r)^2 [\text{var}(\bar{x}_i)] \tag{S3} \\
&= \frac{1}{4} K^2 (\Delta r)^2 \text{var}(\bar{x}_j) + \frac{1}{4} K^2 (\Delta r)^2 \text{var}(\bar{x}_k) + K^2 (\Delta r)^2 \sum_{i=j+1}^{k-1} [\text{var}(\bar{x}_i)] \\
&= (K\Delta r)^2 \left[\frac{\text{var}(\bar{x}_j) + \text{var}(\bar{x}_k)}{4} + \sum_{i=j+1}^{k-1} \text{var}(\bar{x}_i) \right]
\end{aligned}$$

To estimate the variance of the mean CV positions from potentially correlated data (e.g., from MD), we followed Zhu and Hummer's use of block averages. In particular, our block averages are developed based on eight blocks of 200 ps sampling data for each string image. The square roots of the free energy variances are plotted as the error bars in Fig. 3 for estimating the free energy uncertainties due to the mean force fluctuations. Generally, the statistical uncertainties of the string free energy profiles, both before and after we apply the chaperone polarization corrections, are reasonably small, ranging from 0 kcal/mol (in reactant), through ~0.5 kcal/mol (in transition state), to ~2 kcal/mol (in product) along the string MFEPs.

S2. Details of distributed atomic polarizability calculations using *Gaussian16*

To compute the distributed atomic polarizabilities, we adopted a procedure used by Marenich *et al.* (Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *Chem. Sci.* **2013**, *4*, 2349), which involves finite differentiations of electric-field perturbed atomic dipole moments based on the Hirshfeld partitioned population charges. To determine the polarizability for the solution-phase wavefunction, we used the PCM implicit solvation model in *Gaussian16*. An example of the PCM *Gaussian 16* input file for the distributed atomic polarizability calculation is given below for a reactant configuration of the Menshutkin reaction.

```
%chk=AcW0.chk
#b3lyp/aug-cc-pvtz scrf=(pcm, read, solvent=water) Pop=Hirshfeld

calculation 1: Fx=Fy=Fz=0 a.u.

0 1
N 6.8100090027 1.8850193024 -2.0462064743
H 5.8476800919 1.6863788366 -2.2189681530
H 7.2945981026 2.4258623123 -2.7716665268
H 7.2027258873 1.0075981617 -1.7767280340
C 7.0129194260 3.8728010654 0.6494874358
H 6.0221219063 3.3011381626 0.7361856103
H 7.6980867386 3.0438008308 0.6684535742
H 7.1601700783 4.4616789818 -0.2778538764
CL 7.1414313316 4.8599047661 2.0914375782

radii=pauling

--link1--
%chk=AcW0.chk
#b3lyp/aug-cc-pvtz scrf=(pcm, read, solvent=water) Pop=Hirshfeld geom=check Field=X+10

calculation 2: Fx=-0.0010 a.u.

0 1

radii=pauling

--link1--
%chk=AcW0.chk
#b3lyp/aug-cc-pvtz scrf=(pcm, read, solvent=water) Pop=Hirshfeld geom=check Field=Y+10

calculation 3: Fy=-0.0010 a.u.

0 1

radii=pauling

--link1--
%chk=AcW0.chk
#b3lyp/aug-cc-pvtz scrf=(pcm, read, solvent=water) Pop=Hirshfeld geom=check Field=Z+10

calculation 4: Fz=-0.0010 a.u.

0 1

radii=pauling
```

To calculate polarizabilities at the base semiempirical (SE) level, all instances of ‘b3lyp/aug-cc-pvtz’ in the example above are replaced by the SE method of interest (e.g., ‘am1’).

Below are excerpts from the output files of *Gaussian16* with information for calculating the distributed atomic polarizabilities of each atom at the B3LYP/aug-cc-pVTZ level for the reactant configuration of the Menshutkin reaction mentioned above.

Standard orientation:					
Center Number	Atomic Number	Atomic Type	Coordinates (Angstroms)		
			X	Y	Z
1	7	0	3.195484	0.006085	0.006825
2	1	0	3.514291	-0.467032	-0.811652
3	1	0	3.434940	1.002082	0.072255
4	1	0	3.464074	-0.589732	0.761526
5	6	0	-0.159772	0.029390	-0.023562
6	1	0	0.168777	-0.863087	-0.665071
7	1	0	0.264093	-0.265129	0.920182
8	1	0	0.237079	1.020421	-0.321536
9	17	0	-1.911353	-0.003321	0.008111

Calculation 1 (F_x = F_y = F_z = 0):

Hirshfeld charges, spin densities, dipoles, and CM5 charges using IRadAn= 5:							
		Q-H	S-H	Dx	Dy	Dz	Q-CM5
1	N	-0.263445	0.000000	0.154940	0.010284	0.009037	-0.821169
2	H	0.100780	0.000000	0.057581	-0.105695	-0.155544	0.290081
3	H	0.113909	0.000000	0.031103	0.200305	0.010866	0.290216
4	H	0.102944	0.000000	0.044990	-0.131151	0.141374	0.291938
5	C	-0.046956	0.000000	-0.047909	0.020278	0.007963	-0.183473
6	H	0.044458	0.000000	0.036521	-0.130842	-0.068663	0.090626
7	H	0.039225	0.000000	0.054611	-0.051005	0.133763	0.093507
8	H	0.051439	0.000000	0.048678	0.149613	-0.034510	0.101464
9	Cl	-0.142355	0.000000	0.177724	-0.010763	-0.007157	-0.153190
	Tot	-0.000001	0.000000	0.558239	-0.048976	0.037129	-0.000001

Calculation 2 (F_x = -0.001 a.u.):

Hirshfeld charges, spin densities, dipoles, and CM5 charges using IRadAn= 5:							
		Q-H	S-H	Dx	Dy	Dz	Q-CM5
1	N	-0.262741	0.000000	0.148181	0.010118	0.008920	-0.820464
2	H	0.099846	0.000000	0.054794	-0.105793	-0.155608	0.289146
3	H	0.113123	0.000000	0.028370	0.200285	0.010831	0.289430
4	H	0.102127	0.000000	0.042238	-0.131271	0.141333	0.291120
5	C	-0.048031	0.000000	-0.049624	0.020314	0.007935	-0.184548
6	H	0.043222	0.000000	0.035317	-0.130492	-0.068303	0.089390
7	H	0.037999	0.000000	0.053706	-0.050927	0.133351	0.092280
8	H	0.050209	0.000000	0.047601	0.149238	-0.034355	0.100234
9	Cl	-0.135755	0.000000	0.162764	-0.010706	-0.007141	-0.146590
	Tot	-0.000001	0.000000	0.523347	-0.049233	0.036962	-0.000001

Calculation 3 (F_y = -0.001 a.u.):

Hirshfeld charges, spin densities, dipoles, and CM5 charges using IRadAn= 5:							
		Q-H	S-H	Dx	Dy	Dz	Q-CM5
1	N	-0.263533	0.000000	0.154810	0.005740	0.009031	-0.821257
2	H	0.101916	0.000000	0.057562	-0.107402	-0.155544	0.291217
3	H	0.111403	0.000000	0.030820	0.198648	0.010858	0.287710
4	H	0.104376	0.000000	0.045005	-0.132831	0.141367	0.293369
5	C	-0.047330	0.000000	-0.047832	0.015317	0.007896	-0.183846
6	H	0.046747	0.000000	0.036706	-0.133075	-0.068916	0.092915
7	H	0.039860	0.000000	0.054659	-0.052907	0.133808	0.094142
8	H	0.048953	0.000000	0.048513	0.147498	-0.034439	0.098978
9	Cl	-0.142394	0.000000	0.177852	-0.029172	-0.007154	-0.153229
	Tot	-0.000001	0.000000	0.558096	-0.088184	0.036907	-0.000001

Calculation 4 (Fz = -0.001 a.u.):

Hirshfeld charges, spin densities, dipoles, and CM5 charges using IRadAn= 5:							
		Q-H	S-H	Dx	Dy	Dz	Q-CM5
1	N	-0.263428	0.000000	0.154813	0.010273	0.004155	-0.821151
2	H	0.102835	0.000000	0.057632	-0.105628	-0.157303	0.292135
3	H	0.113735	0.000000	0.031043	0.200313	0.009043	0.290042
4	H	0.101038	0.000000	0.044823	-0.131239	0.139588	0.290032
5	C	-0.046916	0.000000	-0.047840	0.020192	0.002391	-0.183433
6	H	0.046191	0.000000	0.036710	-0.130981	-0.070914	0.092359
7	H	0.036888	0.000000	0.054435	-0.050996	0.131545	0.091170
8	H	0.052171	0.000000	0.048770	0.149619	-0.036545	0.102196
9	Cl	-0.142516	0.000000	0.177743	-0.010758	-0.025741	-0.153351
	Tot	-0.000001	0.000000	0.558129	-0.049204	-0.003779	-0.000001

The γ ($\gamma = x, y, z$) component of the distributed dipole (in atomic units) for QM atom i at each value of the electric field F is calculated by,

$$\mu_{\gamma,i} = q_i R_{\gamma,i} + \delta\mu_{\gamma,i} \quad (\text{S4})$$

where q_i are obtained from column 3 in the excerpts given above, $R_{\gamma,i}$ are the standard-orientation Cartesian coordinates converted from Å to atomic units, and $\delta\mu_{x,i}$, $\delta\mu_{y,i}$, and $\delta\mu_{z,i}$ are obtained from columns 5-7 in the above excerpts.

The $\gamma\gamma$ component of atom i 's polarizability tensor (in atomic units) is calculated as,

$$\alpha_{\gamma\gamma,i} = \frac{\mu_{\gamma,i}(F_\gamma) - \mu_{\gamma,i}(0)}{F_\gamma} \quad (\text{S5})$$

where F_γ is the magnitude of an electric field \mathbf{F} along the γ axis, and $F_\gamma = -0.001$ a.u. in this calculation. The values of $\mu_{x,i}(0)$, $\mu_{y,i}(0)$, and $\mu_{z,i}(0)$ are based on calculation 1. The values of $\mu_{x,i}(F_x)$, $\mu_{y,i}(F_y)$, and $\mu_{z,i}(F_z)$ are based on calculation 2, 3, and 4, respectively.

By combining $\mu_{\gamma,i}(0)$, $\mu_{\gamma,i}(F_\gamma)$, and F_γ for each atom i , averaging the quantities of obtained from Eq. (S5) for three Cartesian directions γ , and converting the units to \AA^3 ($1 \text{ Bohr}^3 = 0.14819 \text{ \AA}^3$), we obtain the distributed isotropic atomic polarizability for QM atom i at the AI-PCM level, which is utilized in Eq. (1) in the text for the definition of the chaperone polarizability. Likewise, the distributed polarizabilities for each QM atom at the SE-PCM level can be obtained with the same procedure.

Table S1. Distributed atomic polarizabilities at the AI(B3LYP/aug-cc-pVTZ)-PCM level

i	$\alpha_{xx,i}^{\text{AI-PCM}}$	$\alpha_{yy,i}^{\text{AI-PCM}}$	$\alpha_{zz,i}^{\text{AI-PCM}}$	$\alpha_i^{\text{AI-PCM}} = 1/3(\alpha_{xx,i}^{\text{AI-PCM}} + \alpha_{yy,i}^{\text{AI-PCM}} + \alpha_{zz,i}^{\text{AI-PCM}})$
N	0.372	0.674	0.723	0.590
H ₁	1.332	0.402	0.728	0.820
H ₂	1.161	0.949	0.274	0.794
H ₃	1.200	0.485	0.671	0.786
C	0.206	0.738	0.826	0.590
H ₄	0.237	0.884	0.656	0.592
H ₅	0.225	0.329	0.931	0.495
H ₆	0.241	1.024	0.367	0.544
Cl	5.749	2.728	2.754	3.744
total	10.724	8.212	7.931	8.956

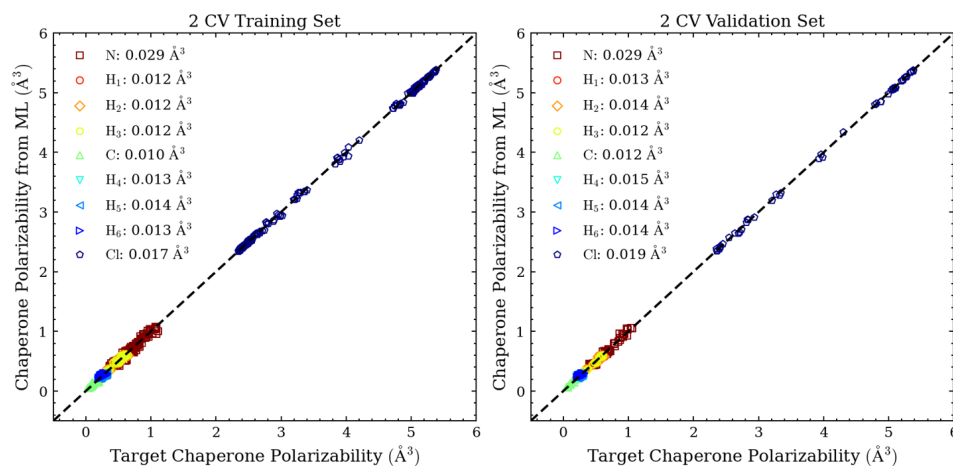
Table S2. Atomic and chaperone polarizabilities for each QM atom between AI(B3LYP/aug-cc-pVTZ)-PCM and SE(AM1)-PCM levels

i	$\alpha_i^{\text{AI-PCM}}$	$\alpha_i^{\text{SE-PCM}}$	$\Delta\alpha_i^{\text{C}}$
N	0.590	-0.020	0.610
H ₁	0.820	0.208	0.612
H ₂	0.794	0.210	0.585
H ₃	0.786	0.198	0.588
C	0.590	0.297	0.293
H ₄	0.592	0.299	0.293
H ₅	0.495	0.243	0.252
H ₆	0.544	0.289	0.255
Cl	3.744	1.382	2.362
total	8.956	3.106	5.850

S3. Input analysis for atomic polarizability corrections from artificial neural network (ANN)

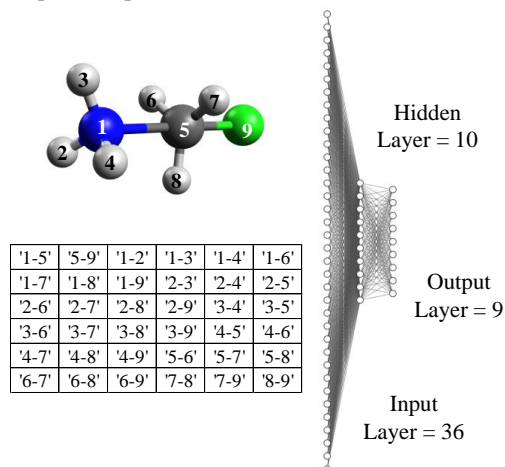
Atomic polarizability differences between B3LYP/aug-cc-pVTZ and AM1 in Fig. 5 are used as the target data set in artificial neural network (ANN) training. Here, training is performed using the collective variables (CVs), defined as the breaking ($r_{\text{C-Cl}}$) and forming ($r_{\text{N-C}}$) bonds, as the input features (Scheme 2). In order to check the quality of the fit, 80% of the samples (192 configurations) are randomly divided into a training set, with the remaining 20% of the samples (48 configurations) being used as the validation set. The correlation for the validation and training sets shows that the fit is reliable, as the error for all atoms is less than 0.03 \AA^3 (Figure S1).

Figure S1. Root mean square error of the ANN for each chaperone polarizability on each atom in training (left panel) and validation (right panel) sample sets using only the two CVs as input features.



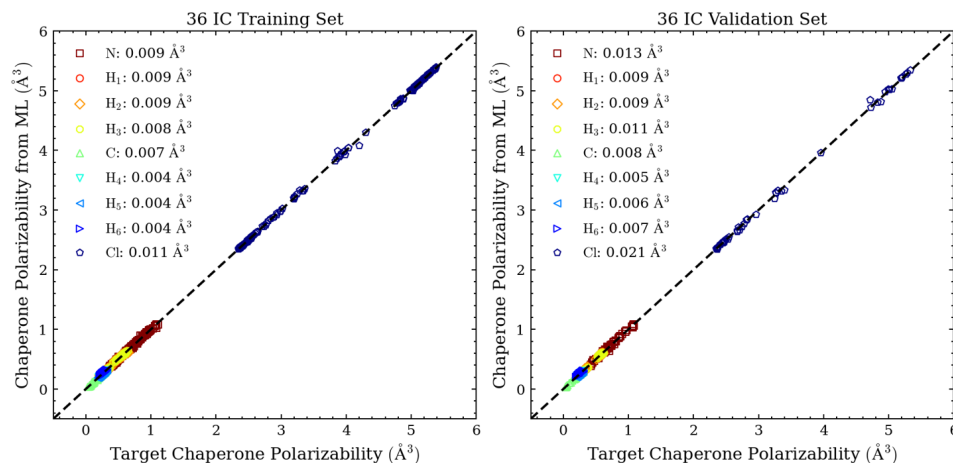
In order to evaluate the input features for the ANN, we expand the input layer to 36 input features (Scheme S1) composed of all possible pairwise distances of the QM solute molecule.

Scheme S1. List of all possible pairwise distances in the Menshutkin reaction for ANN training.



Compared to the two-CV results, the error is only marginally improved with additional input features (Figure S2). Therefore, it is viable to correct nine atomic corrections with two-CV inputs, since the results are consistent with the larger input set.

Figure S2. Root mean square error of the ANN for each chaperone polarizability on each atom in training (left panel) and validation (right panel) sample sets using all possible pairwise distances (referred to as “36 IC”) as input features.



S4. Details of the gradient calculation for distributed atomic polarizabilities from artificial neural network (ANN)

A calculation of the gradients from the distributed atomic polarizabilities involves a differentiation of the ANN. Here, a summary of the ANN which consists of two input features ($k = 2$), ten hidden nodes ($j = 10$), and nine outputs ($i = 9$) is provided below:

```
Model Summary:
<tensorflow.python.keras.optimizer_v2.adam.Adam object at 0x7f67711b2e90>

2 layer(s), 2 input(s), 9 output(s)
10 node(s) in Hidden layer_1 (Activation = tanh), Output layer (Activation = linear)

Model: "sequential_1"

```

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 10)	30
activation_1 (Activation)	(None, 10)	0
dense_2 (Dense)	(None, 9)	99
activation_2 (Activation)	(None, 9)	0

```

Total params: 129
Trainable params: 129
Non-trainable params: 0

```

$W_{j,k}^H$ – dense_1 (Dense) weights:

```
-0.599727 1.027091 2.741373 0.083218 -0.000529 -1.953961 -1.940043 -1.556540 -0.119394 1.573665
0.507001 0.700813 -3.065630 -0.428064 -0.000901 1.870803 -1.614903 -0.141739 -0.697106 -2.794809
```

b_j^H – dense_1 (Dense) biases:

```
0.185776
-1.800465
1.346971
0.953772
0.003120
-0.097830
0.138638
1.475396
0.944476
1.996852
```

$W_{i,j}^O$ – dense_2 (Dense) weights:

```
-0.005499 0.116408 -0.060738 0.090293 -0.052982 0.112268 0.058761 0.144776 0.360778
0.385722 0.053184 -0.433583 -0.273320 -0.245747 0.711497 0.132981 0.101145 1.227100
-0.011195 -0.233242 -0.146342 -0.241261 -0.121331 0.010316 0.091536 0.056560 -1.247315
-0.025083 0.329692 -0.002514 0.273491 0.199837 0.264404 0.272614 0.422577 -0.645320
0.000368 -0.000008 -0.000218 0.000214 0.000201 -0.000240 0.000018 -0.000340 0.000070
-0.824427 -0.692168 -0.547976 -0.720121 0.235040 0.118191 0.254213 0.188730 -0.890170
-0.682484 0.060105 -0.333702 -0.384628 -0.302415 0.101232 -0.076361 0.015166 -1.056963
0.801471 0.093110 0.035777 0.048964 -0.409850 0.302678 0.167761 0.020224 -0.874134
0.102439 -0.636886 -0.357353 -0.717807 0.068446 0.178237 0.021762 -0.146412 -0.055213
-0.656054 -0.219884 -0.207838 -0.221248 0.243635 0.074877 0.119667 0.115799 -0.722511
```

b_i^0 – dense_2 (Dense) biases:

```
0.303051
-0.003902
0.260096
-0.225453
-0.151535
0.058515
0.204238
0.049822
0.792560
```

Combined with the weights and biases of the ANN, an example calculation for the polarizability gradients is provided below using the following coordinate and force files from CHARMM.

CHARMM coordinate file:

```
*STREAM TO REAL FILE GIVEN AS ZZZ=FILENAME ON COMMANDLINE. NOTE THAT THE FILENAM
* CANNOT CONSIST OF A MIXTURE OF UPPER- AND LOWER-CASE LETTERS.
* DATE: 10/26/20 1:10: 5 CREATED BY USER: brykim
*
6465
1 1 AMM NZ 6.81001 1.88502 -2.04621 AMM 1 0.00000
2 1 AMM HZ1 5.84768 1.68638 -2.21897 AMM 1 0.00000
3 1 AMM HZ2 7.29460 2.42586 -2.77167 AMM 1 0.00000
4 1 AMM HZ3 7.20273 1.00760 -1.77673 AMM 1 0.00000
5 2 MECL C1 7.01292 3.87280 0.64949 MECL 1 0.00000
6 2 MECL H1 6.02212 3.30114 0.73619 MECL 1 0.00000
7 2 MECL H2 7.69809 3.04380 0.66845 MECL 1 0.00000
8 2 MECL H3 7.16017 4.46168 -0.27785 MECL 1 0.00000
9 2 MECL CL 7.14143 4.85990 2.09144 MECL 1 0.00000
```

CHARMM force file:

```
COORDINATE FILE MODULE
TITLE> *STREAM TO REAL FILE GIVEN AS ZZZ=FILENAME ON COMMANDLINE. NOTE THAT THE FILENAM
TITLE> * CANNOT CONSIST OF A MIXTURE OF UPPER- AND LOWER-CASE LETTERS.
TITLE> * DATE: 10/26/20 1:10: 5 CREATED BY USER: brykim
TITLE> *
6465
1 1 AMM NZ 0.75231 -9.33158 -6.16389 AMM 1 0.00000
2 1 AMM HZ1 -3.45198 -13.32486 12.12150 AMM 1 0.00000
3 1 AMM HZ2 17.41862 31.01837 -17.59062 AMM 1 0.00000
4 1 AMM HZ3 -15.26584 -9.20622 11.31869 AMM 1 0.00000
5 2 MECL C1 37.69368 6.15070 -16.23114 MECL 1 0.00000
6 2 MECL H1 -8.78734 -23.82849 9.96549 MECL 1 0.00000
7 2 MECL H2 -32.82242 14.84811 -2.85069 MECL 1 0.00000
8 2 MECL H3 6.81549 -3.21911 -4.97618 MECL 1 0.00000
9 2 MECL CL -3.20828 5.10224 16.42111 MECL 1 0.00000
```

During molecular dynamics, the gradient from the polarizability is determined on the fly with the ANN based on the CV input features. An example calculation for the chaperone polarizability and respective gradients are provided below. Here, in this example, the values of

$p_1 = r_{\text{N-C}} = 3.355 \text{ \AA}$ and $p_2 = r_{\text{C-Cl}} = 1.752 \text{ \AA}$ for the CV's, are obtained by calculating the bond lengths of CVs from the coordinate file above. Combined with the weights and biases above, the chaperone polarizabilities are determined with Eq. (3) in the text. To calculate the gradient for each atomic polarizability correction, the ANN is differentiated based on Eq. (6) in the text. By combining the weights and biases from each layer, the atomic polarizability correction in addition to the correcting forces on the CV's, in units of kcal/mol/ \AA are determined below (Table S3).

Table S3. Chaperone polarizabilities and gradients on CVs obtained from ANN

i	$\Delta\alpha_i^C$	$\frac{d\Delta\alpha_i^C}{dr_{N-C}}$	$\frac{d\Delta\alpha_i^C}{dr_{C-Cl}}$
N	0.667	-0.032	0.019
H ₁	0.575	0.034	0.219
H ₂	0.578	0.033	0.16
H ₃	0.578	0.04	0.266
C	0.288	0.031	-0.135
H ₄	0.259	-0.016	-0.145
H ₅	0.267	0.006	-0.097
H ₆	0.267	0.005	-0.053
Cl	2.351	-0.151	0.402
total	5.83	-0.05	0.636

where, the polarizability gradient of each QM atom on CV's are dispersible into nuclear coordinates as follows (Table S4).

Table S4. Polarizability gradients on Cartesian coordinates of QM atoms obtained from ANN

i	$\frac{d\Delta\alpha_i^C}{dx_N}$	$\frac{d\Delta\alpha_i^C}{dy_N}$	$\frac{d\Delta\alpha_i^C}{dz_N}$	$\frac{d\Delta\alpha_i^C}{dx_C}$	$\frac{d\Delta\alpha_i^C}{dy_C}$	$\frac{d\Delta\alpha_i^C}{dz_C}$	$\frac{d\Delta\alpha_i^C}{dx_{Cl}}$	$\frac{d\Delta\alpha_i^C}{dy_{Cl}}$	$\frac{d\Delta\alpha_i^C}{dz_{Cl}}$
N	0.002	0.019	0.026	-0.003	-0.029	-0.041	0.001	0.010	0.015
H ₁	-0.002	-0.020	-0.028	-0.014	-0.103	-0.153	0.016	0.123	0.180
H ₂	-0.002	-0.019	-0.026	-0.010	-0.071	-0.106	0.012	0.090	0.132
H ₃	-0.002	-0.024	-0.032	-0.017	-0.126	-0.187	0.020	0.150	0.219
C	-0.002	-0.018	-0.025	0.012	0.094	0.136	-0.010	-0.076	-0.111
H ₄	0.001	0.009	0.013	0.010	0.072	0.107	-0.011	-0.082	-0.119
H ₅	0.000	-0.003	-0.005	0.007	0.058	0.085	-0.007	-0.055	-0.080
H ₆	0.000	-0.003	-0.004	0.004	0.033	0.047	-0.004	-0.030	-0.043
Cl	0.009	0.089	0.121	-0.039	-0.316	-0.452	0.029	0.226	0.331
total	0.004	0.030	0.040	-0.050	-0.388	-0.564	0.046	0.356	0.524

It is important to note that the gradients listed in the table above are exclusively due to the correction to the atomic polarizability on the CV atoms. More precisely, the forces from the permanent electric field are not included in the table above. Therefore, it is important to be mindful of the fact that the gradients from the distributed atomic polarizability need to be supplemented onto the gradients of the permanent electric field to obtain a consistent gradient from polarization.

S5. Statistical distributions of molecular polarizabilities

The mean molecular polarizabilities and their standard deviations (stdev) sampled in each individual reaction coordinate (RC) windows are tabulated in Table S5; the graphical representation of the same data can be found in Fig. 1 in the text.

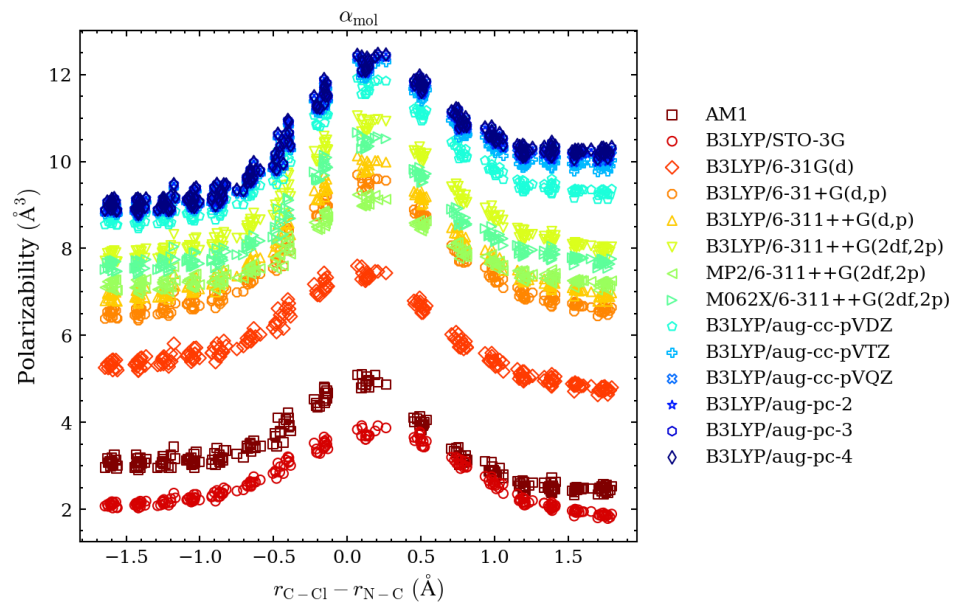
Table S5. Statistical distributions of molecular polarizabilities (in Å³)

RC (in Å): $r_{C-Cl} - r_{N-C}$	AM1/MM		Difference		B3LYP/aug-cc-pVTZ	
	mean	stdev	mean	stdev	mean	stdev
-1.597	3.073	0.081	5.841	0.030	8.913	0.091
-1.411	3.063	0.097	5.835	0.040	8.898	0.110
-1.230	3.104	0.129	5.866	0.036	8.970	0.155
-1.049	3.110	0.095	5.923	0.039	9.033	0.124
-0.870	3.189	0.115	5.977	0.048	9.165	0.152
-0.661	3.417	0.095	6.145	0.091	9.562	0.168
-0.440	3.834	0.227	6.430	0.134	10.264	0.339
-0.168	4.568	0.156	6.824	0.077	11.392	0.201
0.141	4.931	0.108	7.188	0.101	12.119	0.141
0.498	4.000	0.075	7.585	0.042	11.584	0.093
0.767	3.218	0.125	7.535	0.029	10.753	0.144
1.000	2.881	0.108	7.464	0.022	10.345	0.114
1.197	2.548	0.075	7.444	0.024	9.991	0.090
1.381	2.522	0.092	7.463	0.025	9.985	0.096
1.557	2.455	0.043	7.478	0.027	9.932	0.051
1.739	2.457	0.070	7.458	0.038	9.915	0.075
<average>	3.273	0.106	6.779	0.050	10.051	0.134
<stdev>/<mean>		3.23%		0.74%		1.33%

Despite a slight broadening in the transition state (TS) region, the overall distributions of the molecular polarizabilities from individual windows are generally narrow. The standard deviations are found in the range of 0.04~0.23 and 0.05~0.34 Å³ for the AM1 and B3LYP results, respectively. Averaged over all 16 images along the string path, the standard deviations in molecular polarizability are 0.11 and 0.13 Å³ (or 3.23% and 1.33% in percentage deviation relative to the associated means) at the AM1 and B3LYP levels. The corrective chaperone molecular polarizabilities that account for the difference between the AM1 and B3LYP levels display an even smaller averaged standard deviation of 0.05 Å³ (or 0.74% in percentage deviation).

S6. Convergence of molecular polarizability with basis sets and AI methods

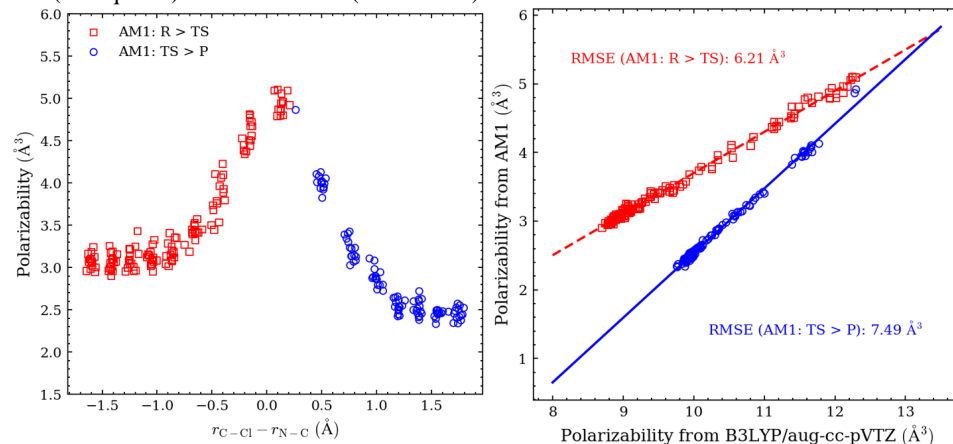
Figure S3. Molecular polarizability of the Menshutkin reaction computed at various AI levels of theory and basis sets.



S7. Distribution of polarizability difference between AM1 and B3LYP/aug-cc-pVTZ

In Figure 2 in the text, a bimodal distribution of the correlation between the molecular polarizability obtained at the AM1 and B3LYP/aug-cc-pVTZ levels is found. In Figure S4 the AM1 molecular polarizability is decomposed to each branch, where data for configurations in the first half of the reaction are plotted with red squares and that for configurations departing the transition state for the product in the second half of the reaction are plotted with blue circles. Overall, the two populations in the AM1 molecular polarizability are related to the TS- and P-formation branches of the minimum free energy path (MFEP), both of which are systematically underestimated. More specifically, the branched deviation in Figure 2 corresponds to a slope change in polarizability regression error for the AM1 method which is inflected at the transition state. In particular, the latter half of the reaction path is associated with a greater error (with an RMSE of 7.49 \AA^3 for the “TS > P” branch, compared to 6.21 \AA^3 for the “R > TS” branch), which indicates a larger issue with the AM1 method with respect to describing the charge-separated species. Note that our ANN model successfully eliminates the branched behavior in the polarizability error distribution, with the B3LYP/aug-cc-pVTZ molecular polarizabilities reproduced throughout the entire MFEP with only a small error of 0.03 \AA^3 (see Fig. 2 in the text).

Figures S4. AM1 polarizability and correlation to AI benchmark (B3LYP/aug-cc-pVTZ) for TS- (red squares) and P-formation (blue circles) branches of the MFEP.

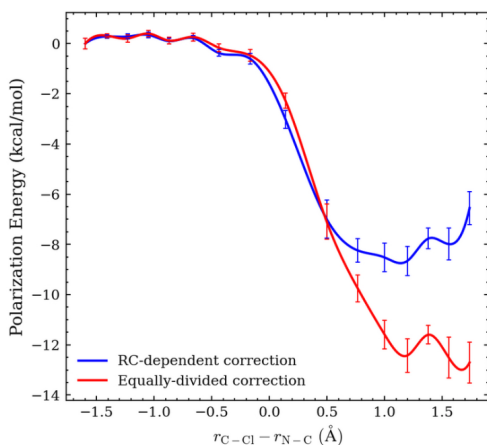


S8. Polarization energy correction using RC-independent chaperone polarizabilities

To examine whether the use of a reaction coordinate (RC)-dependent polarizability correction is essential, we tested a set of equally-divided RC-independent atomic polarizabilities that yield an average molecular polarizability correction of 6.78 \AA^3 , which is the root-mean-square error of the AM1-computed molecular polarizabilities with respect to the B3LYP results (see Fig. 2). In our ML-RC-dependent model, we used the atomic polarizabilities that are quantum mechanically partitioned from the molecular ones. To remove this feature from the RC-independent control scheme, we divided this fixed-valued molecular polarizability correction of 6.78 \AA^3 equally to the nine solute atoms. Note that this control scheme is designed only to estimate the overall effect when the essential features of our ML model (i.e., RC-dependence and QM partitioning) are removed, but the scheme itself can be less physical, especially considering the substantial charge separation during the Menshutkin reaction along its RC. To avoid potential instabilities in the dynamically-corrected free energy simulations, we used the ensemble-averaged polarization energy correction $\langle \Delta E^{\text{pol}} \rangle$ as an estimate (in a free energy perturbation manner) of the change in free energy upon chaperone incorporation. The $\langle \Delta E^{\text{pol}} \rangle$ using the RC-dependent and RC-independent chaperone polarizabilities are compared in Fig. S5.

In Fig. S5, large differences in $\langle \Delta E^{\text{pol}} \rangle$ are found when the RC-independent chaperone polarizabilities (labelled as “Equally-divided correction”) are used to replace the ML-RC-dependent chaperone polarizabilities (labelled as “RC-dependent correction”). The difference in $\langle \Delta E^{\text{pol}} \rangle$ starts relatively small for the reactant branch of the reaction, but becomes more pronounced in the product-forming branch (after $\text{RC} > +0.5 \text{ \AA}$), and ends up with a large magnitude of $\sim 7 \text{ kcal/mol}$ in the final product region ($\text{RC} > \sim +1.5 \text{ \AA}$). The smaller magnitude of $\langle \Delta E^{\text{pol}} \rangle$ for the RC-independent scheme indicates an underestimated induced polarization of the charge-separated product species, when an average “non-reactive” correction is used. This result further demonstrates that “reactive” fitting of the polarizability corrections is important.

Figure S5. Ensemble-averaged polarization energy correction $\langle \Delta E^{\text{pol}} \rangle$ using equally-divided-RC-independent and ML-RC-dependent chaperone polarizabilities.



S9. Polarization energy, polarizability, and radial distribution functions for QM atoms

In Figures S6 to S14, the polarization energy, polarizability, and radial distribution functions are plotted for each QM solute atom. In particular, 240 configurations from three string paths are used to generate panels (a-c), where 1,000 configurations are sampled around the reactant, transition, and product states for panels (d-f). Specifically, panel (a) shows the polarization energy from each water molecule which is represented as a function of the reaction coordinate and the pairwise distance to each water molecule. In panel (b), the solution-phase (PCM) atomic polarizabilities are compared at the AM1 and B3LYP/aug-cc-pVTZ levels; the polarizability results obtained for more basis sets and AI levels of theory are shown in panel (c). Finally, the radial distribution functions (RDF) between each QM atom with water oxygen (O_w) are shown in panels (d-f) for the AM1/MM and the chaperone-corrected dp-AM1/MM simulations.

Analysis of Nitrogen (N; Figure S6), analysis of Hydrogens from NH_3 (H_{1-3} ; Figures S7-S9), analysis of Carbon (C; Figure S10), analysis of Hydrogens from CH_3Cl (H_{4-6} ; Figures S11-S13), and analysis of Chlorine (Cl; Figure S14) are given. In panel (a) the polarization energy of water molecules around the atom of interest is calculated at the AM1 (red circles) and B3LYP/aug-cc-pVTZ (blue circles) levels, the difference is shown with green triangles. In panel (b) the solution-phase (PCM) polarizability of the atom of interest is calculated at the AM1 (red circles) and B3LYP/aug-cc-pVTZ (blue circles) levels, and their difference is shown with green triangles. In panel (c), the polarizability of the atom of interest is computed at various AI levels of theory and basis sets. Radial distribution functions (RDFs) around the atom of interest obtained from AM1/MM (dotted red) and dp-AM1/MM (with polarizabilities corrected to the B3LYP/aug-cc-pVTZ level; solid blue) are shown for: reactant (R; panel d), transition state (TS; panel e), and product (P; panel f).

Figure S6. Polarization energy, polarizability, and radial distribution functions for the N atom

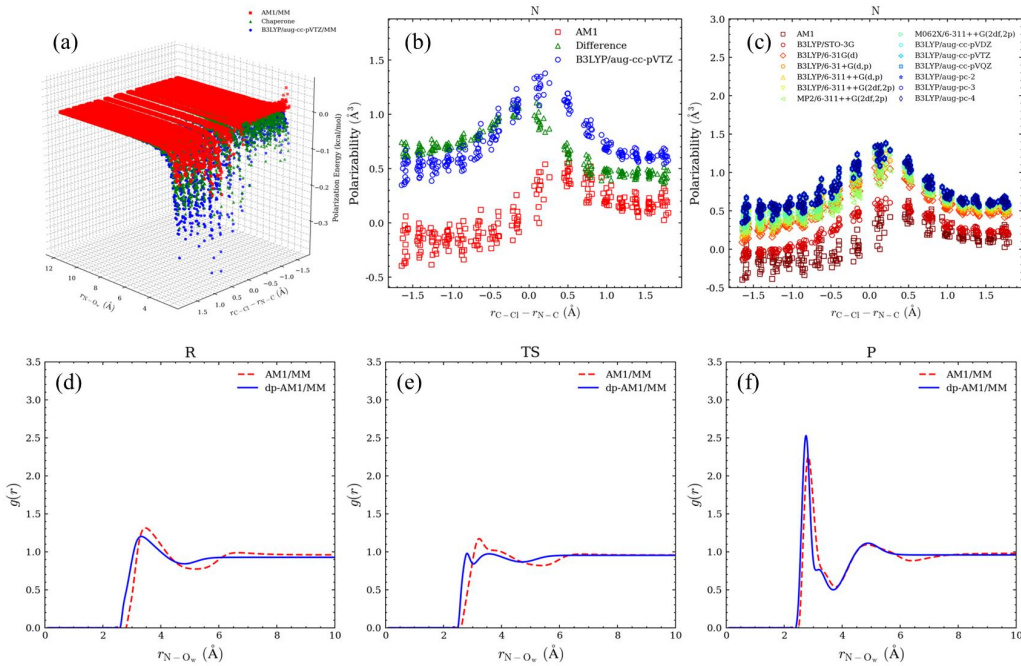


Figure S7. Polarization energy, polarizability, and radial distribution functions for the H_1 atom

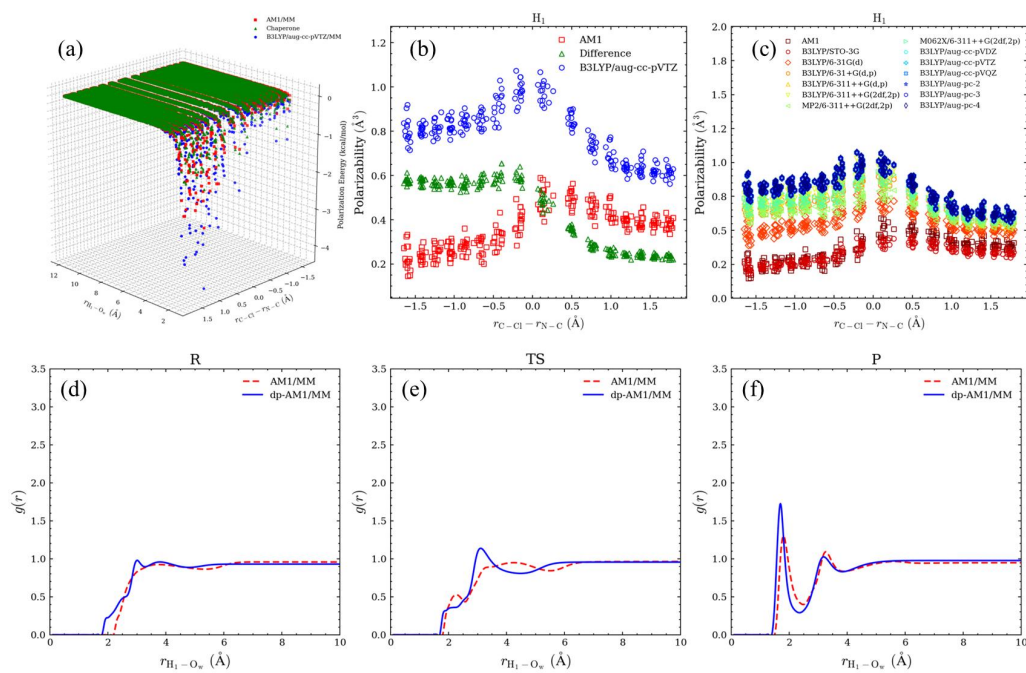


Figure S8. Polarization energy, polarizability, and radial distribution functions for the H₂ atom

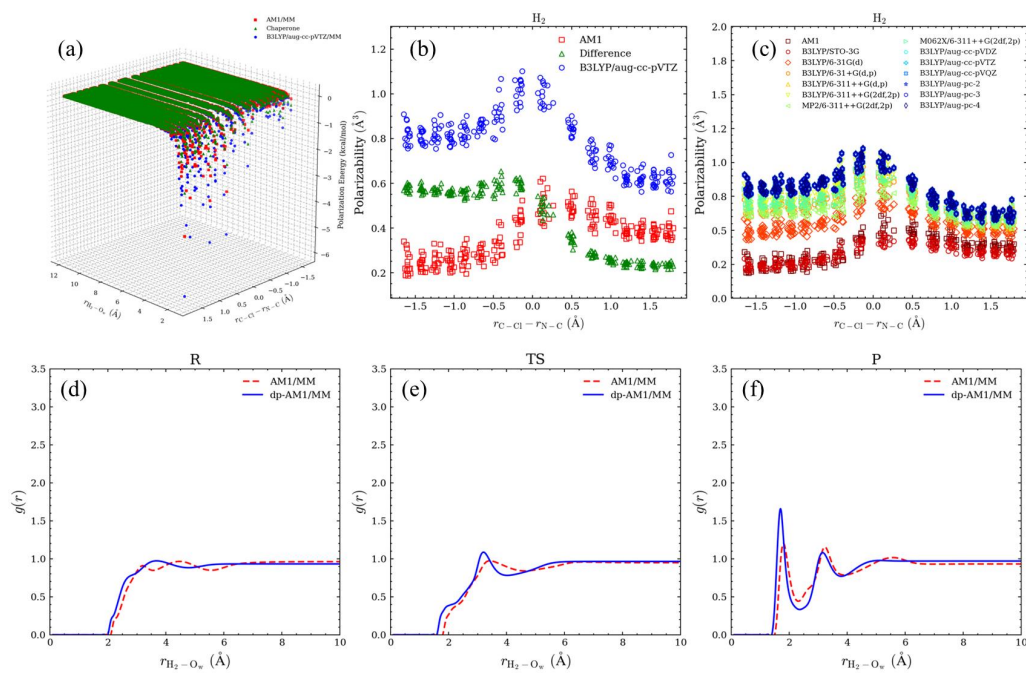


Figure S9. Polarization energy, polarizability, and radial distribution functions for the H_3 atom

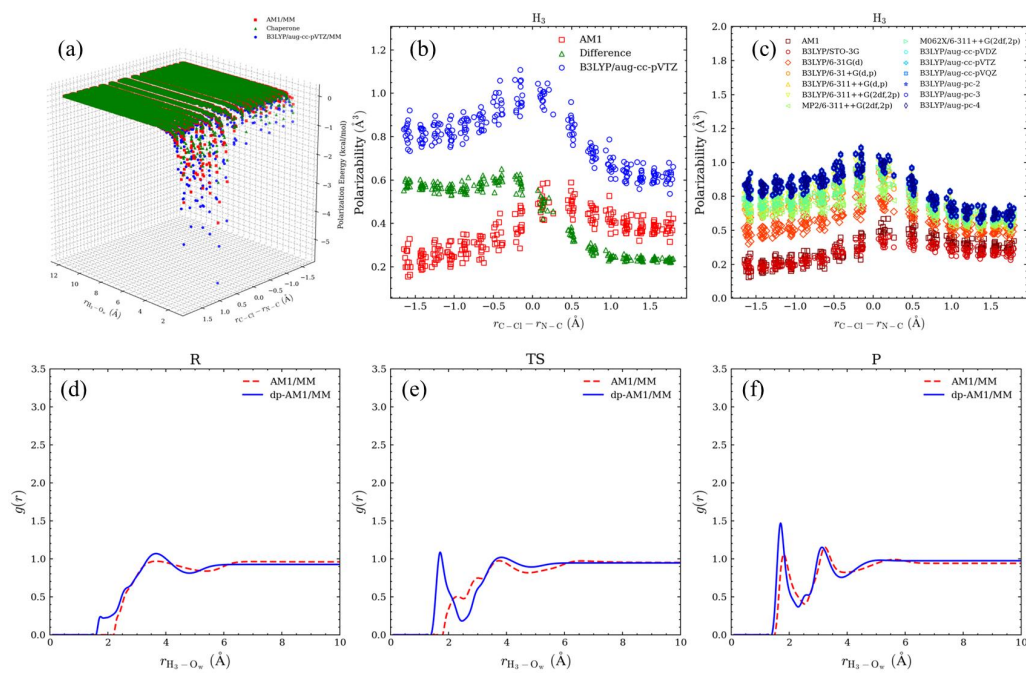


Figure S10. Polarization energy, polarizability, and radial distribution functions for the C atom

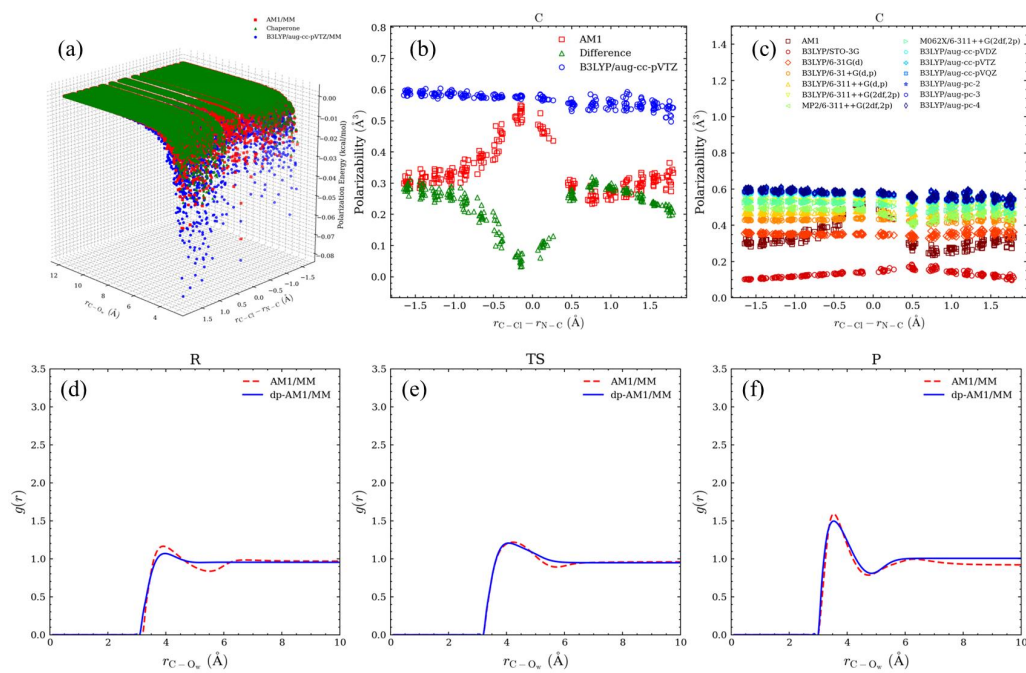


Figure S11. Polarization energy, polarizability, and radial distribution functions for the H_4 atom

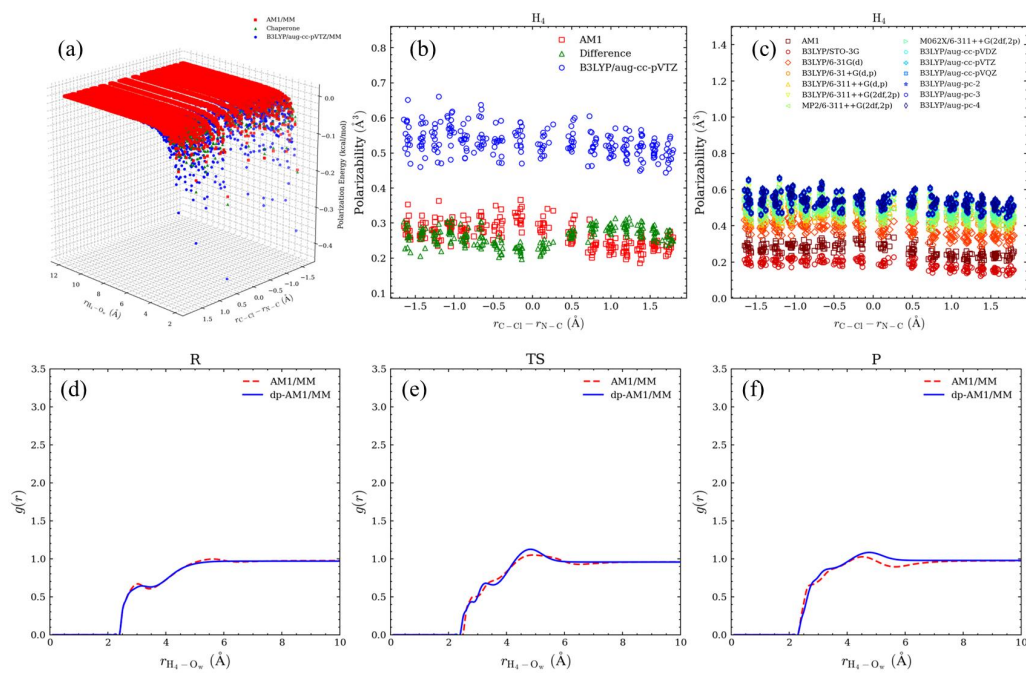


Figure S12. Polarization energy, polarizability, and radial distribution functions for the H_2 atom

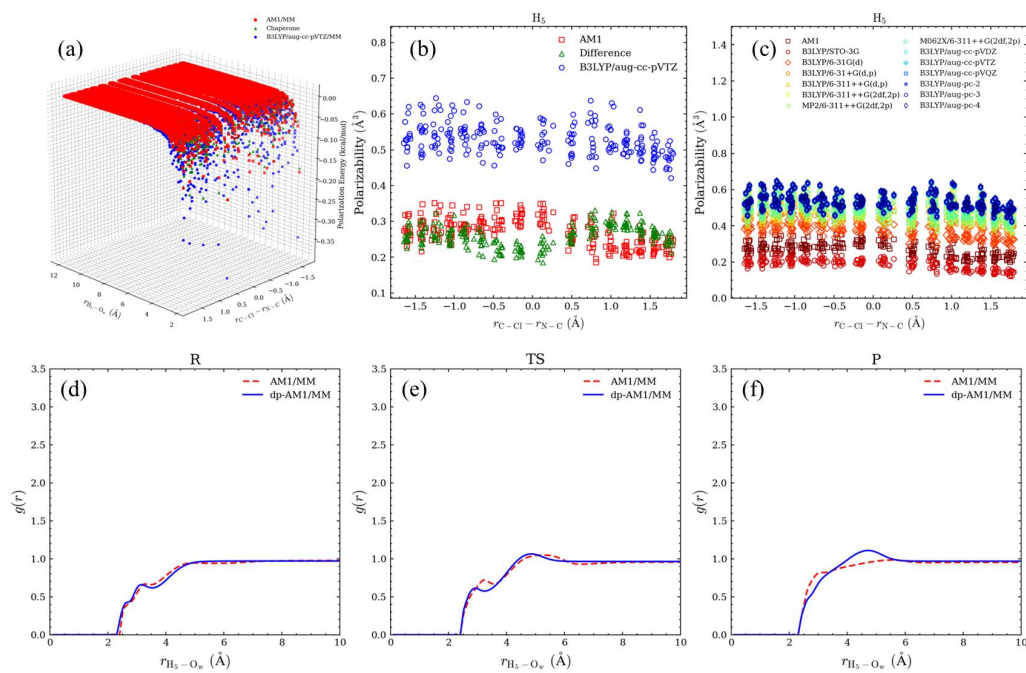


Figure S13. Polarization energy, polarizability, and radial distribution functions for the H_6 atom

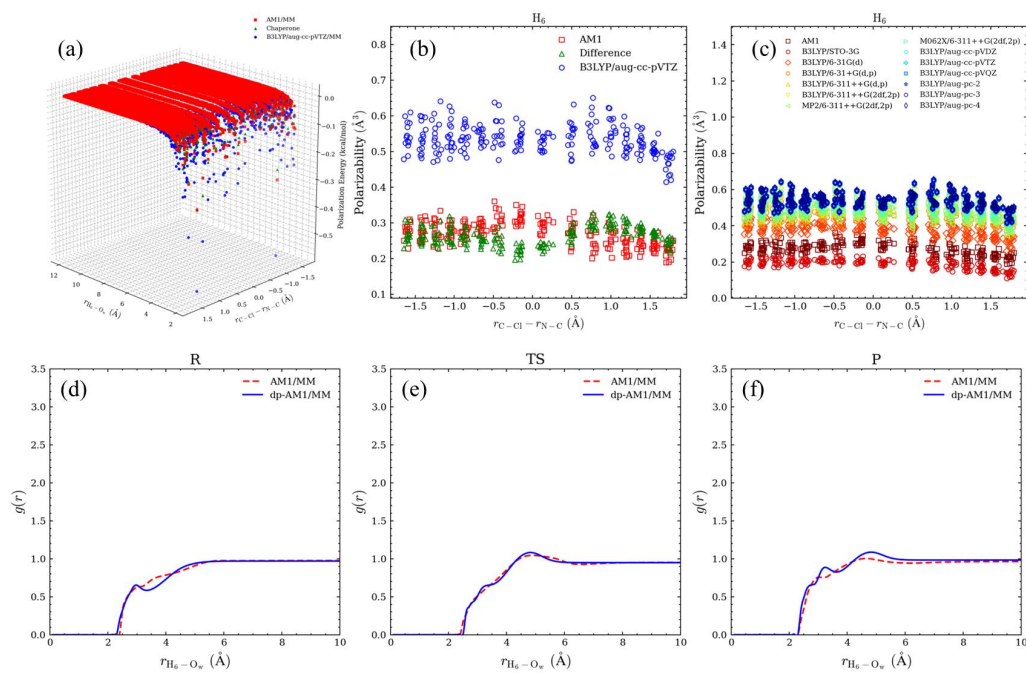
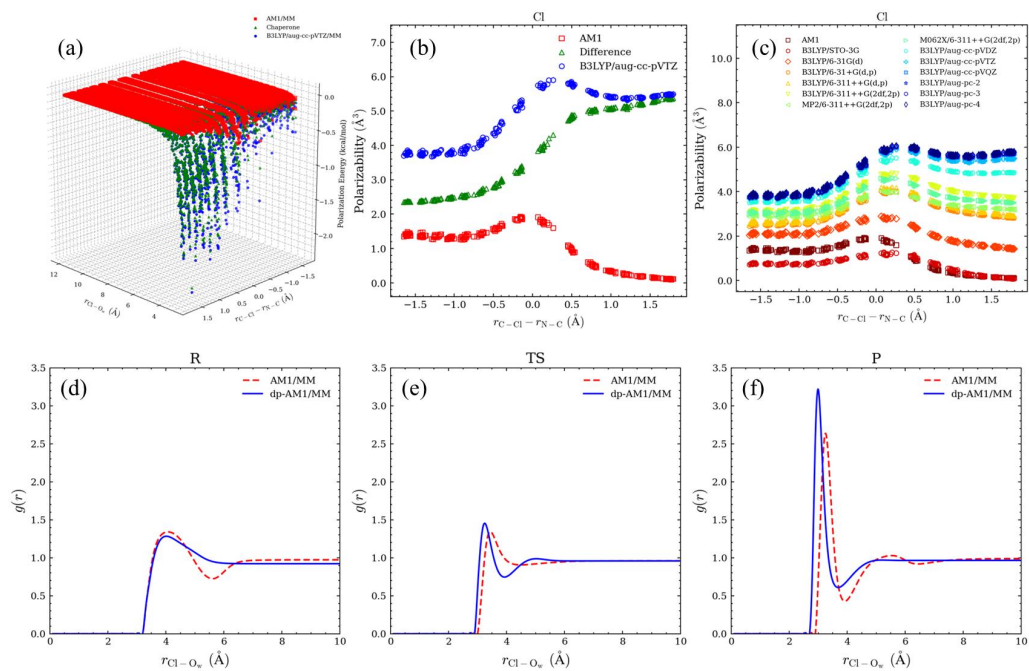


Figure S14. Polarization energy, polarizability, and radial distribution functions for the Cl atom



S10. Mulliken charges on QM atoms

In Figures S15 to S23, we plot the Mulliken charges on the QM solute atoms for the Menshutkin reaction based on 240 solution-phase configurations collected along three string MFEPs.

Mulliken charges on Nitrogen (N; Figure S15), on Hydrogens from NH_3 (H_{1-3} ; Figures S16-S18), on Carbon (C; Figure S19), on Hydrogens from CH_3Cl (H_{4-6} ; Figures S20-S22), and on Chlorine (Cl; Figure S23) computed at the AM1 and B3LYP/aug-cc-pVTZ levels with PCM implicit solvation as well as the differences between the two levels are plotted.

Among all the atomic charges, the Mulliken charge on the Cl atom undergoes the greatest change from $\sim -0.2 e$ in the reactant to $\sim -1.0 e$ in the product state (see Fig. S23), which reflects the formation of a chloride anion resulting from charge separation in the Menshutkin reaction. In other words, the evolution of the Mulliken charge on Cl is well synchronized with the rearrangement of chemical bonds in this case, and therefore this atomic charge may be used as an alternative reaction coordinate for this reaction. When we look at the charge differences between AM1 and B3LYP (also available in Figs. S15-S23), the Cl atom, however, does not need the greatest correction; the charge difference on Cl between the two levels is found in the range of 0 to $0.1 e$ along the reaction coordinate (Fig. S23), which is comparable to or even smaller than some of the differences we see on other heavy atoms in the system (e.g., N and C, shown in Fig. S15 and S19, respectively).

Interestingly, the overall shape of the Mulliken-charge-correction curve for Cl (Fig. S23) looks similar to the trend of the molecular-polarizability correction plotted in Fig. 1. The observed correlation suggests that the chaperone polarizabilities needed may be described equally well by York's charge-depend density-expansion CPE polarization model (Giese & York *JCP* **2005**, *123*, 164108), in which additional electric polarization response is constructed as a function of atomic charges.

Figure S15. Mulliken charges on the N atom

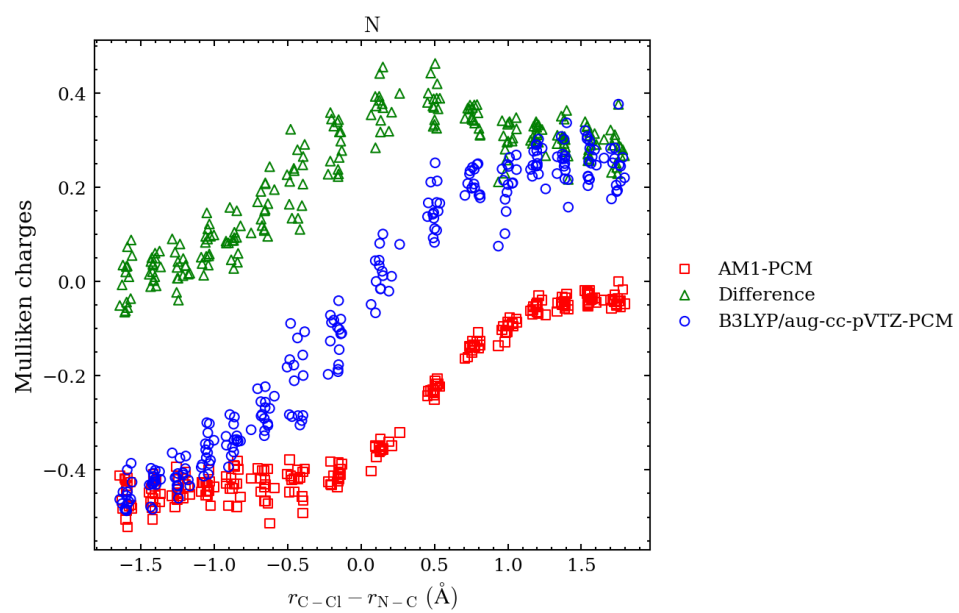


Figure S16. Mulliken charges on the H₁ atom

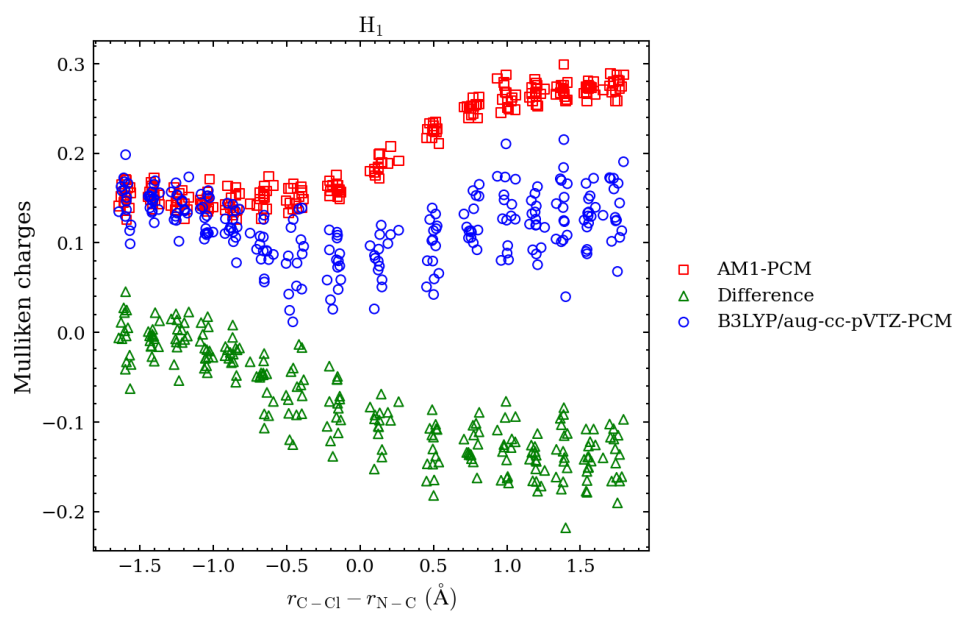


Figure S17. Mulliken charges on the H₂ atom

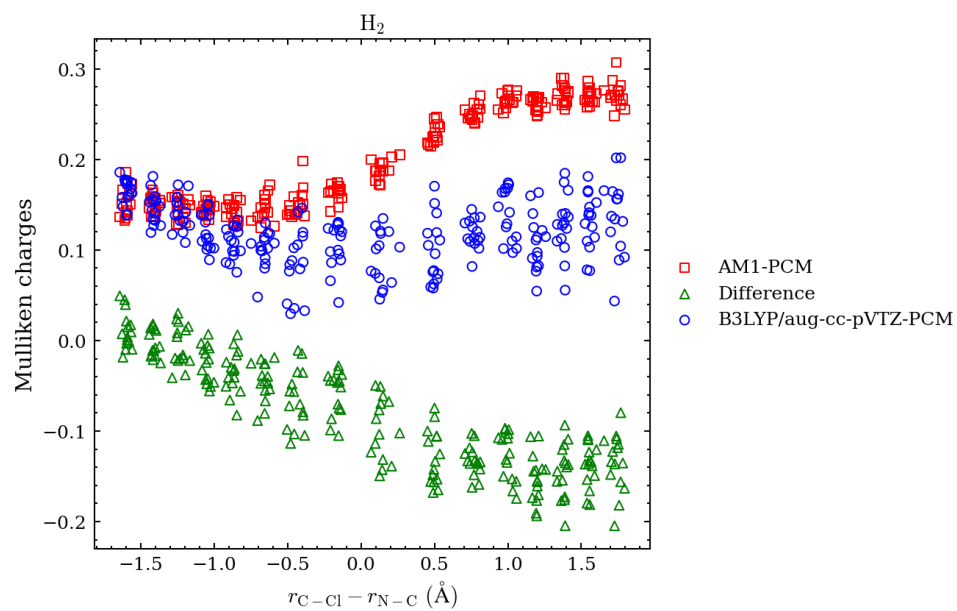


Figure S18. Mulliken charges on the H₃ atom

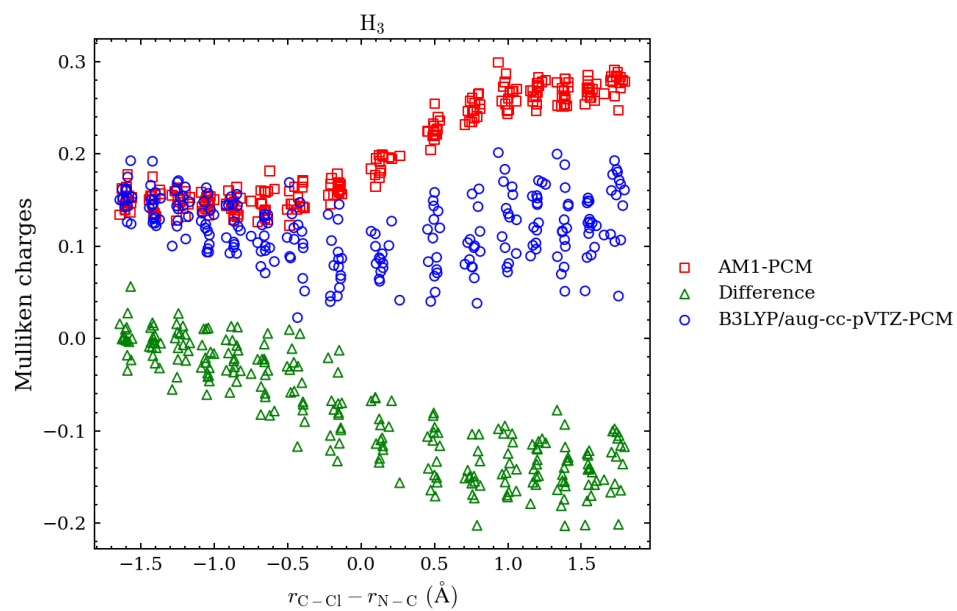


Figure S19. Mulliken charges on the C atom

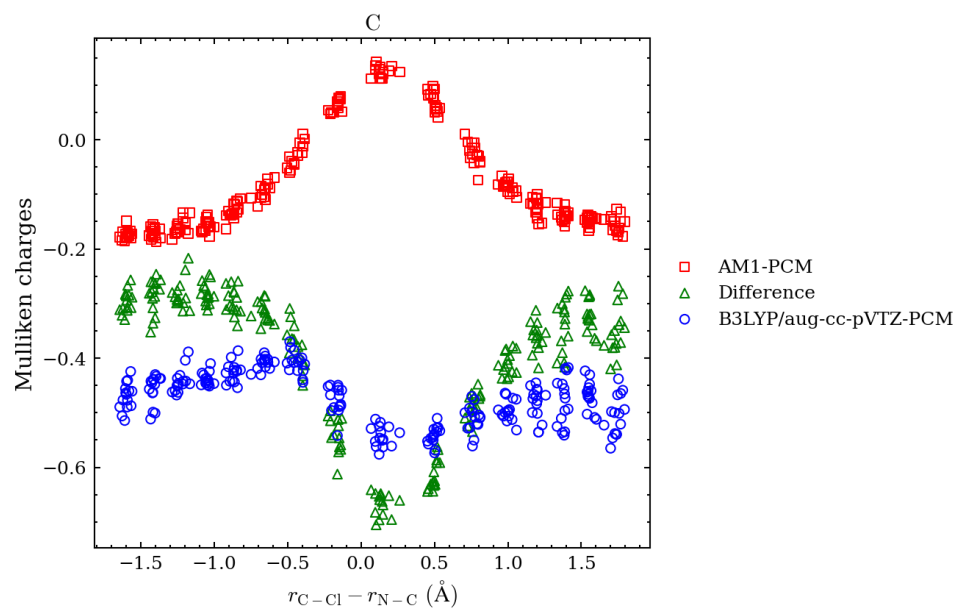


Figure S20. Mulliken charges on the H₄ atom

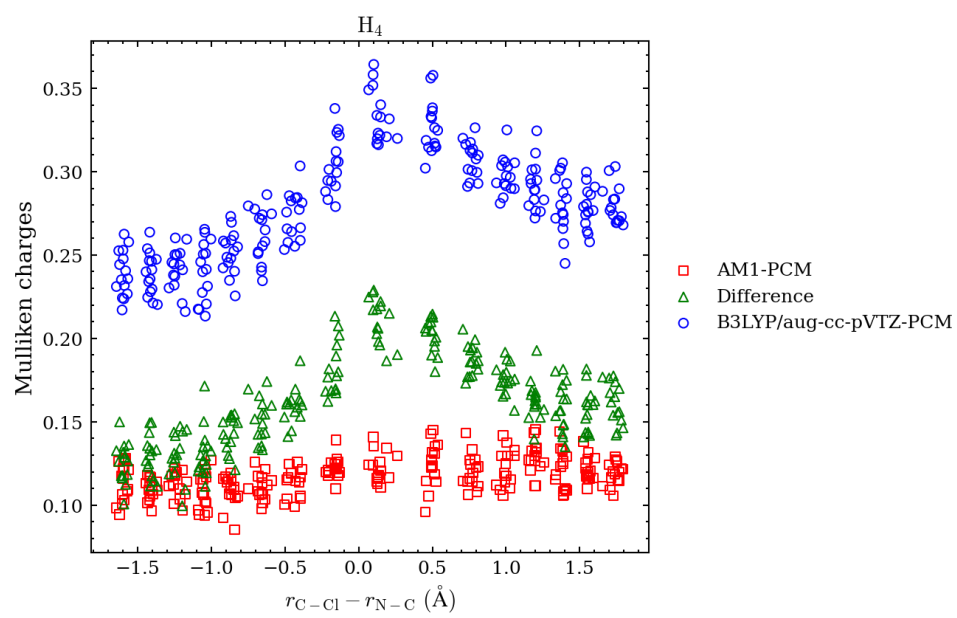


Figure S21. Mulliken charges on the H₅ atom

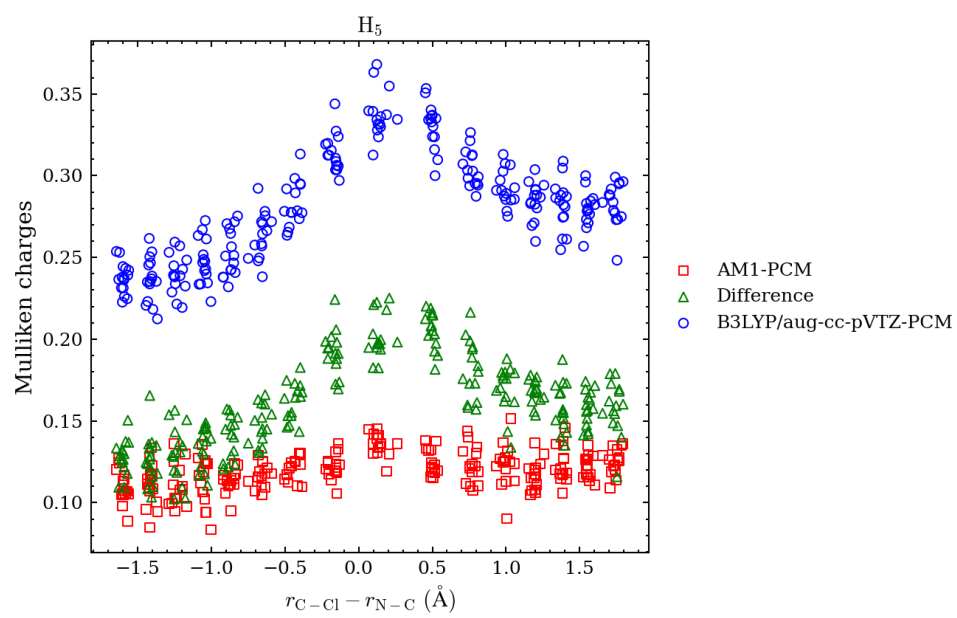


Figure S22. Mulliken charges on the H₆ atom

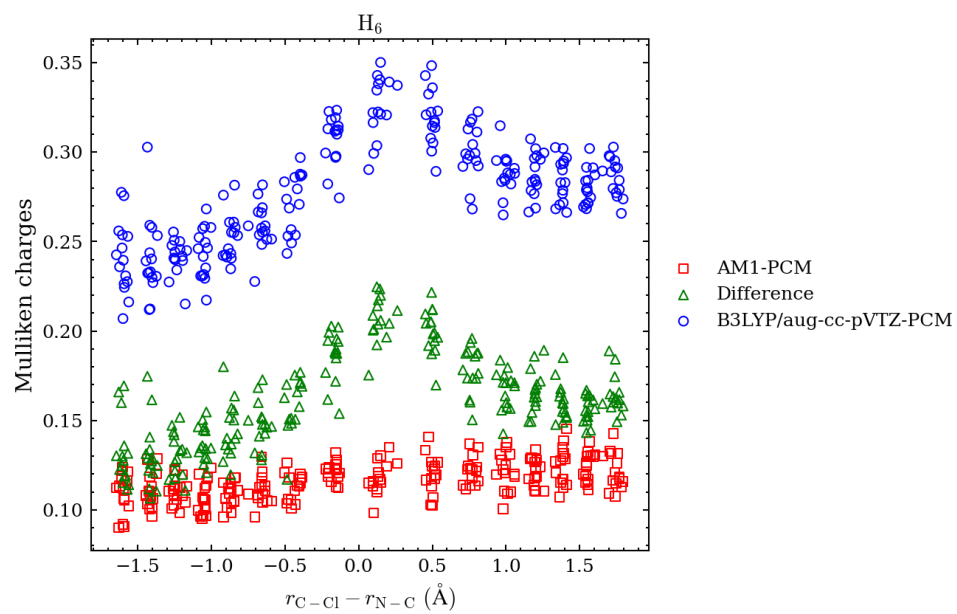
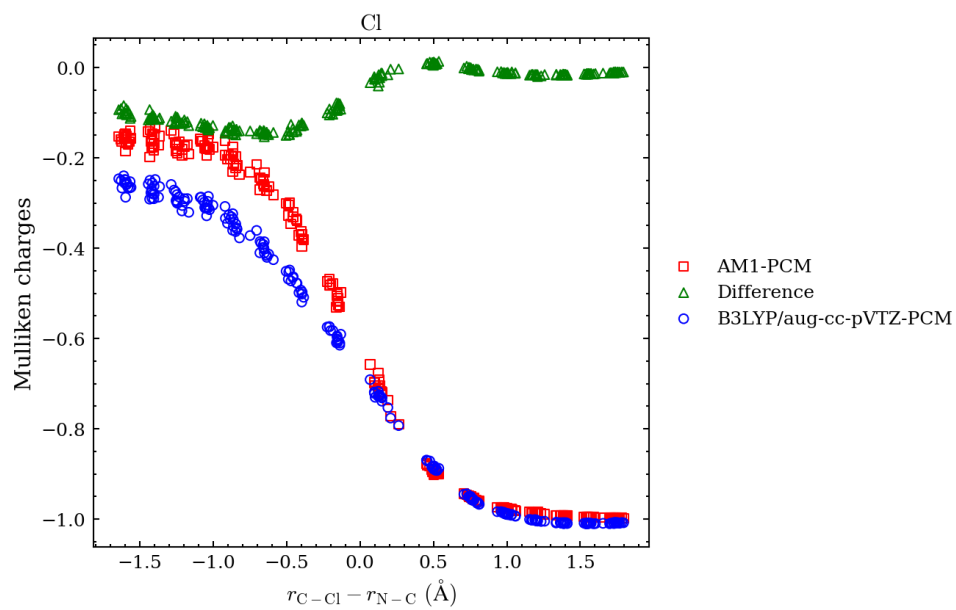


Figure S23. Mulliken charges on the Cl atom



PUBLICATION(S)

1. Zhou, Y., Ojeda-May, P., Nagaraju, M., **Kim, B.**, & Pu, J. (2018). Mapping Free Energy Pathways for ATP Hydrolysis in the E. coli ABC Transporter HlyB by the String Method. *Molecules*, *23*(10), 2652.
<https://www.mdpi.com/1420-3049/23/10/2652>
2. **Kim, B.**, Snyder, R., Nagaraju, M., Zhou, Y., Ojeda-May, P., Keeton, S., Hege, M., Shao, Y., & Pu, J. (2021, 2021/08/10). Reaction Path-Force Matching in Collective Variables: Determining Ab Initio QM/MM Free Energy Profiles by Fitting Mean Force. *Journal of Chemical Theory and Computation*, *17*(8), 4961-4980.
<https://doi.org/10.1021/acs.jctc.1c00245>
3. **Kim, B.**, Shao, Y., & Pu, J. (2021, 2021/12/14). Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability. *Journal of Chemical Theory and Computation*, *17*(12), 7682-7695.
<https://doi.org/10.1021/acs.jctc.1c00567>
4. Snyder, R., **Kim, B.**, Pan, X., Shao, Y., & Pu, J. (2022). Facilitating ab initio QM/MM free energy simulations by Gaussian process regression with derivative observations [10.1039/D2CP02820D]. *Physical Chemistry Chemical Physics*, *24*(41), 25134-25143.
<https://doi.org/10.1039/D2CP02820D>

ORAL PRESENTATION(S)

1. **Bryant Kim**. Determination of Accurate Polarizabilities with Machine Learning Approaches at the Atomic Scale." Literature Seminar - IUPUI. November 10, 2021. Indianapolis, IN.
2. **Bryant Kim** and Jingzhi Pu*. Free Energy Calculations for Solution Phase Reactions Using Artificial Intelligence." 2nd Chemistry Research Day Talk C&CB IUPUI. January 7, 2020. Indianapolis, IN.
3. **Bryant Kim** and Jingzhi Pu*. A QM/MM Solution Phase Study Using Multidimensional Reaction Path Force Matching in Collective Variables with Machine Learning." IUPUI Graduate Student Multidisciplinary Poster Symposium. August 16, 2019. Indianapolis, IN.
4. **Bryant Kim** and Jingzhi Pu*. Reaction Path-Force Matching in Collective Variables: Determining QM/MM free energy profiles for solution-phase reactions." IUPUI Graduate Student Multidisciplinary Poster Symposium. July 20, 2018. Indianapolis, IN.
5. **Bryant Kim** and Jingzhi Pu*. Reaction Path-Force Matching in Collective Variables: Determining QM/MM free energy profiles for solution-phase reactions." Chemistry Research Day Talk C&CB IUPUI. May 9, 2018. Indianapolis, IN.

POSTER PRESENTATION(S)

1. **Bryant Kim**, Yihan Shao, and Jingzhi Pu*. Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability. 265th ACS National Meeting. March 27-28, 2023. Indianapolis, IN.
2. **Bryant Kim**, Yihan Shao, and Jingzhi Pu*. Doubly Polarized QM/MM with Machine Learning Chaperone Polarizability. ACS - Think Like a Molecule Poster Session. April 1, 2022. Indianapolis, IN.
3. **Bryant Kim** and Jingzhi Pu*. A QM/MM Solution Phase Study Using Reaction-Path Force-Matching in Collective Variables with Machine Learning." Beyond Boundaries-Indiana Academies Symposium. April 10, 2021. Indianapolis, IN. (Grand Prize)
4. **Bryant Kim** and Jingzhi Pu*. A QM/MM Solution Phase Study Using Reaction-Path Force-Matching in Collective Variables with Machine Learning." ACS - Think Like a Molecule 2020 Virtual Poster Session (Virtual Presentation). April 18, 2020. Indianapolis, IN. (1st Place)
5. **Bryant Kim** and Jingzhi Pu*. Assessing Force Matching QM/MM Models for Computing Solution Phase Free Energy Profiles." ACS - Think Like a Molecule Poster Session. April 5, 2019. Indianapolis, IN. (3rd Place)
6. **Bryant Kim** and Jingzhi Pu*. Assessing Force Matching QM/MM Models for Computing Solution Phase Free Energy Profiles." Chemical Biology Departmental Poster Session. December 5, 2018. Indianapolis, IN.
7. **Bryant Kim**, Mulpuri Nagaraju, Yan Zhou, Pedro Ojeda-May, Seth Keeton, Mellisa Hege, and Jingzhi Pu*. Reaction Path-Force Matching in Collective Variables: Determining ab initio QM/MM free energy profiles for solution-phase reactions." ACS - Think Like a Molecule Poster Session. April 17, 2018. Indianapolis, IN. (1st Place)
8. **Bryant Kim**, Mulpuri Nagaraju, Yan Zhou, Pedro Ojeda-May, Seth Keeton, Mellisa Hege, and Jingzhi Pu*. Reaction Path-Force Matching in Collective Variables: Deter-

mining ab initio QM/MM free energy profiles for solution-phase reactions. 255th ACS National Meeting. March 26-27, 2018. New Orleans, LA.

VITA

Bryant Kim was born in the suburban town of Wayne, Pennsylvania in 1989. In 2003, he moved with his family to Carmel, Indiana where he spent the rest of his childhood. His parents, both immigrants from South Korea, instilled in him a strong work ethic and a passion for learning. After earning a Bachelor degree in chemistry from IUPUI, Bryant embarked on a career in the pharmaceutical industry. He worked at prominent companies such as Eli Lilly and Mylan Technologies, where he gained experience



in good manufacturing practices, wet lab analytical techniques, and documentation of lab work. Later, he worked at a contract research organization that supported dosing studies for animals and humans. Bryant's interest in computational chemistry led him to pursue a PhD in theoretical and computational chemistry. He is passionate about applying technology to the pharmaceutical industry and is committed to staying up-to-date with the latest computational approaches for solving chemical problems. In his free time, Bryant enjoys playing various sports such as basketball, soccer, baseball, and ultimate Frisbee. He also enjoys playing the guitar and piano whenever he can in addition to science fiction movies and books. One of Bryant's proudest achievements is his decision to change his career path and pursue a PhD in computational chemistry. With the support of his family, advisor, and friends, he overcame his doubts and pursued his passion. Bryant believes in the philosophy of "root to shoot," much like the roots which support a tree, emphasizing the importance of hard work and dedication in achieving success. He hopes to make a positive impact on the world by helping those in need, inspiring future generations, and using his computational expertise to leave a lasting legacy.