

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Designing Machine Learning-Enhanced Tools and Physics-based Techniques for Force Field and Electrostatic Models

Permalink

<https://escholarship.org/uc/item/4f06j6t0>

Author

Guan, Xingyi

Publication Date

2024

Peer reviewed|Thesis/dissertation

Designing Machine Learning-Enhanced Tools and Physics-based Techniques for Force
Field and Electrostatic Models

by

Xingyi Guan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Teresa Head-Gordon, Chair

Professor David Limmer

Professor Ali Mesbah

Summer 2024

Designing Machine Learning-Enhanced Tools and Physics-based Techniques for Force
Field and Electrostatic Models

Copyright 2024
by
Xingyi Guan

Abstract

Designing Machine Learning-Enhanced Tools and Physics-based Techniques for Force Field and Electrostatic Models

by

Xingyi Guan

in Chemistry

University of California, Berkeley

Professor Teresa Head-Gordon, Chair

In recent years, the landscape of molecular science has been profoundly transformed by the integration of data-driven methodologies alongside traditional deterministic and stochastic approaches. Historically, the study of molecular behavior and interactions relied heavily on deterministic algorithms, which follow a fixed sequence of computational steps to simulate molecular dynamics, and stochastic simulations, which incorporate randomness to explore various molecular states and pathways. These methods were complemented by physical models grounded in the established principles of chemistry and physics, forming the backbone of theoretical molecular science. However, these conventional approaches often faced limitations in scalability, computational cost, and generalizability for complex systems. The improvements in computational hardware, coupled with the accumulation of vast amounts of molecular data, have enabled the development of models that can surpass traditional methods in both accuracy and efficiency, leveraging both physics-based and machine learning (ML) approaches. This dissertation focuses on the development of new models utilizing more accessible data, provides guidelines for computational data generation, and explores the synergy between data acquisition strategies and data-driven models. These studies demonstrate that by carefully designing data acquisition strategies and integrating data-driven models with physics-based approaches, it is possible to enhance the predictive capabilities of computational methods in chemistry, particularly in force field development and electrostatic modeling. Through a series of studies, this work illustrates the potential of combining the strengths of both traditional and modern computational techniques to achieve more accurate and efficient predictions in molecular science.

The accurate prediction of electrostatic interactions is a critical aspect of understanding molecular behavior. The electrostatic potential (ESP) is a property of great research interest for understanding and predicting electrostatic charge distributions that drive interactions between molecules. However, traditional approaches often rely on detailed quantum mechanical calculations, which can be computationally expensive. In Chapter 2, I introduce a coarse-grained electron model (C-GEM), whose parameters are fitted to computationally generated high-quality Density Functional Theory (DFT) data, that offers a balance between accuracy and computational efficiency. Extensive validation against high-level quantum mechanical calculations demonstrates that C-GEM can reliably predict electrostatic potentials and interaction energies in proteins. The model’s implementation in large-scale molecular simulations shows significant reductions in computational cost, making it a viable tool for studying complex biological systems.

The generation of reference data for deep learning models poses significant challenges for reactive systems, especially for combustion reactions due to the extreme conditions that produce radical species and alternative spin states. In Chapter 3, intrinsic reaction coordinate (IRC) calculations are extended with *ab initio* molecular dynamics (MD) simulations and normal mode displacement calculations to comprehensively map the potential energy surface (PES) for 19 reaction channels involved in hydrogen combustion. This extensive dataset comprises approximately 290,000 potential energies and 1,260,000 nuclear force vectors, evaluated using a high-quality range-separated hybrid density functional, ω B97X-V. The dataset includes detailed information on transition state configurations as well as geometries along the reactive path way that links reactant to product, providing a robust reference for training deep learning models aimed at studying hydrogen combustion reactions. This benchmark dataset not only serves as a valuable resource for understanding the intricate mechanistic pathways of hydrogen combustion but also provide a paradigm for building dataset that facilitates the development and validation of machine learning models for reactive chemistry.

Building on the extensive benchmark dataset for hydrogen combustion detailed in Chapter 3, an initial machine learning model is trained to predict energies and forces for hydrogen combustion reactive system using NewtonNet, a physics inspired equivariant message passing neural network(MPNN). This reactive gas phase chemistry network is particularly challenging due to the need for comprehensive potential energy surfaces that accurately represent a wide range of molecular configurations. Traditional approaches often rely on chemical intuition to select training data, which can result in incomplete PESs in an ML setting. To address this challenge, I employ an

active learning strategy to systematically explore diverse energy landscapes using metadynamics simulations and continuously adding unseen data for retraining, helping to create a ML model that avoids unforeseen high-energy or unphysical configurations. By integrating metadynamics, the active learning process more rapidly converges the PES, also allowing a hybrid of ML and *ab initio* molecular dynamics (MD) that initiates rare calls to external *ab initio* sources when discrepancies are detected by the query by committee models. This hybrid ML-physics approach reduces computational costs by two orders of magnitude and eliminates the need for excessive ML retraining. The enhanced model accurately predicts free energy changes and transition state mechanisms for several hydrogen combustion reaction channels, demonstrating the efficacy of combining advanced data acquisition strategies with robust ML techniques to achieve high precision and efficiency in molecular simulations.

To summarize, this dissertation underscores the potential of combining data-driven models with physics-based approaches to overcome the limitations of traditional computational methods in molecular science. Through the development of the coarse-grained electron model (C-GeM), the creation of a comprehensive benchmark dataset for hydrogen combustion, and the implementation of an active learning workflow for reactive force field development, insights are provided in developing new computational tools and leveraging them to better understand molecular interactions and reactivity.

Contents

Contents	i
List of Figures	iii
List of Tables	v
1 Introduction	1
1.1 Overview	1
1.2 Modeling Electrostatics through the Course Grained Electron Model .	4
1.3 Building a Benchmark Dataset for Hydrogen Combustion	6
1.4 Active Learning for Machine Learning Force Field Development . . .	8
1.5 REFERENCES	9
2 Protein C-GeM: A coarse-grained electron model for fast and accurate protein electrostatics prediction [†]	19
2.1 INTRODUCTION	19
2.2 THEORY	23
2.3 METHODS	25
2.4 RESULTS AND DISCUSSIONS	29
2.5 CONCLUSIONS AND OUTLOOK	41
2.6 ACKNOWLEDGMENTS	42
2.7 REFERENCES	42
Appendices	50
2.A Results on individual molecules	50
2.B Conformations of LEU_TYR_GLN tripeptide	57
2.C APBS parameters	57
3 A benchmark dataset for Hydrogen Combustion[†]	58
3.1 BACKGROUND & SUMMARY	58

3.2	METHODS	60
3.3	TECHNICAL VALIDATION	62
3.4	DATA RECORDS	66
3.5	USAGE NOTES	66
3.6	DATA AND CODE AVAILABILITY	66
3.7	ACKNOWLEDGMENTS	66
3.8	REFERENCES	67
Appendices		71
3.A	Supplementary Figures	71
4	Using machine learning to go beyond potential energy surface benchmarking for chemical reactivity[†]	76
4.1	INTRODUCTION	77
4.2	RESULTS	78
4.3	DISCUSSIONS	92
4.4	METHODS	93
4.5	DATA AVAILABILITY	102
4.6	CODE AVAILABILITY	102
4.7	ACKNOWLEDGMENTS	102
4.8	AUTHOR CONTRIBUTIONS STATEMENT	103
4.9	REFERENCES	103
Appendices		110
4.A	Proof of Equivariance and Invariance	110
4.B	Supplementary Figures	111
4.C	Supplementary Tables	112

List of Figures

1.1	Data, feature representation, and machine learning methods are three fundamental components of a machine learning study, and their interplay crucially decides the ability to make chemical prediction [†]	2
1.2	A schematic illustration of how C-GeM model initialize core and shells on atom and predict electrostatic potential for a protein.	5
2.1	Schematic illustration of how C-GeM generates the electrostatic potential from given molecular geometry.	22
2.2	a) The mean ESP generated with Gaussian charges aligns perfectly with that generated with point charges. b) The time to compute ESP from core and shell positions with respect to number of grid points for different molecules using point charge and Gaussian charge treatment.	28
2.3	Average mean absolute error electrostatic potential and average dipole error of different atom typed C-GeM models, EEM, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V/def2-qzvpp reference for a) 54 small protein analogs and b) 57 tripeptides, labeled with the average error and standard deviation within the set. . .	32
2.4	Demonstration for C-GeM on charged molecules a) Methylammonium (net -1 charge) b) Tripeptide ARG-LYS-ILE (net +2 charge)	36
2.5	Mean absolute error (MAE) on electrostatic potential (ESP) and dipole error of different atom typed C-GeM models, AM1-BCC, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for a) training set for charged side chains b) testing set for charged side chains.	38
2.6	Predicted ESP figure for crambin(1CRN) with ω B97X-V / cc-pVDZ, HF/6-31G*, EEM, CGem, CGem_CH, CGem_CHN and APBS.	40
2.B.1	Six conformations of LEU_TYR_GLN tripeptide	57
3.3.1	Potential energy surface for the hydrogen transfer reaction 2	63

3.3.2 Representative potential energy surfaces for oxygen transfer, association, and substitution reactions along two reaction coordinates CN1 and CN2.	64
3.3.3 The changes in the PES for reaction channel 12 involving changes in spin state.	65
3.A.1 Additional association/dissociation reaction channels	72
3.A.2 Additional oxygen transfer reaction channel	73
3.A.3 Additional hydrogen transfer reaction channels	74
3.A.4 Additional hydrogen transfer reaction channels (continued)	75
4.2.1 The learning curve of NewtonNet for the hydrogen combustion data	79
4.2.2 Observations of the missing data in the machine-learned model for hydrogen combustion and the addition of dilation data.	82
4.2.3 Schematic illustration of active learning workflow using query by committee and metadynamics.	84
4.2.4 Potential energy surface in collective coordinates (left) and change in energy and force mean absolute error (MAE) as active learning round proceeds (right) for Reaction 18.	86
4.2.5 Schematic illustration of the new workflow for rebuilding the free energy surface.	90
4.2.6 Free energy surface reconstructed from metadynamics using the hybrid model for hydrogen combustion.	91
4.4.1 (a) Newton's laws for the force and displacement calculations for atom i with respect to its neighbors. (b) Schematic view of the NewtonNet message passing layer.	95
4.B.1 Two representative structures that the original ML model predicts with large error.	111
4.B.2 Spot checking the hybrid mode model for Rxn18 for energies and forces.	112

List of Tables

2.1	Parameters for C-GeM models CGem, CGem_CH and CGem_CHN. . . .	30
2.2	Computation time per molecule of C-GeM, CH atomtyped C-GeM, CHN atomtyped C-GeM on small protein analogs and tripeptides.	34
2.3	Mean absolute error (MAE) in eV on electrostatic potential (ESP) of different atom typed C-GeM models with respect to ω B97X-V / def2-qzvpp reference on 6 conformations of tripeptide LEU_TYR_GLN.	35
2.A.1	MAE and RMSE on ESP and dipole error for a) 54 small protein analogs b) 57 tripeptides.	50
2.A.2	MAE and RMSE on ESP and dipole error for charged tripeptide training and testing set	51
2.A.3	MAE on electrostatic potential (ESP) for CGem, CGem_CH, CGem_CHN, AM1-BCC, EEM , Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for individual molecules in small protein analogs set.	52
2.A.4	MAE on electrostatic potential (ESP) for CGem, CGem_CH, CGem_CHN, EEM , Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for individual molecules in tripeptides set.	53
2.A.5	MAE on electrostatic potential (ESP) for CGem, CGem_CH, CGem_CHN, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for individual molecules in charged train set.	55
2.A.6	MAE on electrostatic potential (ESP) for CGem, CGem_CH, CGem_CHN, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for individual molecules in charged test set.	56
3.2.1	Data Summary for the Potential Energy Surface of Hydrogen Combustion.	61

4.2.1 Committer statistics at 500K with (a) the original model (b) the final model after active learning with IRC dilation and active dynamics. . . .	88
4.C.1 The 19 reactions contained in the hydrogen combustion benchmark dataset.	113
4.C.2 Metadynamics collective variables used in the active learning and free energy reconstruction.	114
4.C.3 Total number of data points added in active learning for each reaction. .	114
4.C.4 AIMD Committer Analysis on identified Free Energy Transition State from the hybrid model at 500K.	115

Acknowledgments

As I reflect on the five years I have spent in the Chemistry Ph.D. program at UC Berkeley, I am very grateful for the numerous people and experiences that have profoundly impacted my journey. This dissertation is the culmination of their unwavering support, guidance, and encouragement.

First and foremost, I would like to express my deepest gratitude to my principal investigator, Teresa Head-Gordon, whose guidance, support, and invaluable insights have been instrumental in the successful completion of this dissertation. Your constant belief in my abilities and your mentorship have profoundly shaped my academic journey and professional growth.

I would also like to extend my heartfelt thanks to the members of my research group. Dr. Mojtaba Haghighatlari and Dr. Jie Li led me into the world of machine learning for chemistry. Dr. Farnaz Heidar-Zadeh and Dr. Itai Leven provided invaluable assistance in initiating the C-GeM project. I thoroughly enjoyed working with Oufan Zhang, Oliver Sun and Eric Wang on the drug discovery project. Although the work is not included in this thesis, it played a significant role in helping me getting opportunities for my future career in industry. Furthermore, I would like to extend my many thanks to Dr. Wanlu Li, Dr. Hongxia Hao, Dr. Jagna Witek, Dr. Meili Liu, Eric Yuan and many other group members for the exciting projects we have been working together and the enjoyable discussions we have had in the lab. I would also like to thank my collaborators: Dr. Martin Head-Gordon, Dr. Samuel M. Blau, Dr. Rommie Amaro, and other collaborators from the Anti Viral Drug Discovery Center, as well as their group members. The stimulating discussions, feedback, and collective efforts have made this journey fruitful and rewarding.

I want to say thank you to my friends: Di Gu, Dian Guo, Eve Xu, Ruoxu Xia, Siyuan Niu, Hanyao Tian, Yijia Cui and many others. Together, we have shared unforgettable trips, meals, concerts, and games, bringing vibrancy and joy to my grad school life. We supported each other through the challenges of the pandemic, and you have broadened my perspective on the world.

Finally, and importantly, I want to express my deepest gratitude to my family, especially my parents, Min Zhang and Huaming Guan. Your constant support allowed me to study abroad, and you granted me the freedom to make every important decision in my life. You provided invaluable mental support during difficult times and have always been there for me. Accomplishing this would not have been possible without your unyielding encouragement and love.

Thank you all for your contributions and support, without which this dissertation would not have been possible.

Chapter 1

Introduction

1.1 Overview

Theoretical studies of molecular behavior and interactions have primarily relied on a combination of deterministic algorithms, stochastic simulations, and physical models of the potential energy surface. Deterministic algorithms, such as molecular dynamics (MD), [17, 44] and stochastic simulations, including Monte Carlo methods, [42, 8] are ways to explore the configurational space of molecules and study thermodynamic properties, time-correlation observables, and rare events [8, 40]. These techniques are used to explore the potential energy surface, which can be described by solving the Schrödinger equation for electrons. [76] The combination of *ab initio* methods with molecular dynamics (AIMD) is often a powerful approach for simulating chemical reactivity, [52] but the high computational cost of wavefunction quantum mechanical calculations scales poorly with system size, [32] and lower-cost Density Functional Theory (DFT) [39] is still often impractical for very large molecules or extensive conformational sampling. Physical force fields, while efficient and sometimes quite accurate for thermodynamics and transport properties, [43] are poorly suited to the accuracy needed to capture electronic effects, and popular pairwise additive models fail to generalize well to different chemical environments. [59]

These traditional approaches have formed the backbone of theoretical molecular science, and must balance tradeoffs of computational cost, scalability, and accuracy, particularly when applied to large and complex systems [17]. In Chapter 2 of my dissertation, I will discuss a new electrostatic model that achieves this balance in predicting the electrostatic potential of protein systems. [25] First, I will give an overview of the past efforts to model electrostatic potential surfaces and introduce the coarse-grained electron model (C-GeM) [46, 25], with a detailed explanation of how

the model works and how the model parameters are fitted with selected tripeptide data. This model represents a significant advancement in the efficient and accurate prediction of electrostatic interactions in proteins, peptides, and small molecule drugs.

Recently, modern data-driven methods, particularly those based on machine learning (ML), have started to infiltrate various fields of molecular science[63, 74, 67, 16, 83, 27, 5, 4, 14]. ML models rely on three critical components: data, feature representation, and model architecture, and the interplay of these three factors crucially decides the ability to make chemical prediction, as demonstrated in Figure 1.1.

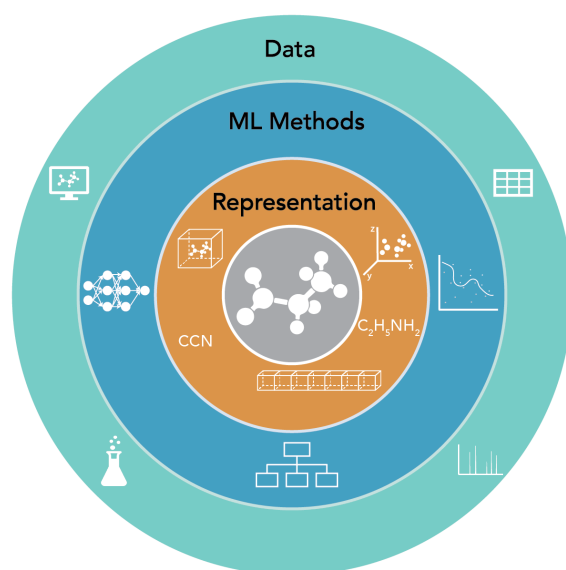


Figure 1.1: Data, feature representation, and machine learning methods are three fundamental components of a machine learning study, and their interplay crucially decides the ability to make chemical prediction[†]

The potential of ML in chemistry was first highlighted by Behler and Parrinello’s work on neural network representations of PES,[7] inspiring a wave of research focused on improving the accuracy and efficiency of these ML models and methods. Graph convolution networks (GCNs) have been used extensively for their ability to handle the connectivity of molecules.[37, 20] The first message passing neural network (MPNN)

[†]Figure reproduced with permission from: Haghghatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **2020**, 6 (7), 1527–1542. <https://doi.org/10.1016/j.chempr.2020.05.014>.

for molecular systems introduced a framework where information is passed along the edges of a graph, effectively capturing the relationships between atoms.[20] SchNet is another notable model that employs continuous-filter convolutional layers to learn molecular representations from data, providing high accuracy for molecular property predictions.[66, 67] In recent years, equivariant models such as PaiNN[68], NequIP[18, 5], and NewtonNet[29] further leverages the principles of equivariance to rotations, translations, and permutations, improving the performance on tasks involving 3D molecular structures.

In addition, the dataset often sets the limit for model performance because ML models interpolate within the space defined by their training data.[66, 82] High-quality and comprehensive datasets are essential for the success of ML models, especially for complex reactive systems.[48, 5] Currently, most ML datasets used for model development are non-reactive, focusing on stable molecular configurations. Prominent examples of common ML model benchmarks include the ANI dataset,[74] which provides a large number of molecular conformations for organic molecules, the MD17 dataset,[11] which offers molecular dynamics trajectories for a small set of organic molecules, and the GDB-17 and QM9 datasets,[62, 60] which contain quantum chemistry calculations for small organic molecules. These datasets have been instrumental in advancing ML models but do not address the complexities of reactive systems.

The hydrogen combustion dataset presented in my thesis work serves as a crucial benchmark for ML model development in reactive chemistry. It captures a wide range of configurations, including high-energy transition states and metastable intermediates, under realistic reaction conditions. This dataset will significantly enhance the training and validation of ML models designed to predict energies and forces in reactive systems, enabling more accurate simulations of combustion and other chemical processes. However, compared with ML in other field such as computer vision and natural language processing, the amount of chemical data is still limited,[49] making it necessary to continue generating new data and explore data efficient ML architectures.

In this introduction chapter of my dissertation I will also describe my research work at the frontline of combining data-driven models with physics-based approaches for ML. I will describe the creation of a comprehensive benchmark dataset for hydrogen combustion,[26, 22] highlighting the challenges of generating high-quality reference data for reactive systems and how this dataset facilitates the development of accurate ML models. Finally, I will discuss the implementation of an active learning workflow for ML-physics force field development,[24] emphasizing the importance of combining ML with traditional *ab initio* methods to achieve robust and efficient molecular simulations of chemically reactive systems.

1.2 Modeling Electrostatics through the Course Grained Electron Model

Electrostatic interactions are crucial for understanding and predicting the behavior of molecules in various chemical and biological contexts.[71, 73] The electrostatic potential (ESP) describes the distribution of electric charges in a molecule, influencing how molecules interact with each other. These interactions are vital for numerous phenomena, including protein-ligand binding, enzyme catalysis, molecular recognition, and material properties. In biochemistry, the ESP is essential for determining the binding sites of proteins and their interactions with ligands or other proteins, which is critical for drug design and understanding cellular processes.[53] In materials science, electrostatics are key to the function of nanoporous materials like zeolites and metal-organic frameworks, which are used for gas storage and separation.[47] In electrochemistry, the efficiency of electrochemical cells relies on the accurate modeling of ion diffusion and double-layer formation at electrode surfaces.[65].

Given the central role of electrostatics, accurate prediction and modeling of the ESP are vital for many applications. Traditional methods for calculating ESP include quantum mechanical calculations, which offer high accuracy but are computationally intensive and limited to small systems.[15] This has led to the development of various models and methods to approximate ESP with greater computational efficiency while maintaining reasonable accuracy, each with its strengths and limitations.[12]

- **Electrostatic Potential Fitted Charges (EPFC):** These approaches involve fitting partial charges on atoms to reproduce the QM-calculated ESP. Methods such as the CHELPG (Charges from Electrostatic Potentials using a Grid) and RESP (Restrained Electrostatic Potential) are commonly used.[9, 6] While these methods can accurately reproduce ESP, they are computationally expensive as they require initial QM calculations.[15]
- **Empirically Derived Charges (EDC):** Models like the AM1-BCC (Austin Model 1-Bond Charge Corrections) [35, 36] use semi-empirical methods and empirical corrections to derive atomic charges. These are less accurate than QM-based methods but more computationally efficient.
- **Density-based Quantum Mechanical Partitioning Techniques:** Techniques such as Mulliken, Hirshfeld, and Bader's Atoms in Molecules (AIM) partition the electron density to assign charges to atoms.[57, 33, 2] These methods provide a physically grounded way to derive atomic charges but still rely on QM calculations, making them computationally demanding for large

systems. More modern methods based on similar principle such as the Additive Variational Hirshfeld (AVH)[30, 31] and the Minimal Basis Iterative Stockholder (MBIS)[79] have been developed to further improve accuracy.

- **Charge Equilibration Methods (CEM):** Methods such as the Electronegativity Equalization Method (EEM) predict partial charges based on atomic electronegativity and hardness.[55] While fast and suitable for large-scale applications, they often suffer from inaccuracies such as unphysical long-range charge transfer and poor representation of out-of-plane polarization.

Each of these methods has been instrumental in advancing our understanding of molecular electrostatics, but they often fall short when applied to large or complex systems due to computational constraints or inherent inaccuracies. The Coarse-Grained Electron Model (C-GeM) [46, 25] represents a significant advancement in the efficient and accurate prediction of electrostatic interactions in molecules. Unlike traditional methods, C-GeM coarse grains the representation of electron density by modeling atoms with a positive core and a negatively charged electron shell, described by Gaussian distributions. This allows C-GeM to capture essential electrostatic properties without the need for computationally expensive QM calculations, and allow it to access larger system such as protein-ligand complexes as shown in Figure 1.2.

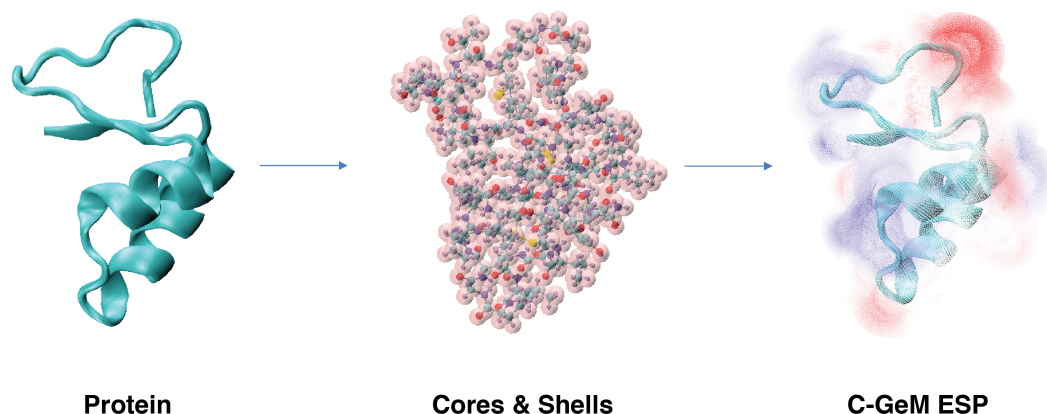


Figure 1.2: A schematic illustration of how C-GeM model initialize core and shells on atom and predict electrostatic potential for a protein.

The accuracy of any electrostatic model heavily depends on the quality and comprehensiveness of the data used for its development and validation. For physics-based models, data must accurately capture the underlying physical interactions, which often involves high-quality QM calculations of the ESP for a diverse set of molecules, covering various chemical environments and configurations.[15] In the development of C-GeM, I curated a dataset of small protein analogs and tripeptides fragmented from real protein PDB structures, and utilized high-fidelity DFT calculations with ω B97X-V functional[51] and def2-QZVPP basis set[80]. The protein C-GeM model is trained and validated with this dataset and further validated with larger protein to ensure accurate model prediction of protein ESP.

Protein C-GeM was benchmarked against a series of common methods including the electronegativity equalization methods (EEM)[56, 55], the original Hirshfeld[33] method, Iterative Hirshfeld (HI)[10], Minimal Basis Iterative Stockholder (MBIS)[79] and Additive Variational Hirshfeld (AVH)[30, 31], and it is found to provide comparable accuracy to *ab initio* charge partitioning methods but with orders of magnitude improvement in computational efficiency, making it suitable for large-scale molecular simulations. Moreover, C-GeM can accurately describe complex electrostatic phenomena such as sigma holes and out-of-plane polarization, which are often inadequately represented in simpler models that are comparable in speed.

Given its computational efficiency, C-GeM can be applied to large biological systems like proteins, making it a valuable tool for biochemistry and drug design. The model can be integrated with molecular dynamics (MD) simulations to provide real-time updates to the electrostatic potential, facilitating more accurate and dynamic simulations of molecular interactions.[45]

Beyond protein ESP prediction, C-GeM also offers a robust and efficient approach to modeling electrostatics in molecular systems in general, bridging the gap between high accuracy and computational feasibility. Fitting the model to energy decomposition analysis (EDA)[38] data can improve upon description of molecular interactions, and enable wider applications in more areas such as protein-ligand docking. By leveraging high-quality training data and innovative modeling techniques, C-GeM provides a powerful tool for advancing molecular ESP representations.

1.3 Building a Benchmark Dataset for Hydrogen Combustion

Accurate modeling of reactive systems poses significant challenges, especially in capturing data near transition states. In hydrogen combustion, these challenges are

compounded by the extreme conditions of high temperature and pressure, leading to the formation of radical species and alternative spin states.[19, 72, 78] In Chapter 3, I provide an approach to create a comprehensive dataset for reactive systems using hydrogen combustion as a prototype example. The dataset aims to facilitate the development of accurate deep learning models for predicting molecular energies and forces, crucial for understanding and optimizing reaction processes.

The PES data for hydrogen combustion were organized into four categories based on the reaction mechanisms involved in the elementary steps: association/dissociation reactions, substitution reactions, oxygen transfer, and hydrogen transfer. This organization facilitates targeted studies of different reaction types and their respective energy landscapes, especially transition states that represent high energy points along a reaction pathway and are pivotal in determining reaction kinetics and mechanisms. However, capturing data near transition states is challenging due to their fleeting nature and high energy, and yet are essential for constructing reliable potential energy surfaces that can be used in machine learning models and molecular simulations.[77, 54, 64] This work addresses this challenge by systematically collecting data off the intrinsic reaction coordinate (IRC)[34] using short ab initio molecular dynamics (AIMD) and normal mode sampling, which ensures that important geometries near the reaction pathway are thoroughly explored for 19 reaction channels of hydrogen combustion.

The short AIMD simulations were performed to sample configurations around the IRC structures, starting from the transition state as the initial configuration for each reaction channel. Simulations were conducted at four different high temperatures (500 K, 1000 K, 2000 K, and 3000 K), generating configurations that capture the dynamic behavior of the system under realistic reaction conditions. In addition to AIMD, normal mode displacement calculations [70, 61] were performed to systematically sample geometries along the IRC. Starting from each IRC structure, vibrational frequencies were calculated, and atoms were displaced along each normal mode to generate additional configurations. This method helps diversify the dataset by including configurations that compress or expand the IRC structures, capturing a wide range of molecular geometries and ensuring comprehensive coverage of the PES. The final data set includes approximately 290,000 potential energies and 1,270,000 nuclear force vectors, evaluated using a high-quality range-separated hybrid density functional, ω B97X-V.[50] This level of theory is known for its accuracy in describing thermochemistry and reactive barriers, making it suitable for capturing the intricate details of hydrogen combustion reactions.[21]

The approach we take to build the hydrogen combustion dataset underscores the importance of systematic data collection in building reliable datasets for reactive systems. However, there are more considerations in the data creation for a ML force

field, which requires a more thoroughly covered data distribution and an inclusion of unphysical geometries, which will be discussed in detail in Chapter 4.

1.4 Active Learning for Machine Learning Force Field Development

Building upon the comprehensive dataset established in Chapter 3 for hydrogen combustion, in Chapter 4 I delve into the application of using this data to develop a machine learning force field (MLFF) for hydrogen combustion. The methods and strategies employed in this study provide a blueprint for future efforts in the field, highlighting the potential of combining advanced computational techniques with high-quality data to advance our understanding of complex chemical reactions. The dataset acquisition strategy introduced in this dissertation also allows us to explore building a complete ML-physics reactive force field for the hydrogen combustion system, to use the model to perform commitor analysis, and to drive metadynamics simulations to determine free energy surfaces of chemical reactivity.[24]

In this work, I utilized our physics-inspired NewtonNet model we published in 2021[28] and 2022[29]. NewtonNet took inspiration from physical principles into its architecture to learn interatomic potentials and forces. It has demonstrated strong predictive capabilities for the energy and forces in the hydrogen combustion system. However, challenges arise when using any ML model for molecular dynamics (MD) simulations, where the ML model can sometimes produce unphysical configurations, a phenomenon known as model hallucination. Machine learning is not physics, and thus ML models require extensive and diverse datasets because they do not inherently understand the underlying physics and must learn all aspects of the system’s behavior from the data.[81] Consequently, ML models need not only ”physical” data along the reactive pathway but also data that seem unreasonable to learn that these are high energy states cannot be accessed at finite temperature and pressure. ML models can also develop ”holes” in the PES, where predictions can be significantly incorrect due to insufficient or biased training data. In contrast, physics-based models, while potentially inaccurate, do not have such holes because they are grounded in physical laws that provide a continuous description of the PES. The uncertainty of a particular point in the PES is not as problematic for physical models as it is for ML potentials, where such uncertainty can significantly affect the model’s reliability.

Active learning is a strategy where the model actively selects the most informative data points for training.[13, 58, 75, 1] Here we use this strategy to identify ”holes” on the PES while improving model performance. Initially, an ML model is trained

using the dataset from Chapter 3, which includes *ab initio* molecular dynamics (AIMD) and normal mode displacement data near the intrinsic reaction coordinate (IRC).[23] To identify areas of uncertainty in the PES, a committee of multiple ML models with different initializations is trained to predict energies and forces, with the variance among these models indicating regions where the model predictions are less reliable.[69] To efficiently explore high-energy configurations and rare events, metadynamics simulations are employed. This technique biases the system to sample new configurations by adding a history-dependent potential, helping to fill gaps in the PES and improve model coverage.[41, 3] Configurations where the committee models disagree are selected for additional *ab initio* calculations, and these new data points are added to the training set, improving the model iteratively. The ML models are retrained with the expanded dataset, enhancing their ability to predict diverse configurations accurately.

The active learning-enhanced ML model demonstrates significant improvements in accuracy and efficiency compared to the initial model trained solely on the Chapter 3 dataset. The iterative process of sampling, data selection, and retraining ensures that the model covers a broad range of configurations, reducing errors in energy and force predictions. However it remains that the PES may never be completed with an active learning process, especially when coupled with enhanced sampling methods that continuously seek to explore unseen regions, which over many iterations is of great expense to continually create new and expensive *ab initio* data. My final solution was the creation of a hybrid approach, where ML predictions are supplemented with infrequent *ab initio* calculations in areas of high uncertainty, ensuring the stability and accuracy of the simulations while reducing the exorbitant hidden cost of data acquisition and constant ML retraining. The resulting ML force field driven by the enhanced dataset and hybrid physics shows stable and accurate performance, even in high-energy regions and near transition states, allowing us to perform reliable committor analysis and free energy transition states of all hydrogen combustion reactions.

1.5 REFERENCES

- [1] Shi Jun Ang, Wujie Wang, Daniel Schwalbe-Koda, Simon Axelrod, and Rafael Gómez-Bombarelli. “Active learning accelerates *ab initio* molecular dynamics on reactive energy surfaces”. English. In: *Chem* 7.3 (Mar. 2021), pp. 738–751. ISSN: 2451-9294, 2451-9308. DOI: 10.1016/j.chempr.2020.12.009.
- [2] Richard F. W. Bader. “A Quantum Theory of Molecular Structure and its Applications”. In: *Chemical Reviews* 91.5 (1991), pp. 893–928. DOI: 10.1021/

- cr00005a013. eprint: <https://doi.org/10.1021/cr00005a013>. URL: <https://doi.org/10.1021/cr00005a013>.
- [3] A. Barducci, G. Bussi, and M. Parrinello. “Well-tempered metadynamics: A smoothly converging and tunable free-energy method”. In: *Physical Review Letters* 100 (2008), p. 020603. DOI: 10.1103/PhysRevLett.100.020603.
 - [4] Ilyes Batatia, David P. Kovacs, Gregor Simm, Christoph Ortner, and Gabor Csanyi. “MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields”. en. In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 11423–11436.
 - [5] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. “E(3)-equivariant graph neural networks for data-efficient and accurate inter-atomic potentials”. en. In: *Nature Communications* 13.11 (May 2022), p. 2453. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29939-5.
 - [6] C.I. Bayly, P. Cieplak, W.D. Cornell, and P.A. Kollman. “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model”. In: *Journal of Physical Chemistry* 97.40 (1993), pp. 10269–10280. DOI: 10.1021/j100142a004.
 - [7] Jörg Behler and Michele Parrinello. “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces”. In: *Phys. Rev. Lett.* 98 (14 Apr. 2007), p. 146401. DOI: <https://doi.org/10.1103/PhysRevLett.98.146401>.
 - [8] Kurt Binder and Dieter W. Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction*. Berlin, Heidelberg: Springer, 2010. ISBN: 978-3-642-03163-2.
 - [9] C.M. Breneman and K.B. Wiberg. “Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis”. In: *Journal of Computational Chemistry* 11.3 (1990), pp. 361–373. DOI: 10.1002/jcc.540110311.
 - [10] Patrick Bultinck, Christian Van Alsenoy, Paul W. Ayers, and Ramon Carbó-Dorca. “Critical Analysis and Extension of the Hirshfeld Atoms in Molecules”. In: *The Journal of Chemical Physics* 126.14 (Apr. 2007), p. 144111. ISSN: 0021-9606. DOI: 10.1063/1.2715563.

- [11] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. “Machine learning of accurate energy-conserving molecular force fields”. In: *Science Advances* 3.5 (2017), e1603015. DOI: 10.1126/sciadv.1603015. URL: <http://advances.sciencemag.org/content/3/5/e1603015.abstract>.
- [12] Minsik Cho, Nitai Sylvetsky, Sarah Eshafi, Golokesh Santra, Irena Efremenko, and Jan M. L. Martin. “The Atomic Partial Charges Arboretum: Trying to See the Forest for the Trees”. In: *ChemPhysChem* 21.8 (2020), pp. 688–696. DOI: <https://doi.org/10.1002/cphc.202000040>. eprint: <https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.202000040>. URL: <https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cphc.202000040>.
- [13] David A. Cohn, Les Atlas, and Richard E. Ladner. “Improving generalization with active learning”. In: *Machine Learning* 15.2 (1994), pp. 201–221. DOI: 10.1007/BF00993277.
- [14] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*. en. Oct. 2022. URL: <https://arxiv.org/abs/2210.01776v2>.
- [15] C.J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. John Wiley Sons, 2004. DOI: 10.1002/0470091808.
- [16] Ralf Drautz. “Atomic cluster expansion for accurate and transferable interatomic potentials”. In: *Physical Review B* 99.1 (Jan. 2019), p. 014104. DOI: 10.1103/PhysRevB.99.014104.
- [17] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. San Diego, CA: Academic Press, 2001. ISBN: 978-0-12-267351-1.
- [18] Mario Geiger and Tess E Smidt. “E3NN: Euclidean Neural Networks”. In: *arXiv preprint arXiv:2207.09453* (2022).
- [19] G. Gerasimov and O. Shatalov. “Kinetic mechanism of combustion of hydrogen–oxygen mixtures”. In: *Journal of Engineering Physics and Thermophysics* 86 (2013), pp. 987–995. DOI: 10.1007/s10891-013-0919-7.
- [20] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. “Neural message passing for quantum chemistry”. In: *arXiv preprint arXiv:1704.01212* (2017).

- [21] L. et al. Goerigk. “A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions”. In: *Physical Chemistry Chemical Physics* 19 (2017), pp. 32184–32215. DOI: 10.1039/C7CP04913G.
- [22] X. Guan et al. “Hydrogen Combustion using IRC, AIMD and normal modes”. In: *Figshare* (2022). URL: <https://doi.org/10.6084/m9.figshare.19601689>.
- [23] X. Guan et al. “Hydrogen Combustion using IRC, AIMD and normal modes”. In: *Figshare* (2022). URL: <https://doi.org/10.6084/m9.figshare.19601689>.
- [24] Xingyi Guan, Joseph P. Heindel, Taehee Ko, Chao Yang, and Teresa Head-Gordon. “Using machine learning to go beyond potential energy surface benchmarking for chemical reactivity”. en. In: *Nature Computational Science* 3.11 (Nov. 2023), pp. 965–974. ISSN: 2662-8457. DOI: 10.1038/s43588-023-00549-5.
- [25] Xingyi Guan, Itai Leven, Farnaz Heidar-Zadeh, and Teresa Head-Gordon. “Protein C-GeM: A Coarse-Grained Electron Model for Fast and Accurate Protein Electrostatics Prediction”. In: *Journal of Chemical Information and Modeling* 61.9 (Sept. 2021), pp. 4357–4369. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.1c00388.
- [26] Xingyi Guan et al. “A benchmark dataset for Hydrogen Combustion”. en. In: *Scientific Data* 9.11 (May 2022), p. 215. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01330-5.
- [27] Mojtaba Haghighatlari, Jie Li, Farnaz Heidar-Zadeh, Yuchen Liu, Xingyi Guan, and Teresa Head-Gordon. “Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods”. In: *Chem* 6.7 (2020), pp. 1527–1542. DOI: <https://doi.org/10.1016/j.chempr.2020.05.014>. URL: <https://doi.org/10.1016/j.chempr.2020.05.014>.
- [28] Mojtaba Haghighatlari et al. “NewtonNet: A Newtonian message passing network for deep learning of interatomic potentials and forces”. In: *arXiv preprint arXiv:2108.02913* (2021). arXiv: 2108.02913 [physics.comp-ph].
- [29] Mojtaba Haghighatlari et al. “NewtonNet: a Newtonian message passing network for deep learning of interatomic potentials and forces”. en. In: *Digital Discovery* 1.3 (2022), pp. 333–343. DOI: 10.1039/D2DD00008C.
- [30] Farnaz Heidar-Zadeh and Paul W. Ayers. “How Pervasive is the Hirshfeld Partitioning?” In: *The Journal of Chemical Physics* 142.4 (2015), p. 044107. DOI: 10.1063/1.4905123. eprint: <https://doi.org/10.1063/1.4905123>. URL: <https://doi.org/10.1063/1.4905123>.

- [31] Farnaz Heidar-Zadeh, Paul W. Ayers, Toon Verstraelen, Ivan Vinogradov, Esteban Vöhringer-Martinez, and Patrick Bultinck. “Information-Theoretic Approaches to Atoms-in-Molecules: Hirshfeld Family of Partitioning Schemes”. In: *The Journal of Physical Chemistry A* 122.17 (May 2018), pp. 4219–4245. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.7b08966.
- [32] Trygve Helgaker, Poul Jørgensen, and Jeppe Olsen. *Molecular Electronic-Structure Theory*. Chichester: John Wiley Sons, 2014. ISBN: 978-0-471-96670-2.
- [33] F. L. Hirshfeld. “Bonded-atom Fragments for Describing Molecular Charge Densities”. In: *Theoretica chimica acta* 44.2 (June 1977), pp. 129–138. ISSN: 1432-2234. DOI: 10.1007/BF00549096.
- [34] Josef Ischtwan and Michael A. Collins. “Determination of the intrinsic reaction coordinate: Comparison of gradient and local quadratic approximation methods”. In: *The Journal of Chemical Physics* 89.5 (1988), pp. 2881–2885. DOI: 10.1063/1.456130.
- [35] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. “Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method”. en. In: *Journal of Computational Chemistry* 21.2 (2000), pp. 132–146. ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(20000130)21:2<132::AID-JCC5>3.0.CO;2-P.
- [36] Araz Jakalian, David B. Jack, and Christopher I. Bayly. “Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation”. In: *Journal of Computational Chemistry* 23.16 (Dec. 2002), pp. 1623–1641. ISSN: 0192-8651. DOI: 10.1002/jcc.10128.
- [37] Steven Kearnes, Kevin McCloskey, Markus Berndl, Vijay Pande, and Patrick Riley. “Molecular graph convolutions: moving beyond fingerprints”. In: *Journal of Computer-Aided Molecular Design* 30 (2016), pp. 595–608. DOI: 10.1007/s10822-016-9938-8.
- [38] Rustam Z. Khaliullin, Erika A. Cobar, Rohini C. Lochan, Alexis T. Bell, and Martin Head-Gordon. “Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals”. eng. In: *The Journal of Physical Chemistry. A* 111.36 (Sept. 2007), pp. 8753–8765. ISSN: 1089-5639. DOI: 10.1021/jp073685z.
- [39] Wolfram Koch and Max C. Holthausen. *Chemist’s Guide to Density Functional Theory*. Weinheim: Wiley-VCH, 2001. ISBN: 978-3-527-30372-4.
- [40] Werner Krauth and Samuel Vionnet. “Concepts in Monte Carlo sampling”. In: *American Journal of Physics* 90 (2022), pp. 60–69. DOI: 10.1119/10.0006198.

- [41] A. Laio and M. Parrinello. “Escaping free-energy minima”. In: *Proceedings of the National Academy of Sciences* 99 (2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399.
- [42] L. D. Landau and E. M. Lifshitz. *Mechanics: Volume 1*. Amsterdam: Elsevier Science, 1982. ISBN: 978-0-08-029141-6.
- [43] Andrew R. Leach. *Molecular Modelling: Principles and Applications*. Harlow, England: Prentice Hall, 2001. ISBN: 978-0-582-38210-7.
- [44] Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Cham: Springer, 2015. ISBN: 978-3-319-16375-8. DOI: 10.1007/978-3-319-16375-8.
- [45] Itai Leven, Hongxia Hao, Akshaya Kumar Das, and Teresa Head-Gordon. “A Reactive Force Field with Coarse-Grained Electrons for Liquid Water”. In: *The Journal of Physical Chemistry Letters* 11.21 (Nov. 2020), pp. 9240–9247. DOI: 10.1021/acs.jpcllett.0c02516.
- [46] Itai Leven and Teresa Head-Gordon. “C-GeM: Coarse-Grained Electron Model for Predicting the Electrostatic Potential in Molecules”. In: *The Journal of Physical Chemistry Letters* 10.21 (Nov. 2019), pp. 6820–6826. DOI: 10.1021/acs.jpcllett.9b02771.
- [47] J.-R. Li, R.J. Kuppler, and H.-C. Zhou. “Gas storage in metal-organic frameworks”. In: *Chemical Society Reviews* 38.5 (2009), pp. 1477–1504. DOI: 10.1039/b802426j.
- [48] Richard Li, Zoe Brown, Chen Zhao, and Zhong Zhan. “The Effects of Data Quality on Machine Learning Performance”. In: *arXiv preprint arXiv:2207.14529* (2021). URL: <https://arxiv.org/abs/2207.14529>.
- [49] Jennifer Listgarten. “The perpetual motion machine of AI-generated data and the distraction of ChatGPT as a ‘scientist’”. en. In: *Nature Biotechnology* 42.3 (Mar. 2024), pp. 371–373. ISSN: 1546-1696. DOI: 10.1038/s41587-023-02103-0.
- [50] N. Mardirossian and M. Head-Gordon. “B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy”. In: *Physical Chemistry Chemical Physics* 16 (2014), pp. 9904–9924. DOI: 10.1039/c3cp54374a.

- [51] Narbe Mardirossian and Martin Head-Gordon. “ ω B97X-V: A 10-parameter, Range-separated Hybrid, Generalized Gradient Approximation Density Functional with Nonlocal Correlation, Designed by a Survival-of-the-fittest Strategy”. In: *Physical Chemistry Chemical Physics* 16.21 (May 2014), pp. 9904–9924. ISSN: 1463-9084. DOI: 10.1039/C3CP54374A.
- [52] Dominik Marx and Jurg Hutter. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge: Cambridge University Press, 2009. ISBN: 978-0-521-89850-6. DOI: 10.1017/CB09780511609633.
- [53] J.A. McCammon and S.C. Harvey. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, 1987. DOI: 10.1017/CB09780511623121.
- [54] William H. Miller. “Ab Initio Quantum Chemistry: Methodologies and Applications”. In: *Journal of Physical Chemistry A* 102 (1998), pp. 7993–8005. DOI: 10.1021/jp9839324.
- [55] Wilfried J. Mortier, Swapan K. Ghosh, and S. Shankar. “Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules”. In: *Journal of the American Chemical Society* 108.15 (1986), pp. 4315–4320. ISSN: 15205126. DOI: 10.1021/ja00275a013.
- [56] Wilfried J. Mortier, Karin Van Genechten, and Johann Gasteiger. “Electronegativity Equalization: Application and Parametrization”. In: *Journal of the American Chemical Society* 107.4 (1985), pp. 829–835. DOI: 10.1021/ja00290a017. eprint: <https://doi.org/10.1021/ja00290a017>. URL: <https://doi.org/10.1021/ja00290a017>.
- [57] R.S. Mulliken. “Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I”. In: *Journal of Chemical Physics* 23.10 (1955), pp. 1833–1840. DOI: 10.1063/1.1740588.
- [58] Dante A. Pertusi, Matthew E. Moura, James G. Jeffryes, Siddhant Prabhu, Bradley Walters Biggs, and Keith E. J. Tyo. “Predicting novel substrates for enzymes with minimal experimental effort with active learning”. In: *Metabolic Engineering* 44 (Nov. 2017), pp. 171–181. ISSN: 1096-7176. DOI: 10.1016/j.ymben.2017.09.016.
- [59] Jay W. Ponder et al. “Current Status of the AMOEBA Polarizable Force Field”. In: *The Journal of Physical Chemistry B* 114.8 (Mar. 2010), pp. 2549–2564. ISSN: 1520-6106, 1520-5207. DOI: 10.1021/jp910674d.
- [60] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. “Quantum chemistry structures and properties of 134 kilo molecules”. In: *Scientific Data* 1.140022 (2014), pp. 1–7. DOI: 10.1038/sdata.2014.22.

- [61] Jeffrey R. Reimers. “A practical method for the use of curvilinear coordinates in calculations of normal-mode-projected displacements and Duschinsky rotation matrices for large molecules”. In: *The Journal of Chemical Physics* 115.20 (Nov. 2001), pp. 9103–9109. ISSN: 0021-9606. DOI: 10.1063/1.1412875.
- [62] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. “Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17”. In: *Journal of Chemical Information and Modeling* 52.11 (2012), pp. 2864–2875. DOI: 10.1021/ci300415d.
- [63] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. “Fast and accurate modeling of molecular atomization energies with machine learning”. In: *Physical Review Letters* 108.5 (2012). DOI: 10.1103/physrevlett.108.058301.
- [64] H. Bernhard Schlegel. “Exploring Potential Energy Surfaces for Chemical Reactions: An Overview of Some Practical Methods”. In: *The Journal of Chemical Physics* 114 (2002), pp. 9758–9765. DOI: 10.1063/1.1461829.
- [65] W. Schmickler and E. Santos. *Interfacial Electrochemistry*. Springer Science Business Media, 2010. DOI: 10.1007/978-3-642-04042-7.
- [66] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller. “SchNet - A deep learning architecture for molecules and materials”. In: *Journal of Chemical Physics* 148.24 (Mar. 2018), p. 241722. ISSN: 0021-9606. DOI: 10.1063/1.5019779.
- [67] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. “SchNetPack: A Deep Learning Toolbox For Atomistic Systems”. In: *Journal of Chemical Theory and Computation* 15.1 (2019), pp. 448–455. DOI: 10.1021/acs.jctc.8b00908. eprint: <https://doi.org/10.1021/acs.jctc.8b00908>.
- [68] Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. “Equivariant message passing for the prediction of tensorial properties and molecular spectra”. In: *arXiv preprint arXiv:2102.03150* (2021). arXiv: 2102.03150.
- [69] H.S. Seung, M. Opper, and H. Sompolinsky. “Query by committee”. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, 1992, pp. 287–294.
- [70] S. W. Shaw and C. Pierre. “Normal Modes for Non-Linear Vibratory Systems”. In: *Journal of Sound and Vibration* 164.1 (June 1993), pp. 85–124. ISSN: 0022-460X. DOI: 10.1006/jsvi.1993.1198.

- [71] F. B. Sheinerman and B. Honig. “On the role of electrostatic interactions in the design of protein–protein interfaces”. In: *Journal of Molecular Biology* 318 (2002), pp. 161–177. DOI: 10.1016/s0022-2836(02)00030-x.
- [72] G. Simm and M. Reiher. “Context-driven exploration of complex chemical reaction networks”. In: *Journal of Chemical Theory and Computation* 13 (2017), pp. 6108–6119. DOI: 10.1021/acs.jctc.7b00945.
- [73] Deborah M. Smith and K. A. Woerpel. “Electrostatic interactions in cations and their importance in biology and chemistry”. In: *Organic Biomolecular Chemistry* 4 (2006), pp. 1195–1201. DOI: 10.1039/B600056H.
- [74] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. “ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost”. In: *Chemical Science* 8.4 (2017), pp. 3192–3203.
- [75] Justin S. Smith, Benjamin Tyler Nebgen, Nicholas Edward Lubbers, Olexandr Isayev, and Adrian E. Roitberg. “Less is more: sampling chemical space with active learning”. In: *Journal of Chemical Physics* 148.24 (2018), p. 241733. DOI: 10.1063/1.5023802.
- [76] Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. New York, NY: Dover Publications, 2012. ISBN: 978-0-486-69186-1.
- [77] Donald G. Truhlar. “Current Status of Transition-State Theory”. In: *The Journal of Physical Chemistry* 100.31 (1996), pp. 12771–12800. DOI: 10.1021/jp953748q.
- [78] Z.W. Ulissi, A.J. Medford, T. Bligaard, and J.K. Nørskov. “To address surface reaction network complexity using scaling relations, machine learning, and DFT calculations”. In: *Nature Communications* 8 (2017), p. 14621. DOI: 10.1038/ncomms14621.
- [79] Toon Verstraelen, Steven Vandenbrande, Farnaz Heidar-Zadeh, Louis Vanduyfhuys, Veronique Van Speybroeck, Michel Waroquier, and Paul W. Ayers. “Minimal Basis Iterative Stockholder: Atoms in Molecules for Force-Field Development”. In: *Journal of Chemical Theory and Computation* 12.8 (2016), pp. 3894–3912. DOI: 10.1021/acs.jctc.6b00456. eprint: <https://doi.org/10.1021/acs.jctc.6b00456>. URL: <https://doi.org/10.1021/acs.jctc.6b00456>.
- [80] Frank Weigend and Reinhart Ahlrichs. “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy”. In: *Physical Chemistry Chemical Physics* 7 (2005), pp. 3297–3305. DOI: 10.1039/b508541a.

- [81] WillardJared, JiaXiaowei, XuShaoming, SteinbachMichael, and KumarVipin. “Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems”. EN. In: *ACM Computing Surveys* (Nov. 2022). DOI: 10.1145/3514228. URL: <https://dl.acm.org/doi/10.1145/3514228>.
- [82] Ying Xu, Hannes Stojic, David McNeill, and Kushal Sharma. “To what extent should we trust AI models when they extrapolate?” In: *arXiv preprint arXiv:2201.11260* (2021). URL: <https://arxiv.org/abs/2201.11260>.
- [83] Jinzhe Zeng, Liqun Cao, Mingyuan Xu, Tong Zhu, and John Z. H. Zhang. “Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation”. en. In: *Nature Communications* 11.1 (Nov. 2020), p. 5713. ISSN: 2041-1723. DOI: 10.1038/s41467-020-19497-z.

Chapter 2

Protein C-GeM: A coarse-grained electron model for fast and accurate protein electrostatics prediction [†]

2.1 INTRODUCTION

The electrostatic potential (ESP) is fundamental for understanding and predicting biomolecular recognition between molecules[60, 48] For proteins in particular, the ESP is often crucial for predicting contact sites of protein-protein association,[32] and the electrostatic complementarity between protein and small molecule ligands or peptide therapeutics is considered critically important to obtain optimal affinity and selectivity in structure-based drug discovery.[45, 8]

An ESP is generated by evaluating the work to move a unit charge probe from infinity to an area of interest on or near the protein surface. Numerically this is achieved by defining a grid, either on the molecular surface of the protein or by drawing equipotential contours in the region around the protein. [38] At each surface point \mathbf{r} , the ESP energy of the probe is calculated and the molecular surface is then displayed to indicate regions of negative or positive electrostatic potential of the protein molecule. An accurate way of obtaining the molecular ESP is through *ab*

[†]Reproduced with permission from: Guan, X.; Leven, I.; Heidar-Zadeh, F.; Head-Gordon, T. Protein C-GeM: A Coarse-Grained Electron Model for Fast and Accurate Protein Electrostatics Prediction. *J. Chem. Inf. Model.* **2021**, 61 (9), 4357–4369.

initio calculations, for which the ESP is defined as

$$V(\mathbf{r}) = \sum_A \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}|} - \int \frac{\rho(\mathbf{r}')d\mathbf{r}'}{|\mathbf{r}' - \mathbf{r}|} \quad (2.1)$$

where Z_A and \mathbf{R}_A are the charge and position of nucleus A, and $\rho(\mathbf{r}')$ is the electronic density at position \mathbf{r}' . However, the computational cost of a full quantum mechanical (QM) ESP increases rapidly with the number of atoms, and becomes prohibitive for systems such as large macromolecules.

Instead a large macromolecule can be partitioned in such a way that the electrostatic potential can be reproduced by assigning partial charges to every atom in a molecule, $\{q_A\}_{A=1}^{N_{\text{atoms}}}$, i.e.,

$$V(\mathbf{r}) \sim \sum_{A=1}^{N_{\text{atoms}}} \frac{q_A}{|\mathbf{r} - \mathbf{R}_A|} \quad (2.2)$$

Atomic charges derived from fitting a classical Coulomb model to reproduce the *ab initio* molecular electrostatic potentials (so called ESP-charges) are frequently used in simulations of macromolecules, [50, 61, 10, 7], and they are the main electrostatic description used for all major fixed charge force fields such as AMBER[59, 47] and CHARMM[24], and utilized in large molecule ESP solvers using the Poisson-Boltzmann equation such as APBS.[22] One widely used ESP charge is the AM1-BCC model[21], which captures the underlying features of the electron distribution including formal charge and delocalization using the semi-empirical AM1 method, and applies bond charge corrections (BCCs) that are fitted to *ab initio* ESP. While more cost-effective than full QM, the ESP-charges are numerically ill-conditioned such as being overly sensitive to conformational changes and restricted to applications where the electron density changes are relatively small.[17] While ESP-fitted charge models such as CHELPG [5] can be more robust, and can accurately reproduce the molecular ESP, they are not competitors to the prediction application because they require the ESP as its input.

Alternatively, QM-based partitioning methods divide a molecule into atomic subsystems by partitioning either the molecular wave-function in Hilbert space (i.e., orbital-based methods) or molecular electron density in real space (i.e., density-based methods). The first and most prevalent orbital-based partitioning method is the Mulliken[37] scheme which divides each molecular orbital into its atomic pieces. The original Mulliken partitioning suffered from excessive basis-set sensitivity, but subsequent refinement alleviated this shortcoming by defining atomic pieces in more sophisticated ways.[46, 29, 23] Unfortunately, the orbital-based charges are generally inferior for reproducing the electrostatic potential, as compared to density-based partitionings.[55]

The density-based QM partitioning exhaustively divide the molecular electron density distribution, $\rho(\mathbf{r})$, between its constituent atoms according to

$$\rho_A(\mathbf{r}) = \sum_A^{N_{\text{atoms}}} w_A(\mathbf{r})\rho(\mathbf{r}) \quad (2.3)$$

$$\sum_A^{N_{\text{atoms}}} w_A(\mathbf{r}) = 1 \quad \text{and} \quad w_A(\mathbf{r}) \geq 0$$

where the electron density of atom A at point \mathbf{r} in space, $\rho_A(\mathbf{r})$, is dictated by its share $w_A(\mathbf{r})$ at that point. Subsequently, the atomic charge of atom A is computed by,

$$q_A = Z_A - \int \rho_A(\mathbf{r})d\mathbf{r} \quad (2.4)$$

The quality of these charges in reproducing the electrostatic potential heavily depends on the definition of atomic weights, $w_A(\mathbf{r})$. The atomic weights used in density-based methods are either binary as in Bader's Quantum Theory of Atoms in Molecules (QTAIM)[2] or fuzzy as developed in the Hirshfeld partitioning schemes and its variants[19, 6, 28, 30, 54, 53, 55, 17]. Among these, the latter results in nearly-spherical atomic regions, so they have rapidly converging atomic multipole expansions and give a good approximation of $V(\mathbf{r})$ based on Eq. (2.2).

The Hirshfeld-family of methods use a set of proatom atomic densities $\{\rho_A^0(\mathbf{r})\}_{A=1}^{N_{\text{atoms}}}$ to assign the atomic weights through [19, 39, 16, 15],

$$w_A(\mathbf{r}) = \frac{\rho_A^0(\mathbf{r})}{\sum_{B=1}^{N_{\text{atoms}}} \rho_B^0(\mathbf{r})} \quad (2.5)$$

The original Hirshfeld[19] method uses neutral proatom densities as the reference; this choice is arbitrary and results in very small atomic charges. To fix these shortcomings, various Hirshfeld-inspired methods have been developed to select optimal proatom densities.[6, 55, 15, 17] The first, and most prevalent, method is Iterative Hirshfeld (HI)[6], which refines the proatoms self-consistently so that they have the same charges as the atoms. Two more recent and promising methods are the Minimal Basis Iterative Stockholder (MBIS)[55] and Additive Variational Hirshfeld (AVH)[15, 17] which variationally optimize the proatom densities so that they best reproduce the molecular density. When the atomic partial charges are determined from the population of these atomic subsystems, the more accurate reproduction of $V(\mathbf{r})$ is a measure of the partitioning scheme's quality and utility[55, 17], and thus we consider them here. Of course there is a very large and extensive number of *ab initio* charge

partitioning methods that we have not considered here, and the interested reader can refer to a recent review by Martin and co-workers[9] to learn more about these approaches. While QM based partitioning approaches have the advantage of being physically grounded and generally applicable to a wide range of systems of interest, and have proven effective for modelling intermolecular interactions[52], they still suffer from the underlying expense of QM calculations and thus are not extensible to large systems such as proteins.

The electronegativity equalization methods (EEM) is an alternative approach that straddles the boundary of empirical fitting but formulated within the QM foundations of atomic hardness and electronegativity.[36, 35] It has been used as the electrostatic model for reactive force fields[12] and has been adapted for fast electrostatic screening applications for large molecular databases as well as protein electrostatics applications due to its relative efficiency.[14, 20, 41] However, although elegant, EEM has some significant shortcomings including unphysical long-range charge transfer, non-integer molecular charge at large molecular separations, lack of out-of-plane polarization, poor parameterization, and lack of transferability that makes EEM methods less accurate than desired but which are analyzed here for completeness.[3, 4, 25]

Hence, an accurate but fast method for protein ESP prediction is still highly desirable. In this study, we evaluate the coarse-grained electron force field model, C-GeM for which atoms are represented by a positive core and an electron shell described by Gaussian charge distributions.[26]. Integration of the Coulombic interactions of the Gaussian densities yields an analytical form for the electrostatic energy between arbitrary core-core, core-shell, and shell-shell interactions. By minimizing the electronic shell positions in the field of atomic core positions, the model can provide accurate electrostatic properties of molecules and their interactions. A schematic of this process is shown in Figure 2.1.

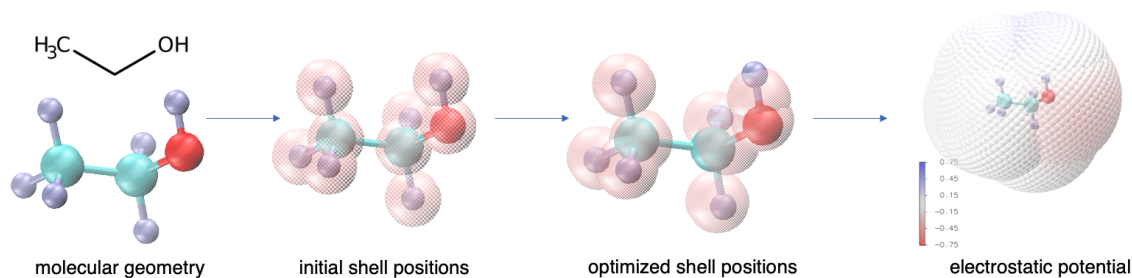


Figure 2.1: Schematic illustration of how C-GeM generates the electrostatic potential from given molecular geometry.

Previous models which share similarities to C-GeM include the core-shell model developed to account for polarization in ionic crystals[33], the PQEq method which utilizes a Gaussian Drude oscillator model together with charge equilibration[40], and the ACP method which partitions the electron density according to the core and valance shell electrons[58]. C-GeM differs from these previous models through its unique ability to predict permanent electrostatics, polarization, and charge transfer without having to perform computationally expensive *ab-initio* calculations. While the C-GeM model has been previously parameterized for the atomic elements carbon, hydrogen, oxygen and chloride[26], in this work we have expanded the C-GeM parameterization for the nitrogen and sulfur atomic elements for a complete protein level chemistry. When optimized with tripeptide and small molecule training data, C-GeM is found to perform better than ESP-fitted charges, EEM, and Hirshfeld charges in reproducing the ESP of the protein test set that is comprised of tripeptides of different sequences and the crambin protein. To improve accuracy of the C-GeM model further we also introduce atom typing, i.e. optimization of different parameters for aliphatic and polar carbon and hydrogen atoms and for primary, secondary and tertiary amines for nitrogen. This atom typing approach thus makes the C-GeM model as accurate as HI charges and competitive with MBIS and AVH density partitioning methods when evaluated on the protein test set. Altogether the C-GeM model offers a new way to do high-throughput electrostatic screening with *ab initio* accuracy with orders of magnitude less computational expense since it does not require the electron density or ESP but instead predicts these quantities.

2.2 THEORY

The C-GeM model divides atoms into positive cores and negative shells, both of which are represented as Gaussian distributed charges. The properties of a core depend on its atom type i , while all of the electrons (shells) are treated equivalently. The charge density of a core of atom type i ($\rho_{i,c}$) and that of a generic shell (ρ_s) is given by the following functional form

$$\rho_{i,c}(\mathbf{r}) = q_{i,c} \left(\frac{\alpha_{i,c}}{\pi} \right)^{3/2} e^{-\alpha_{i,c}(|\mathbf{r}-\mathbf{r}_{i,c}|^2)} \quad (2.6)$$

$$\rho_s(\mathbf{r}) = q_s \left(\frac{\alpha_s}{\pi} \right)^{3/2} e^{-\alpha_s(|\mathbf{r}-\mathbf{r}_s|^2)} \quad (2.7)$$

where (\mathbf{r}) is an arbitrary position in space and $\mathbf{r}_{i,c}$ and \mathbf{r}_s are position vectors for the core and shell centers, respectively. The shell charge (q_s) is always set to -1, while the core charge ($q_{i,c}$) is usually set to +1 but can vary based on the chemical conditions

of charge as we illustrate below. The width of a Gaussian charge is controlled by $\alpha_{i,c}$ for cores and α_s for shells:

$$\alpha_{i,c} = \frac{\lambda}{2R_{i,c}^2} \quad \alpha_s = \frac{\lambda}{2R_s^2} \quad (2.8)$$

where λ is a global fitting parameter, $R_{i,c}$ is the atomic covalent radius [10] of atom type i that is further fine tuned to reflect the atomic radii in actual molecules, and R_s is the effective radius of the shells.

The Coulombic interaction between two elements (core-core, core-shell and shell-shell) can be expressed as the integration over two Gaussian densities, which has the following analytical form:

$$\begin{aligned} E_{ij}^{elec}(r_{ij}) &= \int \int \frac{\rho_i(\mathbf{r}_i)\rho_j(\mathbf{r}_j)}{|\mathbf{r}_i - \mathbf{r}_j|} d\mathbf{r}_i d\mathbf{r}_j \\ &= \frac{q_i q_j}{r_{ij}} \operatorname{erf}\left(\sqrt{\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}} r_{ij}\right) \end{aligned} \quad (2.9)$$

where r_{ij} is the distance between the two elements. In the limit of $r_{ij} \rightarrow 0$, the pairwise Coulombic interaction can be rewritten as

$$\lim_{r_{ij} \rightarrow 0} E_{ij}^{elec}(r_{ij}) = \frac{2q_i q_j}{\sqrt{\pi}} \left(\sqrt{\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}} \right) \quad (2.10)$$

In addition to electrostatics, a Gaussian energy term is used that reflects the strength of core-shell or shell-shell interaction, taking into account the electronegativity of specific atom types:

$$E_{ij}^{gauss}(r_{ij}) = \beta_i e^{-\gamma_i r_{ij}^2} + P(r_{ij}) \quad (2.11)$$

where β_i is a parameter accounting for the magnitude of the interaction energy, $P(r_{ij})$ is a penalty term for shell-shell distances that are too close, and γ is a parameter that controls the radial range of the interaction, which is defined as

$$\gamma_{i,c} = \frac{\omega_c}{2R_{i,c}} \quad (2.12)$$

for core-shell Gaussian interactions, controlled by a global parameter ω_c and atomic parameter $R_{i,c}$ for atom type i . The radial range for shell-shell interaction is controlled by global parameter γ_s .

With the theoretical idea that the C-GeM energy between a core of atom type i and its shell j should match the ionization potential of that atom type (χ_i), we demand that

$$\chi_i = E_{ij}^{elec}(r_{ij} = 0) + E_{ij}^{gauss}(r_{ij} = 0) \quad (2.13)$$

where χ_i is the ionization potential of atom type i . In the case of a shell-shell interaction, we use a global fitting parameter χ_{shell} to represent the effective shell-shell interaction energy that leads to following definition for the magnitude of Gaussian interaction β_i :

$$\begin{aligned}\beta_i &= \lim_{r_{ij} \rightarrow 0} \frac{\chi_i - E_{ij}^{elec}}{e^{-\gamma_i r_{ij}^2}} \\ &= \chi_i - \frac{2q_i q_j}{\sqrt{\pi}} \left(\sqrt{\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}} \right)\end{aligned}\tag{2.14}$$

To avoid shell configurations that optimize to the exact same position and become inseparable, we introduced a penalty term for shell-shell interaction at very short range. This term effectively help shells avoid each other so that they experience distinct forces at all time.

$$P(r_{ij}) = \begin{cases} 10e^{-200r_{ij}}, & \text{if } i \in \text{shells and } j \in \text{shells} \\ 0, & \text{otherwise} \end{cases}\tag{2.15}$$

The total C-GeM energy of a given system with fixed cores involves an optimization of the shell positions to minimize the energy,

$$E_{CGeM} = \sum_i \sum_{j < i} E_{ij}^{elec}(r_{ij}) + E_{ij}^{gauss}(r_{ij})\tag{2.16}$$

as per a usual Born-Oppenheimer assumption. The resulting shell configurations is used to generate the electrostatic potential on a set of given points using the following equation:

$$V(\mathbf{r}) = \sum_{i \in \text{cores}} \frac{q_i}{(|\mathbf{r} - \mathbf{r}_i|)} + \sum_{i \in \text{shells}} \frac{q_i}{(|\mathbf{r} - \mathbf{r}_i|)}\tag{2.17}$$

where all of the core and shell Gaussian charges are approximated by point charges at their center to speed up the ESP evaluation.

2.3 METHODS

To address both neutral and charged systems, we require an identification of the formal charge on each atom. All neutral atoms are initialized with a +1 core and a -1 shell at atomic center; a negatively charged atom receives an additional -1 charge shell based on its formal charge, and these additional shells are randomly displaced within 10^{-3} \AA distance to avoid overlaps; a positively charged atom is initialized with

an incremented core charge ($q_c = 1 + \text{formal charge}$) and a -1 shell at the atomic center.

C-GeM training and testing protocol. There are five global parameters (λ , ω_{core} , γ_{shell} , χ_{shell} and R_{shell}) and two atom-specific parameters per atom type (χ_i and R_i) in the C-GeM model. These parameters are fitted by minimizing the average mean absolute error (MAE_{avg}) over the training set with respect to *ab initio* ESP, where the MAE for one molecule is computed as:

$$MAE = \frac{1}{n} \sum_i^n |V_{C-GeM}(\mathbf{r}_i) - V_{DFT}(\mathbf{r}_i)| \quad (2.18)$$

where n is the total number of grid points, $V_{C-GeM}(\mathbf{r}_i)$ and $V_{DFT}(\mathbf{r}_i)$ are the C-GeM and the DFT ESP computed for a grid point at position \mathbf{r}_i .

The training set consists of 54 small molecules and 38 tripeptides, with an additional 19 tripeptides defining the validation set. The protein analogs are small molecules that represents the chemistry of amino acids, and a list of these molecule is provided in Supplementary Table 2.A.3 and Table 2.A.4. The 57 larger and more complex tripeptides are formulated by fragmentation of larger proteins culled from the PDB, [27] and uniformly sampled by amino acid residue types to capture the diversity of peptides. Three models are trained by minimizing the mean MAE of all of the small protein analogs and 2/3 of the tripeptides using the Nelder-Mead algorithm [13], with one set of 19 tripeptides used as a validation set. The final model is obtained by the average of parameters from these three training models. The parameters for charge related atom types C_{+1} , N_{+1} , H_C , C_C and O_A are optimized while fixing all other parameters obtained from the neutral model with a charged training set (Table 2.A.5) of 17 molecules including small charged molecules and tripeptides, and tested on a charged test set (Table 2.A.6) of 18 tripeptides that are positively charged, negatively charged or zwitterionic. Finally we also test the various models on the crambin protein (PDB ID 1CRN [51]), whose hydrogens are added using the Reduce (3.23) software.[62]

ESP generated by Gaussian charges vs point charges. There are two approaches to generate the ESP from a set of core and shell positions. One is the Gaussian charge approach, where the ESP is computed by

$$V(\mathbf{r}) = \sum_i^{cores} E_{ik}^{elec}(|\mathbf{r} - \mathbf{r}_i|) + \sum_j^{shells} E_{jk}^{elec}(|\mathbf{r} - \mathbf{r}_j|) \quad (2.19)$$

where a point(k) in space is treated as a fictitious core with $q_k = +1$ and $\alpha_k = 1569.8$, which is a Gaussian sharply peaked at position \mathbf{r} . This approach is the natural

approach arise from the Gaussian definition of cores and shells. The other approach is to treat all cores and shells as point charges when calculating the ESP:

$$V(\mathbf{r}) = \sum_{i \in \text{cores}} \frac{q_i}{(|\mathbf{r} - \mathbf{r}_i|)} + \sum_{i \in \text{shells}} \frac{q_i}{(|\mathbf{r} - \mathbf{r}_i|)} \quad (2.20)$$

Note that this treatment of approximate cores and shells as point charges at their Gaussian center is only done in the process of generating the ESP, not in the optimization of shell positions. The two approaches gives essentially indistinguishable prediction in ESP as shown in Figure 2.2a), where the mean ESP generated with Gaussian charges aligns perfectly ($R^2 = 0.99999993$) with that generated with point charges. The average MAE between ESP generated with Gaussian charges and ESP generated with point charges is only $3.97 * 10^{-4}$ eV, which is trivial compared to the average magnitude of ESP at 0.755 eV. However, the point charge approach is advantageous in terms of calculation speed as it avoids the relatively expensive operation of erf function evaluation. This is demonstrated in Figure 2.2b), where the ESP time (the time to compute ESP from fixed core and shell positions) is plotted against the number of grid points for all molecules used in for training and testing process for parameter optimization. The ESP time for the point charge treatment is clearly faster than the Gaussian charge treatment by roughly a factor of 10. When the number of points is a large number, this difference can be significant to influence the efficiency of ESP evaluation. As both methods provide essentially the same accuracy and the point charge treatment is clearly faster, in the following discussion of this paper, we will adopt the point charge approach to calculate the ESP. In the cases where the ESP grid points of interest is closer to the atomic center, we switch back to the Gaussian charge implementation.

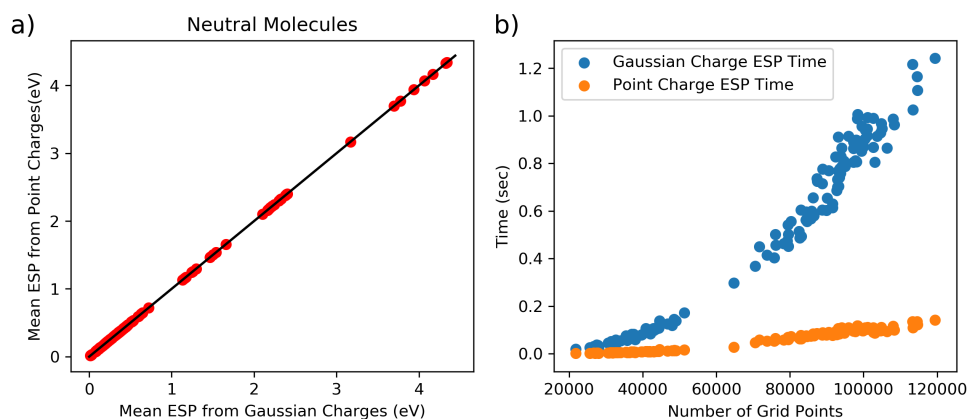


Figure 2.2: **a)** The mean ESP generated with Gaussian charges aligns perfectly with that generated with point charges. **b)** The time to compute ESP from core and shell positions with respect to number of grid points for different molecules using point charge and Gaussian charge treatment.

DFT reference and other methods for computing the ESP. The reference *ab initio* ESP for all molecules except crambin are generated with the Q-Chem 5.2 software package[49] using the ω B97X-V functional [31] with the def2-QZVPP basis set; the ESP of crambin is generated using ω B97X-V with the cc-pVDZ basis set. We also compare the results of C-GeM with other available methods including EEM, AM1-BCC, Hirshfeld, Iterative Hirshfeld, MBIS and AVH. The EEM-derived charges are obtained from the LAMMPS[42] ReaxFF[1] implementation of EEM using the peptide and protein parameters.[34] The AM1-BCC charges are obtained from antechamber tool part of AMBERTools. The Hirshfeld, Iterative Hirshfeld, MBIS and AVH charges are computed with IOData[57], ChemTools[18] and HORTON 2.1.1 software packages.[56]

In addition, the electrostatic potential for crambin has been performed with continuum electrostatic calculations using the Adaptive Poisson–Boltzmann Solver (APBS) v3.0.0. [22] For APBS the hydrogen added crambin structure was prepared with PDB2PQR v3.1.0 [11] using the AMBER force field, and enabling the generation of pqr files with atomic charges and radii. APBS computations were carried out with the linearized PB equation with a 1.0 dielectric constant for solvent and solute (protein) to mimic the vacuum condition in other calculations. Temperature was set to 298.15 K, and a single Debye–Hückel boundary condition was applied. The grid dimension was set to 353 x 353 x 353 such that the grid spacing is 0.149 x 0.125 x 0.150 Å, similar to the grid spacing in the molecular surface grid we used in the *ab*

initio calculations. The ESP generated with APBS was mapped onto the molecular surface grid through the multivalue utility in APBS software package.

Grid resolution and timing metrics. The grid points on which electrostatic potentials are evaluated are generated following the Merz-Singh-Kollman (MK) scheme [50] on 10 evenly distributed layers of range from 1.4-2.54 vdW radii distance. Here we report the ESP generated on an average of 37,000 grid points for small protein analogs (average 12.4 atoms) and 90,000 grid points for tripeptides (average 37.4 atoms). The computation times are measured with the `timeit` python module on a single core Intel XEON Gold 6230 CPU unless otherwise mentioned. The times for DFT calculations are obtained from QChem output files.

2.4 RESULTS AND DISCUSSIONS

Neutral small protein analogs and tripeptides

In this study, we trained three protein C-GeM models that share common global parameters ω_{core} , γ_{shell} , λ , R_{shell} and χ_{shell} : 1) C-GeM without atom typing, where each element (H, C, N, O, S, Cl) has its own atomic parameters for the ionization potential and atomic radius. 2) C-GeM with C and H atom typed, where C is classified into polar carbon (C_A) and aliphatic carbon (C_B) based on whether it has an electronegative neighboring atom (N,O,S,Cl), and H is classified into polar hydrogen (H_A) and aliphatic hydrogen (H_B) in the same way. 3) C-GeM with C, H and N atom typed, where on top of model 2, we further classify nitrogen according to the number of H neighbors it has, into N_A for N with 2 H neighbors, N_B for N with 1 H neighbor and N_C for N with no H neighbor. These three models are referred to as CGem, CGem_CH and CGem_CHN respectively, and their parameters are shown in Table 2.1.

Table 2.1: Parameters for C-GeM models CGem, CGem.CH and CGem.CHN. H_A for polar hydrogen, H_B for aliphatic hydrogen, H_C for hydrogen directly bonded to positive atoms; C_{+1} for carbon with a positive formal charge, C_A for polar carbon, C_B for aliphatic carbon, C_C for carbon directly bonded to positive atoms; N_{+1} for nitrogen with a positive formal charge, N_A for N with 2 H neighbors, N_B for N with 1 H neighbor and N_C for N with no H neighbor; O_A for oxygens in negatively charged acetate group

global parameters						
$\omega_{core}(\text{\AA}^{-1})$	$\gamma_{shell}(\text{\AA}^{-2})$	λ	$R_{shell}(\text{\AA})$	χ_{shell} (eV)		
0.152	5.220	2.103	0.708	19.956		
C-GeM atomic parameters						
atom type	CGem		CGem.CH		CGem.CHN	
	R(\AA)	χ (eV)	R(\AA)	χ (eV)	R(\AA)	χ (eV)
H	0.67	-16.33	-	-	-	-
C	0.59	-19.12	-	-	-	-
N	0.44	-21.85	0.55	-23.08	-	-
O	0.34	-24.26	0.54	-22.83	0.51	-23.35
S	0.66	-21.28	0.86	-18.43	0.84	-19.29
Cl	0.31	-25.43	0.63	-21.73	0.56	-22.87
H_A	-	-	0.22	-12.97	0.20	-13.79
H_B	-	-	0.68	-16.42	0.65	-16.95
H_C	0.81	-15.49	0.52	-13.35	0.57	-13.52
C_A	-	-	0.77	-15.12	0.75	-15.74
C_B	-	-	0.57	-19.48	0.57	-19.49
C_C	0.60	-19.52	0.71	-13.26	0.72	-14.32
N_A	-	-	-	-	0.54	-23.65
N_B	-	-	-	-	0.61	-20.04
N_C	-	-	-	-	0.50	-24.43
O_A	0.62	-22.78	0.58	-23.50	0.60	-23.57
C_{+1}	0.55	-31.93	0.68	-30.15	0.74	-31.50
N_{+1}	0.88	-27.99	0.73	-38.02	0.72	-39.13

To sample the protein chemistry space, we developed the models with data from small protein analogs that covers the basic functional groups and scaffolds for peptides, along with tripeptides that describe actual protein chemistry but are small enough for high quality *ab initio* computation. The performance of these models was evaluated

with respect to MAE_{avg} and RMSE_{avg} between the ESP of the reference $\omega\text{B97X-V}/\text{def2-qzvpp}$ theory and the ESP generated by the various models, as well as the dipole error obtained as the norm of the difference vector by subtracting the *ab initio* reference dipole from the approximate dipole. In Figure 2.3 we present the MAE_{avg} and mean dipole error of the C-GeM models as well as empirically derived partial charge method EEM and QM-calculation-based atomic partial charge methods Hirshfeld, HI, MBIS and AVH. The statistics of the results including RMSE_{avg} are listed in Table 2.A.1. Among the C-GeM models, atom typing hydrogen and carbon improves the MAE_{avg} from 0.067 eV to 0.059 eV and RMSE_{avg} from 0.094 eV to 0.082 eV in terms of ESP quality for small protein analogs, and improves the MAE_{avg} from 0.122 eV to 0.084 eV and RMSE_{avg} from 0.166 eV to 0.117 eV in terms of ESP quality for tripeptides. Nitrogen atom typing further improves the ESP MAE_{avg} and RMSE_{avg} down to 0.053 eV and 0.077 eV for small protein analogs, and 0.070 eV and 0.101 eV for tripeptides respectively. Oxygen atom typing were explored, but it showed minimal improvements on the training molecules while introducing an overfitting problem that degrades the results for the validation set. Therefore, oxygen was kept as its elemental type.

All three C-GeM models significantly outperform the EEM model (0.094 eV MAE_{avg} , for small protein analogs and 0.185 eV MAE_{avg} for tripeptides), which is the only method of comparable computational cost to C-GeM. The C-GeM models are also more accurate for tripeptides than the AM1-BCC charges (0.96 eV MAE_{avg}) that relies on the semi-empirical AM1 method and thus is computationally slower than C-GeM. All C-GeM models are significantly better than Hirshfeld (0.106 eV and 0.176 eV MAE_{avg} respectively), whereas the best CGem_CHN model also outperforms the AVH method (0.086 eV and 0.100 eV) by 30%, and slightly outperforms the HI method (0.058 eV and 0.080 eV), while MBIS (0.040 eV and 0.053 eV) remains the best among all methods, albeit with much greater expense and thus not affordable for proteins.

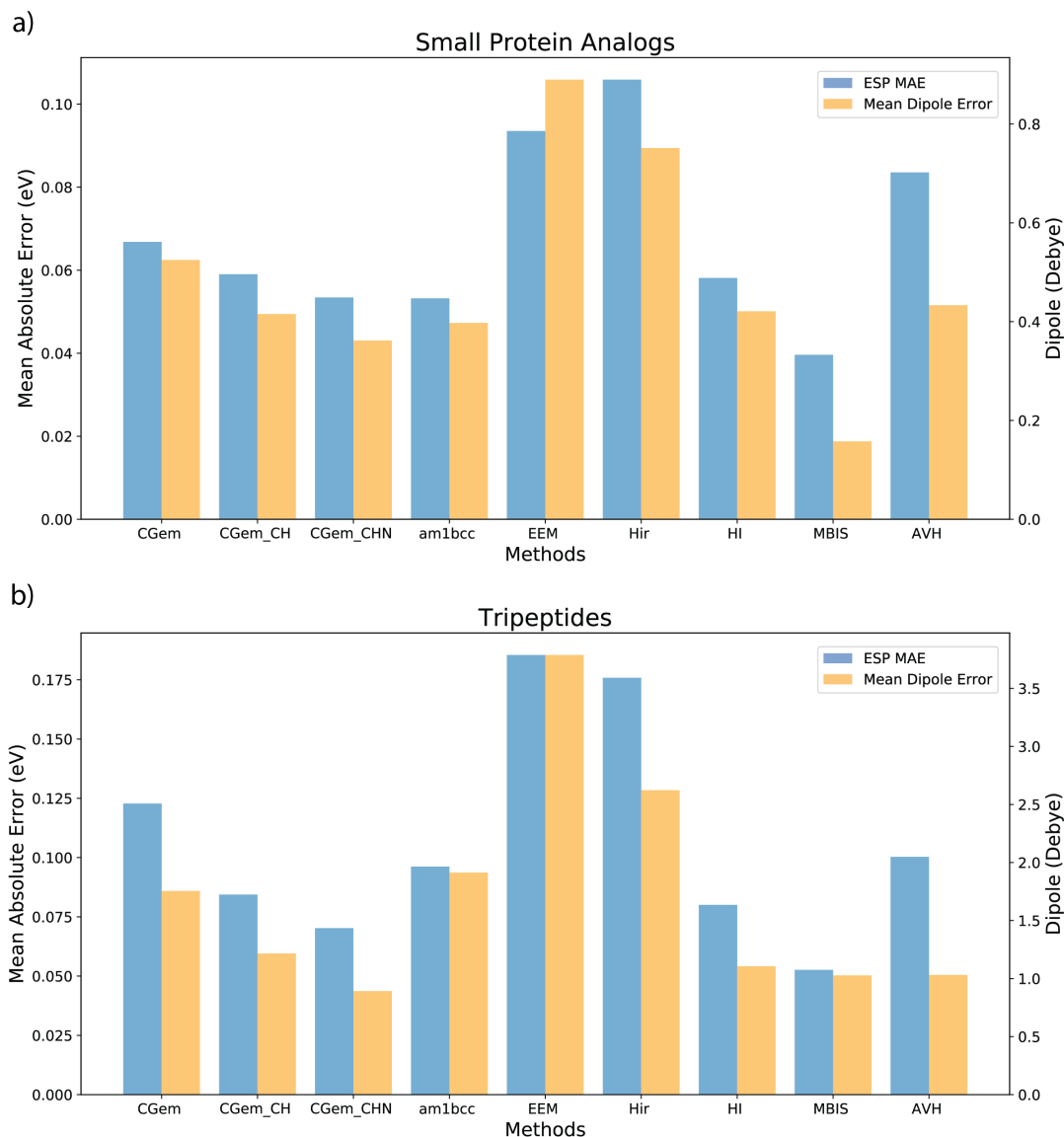


Figure 2.3: Average mean absolute error electrostatic potential and average dipole error of different atom typed C-GeM models, EEM, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V/ def2-qzvpp reference for a) 54 small protein analogs and b) 57 tripeptides, labeled with the average error and standard deviation within the set.

While producing an accurate ESP description that is comparable to *ab initio* cal-

ulation based charges, the C-GeM models produce excellent prediction for molecular dipoles, which is a property that the model was not parameterized for, but arises naturally from off-centered shell positions. Atom typing improves the mean dipole errors for C-GeM models, from 0.525 Debye in CGem to 0.415 Debye in CGem.CH to 0.362 Debye in CGem.CHN for the small protein analogs set, and from 1.755 Debye in CGem to 1.216 Debye in CGem.CH to 0.892 Debye in CGem.CHN for the tripeptide set, similar in trend as to how atom typing improves the ESP MAE_{avg} and $RMSE_{avg}$. In the small protein analogs set, the mean dipole error of CGem.CHN is only inferior to that of MBIS (0.158 Debye), and superior to all other QM based or empirically derived methods including EEM (0.890 Debye), AM1-BCC (0.397 Debye), Hirshfeld (0.751 Debye), HI (0.421 Debye) and AVH (0.433 Debye). For the tripeptide set, CGem.CHN produces the best mean dipole error among all of the methods including MBIS (1.028 Debye), HI (1.106 Debye), AVH (1.032 Debye), whereas EEM (3.788 Debye), Hirshfeld (2.623 Debye) and AM1-BCC (1.914 Debye), have errors larger than all of the C-GeM models.

For all of the methods, the tripeptides are more challenging to predict than the small protein analogs because of their intrinsically larger size. For instance, the mean absolute ESP value is 0.190 eV for small protein analogs and 0.415 eV for tripeptides, and the mean dipole magnitude is 1.569 Debye for small protein analogs but 7.389 Debye for tripeptides. However, the relative performance among the methods we compared is quite stable across the two different datasets. The fact that the performance of the C-GeM models are relatively stable compared to QM based charge partitioning method, which are general methods that does not distinguish sizes or specific protein chemistry, reflects that C-GeM is transferable with respect to system size for protein like molecules.

While having similar accuracy, the C-GeM models are orders of magnitude faster than the QM charge partitioning approaches that are based on high quality *ab initio* calculations. The DFT benchmark calculation (ω B97X-V/def2-qzvpp) on average takes 8.4 minutes and 6.9 hours per molecule for small protein analogs and tripeptides, respectively, even after taking advantage of OpenMP parallelization techniques, and the QM based charge partitioning methods require additional steps to partition atomic densities on top of the QM calculation. The AM1-BCC method takes 8.38 seconds for small protein analogues and 5.77 minutes for tripeptides using antechamber program, and although more efficient than the QM-based methods, is also 2-3 orders of magnitude slower than our C-GeM models.

Table 2.2: Computation time per molecule of C-GeM, CH atomtyped C-GeM, CHN atomtyped C-GeM on small protein analogs and tripeptides. Charge time is the time to initialize and optimize shell positions for C-GeM models, and ESP time is the time to map C-GeM cores and shells onto predefined grid points for electrostatic potential.

Small Protein Analogs			
	Charge Time (sec)	ESP Time (sec)	Total Time (sec)
CGeM	0.053	0.010	0.064
CGeM.CH	0.047	0.009	0.057
CGeM.CHN	0.044	0.009	0.053
Tripeptides			
	Charge Time (sec)	ESP Time (sec)	Total Time (sec)
CGeM	0.126	0.081	0.207
CGeM.CH	0.133	0.080	0.213
CGeM.CHN	0.121	0.080	0.201

By contrast, the C-GeM models can predict the ESP on the order of tenth of a second on a single core of Intel XEON Gold 6230 CPU, and all C-GeM models have very similar computational timings for the ESP, about 0.01 seconds for small protein analogs and about 0.08 seconds for tripeptides (Table 2.2), which is comparable to the EEM class of methods. The actual timing comparisons between EEM and C-GeM models are not directly comparable because the EEM times were obtained with a C++ code in LAMMPS and the C-GeM times were obtained with our in-house Python code, but EEM is the same order of magnitude for the system sizes we’ve investigated until this point; we return to timings again later in the crambin protein case. The internal comparisons among C-GeM models shows that atom typing did not slow down the calculation, despite adding additional step to classify the atoms. The charge time decreases in the order of CGem, CGem.CH, and CGem.CHN in both the small protein analogs set and the tripeptides set, which suggests that atom typing of C,H and N helps the shell optimization process to converge faster.

To demonstrate that the C-GeM model can deal with the conformational variations of a molecule, we compute C-GeM ESP for a tripeptide molecule randomly selected from the PEPCONF[44] dataset. The error of C-GeM models relative to the ω B97X-V / def2-qzvpp reference on the 6 conformations of tripeptide LEU_TYR_GLN(Figure 2.B.1) are shown in Table 2.3, which supports the fact that all C-GeM models yield stable predictions on varied conformations of the same molecule.

Table 2.3: Mean absolute error (MAE) in eV on electrostatic potential (ESP) of different atom typed C-GeM models with respect to ω B97X-V / def2-qzvpp reference on 6 conformations of tripeptide LEU_TYR_GLN.

molecule	CGem MAE	CGem_CH MAE	CGem_CHN MAE
CONF_1	0.087	0.072	0.063
CONF_2	0.108	0.067	0.083
CONF_3	0.105	0.073	0.075
CONF_4	0.109	0.077	0.078
CONF_5	0.088	0.064	0.061
CONF_6	0.114	0.075	0.081

Charged small protein analogs and tripeptides

In the previous section, we demonstrated that C-GeM models can predict the ESP of molecules at accuracy comparable to *ab initio* generated charges but orders of magnitude faster for neutral small protein analogs and tripeptides. However, proteins under physiological conditions have residues that are charged under neutral pH, which would need specialized treatment in the C-GeM models. We considered two residues that are negative under neutral pH, aspartic acid (Asp) and glutamic acid (Glu), and two residues that are positive under neutral pH, arginine (Arg) and lysine (Lys). For the negatively charged residues, an extra shell is added onto the negatively charged atom (O_A for negative oxygen) as shown in Figure 2.4(a), which creates a net charge of -1 localized around the negatively charged atom. For positively charged residues, the idea is to assign the core of the charged atom a +2 charge and mark it as a different atom type (C_{+1} for carbon and N_{+1} for nitrogen as shown in Table 2.1), while still having a shell on that atom to allow for shell movements. As shown in Figure

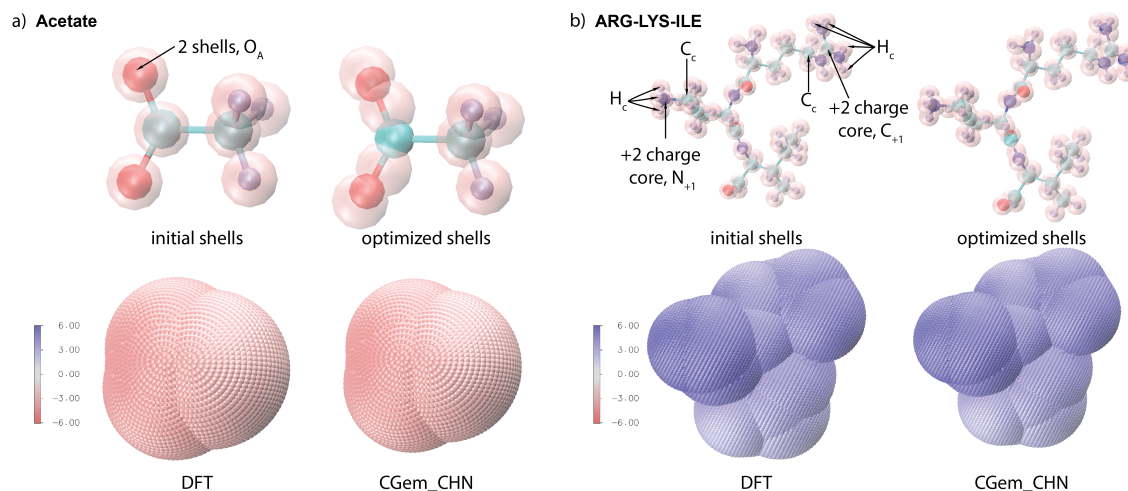


Figure 2.4: Demonstration for C-GeM on charged molecules **a)** Methylammonium (net -1 charge) **b)** Tripeptide ARG-LYS-ILE (net +2 charge)

2.4(b) for Lys, the positive nitrogen carries a +2 core, and for Arg, we placed the +2 core on the guanidino carbon instead of the formally charged nitrogen to account for the equivalence of the two guanidino nitrogens. We also find it useful to have separate atom type for the hydrogens (H_C) and carbons (C_C) that are directly bonded to the positive atoms.

With this protocol, we trained the parameters for the charged atoms, and fixing all of the parameters we obtained from the neutral model, using a training set of 17 molecule consisting of 4 small side chain analog molecules and 13 tripeptides that are positive, negative or zwitterionic (Table 2.A.5). The resulting models were tested on another 18 tripeptides (Table 2.A.6) with charged residues that the model has not seen. The MAE_{avg} and mean dipole errors C-GeM models on the charged dataset compared to QM based charges are presented in Figure 2.5 and Table 2.A.2. EEM charges are not included because the LAMMPS implementation of EEM fails to deal with non-zero charges. The charged molecules in general have larger mean ESP values (2.26 eV for the charged training set and 2.24 eV for the charged test set) and much larger dipoles (78.1 Debye and 137.1 Debye for the charged training and test set, respectively), which could make the prediction more difficult.

In general Figure 2.5 and Table 2.A.2 show that the C-GeM models exhibit a stable performance on these difficult charged molecules that are not too far from their corresponding performance on the neutral molecules. The basic CGeM model yields 0.102 eV MAE_{avg} for the charged training set and 0.116 eV MAE_{avg} while the best

CGem_CHN model yields 0.081 eV MAE_{avg} for the charged training set, and 0.076 eV MAE_{avg} for the charged test set. This is a significant improvement in MAE_{avg} for charge training and test set, respectively, over Hirshfeld charges (0.154 eV and 0.174 eV), and comparable to HI (0.060 eV and 0.071 eV), MBIS (0.060 eV and 0.065 eV) and AVH (0.078 eV and 0.086 eV). The dipole errors exhibit a similar trend: the best C-GeM model CGem_CHN reports a dipole error of 1.34 Debye for the charged training set and 1.15 Debye for charged test set, which is comparable to MBIS (1.33 Debye and 1.36 Debye) and significantly improved over Hirshfeld (1.96 Debye and 2.50 Debye), but worse than HI (0.88 Debye and 1.10 Debye) and AVH (0.88 Debye and 0.88 Debye). Both the trend and the numbers are very similar across the charged training set and testing set, which shows the generality of the models. The AM1-BCC charges are relatively accurate in the neutral molecule case, but clearly have some difficulty in predicting charged protein chemistry molecules, giving rise to a MAE of 0.161 eV and 0.149 eV, and dipole error at 3.67 Debye and 3.36 Debye for the charged training and test set, respectively.

It is worth noting that the overall dipole error is amplified due to the large magnitude, as we are defining dipole error as the norm of the difference vector between C-GeM or partial charge derived dipoles and the DFT dipole $|\boldsymbol{\mu}_{DFT} - \boldsymbol{\mu}_{C-GeM}|$, which captures both the magnitude and directional information. Hence with a large dipole, a small deviation in the angle θ between $\boldsymbol{\mu}_{DFT}$ and $\boldsymbol{\mu}_{C-GeM}$ can result in large errors in the norm of the difference vector, even if the error in magnitude $||\boldsymbol{\mu}_{DFT}| - |\boldsymbol{\mu}_{C-GeM}||$ is small. For instance, the 1.23 Debye dipole error in CGem_CHN can be decomposed into 0.65 Debye error in magnitude and 0.91° in θ .

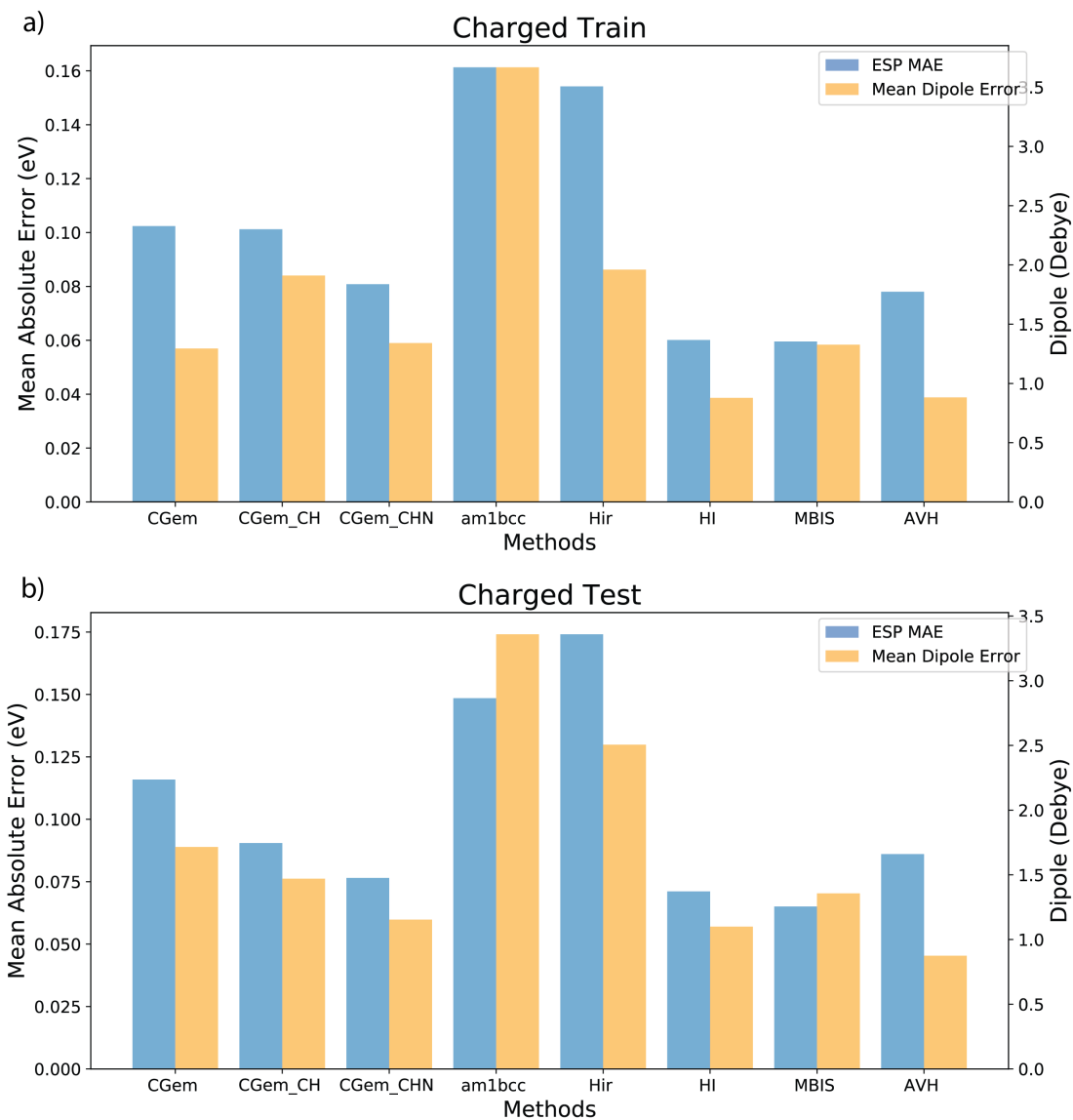


Figure 2.5: Mean absolute error (MAE) on electrostatic potential (ESP) and dipole error of different atom typed C-GeM models, AM1-BCC, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for a) training set for charged side chains b) testing set for charged side chains.

Evaluation of C-GeM model on the crambin protein

As the C-GeM models worked well to reproduce the DFT benchmark for the ESP and dipole directions in both the neutral and the charged cases of small molecules and protein fragments, we precede to examine C-GeM models on a full protein, crambin, which is difficult in terms of resources for the QM based partial charge methods, but totally accessible for C-GeM models as they are orders of magnitudes faster.

The ESP map of crambin is shown in Figure 2.6 with their minimum and maximum ESP value labeled. The C-GeM models give qualitatively correct predictions for the ESP compared to the DFT reference computed ESP with ω B97X-V / cc-pVDZ, with MAE of 0.13, 0.12 and 0.11 for CGem, CGem.CH and CGem.CHN respectively. These predictions are superior to the EEM method (0.49 eV MAE), which fails to describe the ESP qualitatively correctly due to unphysical long-range charged transfer, and APBS (0.25 eV MAE), which essentially is due to the AMBER ESP fitted partial charges (the dielectric constant was set to 1 to account for protein in vacuum of all methods). The CGem.CHN predicts -4.26 eV and 3.02 eV as minimum and maximum on the ESP surface, which is very close to -4.37 eV and 3.03 eV predicted by the DFT reference at the same position in space. By contrast the EEM method yields a more featureless ESP, predicting a minimum and maximum of -1.42 eV and 0.71 eV, whereas the APBS result exaggerates the extremes with -5.15 eV and 3.67 eV for the minimum and maximum, respectively. Finally, the best C-GeM model CGem.CHN also gives a relatively acceptable dipole error of 6.45 Debye compared to the total 37.8 Debye for crambin as determined by the DFT benchmark. In this case the EEM dipole moment is egregiously incorrect.

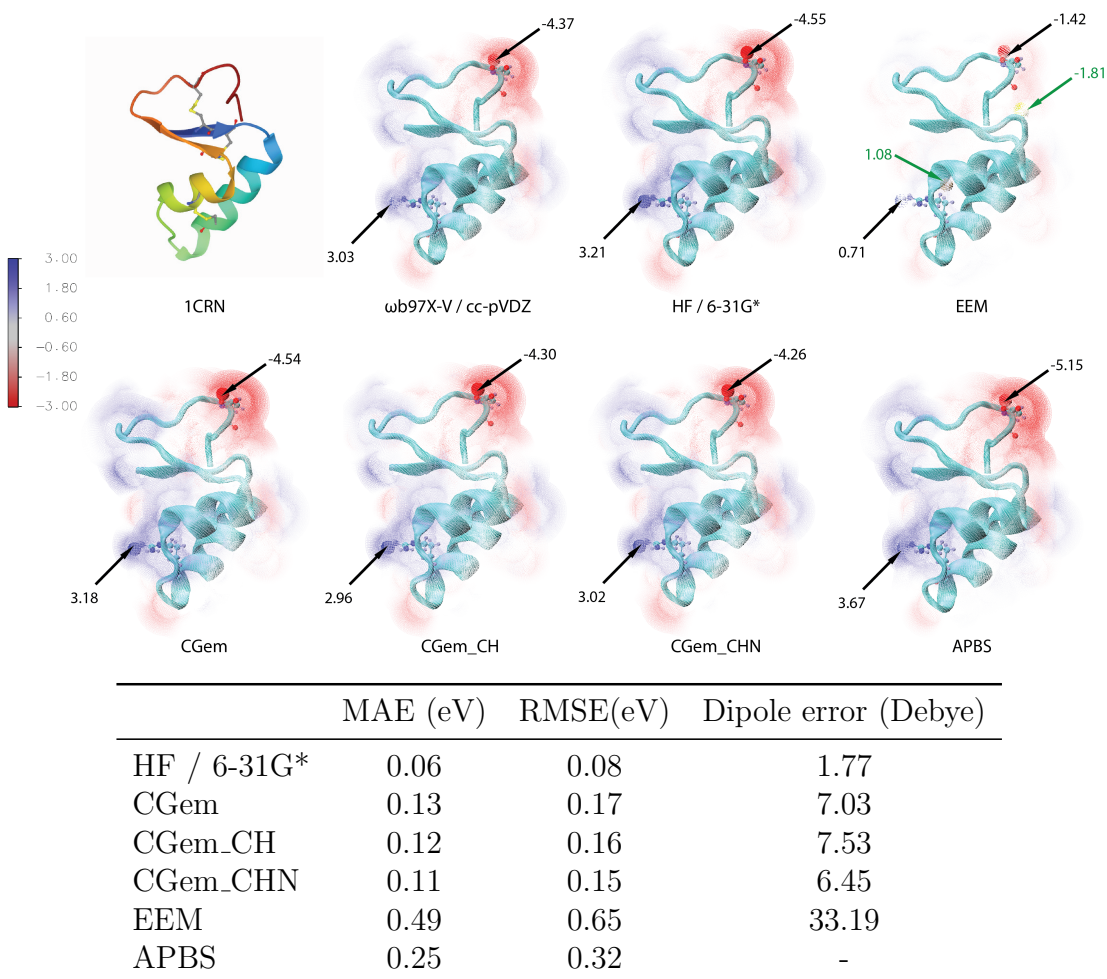


Figure 2.6: Predicted ESP figure for crambin(1CRN) with ω B97X-V / cc-pVDZ, HF/6-31G*, EEM, CGem, CGem.CH, CGem.CHN and APBS. The electrostatic potential (in eV) at points with maximum and minimum ESP value for ω B97X-V / cc-pVDZ are labeled. The table presents the MAE and RMSE on ESP and the dipole error of these methods with respect to ω B97X-V / cc-pVDZ reference for crambin.

The advantage in computational efficiency for the C-GeM models is very significant in the case of this larger molecule of more than 600 atoms. The C-GeM models can predict the ESP on more than 500,000 grid points within 20 seconds, which is five orders of magnitude faster than the ω B97X-V / cc-pVDZ reference. The C-GeM models are also faster than APBS at the same grid resolution, noting that the speed of APBS suffer from first computing the ESP on a full-space grid of similar spacing,

and then interpolation onto the molecular surface grid. The best C-GeM model, CGem_CHN (13.7 sec) is faster than CGem (15.6 sec) and CGem_CH (17.4 sec) despite it requiring additional steps of atom typing, which again shows that atom typing speeds up the convergence of the shell position in the optimization cycles.

2.5 CONCLUSIONS AND OUTLOOK

The ability to generate accurate electrostatic potential surfaces for predicting protein binding motifs with high computational efficiency for high-throughput screening of drug molecules is an important area for structural based drug discovery. At present this dual goal of accuracy and efficiency has been difficult to achieve. Here we have introduced a new method for generating the ESP that is both accurate and fast using the C-GeM approach. We have shown that it offers accuracy comparable to the expensive *ab initio* methods with orders of magnitude reduction in expense, and is far more accurate than cheaper computational alternatives such as EEM or PBE approaches.

We have also shown that the EEM model and the density partitioning Hirshfeld schemes are the least competitive in regards accuracy, which is not surprising, but are compared here because of their continued popularity. The AM1-BCC model, usually thought of as an efficient method, was found to be inferior to the C-GeM models in both efficiency and accuracy, and is found to be unstable when computing charged protein fragments. While more first principle approaches such as HI, MBIS, or AVH are relatively accurate, they are computationally expensive and thus unsuitable for high-throughput computation on large proteins or for the many molecules required for high throughput screening applications.

In summary, the C-GeM force field accuracy comes in part from eliminating unphysical long-range charge transfer, by accounting for out-of-plane polarization, and charges are not required to be centered on atoms, thereby accounting for electrostatic features such as sigma holes that define important binding motifs for biomolecules. The C-GeM model is light-weight in parameters compared to other many-body force fields such as AMOEBA [43], which has many more atom types and many more parameters such as the atomic multipoles up through quadrupoles, atomic polarizability parameters, and damping functions. By contrast the protein C-GeM model has at most 15 atom types each with 2 atomic parameters that represent the electronegativity and ionization potential, and a common set of 5 global parameters for all atoms. In future development work we will advance C-GeM further to account for more complicated solvent environments and physiological salt conditions that are important for biomolecular recognition, and apply the model to more diverse

applications beyond ESP predictions.

2.6 ACKNOWLEDGMENTS

The methods work was supported by the National Science Foundation under grant CHE-1955643 and the application area by the C3.ai Digital Transformation Institute. We also thank the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

2.7 REFERENCES

- [1] H. M. Aktulga, J. C. Fogarty, S. A. Pandit, and A. Y. Grama. “Parallel Reactive Molecular Dynamics: Numerical Methods and Algorithmic Techniques”. In: *Parallel Computing* 38.4 (Apr. 2012), pp. 245–259. ISSN: 0167-8191. DOI: 10.1016/j.parco.2011.08.005.
- [2] Richard F. W. Bader. “A Quantum Theory of Molecular Structure and its Applications”. In: *Chemical Reviews* 91.5 (1991), pp. 893–928. DOI: 10.1021/cr00005a013. eprint: <https://doi.org/10.1021/cr00005a013>. URL: <https://doi.org/10.1021/cr00005a013>.
- [3] L.W. Bertels, L.B. Newcomb, M. Alaghemandi, J.R. Green, and M. Head-Gordon. “Benchmarking the Performance of the ReaxFF Reactive Force Field on Hydrogen Combustion Systems”. In: *Journal of Physical Chemistry A* 124 (2020), pp. 5631–5645. DOI: 10.1021/acs.jpca.0c02734.
- [4] Jacob R. Boes, Mitchell C. Groenenboom, John A. Keith, and John R. Kitchin. “Neural Network and ReaxFF Comparison for Au Properties”. In: *International Journal of Quantum Chemistry* 116.13 (2016), pp. 979–987. DOI: <https://doi.org/10.1002/qua.25115>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.25115>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.25115>.
- [5] Curt M. Breneman and Kenneth B. Wiberg. “Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis”. In: *Journal of Computational Chemistry* 11.3 (1990), pp. 361–373. ISSN: 1096-987X. DOI: 10.1002/jcc.540110311.

- [6] Patrick Bultinck, Christian Van Alsenoy, Paul W. Ayers, and Ramon Carbó-Dorca. “Critical Analysis and Extension of the Hirshfeld Atoms in Molecules”. In: *The Journal of Chemical Physics* 126.14 (Apr. 2007), p. 144111. ISSN: 0021-9606. DOI: 10.1063/1.2715563.
- [7] De-Li Chen, Abraham C. Stern, Brian Space, and J. Karl Johnson. “Atomic Charges Derived from Electrostatic Potentials for Molecular and Periodic Systems”. In: *The Journal of Physical Chemistry A* 114.37 (Sept. 2010), pp. 10225–10233. ISSN: 1089-5639, 1520-5215. DOI: 10.1021/jp103944q.
- [8] Art E. Cho, Victor Guallar, Bruce J. Berne, and Richard Friesner. “Importance of Accurate Charges in Molecular Docking: Quantum Mechanical/Molecular Mechanical (QM/MM) Approach”. In: *Journal of Computational Chemistry* 26.9 (2005), pp. 915–931. ISSN: 1096-987X. DOI: <https://doi.org/10.1002/jcc.20222>.
- [9] Minsik Cho, Nitai Sylvetsky, Sarah Eshafi, Golokesh Santra, Irena Efremenko, and Jan M. L. Martin. “The Atomic Partial Charges Arboretum: Trying to See the Forest for the Trees”. In: *ChemPhysChem* 21.8 (2020), pp. 688–696. DOI: <https://doi.org/10.1002/cphc.202000040>. eprint: <https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.202000040>. URL: <https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cphc.202000040>.
- [10] Beatriz Cordero, Verónica Gómez, Ana E. Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. “Covalent Radii Revisited”. In: *Dalton Transactions* 0.21 (2008), pp. 2832–2838. DOI: 10.1039/B801115J.
- [11] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker. “PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations”. In: *Nucleic Acids Research* 32.Web Server (July 2004), W665–W667. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkh381.
- [12] Adri C. T. van Duin, Siddharth Dasgupta, Francois Lorant, and William A. Goddard. “ReaxFF: A Reactive Force Field for Hydrocarbons”. In: *The Journal of Physical Chemistry A* 105.41 (2001), pp. 9396–9409. DOI: 10.1021/jp004368u. eprint: <https://doi.org/10.1021/jp004368u>.
- [13] Fuchang Gao and Lixing Han. “Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters”. In: *Computational Optimization and Applications* 51.1 (Jan. 2012), pp. 259–277. ISSN: 1573-2894. DOI: 10.1007/s10589-010-9329-3.

- [14] Stanislav Geidl, Tomáš Bouchal, Tomáš Raček, Radka Svobodová Vařeková, Václav Hejret, Aleš Křenek, Ruben Abagyan, and Jaroslav Koča. “High-quality and Universal Empirical Atomic Charges for Chemoinformatics Applications”. In: *Journal of Cheminformatics* 7.1 (Dec. 2015), p. 59. ISSN: 1758-2946. DOI: 10.1186/s13321-015-0107-1.
- [15] Farnaz Heidar-Zadeh and Paul W. Ayers. “How Pervasive is the Hirshfeld Partitioning?” In: *The Journal of Chemical Physics* 142.4 (2015), p. 044107. DOI: 10.1063/1.4905123. eprint: <https://doi.org/10.1063/1.4905123>. URL: <https://doi.org/10.1063/1.4905123>.
- [16] Farnaz Heidar-Zadeh, Paul W. Ayers, and Patrick Bultinck. “Deriving the Hirshfeld Partitioning using Distance Metrics”. In: *The Journal of Chemical Physics* 141.9 (2014), p. 094103. DOI: 10.1063/1.4894228. eprint: <https://doi.org/10.1063/1.4894228>. URL: <https://doi.org/10.1063/1.4894228>.
- [17] Farnaz Heidar-Zadeh, Paul W. Ayers, Toon Verstraelen, Ivan Vinogradov, Esteban Vöhringer-Martinez, and Patrick Bultinck. “Information-Theoretic Approaches to Atoms-in-Molecules: Hirshfeld Family of Partitioning Schemes”. In: *The Journal of Physical Chemistry A* 122.17 (May 2018), pp. 4219–4245. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.7b08966.
- [18] Farnaz Heidar-Zadeh et al. “An Explicit Approach to Conceptual Density Functional Theory Descriptors of Arbitrary Order”. In: *Chemical Physics Letters* 660 (2016), pp. 307–312. ISSN: 0009-2614. DOI: <https://doi.org/10.1016/j.cplett.2016.07.039>. URL: <https://www.sciencedirect.com/science/article/pii/S0009261416305280>.
- [19] F. L. Hirshfeld. “Bonded-atom Fragments for Describing Molecular Charge Densities”. In: *Theoretica chimica acta* 44.2 (June 1977), pp. 129–138. ISSN: 1432-2234. DOI: 10.1007/BF00549096.
- [20] Crina-Maria Ionescu, Stanislav Geidl, Radka Svobodová Vařeková, and Jaroslav Koča. “Rapid Calculation of Accurate Atomic Charges for Proteins via the Electronegativity Equalization Method”. In: *Journal of Chemical Information and Modeling* 53.10 (Oct. 2013), pp. 2548–2558. ISSN: 1549-9596, 1549-960X. DOI: 10.1021/ci400448n.
- [21] Araz Jakalian, David B. Jack, and Christopher I. Bayly. “Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation”. In: *Journal of Computational Chemistry* 23.16 (Dec. 2002), pp. 1623–1641. ISSN: 0192-8651. DOI: 10.1002/jcc.10128.

- [22] Elizabeth Jurrus et al. “Improvements to the APBS Biomolecular Solvation Software Suite”. In: *Protein Science* 27.1 (2018), pp. 112–128. ISSN: 1469-896X. DOI: <https://doi.org/10.1002/pro.3280>.
- [23] Gerald Knizia. “Intrinsic Atomic Orbitals: An Unbiased Bridge between Quantum Theory and Chemical Concepts”. In: *Journal of Chemical Theory and Computation* 9.11 (2013), pp. 4834–4843. DOI: 10.1021/ct400687b. eprint: <https://doi.org/10.1021/ct400687b>. URL: <https://doi.org/10.1021/ct400687b>.
- [24] Anmol Kumar, Ozge Yoluk, and Alexander D. MacKerell Jr. “FFParam: Standalone Package for CHARMM Additive and Drude Polarizable Force Field Parametrization of Small Molecules”. In: *Journal of Computational Chemistry* 41.9 (2020), pp. 958–970. ISSN: 0192-8651. DOI: <https://doi.org/10.1002/jcc.26138>. URL: <https://doi.org/10.1002/jcc.26138>.
- [25] G. Lee Warren, Joseph E. Davis, and Sandeep Patel. “Origin and Control of Superlinear Polarizability Scaling in Chemical Potential Equalization Methods”. In: *The Journal of Chemical Physics* 128.14 (Apr. 2008), p. 144110. ISSN: 0021-9606. DOI: 10.1063/1.2872603.
- [26] Itai Leven and Teresa Head-Gordon. “C-GeM: Coarse-Grained Electron Model for Predicting the Electrostatic Potential in Molecules”. In: *The Journal of Physical Chemistry Letters* 10.21 (Nov. 2019), pp. 6820–6826. DOI: 10.1021/acs.jpcllett.9b02771.
- [27] Jie Li, Kochise C. Bennett, Yuchen Liu, Michael V. Martin, and Teresa Head-Gordon. “Accurate Prediction of Chemical Shifts for Aqueous Protein Structure on “Real World” Data”. In: *Chemical Science* 11.12 (2020), pp. 3180–3191. ISSN: 2041-6520. DOI: 10.1039/C9SC06561J. URL: <http://dx.doi.org/10.1039/C9SC06561J>.
- [28] Timothy C. Lillestolen and Richard J. Wheatley. “Atomic Charge Densities Generated using an Iterative Stockholder Procedure”. In: *The Journal of Chemical Physics* 131.14 (2009), p. 144101. DOI: 10.1063/1.3243863. eprint: <https://doi.org/10.1063/1.3243863>. URL: <https://doi.org/10.1063/1.3243863>.
- [29] W. C. Lu, C. Z. Wang, M. W. Schmidt, L. Bytautas, K. M. Ho, and K. Ruedenberg. “Molecule Intrinsic Minimal Basis Sets. I. Exact Resolution of ab initio Optimized Molecular Orbitals in Terms of Deformed Atomic Minimal-Basis Orbitals”. In: *Journal of Chemical Physics* 120.6 (2004), pp. 2629–2637. DOI: 10.1063/1.1638731. eprint: <https://doi.org/10.1063/1.1638731>. URL: <https://doi.org/10.1063/1.1638731>.

- [30] Thomas A. Manz and David S. Sholl. “Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials”. In: *Journal of Chemical Theory and Computation* 6.8 (2010), pp. 2455–2468. DOI: 10.1021/ct100125x. eprint: <https://doi.org/10.1021/ct100125x>. URL: <https://doi.org/10.1021/ct100125x>.
- [31] Narbe Mardirossian and Martin Head-Gordon. “ ω B97X-V: A 10-parameter, Range-separated Hybrid, Generalized Gradient Approximation Density Functional with Nonlocal Correlation, Designed by a Survival-of-the-fittest Strategy”. In: *Physical Chemistry Chemical Physics* 16.21 (May 2014), pp. 9904–9924. ISSN: 1463-9084. DOI: 10.1039/C3CP54374A.
- [32] Neil Q. McDonald, Risto Lapatto, Judith Murray Rust, Jennifer Gunning, Alexander Wlodawer, and Tom L. Blundell. “New Protein Fold Revealed by a 2.3-Å Resolution Crystal Structure of Nerve Growth Factor”. In: *Nature* 354.63526352 (Dec. 1991), pp. 411–414. ISSN: 1476-4687. DOI: 10.1038/354411a0.
- [33] P J Mitchell and D Fincham. “Shell Model Simulations by Adiabatic Dynamics”. In: *Journal of Physics: Condensed Matter* 5.8 (Feb. 1993), pp. 1031–1038. DOI: 10.1088/0953-8984/5/8/006. URL: <https://doi.org/10.1088/0953-8984/5/8/006>.
- [34] Susanna Monti, Alessandro Corozzi, Peter Fristrup, Kaushik L. Joshi, Yun Kyung Shin, Peter Oelschlaeger, Adri C. T. van Duin, and Vincenzo Barone. “Exploring the Conformational and Reactive Dynamics of Biomolecules in Solution using an Extended Version of the Glycine Reactive Force Field”. In: *Physical Chemistry Chemical Physics* 15.36 (Aug. 2013), pp. 15062–15077. ISSN: 1463-9084. DOI: 10.1039/C3CP51931G.
- [35] Wilfried J. Mortier, Swapan K. Ghosh, and S. Shankar. “Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules”. In: *Journal of the American Chemical Society* 108.15 (1986), pp. 4315–4320. ISSN: 15205126. DOI: 10.1021/ja00275a013.
- [36] Wilfried J. Mortier, Karin Van Genechten, and Johann Gasteiger. “Electronegativity Equalization: Application and Parametrization”. In: *Journal of the American Chemical Society* 107.4 (1985), pp. 829–835. DOI: 10.1021/ja00290a017. eprint: <https://doi.org/10.1021/ja00290a017>. URL: <https://doi.org/10.1021/ja00290a017>.
- [37] R.S. Mulliken. “Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I”. In: *Journal of Chemical Physics* 23.10 (1955), pp. 1833–1840. DOI: 10.1063/1.1740588.

- [38] Jane S. Murray and Peter Politzer. “The Electrostatic Potential: an Overview”. In: *WIREs Computational Molecular Science* 1.2 (2011), pp. 153–163. ISSN: 1759-0884. DOI: 10.1002/wcms.19.
- [39] Roman F. Nalewajski and Robert G. Parr. “Information Theory, Atoms in Molecules, and Molecular Similarity”. In: *Proceedings of the National Academy of Sciences* 97.16 (2000), pp. 8879–8882. ISSN: 0027-8424. DOI: 10.1073/pnas.97.16.8879. eprint: <https://www.pnas.org/content/97/16/8879.full.pdf>. URL: <https://www.pnas.org/content/97/16/8879>.
- [40] Saber Naserifar, Daniel J. Brooks, William A. Goddard, and Vaclav Cvicek. “Polarizable Charge Equilibration Model for Predicting Accurate Electrostatic Interactions in Molecules and Solids”. In: *The Journal of Chemical Physics* 146.12 (2017), p. 124117. DOI: 10.1063/1.4978891. eprint: <https://doi.org/10.1063/1.4978891>. URL: <https://doi.org/10.1063/1.4978891>.
- [41] Yongzhong Ouyang, Fei Ye, and Yizeng Liang. “A Modified Electronegativity Equalization Method for Fast and Accurate Calculation of Atomic Charges in Large Biological Molecules”. In: *Physical Chemistry Chemical Physics* 11.29 (2009), pp. 6082–6089. DOI: 10.1039/B821696G.
- [42] Steve Plimpton. “Fast Parallel Algorithms for Short-Range Molecular Dynamics”. In: *Journal of Computational Physics* 117.1 (Mar. 1995), pp. 1–19. ISSN: 0021-9991. DOI: 10.1006/jcph.1995.1039.
- [43] Jay W. Ponder et al. “Current Status of the AMOEBA Polarizable Force Field”. In: *The Journal of Physical Chemistry B* 114.8 (Mar. 2010), pp. 2549–2564. ISSN: 1520-6106, 1520-5207. DOI: 10.1021/jp910674d.
- [44] Viki Kumar Prasad, Alberto Otero-de-la-Roza, and Gino A. DiLabio. “PEP-CONF, a Diverse Data Set of Peptide Conformational Energies”. In: *Scientific Data* 6.1 (2019), p. 180310. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.310.
- [45] Prakash Chandra Rathi, R. Frederick Ludlow, and Marcel L. Verdonk. “Practical High-Quality Electrostatic Potential Surfaces for Drug Discovery Using a Graph-Convolutional Deep Neural Network”. In: *Journal of Medicinal Chemistry* 63.16 (Aug. 2020), pp. 8778–8790. ISSN: 0022-2623, 1520-4804. DOI: 10.1021/acs.jmedchem.9b01129.
- [46] Alan E. Reed, Robert B. Weinstock, and Frank Weinhold. “Natural Population Analysis”. In: *The Journal of Chemical Physics* 83.2 (July 1985), pp. 735–746. ISSN: 0021-9606. DOI: 10.1063/1.449486.

- [47] Romelia Salomon-Ferrer, David A. Case, and Ross C. Walker. “An Overview of the Amber Biomolecular Simulation Package”. In: *WIREs Computational Molecular Science* 3.2 (2013), pp. 198–210. ISSN: 1759-0884. DOI: <https://doi.org/10.1002/wcms.1121>.
- [48] Eolo Scrocco and Jacopo Tomasi. “The Electrostatic Molecular Potential as a Tool for the Interpretation of Molecular Properties”. In: *New Concepts II. Topics in Current Chemistry Fortschritte der Chemischen Forschung*. Springer, 1973, pp. 95–170. ISBN: 978-3-540-37729-0. DOI: 10.1007/3-540-06399-4_6.
- [49] Yihan Shao et al. “Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package”. In: *Molecular Physics* 113.2 (Jan. 2015), pp. 184–215. ISSN: 0026-8976. DOI: 10.1080/00268976.2014.952696.
- [50] U. Chandra Singh and Peter A. Kollman. “An Approach to Computing Electrostatic Charges for Molecules”. In: *Journal of Computational Chemistry* 5.2 (1984), pp. 129–145. ISSN: 1096-987X. DOI: <https://doi.org/10.1002/jcc.540050204>.
- [51] M. M. Teeter. “Water Structure of a Hydrophobic Protein at Atomic Resolution: Pentagon Rings of Water Molecules in Crystals of Crambin”. In: *Proceedings of the National Academy of Sciences* 81.19 (Oct. 1984), pp. 6014–6018. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.81.19.6014.
- [52] Steven Vandenbrande, Michel Waroquier, Veronique Van Speybroeck, and Toon Verstraelen. “The Monomer Electron Density Force Field (MEDFF): A Physically Inspired Model for Noncovalent Interactions”. In: *Journal of Chemical Theory and Computation* 13.1 (2017), pp. 161–179. DOI: 10.1021/acs.jctc.6b00969. eprint: <https://doi.org/10.1021/acs.jctc.6b00969>. URL: <https://doi.org/10.1021/acs.jctc.6b00969>.
- [53] T. Verstraelen, P. W. Ayers, V. Van Speybroeck, and M. Waroquier. “Hirshfeld-E Partitioning: AIM Charges with an Improved Trade-off between Robustness and Accurate Electrostatics”. In: *Journal of Chemical Theory and Computation* 9.5 (2013), pp. 2221–2225. DOI: 10.1021/ct4000923. eprint: <https://doi.org/10.1021/ct4000923>. URL: <https://doi.org/10.1021/ct4000923>.
- [54] T. Verstraelen, P.W. Ayers, V. Van Speybroeck, and M. Waroquier. “The Conformational Sensitivity of Iterative Stockholder Partitioning Schemes”. In: *Chemical Physics Letters* 545 (2012), pp. 138–143. ISSN: 0009-2614. DOI: <https://doi.org/10.1016/j.cplett.2012.07.028>. URL: <https://www.sciencedirect.com/science/article/pii/S0009261412008184>.

- [55] Toon Verstraelen, Steven Vandenbrande, Farnaz Heidar-Zadeh, Louis Vanduyfhuys, Veronique Van Speybroeck, Michel Waroquier, and Paul W. Ayers. “Minimal Basis Iterative Stockholder: Atoms in Molecules for Force-Field Development”. In: *Journal of Chemical Theory and Computation* 12.8 (2016), pp. 3894–3912. DOI: 10.1021/acs.jctc.6b00456. eprint: <https://doi.org/10.1021/acs.jctc.6b00456>. URL: <https://doi.org/10.1021/acs.jctc.6b00456>.
- [56] Toon Verstraelen et al. HORTON 2.1.1, 2017. URL: <http://theochem.github.com/horton/>,.
- [57] Toon Verstraelen et al. “IOData: A python library for reading, writing, and converting computational chemistry file formats and generating input files”. In: *Journal of Computational Chemistry* 42.6 (2021), pp. 458–464. DOI: <https://doi.org/10.1002/jcc.26468>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.26468>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26468>.
- [58] Alexander A. Voityuk, Anton J. Stasyuk, and Sergei F. Vyboishchikov. “A Simple Model for Calculating Atomic Charges in Molecules”. In: *Phys. Chem. Chem. Phys.* 20 (36 2018), pp. 23328–23337. DOI: 10.1039/C8CP03764G. URL: <http://dx.doi.org/10.1039/C8CP03764G>.
- [59] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. “Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations”. In: *J. Mol. Graph. Model.* 25.2 (2006), pp. 247–260.
- [60] P. K. Weiner, R. Langridge, J. M. Blaney, R. Schaefer, and P. A. Kollman. “Electrostatic Potential Molecular Surfaces.” In: *Proceedings of the National Academy of Sciences* 79.12 (June 1982), pp. 3754–3758. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.79.12.3754.
- [61] R. J. Woods and R. Chappelle. “Restrained Electrostatic Potential Atomic Partial Charges for Condensed-Phase Simulations of Carbohydrates”. In: *Theochem* 527.1–3 (Aug. 2000), pp. 149–156. ISSN: 0166-1280. DOI: 10.1016/S0166-1280(00)00487-5.
- [62] J. Michael Word, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. “Asparagine and Glutamine: using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation; Edited by J. Thornton”. In: *Journal of Molecular Biology* 285.4 (Jan. 1999), pp. 1735–1747. ISSN: 0022-2836. DOI: 10.1006/jmbi.1998.2401.

Appendix

2.A Results on individual molecules

Table 2.A.1: Mean absolute error (MAE) and root mean square error (RMSE) on electrostatic potential (ESP) and dipole error ($|\boldsymbol{\mu}_{DFT} - \boldsymbol{\mu}_{method}|$) of different atom typed C-GeM models, EEM, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for a) 54 small protein analogs b) 57 tripeptides.

	Small Protein Analogs			Tripeptides		
	MAE	RMSE	Dipole error	MAE	RMSE	Dipole error
CGem	0.065	0.091	0.504	0.119	0.162	1.743
CGem_CH	0.058	0.082	0.383	0.084	0.117	1.177
CGem_CHN	0.053	0.077	0.362	0.070	0.101	0.892
AM1-BCC	0.053	0.073	0.397	0.096	0.125	1.914
EEM	0.094	0.125	0.890	0.185	0.236	3.788
Hir	0.106	0.143	0.751	0.176	0.224	2.623
HI	0.058	0.077	0.421	0.080	0.104	1.106
MBIS	0.040	0.058	0.158	0.053	0.073	1.028
AVH	0.084	0.113	0.433	0.100	0.136	1.032

Table 2.A.2: Mean absolute error (MAE) and root mean square error (RMSE) on electrostatic potential (ESP) and dipole error ($|\boldsymbol{\mu}_{DFT} - \boldsymbol{\mu}_{method}|$) of different atom typed C-GeM models, EEM, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for a) 18 molecules in charge training set b) 17 tripeptides in charge testing set.

	Charged Train			Charged Test		
	MAE	RMSE	Dipole error	MAE	RMSE	Dipole error
CGem	0.099	0.140	1.228	0.107	0.151	1.474
CGem_CH	0.105	0.137	1.930	0.097	0.132	1.558
CGem_CHN	0.082	0.110	1.310	0.080	0.110	1.234
AM1-BCC	0.161	0.206	3.667	0.149	0.193	3.358
Hir	0.154	0.200	1.960	0.174	0.222	2.506
HI	0.060	0.082	0.878	0.071	0.094	1.099
MBIS	0.060	0.079	1.327	0.065	0.086	1.357
AVH	0.078	0.109	0.882	0.086	0.120	0.875

molecule	CGem	CGem_CH	CGem_CHN	AM1-BCC	EEM	Hir	HI	MBIS	AVH
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
TRP-PHE-PRO	0.134	0.087	0.072	0.075	0.183	0.157	0.059	0.056	0.123
GLN-ASN-ILE	0.134	0.092	0.112	0.069	0.227	0.174	0.079	0.046	0.075
GLY-SER-MET	0.135	0.089	0.073	0.071	0.159	0.191	0.107	0.049	0.097
GLN-PHE-TYR	0.113	0.065	0.077	0.080	0.214	0.177	0.077	0.054	0.117
GLN-LEU-ILE	0.091	0.077	0.056	0.077	0.142	0.161	0.071	0.061	0.088
TRP-THR-VAL	0.128	0.077	0.057	0.096	0.129	0.144	0.076	0.051	0.099
LEU-SER-LEU	0.085	0.068	0.045	0.119	0.126	0.165	0.055	0.055	0.064
PHE-THR-ASN	0.119	0.068	0.058	0.157	0.189	0.201	0.079	0.066	0.118
LEU-SER-HIS	0.140	0.091	0.052	0.133	0.219	0.220	0.067	0.059	0.114
ILE-TRP-THR	0.131	0.085	0.062	0.099	0.185	0.152	0.066	0.056	0.095
ASN-TRP-ALA	0.123	0.087	0.081	0.106	0.202	0.194	0.087	0.052	0.128
LEU-ASN-TRP	0.109	0.073	0.075	0.096	0.197	0.163	0.084	0.058	0.108
VAL-VAL-ASN	0.117	0.073	0.064	0.122	0.195	0.176	0.092	0.054	0.080
ILE-TRP-THR	0.108	0.076	0.067	0.124	0.156	0.184	0.085	0.065	0.110
PRO-GLN-ILE	0.107	0.078	0.080	0.088	0.177	0.142	0.074	0.044	0.081
PRO-SER-MET	0.138	0.086	0.092	0.078	0.218	0.136	0.107	0.049	0.089
TRP-GLY-LEU	0.128	0.095	0.076	0.111	0.172	0.158	0.060	0.053	0.103
VAL-LEU-PRO	0.109	0.066	0.076	0.067	0.147	0.129	0.069	0.042	0.076
ALA-TYR-TRP	0.140	0.069	0.065	0.093	0.154	0.183	0.070	0.056	0.126
CYS-SER-VAL	0.131	0.082	0.072	0.086	0.137	0.173	0.104	0.056	0.113
GLY-THR-CYS	0.137	0.079	0.069	0.075	0.175	0.181	0.095	0.048	0.106
LEU-GLY-ALA	0.098	0.079	0.061	0.107	0.246	0.150	0.061	0.059	0.074
GLY-ILE-PRO	0.113	0.092	0.100	0.087	0.162	0.191	0.074	0.053	0.088
THR-LEU-ILE	0.103	0.062	0.046	0.092	0.118	0.125	0.066	0.040	0.076
HIS-GLY-ALA	0.167	0.144	0.081	0.078	0.186	0.166	0.078	0.046	0.099
THR-MET-ALA	0.098	0.070	0.062	0.078	0.130	0.169	0.100	0.052	0.091
THR-CYS-SER	0.131	0.087	0.074	0.085	0.210	0.183	0.092	0.041	0.096
ILE-SER-CYS	0.111	0.095	0.070	0.071	0.212	0.177	0.089	0.047	0.108
LEU-HIS-HIS	0.222	0.137	0.072	0.115	0.349	0.207	0.070	0.051	0.112
VAL-SER-PHE	0.108	0.061	0.051	0.072	0.178	0.157	0.061	0.040	0.114
PHE-TYR-ASN	0.092	0.057	0.044	0.084	0.146	0.195	0.077	0.061	0.125
ASN-GLN-SER	0.118	0.088	0.082	0.104	0.145	0.193	0.073	0.057	0.102
SER-CYS-THR	0.127	0.096	0.083	0.108	0.120	0.219	0.102	0.055	0.132
Avg	0.123	0.084	0.070	0.096	0.185	0.176	0.080	0.053	0.100

Table 2.A.5: MAE on electrostatic potential (ESP) for CGem, CGem.CH, CGem.CHN, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for individual molecules in charged train set.

tripeptide	CGem	CGem.CH	CGem.CHN	AM1-BCC	Hir	HI	MBIS	AVH
Methylammonium_+1	0.037	0.033	0.019	0.032	0.043	0.034	0.040	0.039
Carbamimidoyl-propylazanium_+1	0.129	0.093	0.180	0.048	0.125	0.046	0.048	0.086
Ethylammonium_+1	0.031	0.049	0.037	0.036	0.038	0.036	0.047	0.046
Acetate_-1	0.071	0.072	0.068	0.046	0.167	0.032	0.041	0.031
ILE-ARG-ASP	0.106	0.107	0.071	0.141	0.170	0.058	0.059	0.082
ILE-ASP-ARG	0.097	0.089	0.067	0.150	0.169	0.068	0.065	0.103
LYS-ALA-GLU	0.100	0.128	0.091	0.867	0.139	0.054	0.055	0.073
LYS-ALA-GLU	0.105	0.146	0.087	0.240	0.175	0.075	0.073	0.080
GLU-ILE-LYS	0.118	0.078	0.073	0.101	0.168	0.056	0.057	0.076
HIS-GLY-VAL	0.180	0.237	0.149	0.155	0.158	0.054	0.071	0.098
ARG-LYS-ILE	0.121	0.113	0.085	0.093	0.151	0.058	0.044	0.096
ASP-ARG-ASN	0.100	0.069	0.060	0.227	0.220	0.071	0.084	0.101
LEU-ASP-GLU	0.120	0.109	0.086	0.139	0.204	0.057	0.067	0.061
LYS-ASP-ALA	0.103	0.091	0.060	0.117	0.157	0.059	0.069	0.081
LYS-VAL-ALA	0.082	0.081	0.058	0.116	0.152	0.056	0.057	0.081
MET-LYS-ASN	0.102	0.077	0.073	0.138	0.162	0.120	0.067	0.087
MET-ARG-GLU	0.104	0.130	0.089	0.150	0.201	0.083	0.069	0.090
GLU-THR-ARG	0.139	0.117	0.102	0.106	0.176	0.065	0.058	0.094
AVG	0.102	0.101	0.081	0.161	0.154	0.060	0.060	0.078

Table 2.A.6: MAE on electrostatic potential (ESP) for CGem, CGem.CH, CGem.CHN, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ω B97X-V / def2-qzvpp reference for individual molecules in charged test set.

tripeptide	CGem	CGem.CH	CGem.CHN	AM1-BCC	Hir	HI	MBIS	AVH
GLU-ASP-PHE	0.118	0.075	0.066	0.209	0.208	0.078	0.097	0.095
ARG-ALA-ALA	0.088	0.103	0.065	0.145	0.174	0.072	0.049	0.092
ARG-VAL-THR	0.102	0.087	0.060	0.131	0.160	0.065	0.061	0.089
SER-ILE-ARG	0.105	0.097	0.078	0.119	0.144	0.072	0.060	0.085
TRP-GLU-LYS	0.143	0.087	0.082	0.141	0.171	0.069	0.069	0.097
ASP-THR-ARG	0.131	0.099	0.087	0.102	0.197	0.068	0.055	0.097
ARG-HIS-LYS	0.144	0.100	0.078	0.197	0.167	0.063	0.063	0.099
THR-ASP-ALA	0.122	0.090	0.061	0.126	0.199	0.072	0.059	0.087
GLN-GLY-LYS	0.116	0.079	0.072	0.199	0.159	0.068	0.076	0.076
LYS-ASP-THR	0.109	0.115	0.099	0.152	0.165	0.057	0.065	0.073
ILE-GLU-ILE	0.111	0.073	0.057	0.088	0.155	0.064	0.056	0.059
PHE-LEU-GLU	0.099	0.093	0.059	0.210	0.201	0.075	0.068	0.095
ASP-VAL-ALA	0.123	0.076	0.077	0.094	0.160	0.098	0.077	0.079
GLN-GLU-ILE	0.101	0.083	0.056	0.221	0.208	0.069	0.072	0.095
ASN-GLN-ASP	0.123	0.104	0.124	0.156	0.190	0.099	0.068	0.091
PRO-VAL-LYS	0.123	0.075	0.099	0.114	0.118	0.062	0.055	0.071
LYS-GLU-GLY	0.111	0.102	0.078	0.120	0.183	0.058	0.058	0.080
AVG	0.116	0.090	0.076	0.149	0.174	0.071	0.065	0.086

2.B Conformations of LEU_TYR_GLN tripeptide

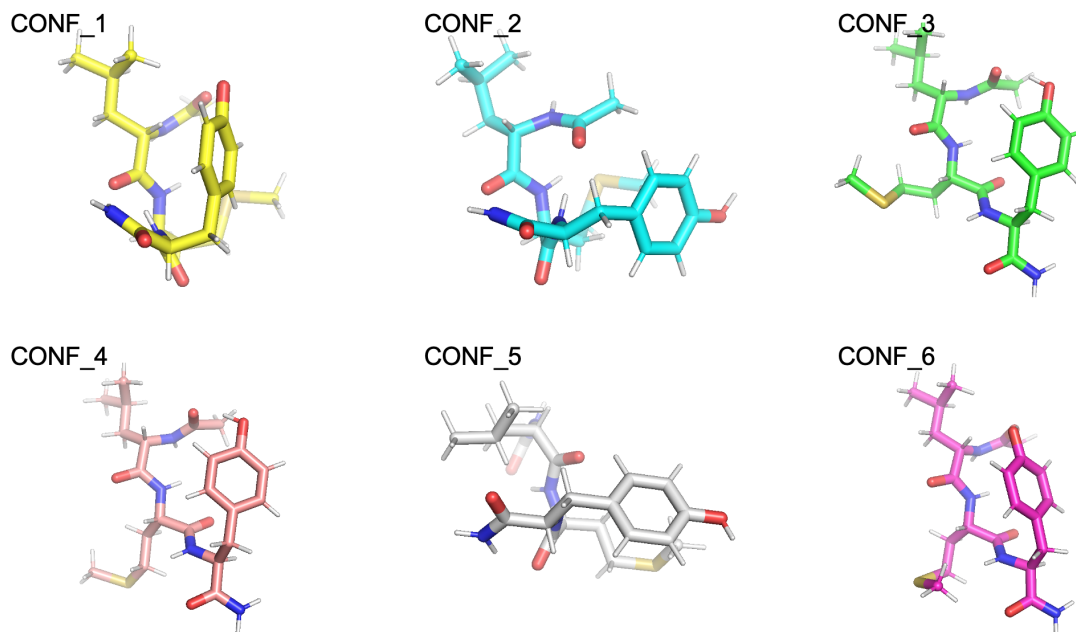


Figure 2.B.1: Six conformations of LEU_TYR_GLN tripeptide

2.C APBS parameters

APBS computation uses the following set of parameters:

mg-auto, dime 353 353 353, cglen 52.4501 44.1626 52.7887, fglen 50.8530 44.1626 51.0522, cgcent mol 1, fgcent mol 1, mol 1, lpbe, bcf1 sdh, pdie 1.0, sdie 1.0, srfm mol, chgm spl2, srad 0.0, swin 0.30, temp 298.15

Chapter 3

A benchmark dataset for Hydrogen Combustion[†]

3.1 BACKGROUND & SUMMARY

The expectation behind training deep learning models to predict molecular energies and atomic forces of molecules is the requirement of large data sets. However, very recently it has become recognized that deep learning methods that are designed with rotationally equivariant operators offer a significant reduction in data needed for training relative to invariant ML models[2, 20, 19, 12], and often outcompete even kernel methods that have traditionally been considered advantageous due to their low data requirements[11]. However, the promise in regards equivariant deep learning models must be further validated by construction of more challenging data sets than encountered up until now. For example, the recent SN2 data set provides reference energy and forces for more than 450,000 structures calculated using Density Functional Theory (DFT), but ultimately is data on highly similar individual reactions of methyl halides with one of four substituted halogens, F, Cl, Br, and I.[27]

Capturing the energy release in hydrogen combustion is a proposed energy solution for zero CO₂ emissions, and many of the elementary reactions of H₂ combustion are also present in other types of fuel generation.[5] Under realistic reaction conditions of very high temperature and high pressure make it extremely difficult to study H₂ combustion reactions experimentally. Because hydrogen combustion is difficult to study experimentally under these extremes[13], theoretical models must play an

[†]Reproduced with permission from: Guan, X.; Das, A.; Stein, C. J.; Heidar-Zadeh, F.; Bertels, L.; Liu, M.; Haghightalari, M.; Li, J.; Zhang, O.; Hao, H.; Leven, I.; Head-Gordon, M.; Head-Gordon, T. A Benchmark Dataset for Hydrogen Combustion. *Sci Data* 2022, 9 (1), 215.

active role in filling the breach, but fundamentally relies on an accurate potential energy model of not only the elementary reactions[9] but the excursions away from the reaction coordinate.

Hydrogen combustion, despite being the simplest combustion system, is nonetheless still quite chemically complicated because it can encounter one or more 19 reaction channels during the combustion event depending on the physical conditions of high temperatures and pressures.[13] This compounds the need for high quality data that is expensive to generate given the need for extensive sampling and the presence of metastable points such as transition states. For non-reacting chemical systems, conventional MD simulations are well-suited for generating a large number of configurations, which are then used as input into single point quantum-chemical energy and force calculations.[3, 23, 24] However, for reactive systems, conventional force-field based MD simulations are not useful as they don't allow breaking and forming of chemical bonds. Recent work has attempted to address this deficiency through graph-based methods that generate reference data for reactive systems,[18, 25] but they are also prone to produce large numbers of specious chemical states and unrealistic intermediates such as highly unstable radicals. Therefore fully *ab initio* sampling methods are a necessity for creation of the many molecular fragments involved in combustion chemistry, including the presence of stable and unstable intermediates, high energy transition states, and a variety of product molecules that can be formed during the reaction that is dependent on the reactive channel.[13, 7, 22, 26, 30, 9].

Our goal here is to characterize the potential energy surface (PES) of hydrogen combustion through the reaction channels proposed by Li et al.[14] using a systematic approach in *ab initio* data generation that samples off the intrinsic reaction coordinate (IRC). This study provides a data set of $\sim 290,000$ potential energies and $\sim 1,270,000$ nuclear force vectors for structures that are sampled near and far from the IRC for 19 hydrogen combustion sub-reactions, some of which are barrierless transitions, others that are dominated by large activation barriers, and even reactions involving changes in spin state.[14] This data set offers a new ML benchmark set that allows systematic investigation of data reduction when using emerging equivariant deep learning model, as well as being of interest in its own right as a source of data for machine learning of energy and forces that drive an MD engine for combustion under extreme thermodynamic conditions.

3.2 METHODS

We have used fully *ab initio* methods for sampling 19 reactive channels for hydrogen combustion as summarized in Table 3.2.1. For each reaction we used the ω B97X-V DFT functional[16] with the cc-pVTZ basis set. All calculations were performed as unrestricted open shell, using an ultrafine integration grid of 99 radial points and 590 angular points, with an SCF convergence of 10^{-8} using the GDM method[29]. All potential energies for each configuration of the 19 reactions are reported as ΔE

$$\Delta E = E_{total} - \sum_i E_{atom}, \quad (3.1)$$

using the atomic energies $E_H = -0.5004966690$ a.u. and $E_O = -75.0637742413$ a.u., and with ΔE converted to units of kcal/mole. All calculations were performed using the Q-Chem program.[21, 6]

We have organized the PES data into four categories that classify the reaction mechanism involved in the elementary steps for each reactive channel: association/dissociation reactions (channels 5-9 and 15), substitution reactions (channel 16), oxygen transfer (channels 1, 11, and 12), and hydrogen transfer (channels 2-4, 10, 13, 14, 17-19). We have kept the same numbering scheme as Li and co-workers[14] in these categories so that readers can refer back to any particular IRC of that work if desired.

The PES for each reaction channel are visualized by means of two collective variables of coordination numbers (CN) represented by

$$CN = \sum_i \frac{2.0}{1 + \exp(\sigma * (r_i - r_{0,i}))}, \quad (3.2)$$

where r_0 is the equilibrium distance and $\sigma = 3.0$ controls the sharpness of the function. Reaction channels 5-7 involve only two atoms, and thus only a 1-D distance scan is performed.

Finally, we developed a strategy for extensive sampling of the PES for the 19 reaction channels for hydrogen combustion as follows:

1. *Transition States and IRCs.* Approximate TS structures were found using the freezing string method[4, 15], and refined by the partitioned-rational function optimization eigenvector following method (P-RFO).[1] An IRC scan is then generated, and vibrational frequency analysis was performed to confirm that reactants and products have no imaginary frequencies and the TS has only one imaginary frequency. As the IRC configurations connect the minimum energy pathway, and therefore span a meaningful fraction of the configurational

Table 3.2.1: *Data Summary for the Potential Energy Surface of Hydrogen Combustion.* Tabulated are the number of structures generated for each hydrogen combustion reaction channel using different methods: IRC, normal mode displacements, and MD simulations at various temperatures. All 19 reaction channels are classified into four mechanistic groups: association/dissociation, substitution, O-transfer and H-transfer. For each configuration, energies and nuclear force vectors were computed and their numbers are tabulated.

No. Reaction	Atoms	IRC	MD simulations	Normal mode	Total energies	Total forces
Association/Dissociation						
5. $H_2 \rightarrow 2H$	2	53			53	318
6. $O_2 \rightarrow 2O$	2	71			71	426
7. $OH \rightarrow O + H$	2	71			71	426
8. $H + OH \rightarrow H_2O$	3	137	10000	5754	15891	143019
9. $H + O_2 \rightarrow HO_2$	3	60	10000	2520	12580	113220
15. $H_2O_2 \rightarrow 2OH$	4	105	10000	8820	18925	227100
Substitution						
16. $H_2O_2 + H \rightarrow H_2O + OH$	5	81	10000	10206	20287	304305
O-transfer						
1. $H + O_2 \rightarrow OH + O$	3	58	10000	3248	13306	119754
11. $HO_2 + H \rightarrow 2OH$	4	94	10000	7896	17990	215880
12. $HO_2 + O \rightarrow OH + O_2$	4	49	10000	4116	14165	169980
H-transfer						
2. $O + H_2 \rightarrow OH + H$	3	29	10000	1624	11653	104877
3. $H_2 + OH \rightarrow H_2O + H$	4	336	10000	30492	40828	489936
4. $H_2O + O \rightarrow 2OH$	4	51	10000	4284	14335	172020
10. $HO_2 + H \rightarrow H_2 + O_2$	4	58	10000	4872	14930	179160
13. $HO_2 + OH \rightarrow H_2O + O_2$	5	51	10000	6426	16477	247155
14. $2HO_2 \rightarrow H_2O_2 + O_2$	6	71	10000	11928	21999	395982
17. $H_2O_2 + H \rightarrow HO_2 + H_2$	5	58	10000	7308	17366	260490
18. $H_2O_2 + O \rightarrow HO_2 + OH$	5	55	10000	6930	16985	254775
19. $H_2O_2 + OH \rightarrow H_2O + HO_2$	6	74	10000	12432	22506	405108
Total					290418	1267977

space of a given reaction, they serve as useful starting geometries for systematic normal mode displacements and stochastic generation of structures using AIMD at finite temperatures to explore the PES for each reaction channel in more detail.

2. *AIMD Simulations.* We employed AIMD simulations to sample configurations around the IRC structures using the TS as the initial configuration for each of the reaction channels. The AIMD simulations were performed at four different high temperatures by initializing the Maxwell-Boltzmann distribution of velocities at temperatures of 500 K, 1000 K, 2000 K and 3000 K. Furthermore at each temperature three different simulation timescales are performed using a 1.21 fs (1.a.u.) time step: 10 independent (i.e. reinitialized velocities) long simulations of 121 fs, 20 independent short trajectories of 60.5 fs, and finally 25 very short

simulations of 24.2 fs. In summary, the AIMD calculations yielded a total of 10000 configurations along with their potential energies and nuclear forces for each reaction channel (see Table 3.2.1).

3. *Normal Mode Displacements.* Systematic normal mode displacements along the IRC is performed. Starting from each IRC structure, the frequencies were calculated and all atoms were displaced along each normal mode (NM) with a ± 0.01 , ± 0.025 , ± 0.05 , ± 0.075 , ± 0.1 , ± 0.125 , and ± 0.15 increment. These sampled structures that compress or expand the IRC structures help to diversify the AIMD geometries for each reaction, yielding $\sim 130,000$ configurations as summarized in Table 3.2.1. The IOData Python library was used for parsing the Q-Chem output files in generating these geometries. [28]

3.3 TECHNICAL VALIDATION

Figure 1 provides a representative *ab initio* sampling of one of the hydrogen transfer reactions, $O + H_2 \rightarrow OH + H$, in which two collective coordinates reasonably capture the potential energy surface of this reaction channel. Upon analyzing the AIMD generated geometries and their energies, it is noticed that both the reactant and product endpoint regions are well sampled (Figure 1(a)). However, near the transition state or in regions of high slope on the potential energy surface, data points from the AIMD simulations are more sparse. The addition of normal mode displacement points greatly improves sampling the configuration space of the PES along the IRC path (Figure 1(b)).

Figure 2 shows that the AIMD and NM calculations are complementary for sampling different areas away from the IRC, particularly evident for reaction channel 1 involving oxygen transfer (Figure 2(a)), reaction 8 that probes the association reaction mechanism (Figure 2(b)), and for reaction channel 16 pertaining to a substitution mechanism (Figure 2(c)). In all cases the use of two collective coordinates is sufficient to capture the IRC and its AIMD and NM extensions, borne out in the supplementary information Figure S1-S4 that provides the potential energy surfaces generated for the remaining reaction channels for these classes of hydrogen combustion reactions.

Figure 3.3.3 shows the nature of the alternative potential energy surfaces that are represented by the changes in spin state from doublet to quartet for the oxygen transfer reaction channel 12. Fig. 3.3.3(a) shows that the energy difference between the two spin states is very small near the reactant, less than 0.2 kcal/mol, but favors the quartet state substantially around the product. Fig. 3.3.3(b) plots the IRC using either the doublet or quartet spin state energies using the quartet spin state static

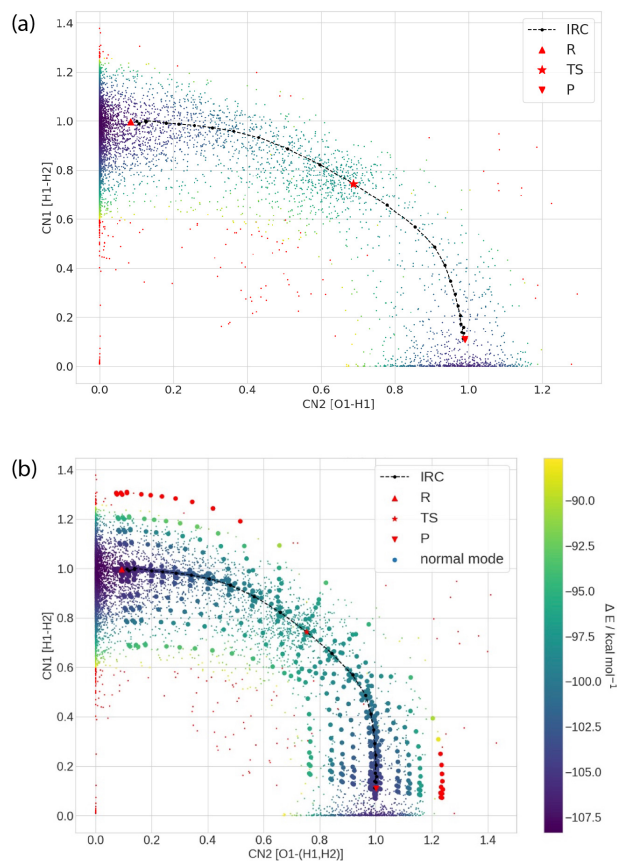


Figure 3.3.1: *Potential energy surface for the hydrogen transfer reaction 2 ($O + H_2 \rightarrow OH + H$).* (a) showing IRC and AIMD sample data only and (b) including normal mode data. CN1 represents the breaking of the H-H bond and CN2 represents the formation of the O-H bond. All energies are reported with respect to the atomization energies as given in Eq. (1) in units of kcal/mole. The red dots on the energy surface are configurations with energies larger than 10 kcal/mol of the energy of the TS structure. The points denoted with R, TS and P are corresponds to the reactant, transition state and product, respectively.

structures, and similarly Figure 3.3.3(c) represents the two spin state energies using the doublet energy configurations. Figure 3.3.3(d) shows the minimum energy of the two spin states along a single generated IRC. These differences indicate that while the geometric effects may be small, the electronic energy differences between spin states are significant. In the supplementary information we also provides the potential energy surfaces generated for reaction channel 6 which also undergoes a

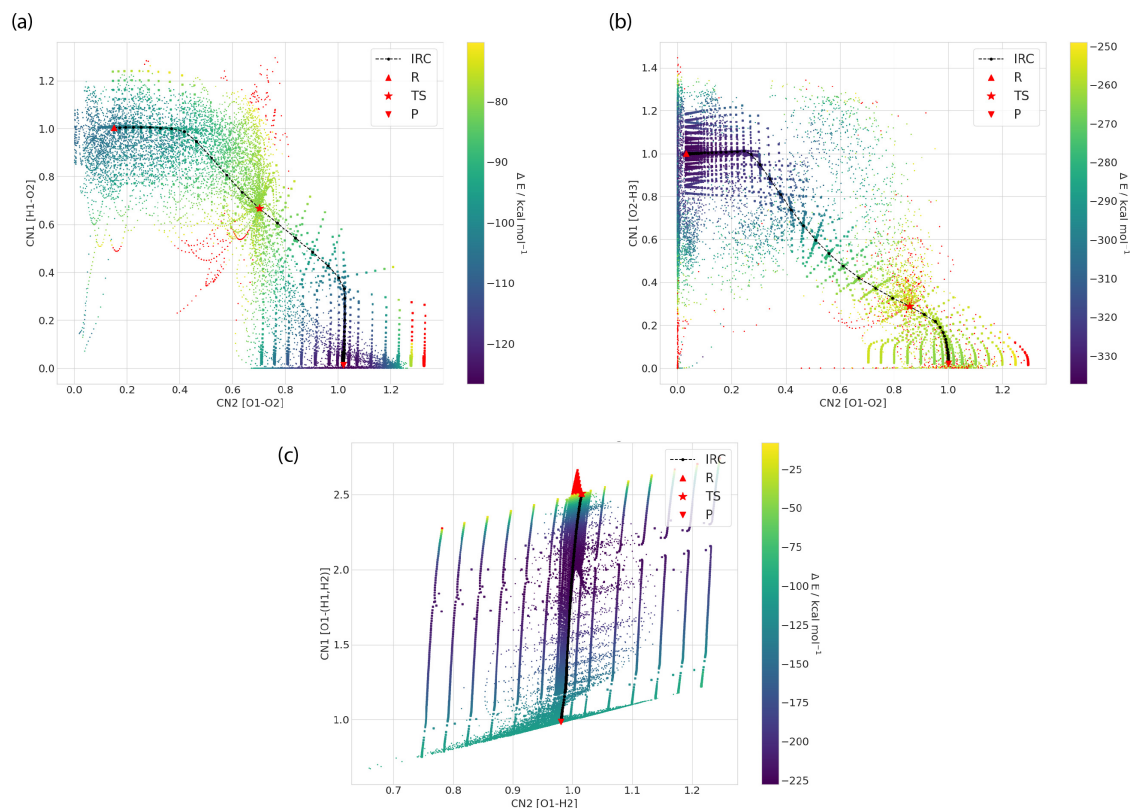


Figure 3.3.2: *Representative potential energy surfaces for oxygen transfer, association, and substitution reactions along two reaction coordinates CN1 and CN2.* (a) oxygen transfer reaction 1 ($H + O_2 \rightarrow OH + O$), (b) association reaction 8 ($H + OH \rightarrow H_2O$), and (c) substitution reaction 16 ($H_2O_2 + H \rightarrow H_2O + OH$). Each CN represents the formation or breaking of respective bond involved in the reaction process mentioned in the axes.

spin state change.

In summary, we generated high quality DFT data for hydrogen combustion reaction channels using range separated hybrid meta-GGA functional ω B97X-V with the cc-pVTZ basis set. This level of theory is considered highly accurate for thermochemistry and reactive barriers[17, 8], and the IRC profiles compared against the gold standard CCSD(T)/cc-pVTZ methods determined very small errors with the DFT level of theory.[5] This work moves beyond benchmarks such as the IRC for H_2 combustion by extensive sampling off the reaction coordinate using ab initio MD simulation and normal mode analysis for each of the 19 reaction channels. Furthermore, we

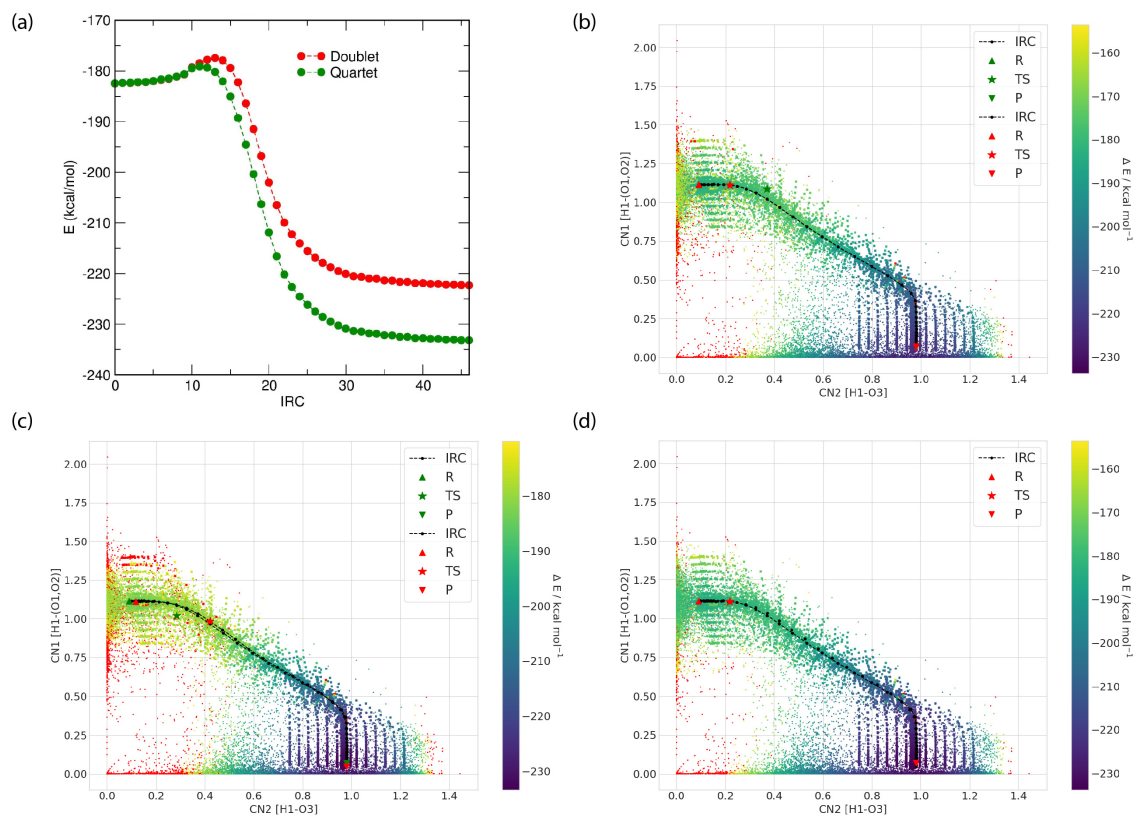


Figure 3.3.3: *The changes in the PES for reaction channel 12 involving changes in spin state.* (a) the spin cross over between the two closely spaced doublet and quartet spin state energy levels around the reactant region with widening differences progressing to product. (b) the IRC path defined by the doublet energy but geometries from the quartet (green), and from the doublet energy and geometries (red). (c) the IRC path defined by the quartet energy but geometries from the doublet (green) and from the quartet energies and geometries (red). (d) Resultant PES obtained reaction channel 12 ($HO_2 + O \rightarrow OH + O_2$) by choosing the minimum energy between the two spin states. Each CN represents the formation or breaking of respective bond involved in the reaction process mentioned in the axes.

also consider multiple spin states of the species formed in the hydrogen combustion process. This high quality data is now available to benchmark deep learning models for chemical reactivity, and as a model of the PES for generating kinetic models of H_2 combustion, especially at high pressure.

3.4 DATA RECORDS

All data can be found in the figshare repository.[10] For each reaction channel the IRC, AIMD and NM generated configurations and corresponding energies and atomic forces are provided in .npz file format; for reaction channel 5, 6 and 7 only IRC generated data are provided as discussed above. Each .npz file contains six keys including, "R" (atomic Cartesian coordinates), "Z" (atomic numbers), "N" (number of atoms), " ΔE " (reference potential energy), "F" (atomic force vectors), and "RXN" (reaction number). All the atomic position are in Å and energy and force vectors are provided in kcal/mol and kcal/mol/Å, respectively. Reaction channels such as 6 and 12 involve nuclear spin changes during the reaction, and therefore IRC calculations are performed for both spin states, with the data sorted to either (1) retain energies and forces consistent with one spin state, or (2) retaining the lowest energy spin state along the IRC for each channel. Furthermore, for reactions 6 and 12 two sets of data are provided namely 06a/06b and 12a/12b corresponding to two different spin states involved in the reaction process.

3.5 USAGE NOTES

The data set contains 19 folders corresponding to each of the reaction channels. Each reaction channel has three .npz files storing the geometries and corresponding potential energies energies and atomic force vectors obtained from IRC, AIMD and NM simulations separately. Each .npz file contains the "R" (atomic Cartesian coordinates), "Z" (atomic numbers), "N" (number of atoms), " ΔE " (reference potential energy), "F" (atomic forces), and "RXN" (reaction number) keys and the corresponding values for each configuration.

3.6 DATA AND CODE AVAILABILITY

All the data and python scripts used to generate coordination number based PES surface to analyze the data for each reaction channel is provided at <https://doi.org/10.34974/0jr1-pb24>.

3.7 ACKNOWLEDGMENTS

We thank the National Science Foundation under grant CHE-1955643. F.H-Z. acknowledges financial support from Natural Sciences and Engineering Research

Council (NSERC) of Canada. M. Liu thanks the China Scholarship Council for a visiting scholar fellowship. C.J.S. acknowledges funding by the Ministry of Innovation, Science and Research of North Rhine-Westphalia (“NRW Rückkehrerprogramm”) and an Early Postdoc Mobility fellowship from the Swiss National Science Foundation. This research used computational resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

X.G., A.D., C.J.S., F.H-Z., L.B., M.H., M.H-G. and T.H-G. conceived the scientific direction for the hydrogen combustion data set, and wrote the complete manuscript. All authors provided comments on the results and manuscript.

3.8 REFERENCES

- [1] J. Baker. “An algorithm for the location of transition states”. In: *Journal of Computational Chemistry* 7 (1986), pp. 385–395.
- [2] Simon Batzner, Tess E. Smidt, Lixin Sun, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, and Boris Kozinsky. “SE(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials”. In: *arXiv preprint arXiv:2101.03164* (2021). arXiv: 2101.03164 [physics.comp-ph].
- [3] Jörg Behler and Michele Parrinello. “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces”. In: *Phys. Rev. Lett.* 98 (14 Apr. 2007), p. 146401. DOI: <https://doi.org/10.1103/PhysRevLett.98.146401>.
- [4] Andrew Behn, Paul Zimmerman, Alexis Bell, and Martin Head-Gordon. “Efficient exploration of reaction paths via a freezing string method”. In: *The Journal of chemical physics* 135 (Dec. 2011), p. 224108. DOI: <https://doi.org/10.1063/1.3664901>.
- [5] L.W. Bertels, L.B. Newcomb, M. Alaghemandi, J.R. Green, and M. Head-Gordon. “Benchmarking the Performance of the ReaxFF Reactive Force Field on Hydrogen Combustion Systems”. In: *Journal of Physical Chemistry A* 124 (2020), pp. 5631–5645. DOI: [10.1021/acs.jpca.0c02734](https://doi.org/10.1021/acs.jpca.0c02734).

- [6] Evgeny Epifanovsky, Andrew TB Gilbert, Xintian Feng, Joonho Lee, Yuezhi Mao, Narbe Mardirossian, Pavel Pokhilko, Alec F White, Marc P Coons, Adrian L Dempwolff, et al. “Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package”. In: *The Journal of Chemical Physics* 155.8 (2021), p. 084801.
- [7] G. Gerasimov and O. Shatalov. “Kinetic mechanism of combustion of hydrogen–oxygen mixtures”. In: *Journal of Engineering Physics and Thermophysics* 86 (2013), pp. 987–995. DOI: [10.1007/s10891-013-0919-7](https://doi.org/10.1007/s10891-013-0919-7).
- [8] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme. “A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions”. In: *Phys. Chem. Chem. Phys.* 19 (2017), pp. 32184–32215. DOI: <https://doi.org/10.1039/C7CP04913G>.
- [9] Colin Grambow, Lagnajit Pattanaik, and William Green. “Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry”. In: *Scientific Data* 7 (May 2020), p. 137. DOI: <https://doi.org/10.1038/s41597-020-0460-4>.
- [10] X. Guan et al. “Hydrogen Combustion using IRC, AIMD and normal modes”. In: *Figshare* (2022). URL: <https://doi.org/10.6084/m9.figshare.19601689>.
- [11] Mojtaba Haghighatlari, Jie Li, Farnaz Heidar-Zadeh, Yuchen Liu, Xingyi Guan, and Teresa Head-Gordon. “Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods”. In: *Chem* 6.7 (2020), pp. 1527–1542. ISSN: 2451-9294. DOI: <https://doi.org/10.1016/j.chempr.2020.05.014>.
- [12] Mojtaba Haghighatlari et al. “NewtonNet: A Newtonian message passing network for deep learning of interatomic potentials and forces”. In: *arXiv preprint arXiv:2108.02913* (2021). arXiv: 2108.02913 [physics.comp-ph].
- [13] Juan Li, Zhenwei Zhao, Andrei Kazakov, and Frederick Dryer. “An updated comprehensive kinetic model of hydrogen combustion”. In: *International Journal of Chemical Kinetics* 36 (Oct. 2004), pp. 566–575. DOI: <https://doi.org/10.1002/kin.20026>.
- [14] Juan Li, Zhenwei Zhao, Andrei Kazakov, and Frederick L. Dryer. “An updated comprehensive kinetic model of hydrogen combustion”. In: *International Journal of Chemical Kinetics* 36.10 (2004), pp. 566–575. DOI: <https://doi.org/10.1002/kin.20026>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/kin.20026>.

- 1002/kin.20026. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/kin.20026>.
- [15] Shaama Mallikarjun Sharada, Paul Zimmerman, Alexis Bell, and Martin Head-Gordon. “Automated Transition State Searches without Evaluating the Hessian”. In: *Journal of Chemical Theory and Computation* 8 (Oct. 2012), pp. 5166–5174. DOI: <https://doi.org/10.1021/ct300659d>.
- [16] N. Mardirossian and M. Head-Gordon. “ ω B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy”. In: *Phys. Chem. Chem. Phys.* 16 (2014), pp. 9904–9924. DOI: <https://doi.org/10.1039/c3cp54374a>.
- [17] Narbe Mardirossian and Martin Head-Gordon. “Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals”. In: *Molecular Physics* 115.19 (2017), pp. 2315–2372. DOI: <https://doi.org/10.1080/00268976.2017.1333644>. eprint: <https://doi.org/10.1080/00268976.2017.1333644>.
- [18] Johannes Margraf and Karsten Reuter. “Systematic Enumeration of Elementary Reaction Steps in Surface Catalysis”. In: *ACS Omega* 4 (Feb. 2019), pp. 3370–3379. DOI: <https://doi.org/10.1021/acsomega.8b03200>.
- [19] Z. et al. Qiao. “Unite: Unitary n-body tensor equivariant network with applications to quantum chemistry”. In: *arXiv preprint* (2021). DOI: 10.48550/arXiv.2105.14655.
- [20] Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. “Equivariant message passing for the prediction of tensorial properties and molecular spectra”. In: *arXiv preprint arXiv:2102.03150* (2021). arXiv: 2102.03150.
- [21] Yihan Shao et al. “Advances in molecular quantum chemistry contained in the Q-Chem 4 program package”. In: *Molecular Physics* 113.2 (2015), pp. 184–215. DOI: <https://doi.org/10.1080/00268976.2014.952696>. eprint: <https://doi.org/10.1080/00268976.2014.952696>.
- [22] G. Simm and M. Reiher. “Context-driven exploration of complex chemical reaction networks”. In: *Journal of Chemical Theory and Computation* 13 (2017), pp. 6108–6119. DOI: 10.1021/acs.jctc.7b00945.
- [23] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. “ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost”. In: *Chemical Science* 8.4 (2017), pp. 3192–3203.

- [24] Peter St. John, Yanfei Guan, Yeonjoon Kim, Brian Etz, Seonah Kim, and Robert Paton. “Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules”. In: *Scientific Data* 7 (July 2020), p. 244. DOI: <https://doi.org/10.1038/s41597-020-00588-x>.
- [25] S. Stocker, G. Csányi, K. Reuter, and J.T. Margraf. “Machine learning in chemical reaction space”. In: *Nature Communications* 11 (2020), p. 10. DOI: [10.1038/s41467-020-19267-x](https://doi.org/10.1038/s41467-020-19267-x).
- [26] Z.W. Ulissi, A.J. Medford, T. Bligaard, and J.K. Nørskov. “To address surface reaction network complexity using scaling relations, machine learning, and DFT calculations”. In: *Nature Communications* 8 (2017), p. 14621. DOI: [10.1038/ncomms14621](https://doi.org/10.1038/ncomms14621).
- [27] Oliver T. Unke and Markus Meuwly. “PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges”. In: *J. Chem. Theory Comput.* 15.6 (2019), pp. 3678–3693. DOI: <https://doi.org/10.1021/acs.jctc.9b00181>. arXiv: 1902.08408.
- [28] Toon Verstraelen et al. “IOData: A python library for reading, writing, and converting computational chemistry file formats and generating input files”. In: *Journal of Computational Chemistry* 42.6 (2021), pp. 458–464. DOI: <https://doi.org/10.1002/jcc.26468>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.26468>.
- [29] Troy Van Voorhis and Martin Head-Gordon. “A geometric approach to direct minimization”. In: *Molecular Physics* 100.11 (2002), pp. 1713–1721. DOI: <https://doi.org/10.1080/00268970110103642>. eprint: <https://doi.org/10.1080/00268970110103642>.
- [30] Jinzhe Zeng, Liqun Cao, Mingyuan Xu, Tong Zhu, and John Zhang. “Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation”. In: *Nature Communications* 11 (Nov. 2020), p. 5713. DOI: <https://doi.org/10.1038/s41467-020-19497-z>.

Appendix

3.A Supplementary Figures

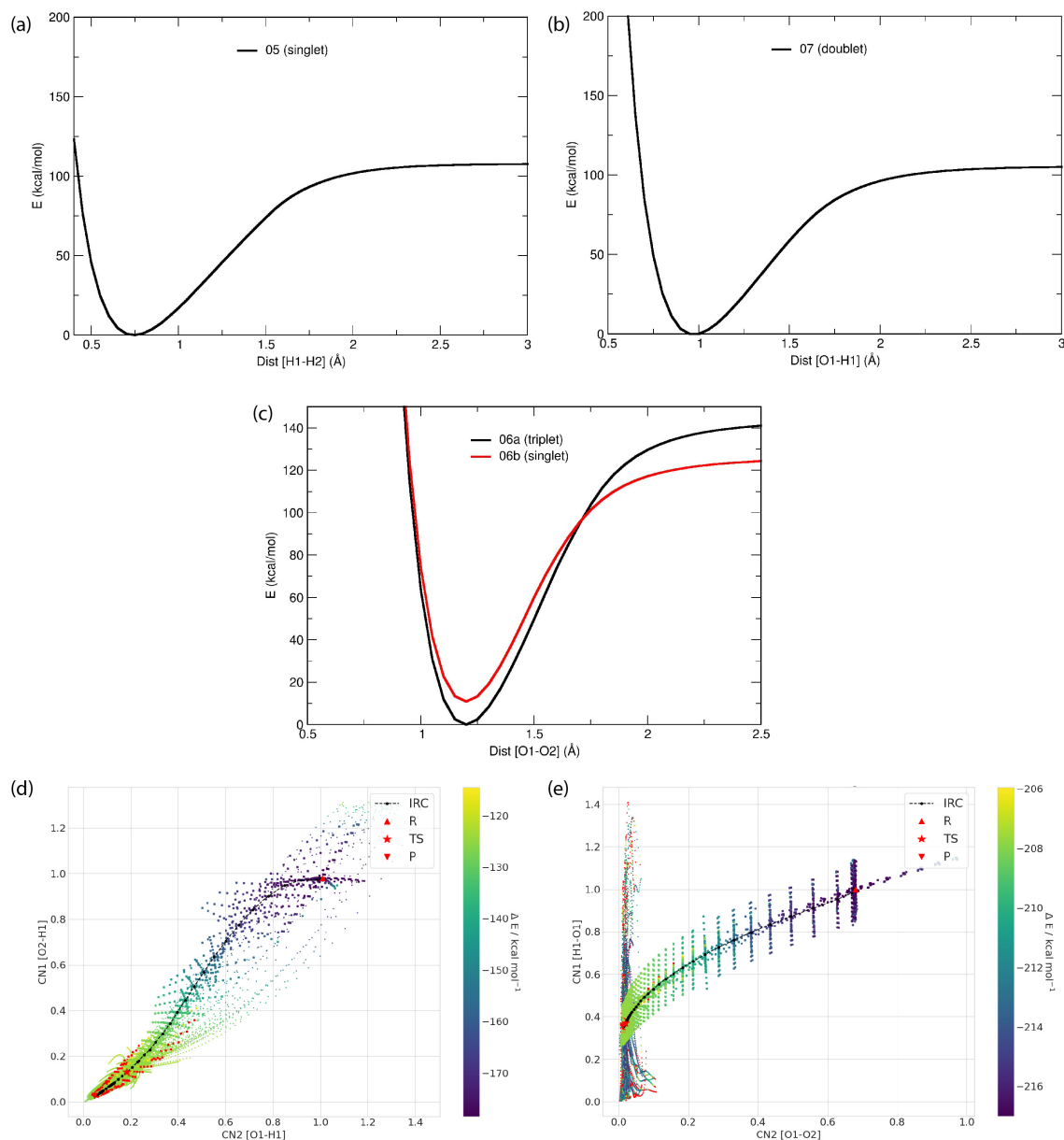


Figure 3.A.1: *Additional association/dissociation reaction channels.* (a) 1D pathway for reaction channel 5 ($H_2 \rightarrow H + H$), (b) reaction channel 7 ($OH \rightarrow O + H$), (c) for reaction channel 6 ($O_2 \rightarrow 2O$) along O-O bond for two spin states triplet and singlet. (d) reaction channel 9 ($H + O_2 \rightarrow HO_2$) and (e) reaction channel 15 ($H_2O_2 \rightarrow 2OH$) are described well along the two reaction coordinates $CN1$ and $CN2$, defined as the coordination numbers for the formation or breaking of relevant bond as per the axis label (bottom). All the red dots on the PES are the configuration having 10 kcal/mol larger energy than the TS structure.

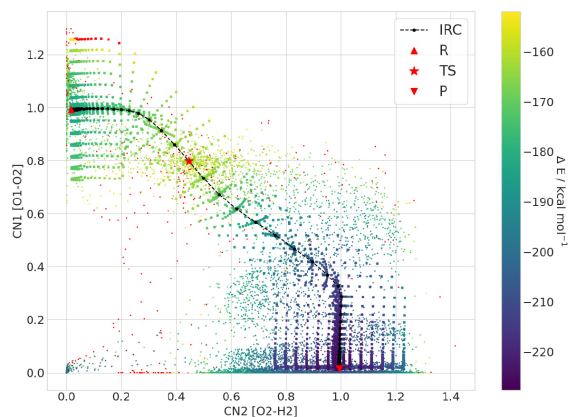


Figure 3.A.2: *Additional oxygen transfer reaction channel.* PES for reaction channel 11, $HO_2 + H \rightarrow 2OH$, along the two reaction coordinates $CN1$ and $CN2$ representing the formation or breaking of respective bond labeled in the axes. All the red dots on the PES are the configuration having 10 kcal/mol larger energy than the TS structure.

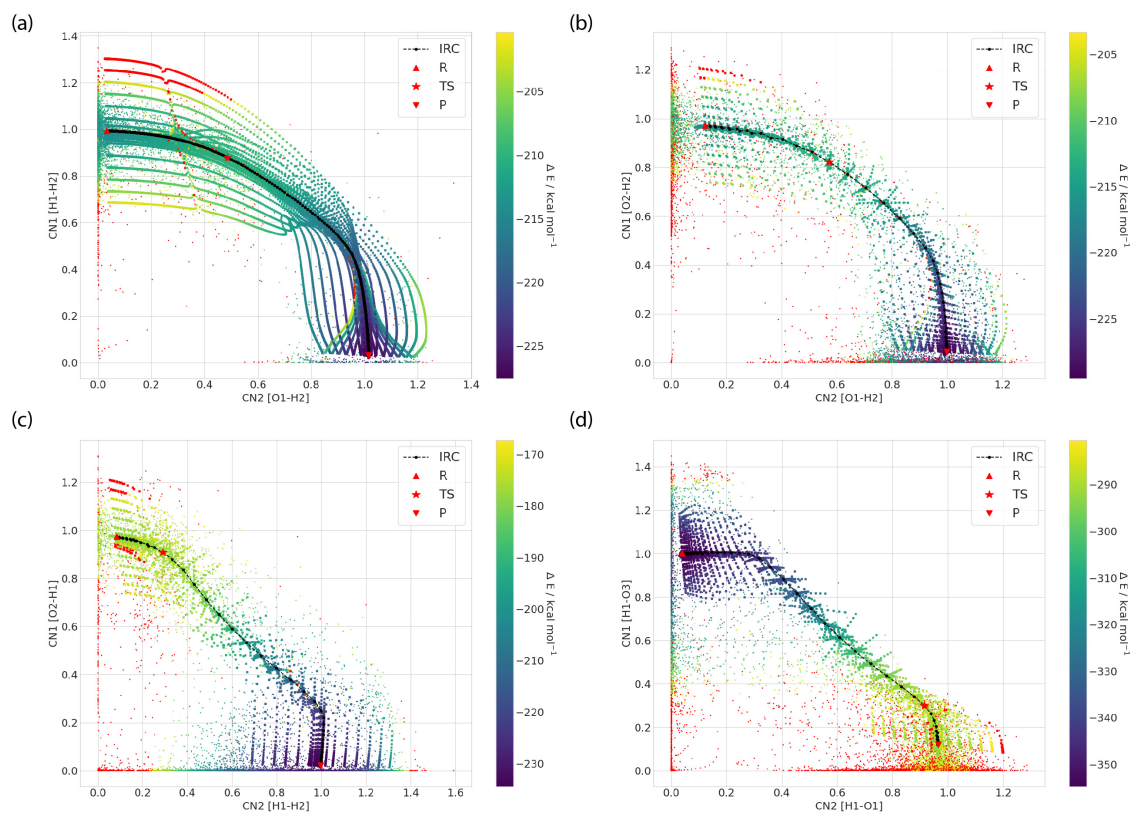


Figure 3.A.3: *Additional hydrogen transfer reaction channels.* PES for (a) reaction channel 3 ($H_2 + OH \rightarrow H_2O + H$), (b) reaction channel 4 ($H_2O + O \rightarrow 2OH$), (c) reaction channel 10 ($HO_2 + H \rightarrow H_2 + O_2$) and (d) reaction channel 13 ($HO_2 + OH \rightarrow H_2O + O_2$) along the two reaction coordinates $CN1$ and $CN2$ defined by the formation or breaking of bonds described in the axes. All the red dots on the PES are the configuration having 10 kcal/mol larger energy than the TS structure.

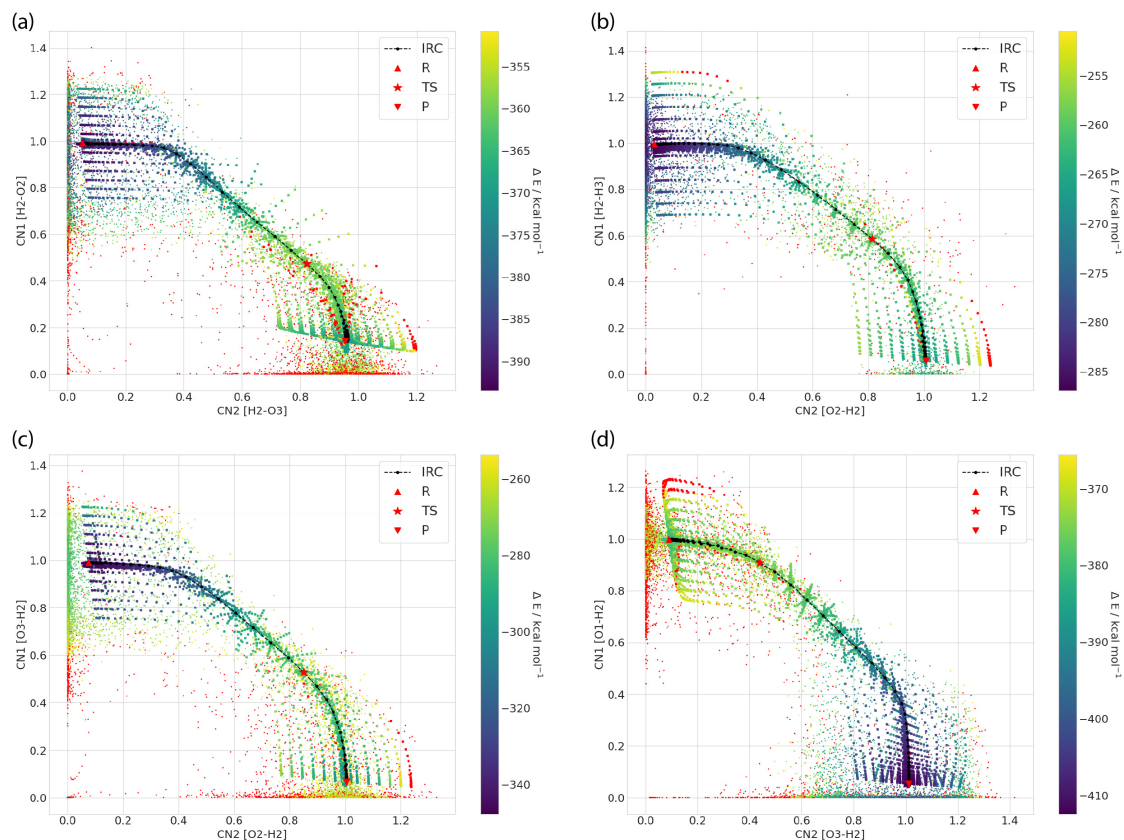


Figure 3.A.4: *Additional hydrogen transfer reaction channels (continued)*. PES for (a) reaction channel 14 ($2HO_2 \rightarrow H_2O_2 + O_2$), (b) reaction channel 17 ($H_2O_2 + H \rightarrow HO_2 + H_2$), (c) reaction channel 18 ($H_2O_2 + O \rightarrow HO_2 + OH$) and (d) reaction channel 19 ($H_2O_2 + OH \rightarrow H_2O + HO_2$) along the two reaction coordinates $CN1$ and $CN2$ defined by the formation or breaking of bonds described in the axes. All the red dots on the PES are the configuration having 10 kcal/mol larger energy than the TS structure.

Chapter 4

Using machine learning to go beyond potential energy surface benchmarking for chemical reactivity[†]

[†]Reproduced with permission from: Guan, X.; Heindel, J. P.; Ko, T.; Yang, C.; Head-Gordon, T. Using Machine Learning to Go beyond Potential Energy Surface Benchmarking for Chemical Reactivity. *Nat Comput Sci* **2023**, 3 (11), 965–974.

4.1 INTRODUCTION

Machine learning (ML) methods are emerging as an alternative to ab initio molecular dynamics (AIMD) and physical-based potential energy functions (or force fields), once trained on high quality ab initio energies and forces associated with a given conformation of nuclei. Starting with the generalized neural network representation of high dimensional potential energy surfaces (PESs) proposed by Behler and Parrinello[5], their work inspired additional state-of-the-art approaches for chemical systems[43, 44, 46, 8, 10, 37]. A key development more recently are ML models that are equivariant to translations and rotations through their architectures, showing testing superiority in accuracy benchmarks with greatly reduced quantities of reference data for training.[47, 12, 1, 33, 4, 16, 38, 49, 20] In what follows we use NewtonNet [20], a physics inspired message passing equivariant neural network, as the underlying deep learning model for energy and force prediction for the surprisingly difficult case of hydrogen combustion. We have recently developed the HCombustion dataset [19] of energies and forces generated using the ω B97X-V [30] density functional theory (DFT) functional with the cc-pVTZ basis set. In the case of hydrogen combustion there are at least 19 reaction channels, multiple stable and unstable intermediates that are dependent on a given reactive channel, and complications that can arise due to creation of radical species and alternative spin states during the combustion process[6, 29].

Regardless of the architecture, the reliability of the ML model still heavily relies on the diversity of the training data, especially for chemically reactive systems that must visit high energy states when undergoing chemical transformations. ML models by their nature interpolate between known training data, but its extrapolation capability is limited, thus predictions can be unreliable when molecular configurations are dissimilar to those in the training set, and are error-prone when applied to MD simulations, especially in the case for gas phase chemical reactivity in which energy configurations are highly diverse. Thus, in order to achieve meaningful chemically reactive simulations with ML potentials, the training sets should cover a wide range of structural space with variable energy stability.[26] However, it is challenging to formulate *a priori* a dataset that is balanced and diverse for a given reactive system. Traditionally, active learning (AL) is a powerful strategy for reducing the amount of labeled data required to develop an ML potential, and selecting informative configurations for labeling.[41, 39, 44, 36, 2, 52, 22, 51, 26] Typically AL uses a query by committee strategy[39, 44], in which variance among a set of identical architectures but stochastically initiated ML models select the most informative data points for labeling, reducing the data generation effort and improving the accuracy of the ML model. However AL informed ML is still not a panacea without having more information.

In this study, we propose an active learning workflow for chemical reactivity that utilizes a different information source - namely the sampling efficiencies that are inherent in statistical mechanics methods for rare events. In particular, we have formulated an AL workflow that expands on the original HCombustion dataset [19] by formulating collective variables (CVs) to first systematically sample a lower manifold of all the intrinsic reaction coordinates (IRCs) of the 19 reactive channels, and then to stochastically sample with metadynamics to take advantage of its known better ergodicity.[27, 3] The idea of using metadynamics to sample physically relevant but high energy transition states was also explored by Yang and co-workers[51], which aimed to select a good selection of CVs to quickly learn the PES in physically relevant regions for reactivity. While the smaller subspace metadynamics fills the free energy wells to more rapidly find transition states, we require more such that the ML model learns about configurations which are not relevant or whose energies are simply inaccessible. In this work we show that in early to mid-stages of active learning using metadynamics it is mostly irrelevant whether the CVs are "good". That is, even bad CVs can help systematically sample in directions orthogonal to the IRC and hence determine regions to avoid. After this AL exploration stage, we can more rapidly build labelled data in the physically relevant regions near the IRC.

This AL-metadynamics strategy allows us to reach a final hydrogen combustion ML model that is more diverse and balanced, and more importantly, an energy surface that is relatively smooth. We then utilize the variance of the ML committee models to report back on additional but much fewer PES rough spots, for which we substitute a direct call to an ab initio force to complete a local in time molecular dynamics update, until the ML models recover accurate forces to continue the trajectory, all without further and expensive retraining. We illustrate the completeness of the ML hydrogen combustion model with a metadynamics application using well formulated CVs inspired by diffusion maps[25] to discover the entropic contribution of free energy transition states for hydrogen combustion reaction channels.

4.2 RESULTS

NewtonNet Initial Training[†]

[†]Partly reproduced with permission from: Haghghatlari, M.; Li, J.; Guan, X.; Zhang, O.; Das, A.; J. Stein, C.; Heidar-Zadeh, F.; Liu, M.; Head-Gordon, M.; Bertels, L.; Hao, H.; Leven, I.; Head-Gordon, T. NewtonNet: A Newtonian Message Passing Network for Deep Learning of Interatomic Potentials and Forces. *Digital Discovery* **2022**, 1 (3), 333–343.

We train NewtonNet on the complete reaction network by sampling training, validation, and test sets randomly formulated from the total Hydrogen Combustion data in Chapter 3. The validation and test sizes are fixed to 1000 data per reaction, and the size of training data varies in a range of 100 to 5000 data points per reaction. The resulting model accuracy on the hold-out test set for both energy and forces is reported in Figure 4.2.1. It is seen that NewtonNet can outperform the best invariant SchNet model[37] with slightly more than one order of magnitude smaller training data (500 vs 5000 samples per reaction), and is capable of achieving the chemical accuracy goal with as little as 500 data points per reaction.

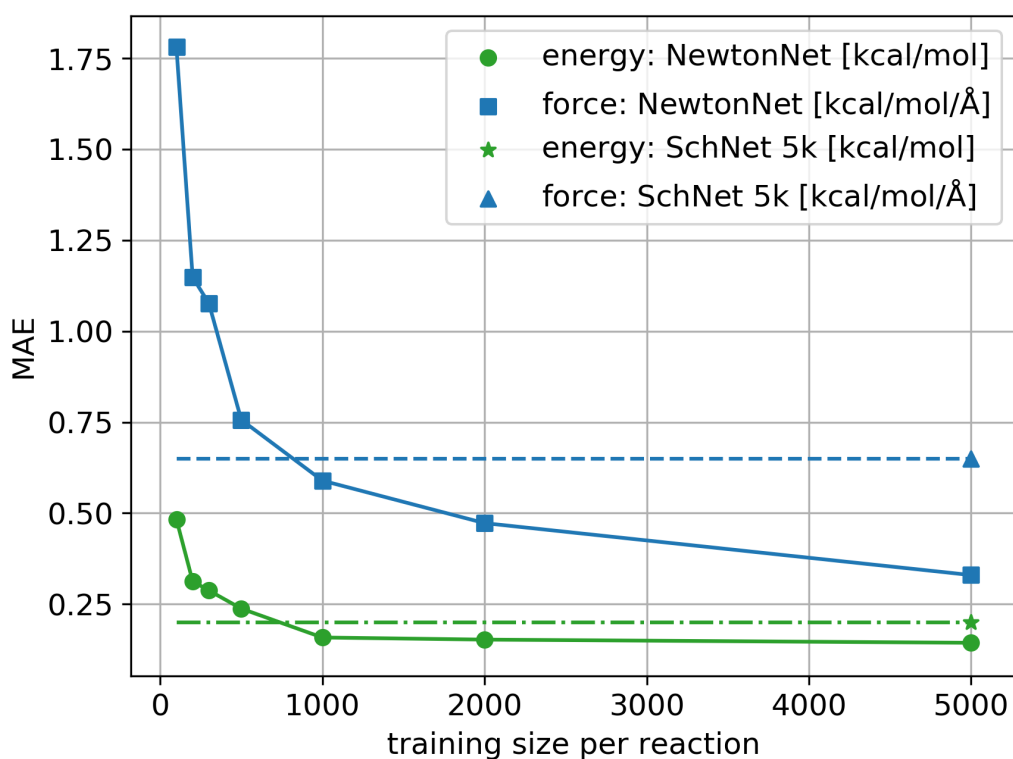


Figure 4.2.1: The learning curve of NewtonNet for the hydrogen combustion data, with MAEs of energy and forces averaged over the 16 independent reactions and with respect to the number of training samples used for each reaction. The dashed lines show the performance of SchNet when trained on all 5k data per sub-reaction.

In conventional deep learning approaches for reactive chemistry, abrupt changes in the force magnitudes can give rise to multimodal distributions of data, which can introduce covariate shift in the training of the models. Here we posit that a better representation of atomic environments using the latent force directions can increase the amount of attention that one atom gives to its immediate neighbors. As a result the performance of NewtonNet in prediction of forces for hydrogen combustion reactive systems, compared to other single small molecule dataset[42, 45, 8, 9, 11] we tested, benefit most from the directional information provided by atoms that break or form new bonds.

Active Learning Workflow

NewtonNet was trained on the HCombustion dataset[19](which we refer to as the original dataset) comprised of AIMD sampling around the IRC at 0K, near the transition state, as well as systematic normal modes with respect to the IRC to produce PES curvature data, all specific to 19 elementary reactions (Supplementary Table 1).[6, 29] Using $\sim 5k$ data per reaction, the trained NewtonNet model reaches very good accuracy for energies (mean absolute error (MAE) of 0.14 kcal/mol) and forces (MAE of 0.33 kcal/mol/Å).

However, when the ML model is applied to molecular simulation, we find it to be highly error-prone, requiring for example new data corresponding to the atomization process to understand the inherent stability of molecules. Furthermore, while the model is trained on relatively stable geometries as well as metastable transition states, it lacks the knowledge of any highly unstable state, and thus predicts configurations with the wrong energy ordering as well as unphysical geometries that were predicted by the ML model to be energetically stable. In Supplementary Figure 1, we show representative structures with geometries close to dissociation, with a DFT energy of -129.16 kcal/mol, but the ML model predicts -324.05 kcal/mol, or the appearance of the hydronium ion with an artificially low energy of -295.56 kcal/mol, while its actual DFT energy is -189.98 kcal/mol. As pointed out by previous active learning studies[32], the MD trajectory can get trapped in these unphysical states or even become unstable numerically for normal step sizes in time. Figure 4.2.2(a) shows a trajectory that started from the transition state of reaction(rxn) 16, proceeding through a series of unphysical states, and eventually ending in configurations where all the bonds are broken.

Hence an MD simulation driven with the ML model will quickly generate trajectories that go to unphysical regions because the QM data will always be small-scale and insufficient for generating a complete PES. It also emphasizes how unintuitive data acquisition is in regards creating a robust ML model of the PES, especially for

chemical reactivity, which requires negative design principles to add unphysically high energy species to the training of the ML model. To make negative design data acquisition systematic and tractable, we formulate an IRC dilation dataset within a lower manifold of collective coordinates that encapsulates the IRC for each reaction channel. Using the example of Rxn16 whose AIMD and normal mode data are shown in Figure 4.2.2b, each geometry along the IRC curve is proportionally scaled with multiple ratios of CVs to generate new high energy species as shown in Figure 4.2.2c. This addition to the dataset helps the models to learn bond dissociation and contraction, as well as spanning a more diverse chemical space than the original dataset.

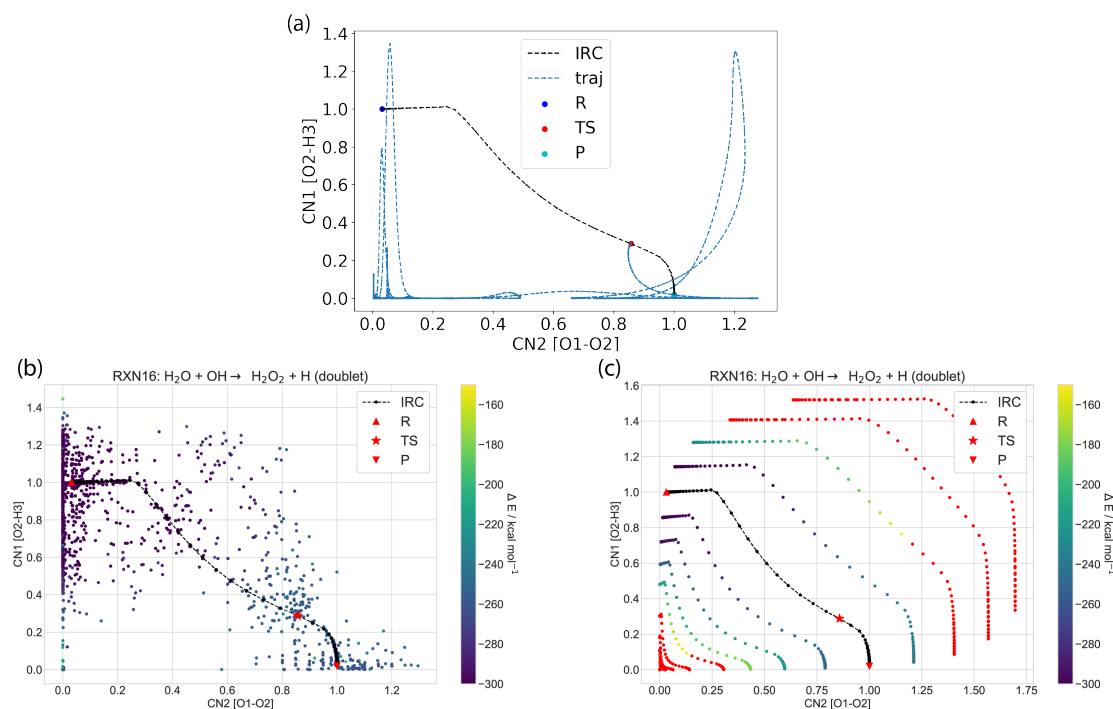


Figure 4.2.2: Observations of the missing data in the machine-learned model for hydrogen combustion and the addition of dilation data. (a) A trajectory for Rxn16 driven with the original ML model visualized on two reaction coordinates CV1: CN1 (O2-H3) and CV2: CN2(O1-O2). (b,c) Rxn16 DFT potential energy surfaces visualized on two reaction coordinates CV1: CN1 (H4-H5) and CV2: CN2(O2-H5). (b) the original HCombustion dataset with AIMD and normal mode data (c) the dialiation dataset. We use the coordination number(CN) between atom i and j $CN(ij) = \frac{2}{1+\exp(3(r_{ij}-r_{eq}))}$ to denote the CV, where r_{ij} is the actual distance between atom i and j and r_{eq} is equilibrium distance between i and j , usually selected as the equilibrium bond length unless otherwise specified. Energy of each point is color coded in (b) and (c), with red color meaning points with energy higher than the Boltzmann weighting threshold.

Although the dilation data is helpful, the PES for hydrogen combustion remains incomplete. We thus introduce an active learning workflow in which we trained four models with 1000 structures from the HCombustion dataset and 200 structures from the dialiation dataset for each reaction using the same architecture but different initial parameters. These four models serve as a starting point of an iterative

process, outlined in Figure 4.2.3, to systematically improve the ML model through metadynamics to more efficiently sample previously unseen and unstable structures through an external biasing potential that forces the system to explore regions of high (free) energy. While in the usual context poor selection of the low-dimensional descriptors affects the rate at which transitions are enhanced, in this particular context we are using metadynamics as a tool to fill in the holes in the ML PES in which the goodness of the CVs is less important than the ability to sample diverse conformations of high energy variance.

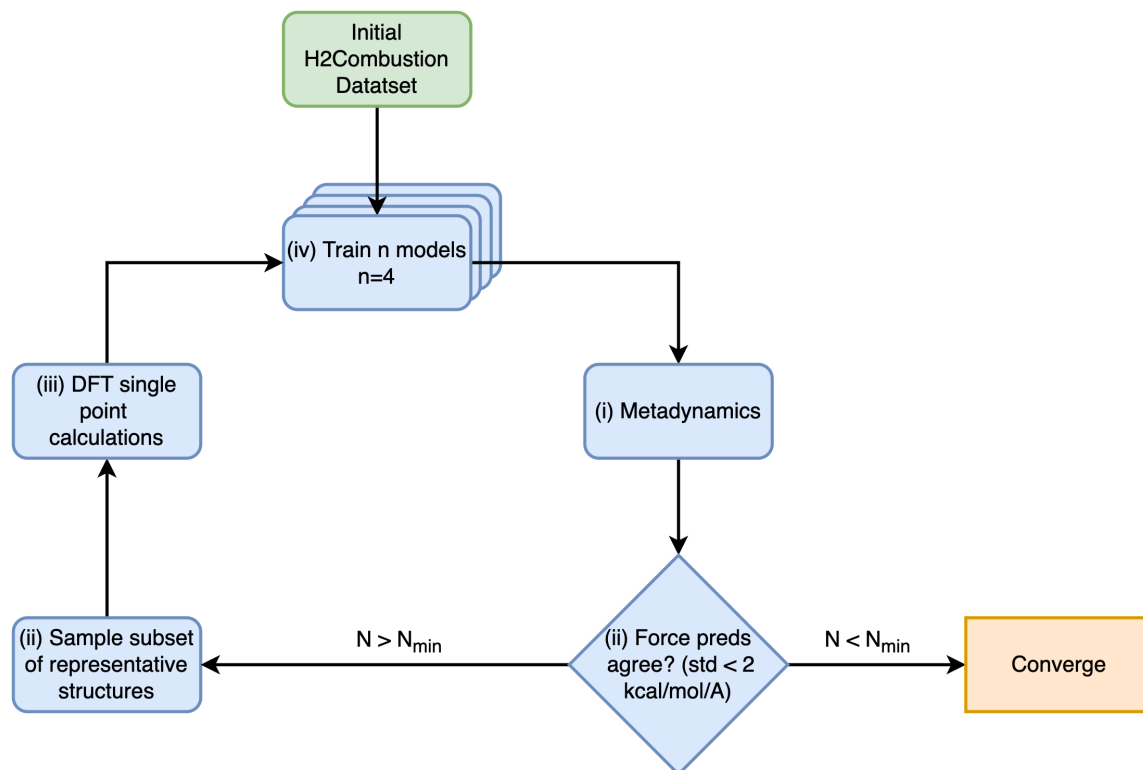


Figure 4.2.3: *Schematic illustration of active learning workflow using query by committee and metadynamics.* The four NewtonNet models serve as a starting point of an iterative process, where each round of active learning consists of the following steps: (i) Perform several short metadynamic simulations to explore the configuration space in a lower dimension. (ii) When the four models disagree outside standard deviation, collect a representative subset of structures to be included in the training set through downsampling. (iii) Perform DFT calculation of energies and atomic forces. (iv) Retrain the ensemble of ML models with the updated training set. N in the figure refer to the number of frames above the standard deviation (std) threshold. To allow for a fast turn around time, we use relatively small epoch size and large learning rate in the training through the active learning rounds. The details of each of the steps are described in the method section 4.2-4.6. The CVs used during the active learning phase is given in Supplementary Table 2.

To illustrate the active learning approach, we consider Rxn18 as an example in which the potential energy surface is projected onto CV1: CN(O2-O5) and CV2: CN(O5-H4), with results shown in Figure 4.2.4. The ML model performance was tracked by analyzing both the original data points derived from AIMD and normal

modes calculations, and the newly added data points accumulated during the AL procedure. Because we used longer metadynamics simulations for sampling as the active learning rounds proceeded and as errors decreased, we show this by dividing the active learning data into four batches: the original data derived from normal mode and AIMD near the transition state of the IRC (Figure 4.2.4(a,b)), data sampled with 2ps metadynamics between active learning round 1-20 (Figure 4.2.4(c,d)), data sampled with 5ps metadynamics between active learning round 21-33 (Figure 4.2.4(e,f)), data sampled with 10ps metadynamics between active learning round 34-48 (Figure 4.2.4(g,h)). Furthermore, these points were Boltzmann weighted in training to make sure the models have some information of higher energy states while focusing most on the regions with physically relevant energies. The decrease in model prediction error on both energy and forces for the reactive pathway data confirms the validity of this approach as shown in Figure4.2.4(b).

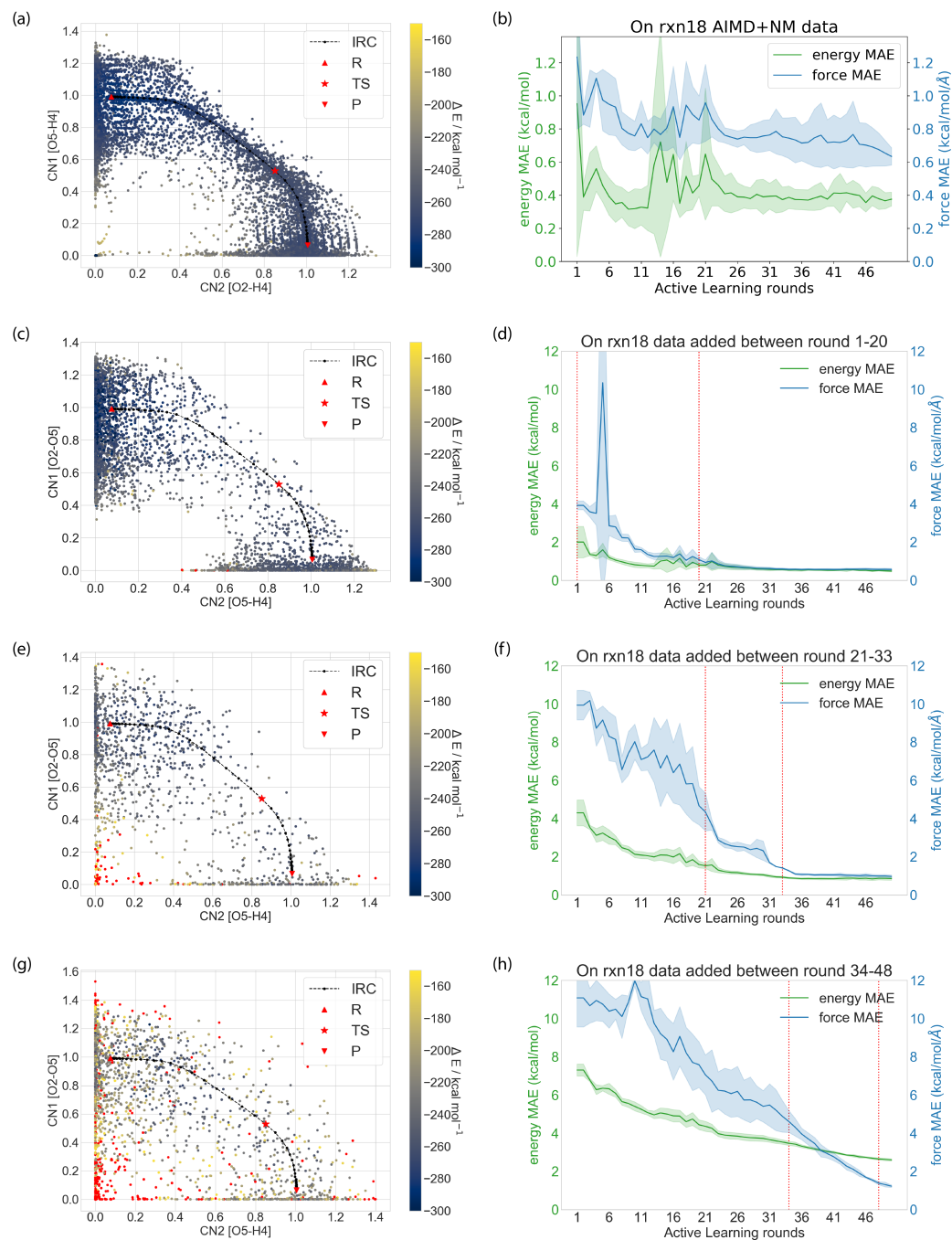


Figure 4.2.4: Potential energy surface in collective coordinates (left) and change in energy and force mean absolute error (MAE) as active learning round proceeds (right) for Reaction 18. The potential energy surface projected onto CV1: CN(O2-H4) and CV2: CN(O5-H4).

Figure 4.2.4: (continued) Energy of each point is color coded; red points have energy higher than the Boltzmann weighting threshold. Shaded regions show one standard deviation above and below the mean. (a,b) Original data test set of rxn18. The IRC dilation data are not shown in (a) because of their high energy nature. (c,d) Data added between AL rounds 1-20, sampled with 2ps metadynamics. (e,f) Data added between AL rounds 21-33, sampled with 5ps metadynamics. (g,h) Data added between AL rounds 34-48, sampled with 10ps metadynamics. The total number of data points generated during AL for training is given in Supplementary Table 3.

As shown on the left side of Figure 4.2.4, active learning spans more of the reaction coordinate space; the lack of points in the upper right quadrant in early stages of active learning indicates that the dilation dataset was sufficient (Figure 4.2.4c) as the model learned to avoid the regions with high energy due to nuclear repulsion. The model also became more accurate near the reaction pathway basins reflected by reduced sampling in this region in later rounds of the active learning workflow (Figure 4.2.4(e,g)). The right side of Figure 4.2.4 shows that there is a very substantial improvement in model error when predicting on the newly added data. Using the original ML model these new sampled points from active learning created large energy and force errors as high as ~ 8 kcal/mol and ~ 12 kcal/mol/Å respectively, which of course would create an untenable MD trajectory if those geometries were visited during an actual application. But after ~ 50 rounds of active learning, the model prediction on new geometries are hugely improved, with 0.97 kcal/mol/Å error in forces on the data points collected between round 21-33, and 1.24 kcal/mol/Å force error on the data points collected between round 34-48.

The somewhat larger errors are still excellent given the much larger energy range required to generate physically meaningful trajectories, which derives from the more substantial improvements of the new data points added with the active learning workflow. It also reflects another important strategy in AL, which is that high energy configurations must be learned but they do not have to be as accurate, and can be manifested in the loss function through Boltzmann weighting (see Methods section 4.3). This is supported by the fact that the original data points are still predicted with relatively small errors in energy and forces with the active learning model (~ 0.4 kcal/mol energy MAE and ~ 0.8 kcal/mol/Å force MAE).

Committer Analysis and the Free Energy Surface

To investigate the outcome of the active learning to create an ML potential for hydrogen combustion, we performed committer analysis on the elementary reaction channels for hydrogen combustion, starting the trajectories from the IRC transition state of the reaction with temperature set to 300 K. Table 4.2.1 presents the committer results for both the original model without active learning and the final model after all active learning rounds.

Table 4.2.1: *Committer statistics at 500K* with (a) the original model (b) the final model after active learning with IRC dilation and active dynamics. The committer ratio to reactant and product for each reaction are reported. Diatomic reactions (reactions 05, 06, 07 and 08) and barrierless reaction 15 were not considered. Reaction 12 has two spin states doublet and quartet in the original dataset, and only the transition state with lower energy (12b quartet) was considered because the model was trained with energy and forces from the lower energy spin state.

rxn	Original Model		Final Model	
	reactant(%)	product(%)	reactant(%)	product(%)
01	12	88	25	75
02	0	100	44	56
03	48	49	50	50
04	100	0	44	56
09	86	14	62	38
10	58	41	55	45
11	91	9	55	45
12	28	72	45	55
13	55	44	65	35
14	7	93	40	60
16	0	100	48	52
17	96	4	49	51
18	22	78	43	57
19	95	5	52	48

The final AL model gives quite different committer statistics than the original model, in which the AL model shows a more even chance to commit to reactant and product for a majority of the reactions. Hence although the AIMD and normal modes sampling provided data representation in the tube around the IRC, it is insufficient such that the commitor analysis is qualitatively different in a majority of the reactions. But once the model is more complete using our AL strategy, there is a shift in the

forward/backward committer distribution in most of the reaction channels in which the 0 K IRC is still a good estimator, meaning the transition state is mostly enthalpic. Even so, for reactions 01, 09, 14, and 18 simulated at finite temperature, the commitor analysis provides evidence that there is either a non-negligible entropy component to the free energy transition state, the CVs are wrong, and/or the ML model is incorrect.

With a more complete ML PES obtained through the active learning procedure, we ran long metadynamics trajectories and reconstructed their free energy surface to determine the transition state free energy for these reactions. Unlike in the exploration phase where any CV can help discover new structures, in the real application to reconstruct the free energy surface, a good choice of CVs is important. Due to the diffusive nature of gas phase reactions, the distance between two fragments can be arbitrary and lead to convergence problem if we directly use distance as the collective variable. Here we use CVs inspired by our recent study of diffusion maps in order to assess CVs suitable for each reaction as evaluated through their correlations with the diffusion coordinates.[25] We found that the coordination number (CN) is often a good CV, and restraints on certain bond distance and bond angles are also necessary to keep the metadynamics trajectory inside the relevant chemical space defined by the CVs.

We also exploit the strength of having an ensemble of ML models in which the standard deviation between the models measures the reliability in the prediction of ML forces on a given configuration. We therefore devise a hybrid mode of running a simulation in which we use ML forces when the standard deviation among models is relatively small, and with *ab initio* forces at the same level of theory used for ML training when the model predictions are not as reliable (Figure 4.2.5). This hybrid mode can be (1) much less expensive than fully retraining the ML models with the new DFT data, and (2) can be more accurate and maintain stability when uncertainties arise, and involves a tradeoff between how complete is the AL process vs. the relative cost of the ML and DFT forces. Because retraining the model varies between 10 hours to 2 days, in late stages of AL the tradeoff favors the ~ 12 seconds to generate the *ab initio* forces instead. As we support further below, a vast majority of the MD trajectory is advanced through ML forces that only cost ~ 0.11 seconds/step and thus the promised efficiencies of ML are still realized. While there is still a possibility that the query-by-committee models agree but give a wrong prediction, in Supplementary Figure 2 we see that the error is in acceptable range when models agree, while the points of disagreement recognized by the procedure are indeed error-prone. Another way to mitigate this concern would be to increase the number of ML models, so that there's smaller chance for models to agree accidentally on unseen data.

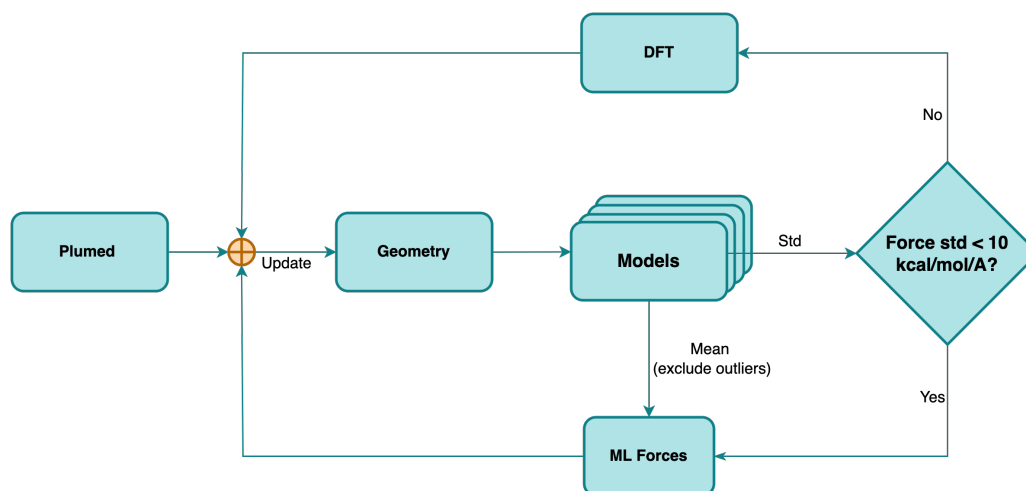


Figure 4.2.5: *Schematic illustration of the new workflow for rebuilding the free energy surface.* The metadynamics proceeds with a model that utilizes query by committee ML trajectories to signal when forces have degraded. In lieu of retraining, the AIMD forces are directly substituted to complete the MD time step. Plumed calculator updates the force according to existing bias to keep the system away from the kinetic traps in the potential energy surface and out into the unexplored parts of the energy landscape.

Figure 4.2.6 shows the reconstructed free energy surfaces using the hybrid model for reaction channels where the reactant-to-product ratios are uneven. For reaction 09, when we use $CN(O1-H3)$ and $CN(O2-H3)$ as the collective variables we find that the IRC transition state leans towards the reactant side on the free energy surface, consistent with the 62:38 committer ratio, and thus shifts quite significantly on the free energy surface toward the product well (Figure 4.2.6a). For Rxn 18 shown in Figure 4.2.6b we can clearly see that the IRC TS is leaning towards the product side, which explains the 43:57 reactant-to-product ratio in the committer analysis. However a lower transition pathway also exists on the free energy surface that resides closer to the reactant. The IRC transition state of both Rxn14 and Rxn01 lean very heavily toward the product side (Figure 4.2.6c, d), as is consistent with the more skewed committer reactant-product ratios (Table 1). Accordingly, the free energy transition state shifts dramatically in both reactions, with saddle point regions that exhibit very large free energy stabilization of 8-12 kcal/mole relative to the IRC transition state arising from entropic effects.

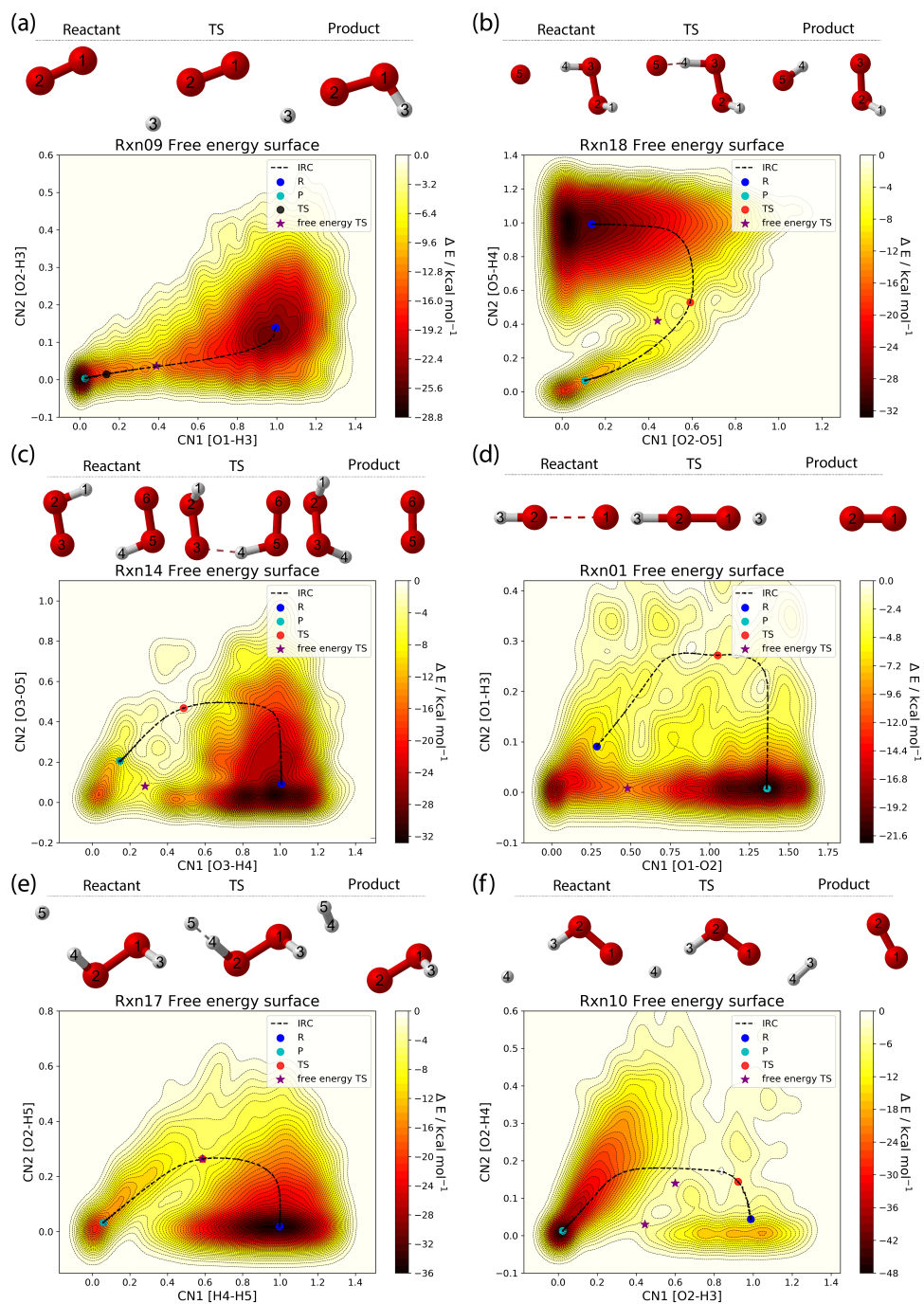


Figure 4.2.6: Free energy surface reconstructed from metadynamics using the hybrid model for hydrogen combustion. (a) Reaction 09, (b) Reaction 18, (c) Reaction 14, (d) Reaction 01, (e) Reaction 17, (f) Reaction 10. Reactant, transition state and product geometry are shown above the free energy surface, with oxygen in red and hydrogen in silver and atomic index labeled. The original IRC pathway is labeled with a red dot for the transition state and the free energy transition state is labeled as a purple star. The CVs used for the free energy are given in Supplementary Table 2.

As a control we tested reactions 17 and 10 in which the reactant-to-product ratio in the committer analysis is close to 50-50 (Figure 4.2.6e, f). For Rxn 17 the IRC and free energy transition state are consistent, indicating that the coordination numbers tend to be good collective variables and that the entropic factors governing these reactions are small. For Rxn 10 we found $CN(O2-O3)$ and $CN(O2-H4)$ to be good collective variables, where $CN(O2-O3)$ monitors the reaction progress and $CN(O2-H4)$ helps separates the transition state from the stable wells. Consistent with the 55:45 reactant-to-product committer ratio the IRC transition state leans towards the reactant side, but on the free surface two possible transition pathways are evident (Figure 4.2.6f). One pathway is very similar to the IRC but with the transition state moving closer to the product well, while the other pathway demarcated by a smaller $CN(O2-H4)$ value involves breaking the O2-H3 and forming the H3-H4 molecule in an asynchronous manner. It is important to consider whether the hybrid model is generating sensible free energy transition states. Therefore we have taken the putative transition states for reactions 01, 09, 10, and 14 generated by the hybrid model (shown in Figure 5), and ran committor analysis using DFT as reported in Supplementary Table 4. It is evident that the AIMD confirms the predictions of the hybrid model.

Through this process, we find that the number of unreliable region of PES are relatively infrequent. For Rxn 18 the full DFT force was called 1364 times over a total of 1 million steps, which means 99.86% of the trajectory is driven by the ML model, and only 0.14% of steps calculated through the *ab initio* update. For Rxn10 the DFT force was called only 136 times, which means that 99.99% of the total MD steps were generated by the ML forces. The small chance to go through *ab initio* updates reflects that the learned PES after our negative design and metadynamics active learning procedures is largely complete. But it is important to emphasize that without the *ab initio* calls the MD trajectories would become corrupted beyond repair and this hybrid ensemble approach allows for highly efficient simulation with a time scale similar to the pure ML run but with more stability, offering a preventative and alternative to occasional model hallucinations.

4.3 DISCUSSIONS

ML methods have traditionally aimed to train a faithful surrogate model of *ab initio* energy and forces for different configurational arrangements of a system in order to explore reaction chemistry more efficiently than AIMD.[43, 31, 20] For real applications it is likely that the interpolative nature of ML will always suffer from data insufficiencies that will limit its total replacement of physics based methods.

While AL approaches can mitigate data insufficiency by sampling and adding new data, it is still hard to converge to a complete ML-PES as the cost of retraining is far from trivial. Ultimately to address real world application studies using ML potentials, any *a priori* formulated reference dataset, even with active learning approaches, will be too small-scale for true PES completeness.

One alternative approach could be delta learning that learns the difference between a lower level theory and a higher level theory[34, 7]. This type of simulation would require calling both the lower theory method and the ML correction at each time step, and can also avoid the detrimental effect from ML-PES rough spots. However, the quality and cost of this type of simulation would highly depend on the underlying lower level theory method, and switching between the two surface can affect the energy conservation in the trajectory.

Alternatively we have shown that any large variance among ML models provides a nice strategy to invoke a hybrid model, which smooths over rough spots on the ML PES with calls to the original ab initio data source. We envision that further improvements to the hybrid ML-physics model would be warranted. First is that the energy drift rate in the microcanonical(NVE) ensemble for a single ML PES is quite acceptable at $7e-5$ kcal/mol/step, but averaging across the 4 member committee, which is common practice in this area currently, gives a relatively poor drift rate of $3e-3$ kcal/mol/step. We believe dynamics and transport properties would improve if we increased the number of ML committee members N , with statistical errors decreasing by $N^{1/2}$, and made feasible by trivial parallelization. Although we showed that the hybrid ML-AIMD model allowed us to realize hoped for ML computational efficiencies by reaching two orders of magnitude improvement over AIMD, we note that we could have optimized this tradeoff by stopping AL training sooner with an increase in calls to the ab initio source from 1% to 10% without impacting efficiency.

4.4 METHODS

NewtonNet Model[†]

Given a molecular graph \mathcal{G} with atomic features $a_i \in \mathbb{R}^{\text{nf}}$ (where nf is the number of features) and interatomic attributes $e_{ij} \in \mathbb{R}^{\text{b}}$, a message passing layer can be defined as[15]:

[†]Partly reproduced with permission from: Haghghatlari, M.; Li, J.; Guan, X.; Zhang, O.; Das, A.; J. Stein, C.; Heidar-Zadeh, F.; Liu, M.; Head-Gordon, M.; Bertels, L.; Hao, H.; Leven, I.; Head-Gordon, T. NewtonNet: A Newtonian Message Passing Network for Deep Learning of Interatomic Potentials and Forces. *Digital Discovery* **2022**, 1 (3), 333–343.

$$m_{ij} = M_l(a_i^l, a_j^l, e_{ij}) \quad (4.1)$$

$$m_i = \sum_{j \in \mathcal{N}(i)} m_{ij} \quad (4.2)$$

$$a_i^{t+1} = U_l(a_i^t, m_i) \quad (4.3)$$

where M_l is the message function and U_l is called the update function, and the sub-/super-script l accounts for the number of times the layer operates iteratively. A combination of explicit differentiable functions and operators with trainable parameters are the common choice for M_l and U_l . The core idea behind the iterative message passing of the atomic environments is to update the feature array a_i^t that represent each atom in its immediate environment.

NewtonNet considers a molecular graph defined by atomic numbers $Z_i \in \mathbb{R}^1$ and relative position vectors $\vec{r}_{ij} = \vec{r}_j - \vec{r}_i \in \mathbb{R}^3$, as input and applying operations that are inspired by Newton’s equations of motion to create features arrays $a_i \in \mathbb{R}^{\text{nf}}$ that represent each atom in its immediate environment with edges defined by force and displacement vectors, \mathbf{f} and $d\mathbf{r}$, respectively, (Fig. 1a). NewtonNet takes advantage of multiple layers of message passing which are rotationally equivariant, described in detail below, in which each layer consists of multiple modules that include operators to construct force and displacement feature vectors, which are contracted to the feature arrays via the energy calculator module (Fig. 1b). We emphasize the critical role of projecting equivariant feature vectors to invariant arrays since one goal of the model is to predict potential energies, which are invariant to the rotations of atomic configurations. We also provide the proof of equivariance of the NewtonNet model in the Supplementary Information as well.

Atomic Feature Aggregator. We initialize the atomic features based on trainable embedding of atomic numbers Z_i , i.e., $a_i^0 = g(Z_i)$ and $g : \mathbb{R}^1 \rightarrow \mathbb{R}^{\text{nf}}$. We next use the edge function $e : \mathbb{R}^3 \rightarrow \mathbb{R}^{\text{nb}}$ to represent the interatomic distances using radial Bessel functions as introduced by Klicpera et al.[24]

$$e(\vec{r}_{ij}) = \sqrt{\frac{2}{r_c}} \frac{\sin(\frac{n\pi}{r_c} \|\vec{r}_{ij}\|)}{\|\vec{r}_{ij}\|} \quad (4.4)$$

where r_c is the cutoff radius and $\|\vec{r}_{ij}\|$ returns the interatomic distance between any atom i and j . We follow Schutt et al.[38] in using a self-interaction linear layer

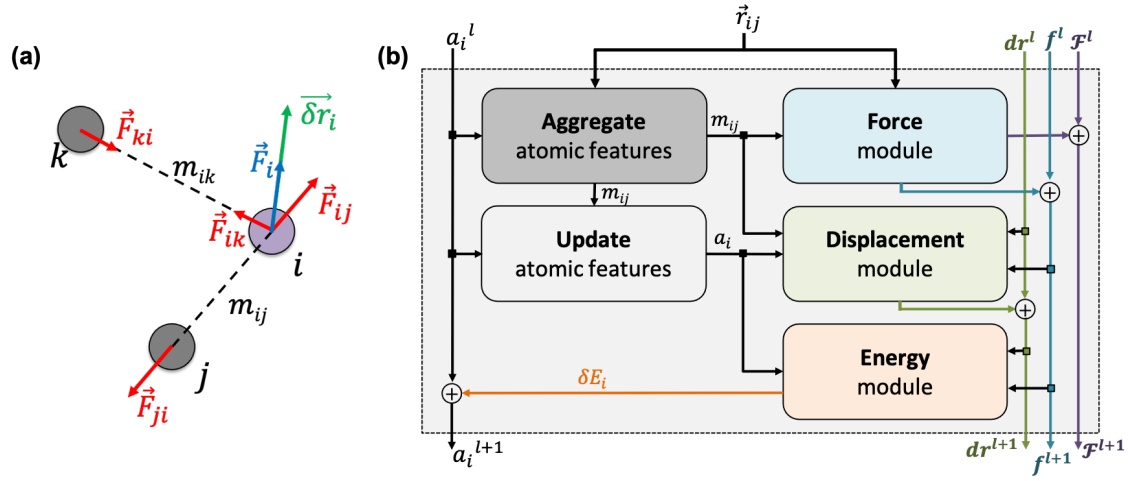


Figure 4.4.1: (a) Newton's laws for the force and displacement calculations for atom i with respect to its neighbors. (b) Schematic view of the NewtonNet message passing layer. At each layer four separate components are updated: atomic feature arrays a_i , latent force vectors \mathcal{F} , and force and displacement feature vectors (\mathbf{f} and \mathbf{dr}).

$\phi_{rbf} : \mathbb{R}^{\text{nb}} \rightarrow \mathbb{R}^{\text{nf}}$ to combine the output of radial basis functions with each other. This operation is followed by using an envelop function to implement a continuous radial cutoff around each atom. For this purpose, we use the polynomial function e_{cut} introduced by Klicpera et al.[24] with the choice of degree of polynomial $p = 7$. Thus, the edge operation $\phi_e : \mathbb{R}^3 \rightarrow \mathbb{R}^{\text{nf}}$ is defined as a trainable transformation of relative atom position vectors in the cutoff radius r_c

$$\phi_e(\vec{r}_{ij}) = \phi_{rbf}(e(\vec{r}_{ij})) e_{cut}(r_c, \|\vec{r}_{ij}\|) \quad (4.5)$$

The output of ϕ_e is rotationally invariant as it only depends on the interatomic distances. Following the notation of neural message passing, we define a message function to collect the neighboring information and update atomic features. Here, we tend to pass a symmetric message between any pair of atoms, i.e., the message that is passed between atom i and atom j are the same in both directions. Thus, we introduce our symmetric message passing m_{ij} by element-wise product between all feature arrays involved in any two-body interaction,

$$m_{ij} = \phi_a(a_i^l) \phi_a(a_j^l) \phi_e(\vec{r}_{ij}) \quad (4.6)$$

where $\phi_a : \mathbb{R}^{\text{nf}} \rightarrow \mathbb{R}^{\text{nf}}$ indicates a trainable and differentiable network with a nonlinear activation function SiLU [13] after the first layer. Note that the ϕ_a is the same function applied to all atoms. Thus, due to the weight sharing and multiplication of output features of both heads of the two-body interaction, the m_{ij} remain symmetric at each layer of message passing. To complete the feature array aggregator, we use the equation 4.2 to simply sum all messages received by central atom i from its neighbors $\mathcal{N}(i)$. Finally, we update the atomic features at each layer using the sum of received messages,

$$a_i^{l+1} = a_i^l + \sum_{j \in \mathcal{N}(i)} m_{ij}. \quad (4.7)$$

Force Calculator. So far, we have followed a standard message passing that is invariant to the rotation. We begin to take advantage of directional information starting from the force calculator module. The core idea behind this module is to construct latent force vectors using the Newton’s third law. The third law states that the force that atom i exerts on atom j is equal and in opposite direction of the force that atom j exerts on atom i . This is the reason that we intended to introduce a symmetric message passing operator. Thus, we can estimate the symmetric force magnitude as a function of m_{ij} , i.e., $\|\vec{F}_{ij}^l\| = \phi_F(m_{ij})$. The product of the force magnitude by unit distance vectors $\hat{r}_{ij} = \vec{r}_{ij}/\|\vec{r}_{ij}\|$ gives us antisymmetric interatomic forces that obey the Newton’s third law (note that $\vec{r}_{ij} = -\vec{r}_{ji}$),

$$\vec{F}_{ij}^l = \phi_F(m_{ij}) \hat{r}_{ij} \quad (4.8)$$

where $\phi_F : \mathbb{R}^{\text{nf}} \rightarrow \mathbb{R}^1$ is a differentiable learned function, and $\vec{F}_{ij}^l \in \mathbb{R}^3$. The total force at each layer \vec{F}_i^l on atom i is the sum of all the forces from the neighboring atoms j in the atomic environment,

$$\vec{F}_i^l = \sum_{j \in \mathcal{N}(i)} \vec{F}_{ij}^l, \quad (4.9)$$

and updating the latent force vectors at each layer,

$$\mathcal{F}_i^{l+1} = \mathcal{F}_i^l + \vec{F}_i^l. \quad (4.10)$$

We ultimately use the latent force vector from the last layer L , $\mathcal{F}_i^L \in \mathbb{R}^3$ in the loss function to ensure this latent space truly mimics the underlying physical rules.

To complete the force calculator module, we borrow the idea of continuous filter from Schut et al.[37] to decompose and scale latent force vectors along each dimension

using another learned function $\phi_f : \mathbb{R}^{\text{nf}} \rightarrow \mathbb{R}^{\text{nf}}$. This way we can featurize the vector field to avoid too much of abstraction in the structural information that they carry with themselves,

$$\Delta \mathbf{f}_i = \sum_{j \in \mathcal{N}(i)} \phi_f(m_{ij}) \vec{F}_{ij}^l. \quad (4.11)$$

As a result, the constructed latent interatomic forces are decomposed by rotationally invariant features along each dimension, i.e., $\Delta \mathbf{f}_i \in \mathbb{R}^{3 \times \text{nf}}$. We call this type of representation feature vectors. Following the message passing strategy, we update the force feature vectors with $\Delta \mathbf{f}_i$ after each layer, while they are initialized with zero values, $\mathbf{f}_i^0 = \mathbf{0}$,

$$\mathbf{f}_i^{l+1} = \mathbf{f}_i^l + \Delta \mathbf{f}_i. \quad (4.12)$$

Momentum Calculator. This is the step that we try to estimate a measure of atomic displacement due to the forces that are exerted on them. We accumulate their displacements at each layer without updating the position of each atom. The main idea in this module is that the displacement must be along the updated force features in the previous step. Inspired by Newton's second law, we approximate the displacement factor using a learned function $\phi_r : \mathbb{R}^{\text{nf}} \rightarrow \mathbb{R}^{\text{nf}}$ that acts on the current state of each atom presented by its atomic features a_i^l ,

$$\delta \mathbf{r}_i = \phi_r(a_i^{l+1}) \mathbf{f}_i^{l+1}. \quad (4.13)$$

We finally update the displacement feature vectors by $\delta \mathbf{r}_i$ and a weighted sum of all the atomic displacements from the previous layer. The weights are estimated based on a trainable function of messages ($\phi'_r : \mathbb{R}^{\text{nf}} \rightarrow \mathbb{R}^{\text{nf}}$) between atoms,

$$\mathbf{d}\mathbf{r}_i^{l+1} = \sum_{j \in \mathcal{N}(i)} \phi'_r(m_{ij}) \mathbf{d}\mathbf{r}_i^l + \delta \mathbf{r}_i. \quad (4.14)$$

The weight component in this step works like attention mechanism to concentrate on the two-body interactions that cause maximum movement in the atoms. Since forces at $l = 0$ are zero, the displacements are also initialized with zero values, i.e., $\mathbf{d}\mathbf{r}_i^0 = \mathbf{0}$.

Energy Calculator. The last module contracts the directional information to the rotationally invariant atomic features. Since we developed the previous steps based on the Newton's equations of motion, one immediate idea is to approximate the potential energy change for each atom using f_i^l and δr_i^l , resembling $\mathbf{f}_i^l \approx -\delta U / \delta \mathbf{r}_i^l$ in the higher dimensional space (\mathbb{R}^{nf}). Thus, we find energy change for each atom by

$$\delta U_i = -\phi_u(a_i^{l+1}) \langle \mathbf{f}_i^{l+1} \cdot \mathbf{dr}_i^{l+1} \rangle, \quad (4.15)$$

where $\delta U_i \in \mathbb{R}^{\text{nf}}$ and $\phi_u : \mathbb{R}^{\text{nf}} \rightarrow \mathbb{R}^{\text{nf}}$ is a differentiable learned function that operates on the atomic features and predicts the energy coefficient for each atom. The dot product of two feature vectors contracts the features along each dimension to a single feature array. We finally update the atomic features once again using the contracted directional information presented as atomic potential energy change,

$$a_i^{l+1} = a_i^{l+1} + \delta U_i. \quad (4.16)$$

This approach is both physically and mathematically consistent with the rotational equivariance operations and the goals of our model development. Physically, the energy change is the meaningful addition to the atomic feature arrays as they are used to predict the atomic energies eventually. Mathematically, the dot product of two feature vectors contracts the rotationally equivariant features to invariant features similar to euclidean distance that we used in the *atomic feature aggregator* module. Note that none of the force, displacement or energy modules are directly mapped to the final energy and force predictions. These are intermediate steps that update atomic features iteratively beyond the immediate neighborhood of each atom.

Dilation Data Preparation

To address the problem that the initial Dataset from Ref. [19] is missing high energy states, we prepared additional data by proportionally scaling each geometric coordinate in the IRC with multiple ratios (0.6, 0.7, 0.8, 0.9, 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, 2.0, 2.4, 2.8, 3.2). The energy and forces of these new geometries were obtained with QChem[40] 5.2 using range separated hybrid meta-GGA functional ω B97X-V [30] with the cc-pVTZ basis set.

Active Learning: Query by Committee

Here we exploit the query by committee active learning method to find a label with the most informative data points through this iterative process:

1. Perform short metadynamic simulations to explore the configuration space in a lower dimension.

2. When the four models disagree outside standard deviation, collect a representative subset of structures to be included in the training set through downsampling.
3. Perform DFT calculation of energies and atomic forces to label the new data.
4. Retrain the ensemble of ML models with the updated training set.

The details of each of the steps are described in the following subsections.

Training of the NN PES

The NewtonNet model was used to train the NN potential in this work. Details of this method can be found in Ref. [20]. Four NewtonNet models, initialized with different weights, were trained for 2000 epochs with 1000 data points per reaction from the original dataset with an additional 200 data points from dilation (Section 4.1). The cutoff radius was set to 5.0 Å. The initial learning rate was set to 0.001 and with 0.7 learning rate decay.

We train the model using small batches of data with batch size M . The loss function penalizes the model for predicted energy values E_m , force components F_{mi} , and the direction of latent force vectors from last message passing layer \mathcal{F}_i^L . These three terms of the loss function \mathcal{L} are formulated as:

$$\begin{aligned} \mathcal{L} = & \frac{\lambda_E}{M} \sum_m^M w_m \left(\tilde{E}_m - E_m \right)^2 \\ & + \frac{\lambda_F}{M} \sum_m^M \frac{w_m}{3N_m} \sum_i^{N_m} \left\| \tilde{\mathbf{F}}_{mi} - \mathbf{F}_{mi} \right\|^2 \\ & + \frac{\lambda_D}{M \times N_m} \sum_m^M w_m \sum_i^{N_m} \left(1 - \frac{\mathcal{F}_{mi}^L \cdot \mathbf{F}_{mi}}{\left\| \mathcal{F}_{mi}^L \right\| \left\| \mathbf{F}_{mi} \right\|} \right) \end{aligned}$$

where N_m is the total number of atoms. The prefactor of the energy error λ_E is set to 1, the prefactor of force error λ_F is set to 20, and the prefactor λ_D for latent force direction is set to 1. We also use the following Boltzmann weighting factor w_m , defined as

$$w_m = \begin{cases} 1 & \text{if } E_m \leq E_{thresh} \\ \exp\left(\frac{-(E_m - E_{thresh})}{k_{BT}}\right) & \text{if } E_m > E_{thresh}, \end{cases}$$

to bias the training towards data points within a relevant energy scale. E_{thresh} is a per-atom quantity that puts less weighting on all data points with energy higher

than 16.744 kcal/mol/atom, which is 10 kcal/mol/atom higher than the highest per atom energy among all reaction channels of the IRCs. To completely converge the ML model, we added one final training step with all previously added data, using a larger epoch size (5000 steps), and a more patient learning rate decay to give the final model that we later use for determining predictions on the free energy surface. Subsequent ML models are trained with exactly the same protocol but with additional data sampled from short metadynamics simulations (next Section).

Metadynamics

New structures are sampled through short metadynamics trajectories for 6 reaction channels: rxn09, rxn10, rxn13, rxn16, rxn17 and rxn18. For each step in the metadynamics simulation, the atomic forces are evaluated by four NN PES models simultaneously. Outliers among the 4 predictions are removed if the absolute difference between the outlier in question and the closest number is larger than the 95% confidence limit value of the Dixon Q’s test. Then the mean of model predicted forces is modified by plumed [48] to allow for enhanced sampling. In this work, all enhanced sampling simulations are performed with the well-tempered metadynamics, in which a Gaussian centered at the visited point is periodically added to the potential. The simulation is driven through the atomic simulation environment(ASE)[28] with a specifically tailored calculator that provide energy and forces for a given structure through the above protocol. The simulation is conducted at 300K with increasing length as we obtain more active learning rounds: 2ps between active learning round 1-20, 5ps round 21-33 and 10ps between round 34-48.

During the simulation, standard deviation on atomic forces over an ensemble of NN potentials is monitored. Whenever the maximum over the atoms exceeds a predefined threshold (2 kcal/mol/Å), the configuration is selected for further downsampling.

Clustering and Down Selection

Each molecule is first represented as a Coulomb matrix[35] that includes the nuclear charges (Z_i and Z_j) of atom i and j along with their Cartesian coordinates (R_i and R_j).

$$C_{ij} = \begin{cases} 0.5Z_i^{2.4}, & i = j \\ \frac{Z_i Z_j}{|R_i - R_j|}, & i \neq j \end{cases} \quad (4.17)$$

To reduce the dimension of the dataset while retaining the majority of structural information, the Coulomb matrix was transformed into the eigen-spectrum by solving the eigenvalue problem $C\mathbf{v} = \lambda\mathbf{v}$ subject to the constraint $\lambda_i \geq \lambda_{i+1}$. The Mini Batch

KMeans clustering algorithm was then applied to categorize the sub-datasets into smaller clusters based on the eigen-spectrum. The value of K (the number of clusters) is chosen automatically with a scaled inertia approach [21]. The scaled inertia is formulated as

$$\text{Scaled Inertia} = \frac{I(K)}{I(K=1)} + \alpha K \quad (4.18)$$

where the inertia(I) is the sum of squared distance of samples to their closest cluster center:

$$I(K) = \sum_{i=1}^N (x_i - C_k)^2 \quad (4.19)$$

where N is the number of samples and C_k is the centroid of a cluster. α is a manually tuned factor that gives penalty to the number of clusters, here we chose $\alpha = 0.0002$. We chose the K value that gives the minimum scaled inertia among all K;300 to do the mini batch K-means clustering on all molecules from a given reaction channel. Afterward, we randomly picked a structure from each cluster, whose energy and forces will be calculated and then be included into the new training set. The final data set had 48,582 data points sampled in the AL procedure.

DFT Single Point Force Calculations

The structures that pass the clustering and downselection process are gathered for labeling and retraining. They are labeled via DFT force calculations using the ω B97X-V functional[30] with the cc-pVTZ basis set as generated from the Q-Chem 5.2 software package.[40, 14] All calculations were performed as unrestricted open shell, using an ultrafine integration grid of 99 radial points and 590 angular points, with an SCF convergence of 10^{-8} using the Gometric Direct Minimization method[50] All potential energies for each configuration of the 19 reactions are reported as ΔE

$$\Delta E = E_{total} - \sum_i E_{atom}, \quad (4.20)$$

using the atomic energies $E_H = -0.5004966690$ a.u. and $E_O = -75.0637742413$ a.u., and with ΔE converted to units of kcal/mole. Some of the structures are tricky to deal with due to the diffusive nature of a gas phase simulation. Therefore, for all structures with atomic oxygen that is separated from the rest of the structure more than 2 Å, we introduced an additional run with FRAGMO initial guess for SCF calculation. [23] We compared the final energy obtained from a calculation with fragmo initial guess and the one with the usual superposition of atomic densities(SAD)

initial guess, and take whichever one with lower energy to give the final energy and forces for labeling the data.

Reconstructing the Free Energy Surface with Metadynamics

The free energy surfaces were calculated from longer metadynamic simulations. The Langevin thermostat was used to maintain temperature at 300K, with a friction coefficient of 0.002 a.u. In the metadynamics simulations, the Gaussians adopted have an initial height of 5 kJ/mol and width of 0.05 for the CVs. A Gaussian was deposited every 100 step with a bias factor equal to 10. The simulations were 200 ps long with step size of 0.2 fs. The free energy were calculated using the `sum_hills` utility in `Plumed`[48] and its surface plotted with python `matplotlib`.

4.5 DATA AVAILABILITY

Coordinates of geometries, energy and forces for hydrogen combustion original dataset[19] is available at <https://doi.org/10.6084/m9.figshare.19601689>. IRC dialation data and active learning generated data[18] used in the training are available at:

<https://doi.org/10.6084/m9.figshare.23290115.v1>. Source data for Figures 1, 3, and 5 is available with this manuscript.

4.6 CODE AVAILABILITY

The full workflow code[17] can be found in https://github.com/THGLab/H2Combustion_AL.

4.7 ACKNOWLEDGMENTS

X.G, J. H. and T.H-G. thank the CPIMS program, Office of Science, Office of Basic Energy Sciences, Chemical Sciences Division of the U.S. Department of Energy under Contract DE-AC02-05CH11231 for support of the machine learning approach to hydrogen combustion. T. Ko and C. Y thank the U.S. Department of Energy via the Scientific Discovery through Advanced Computing (SciDAC) program for the collective variables. This work used computational resources provided by the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract DE-AC02-05CH11231.

4.8 AUTHOR CONTRIBUTIONS STATEMENT

X.G. and T.H.G. designed the project. X.G. carried out the AIMD simulations, metadynamic calculations, and active learning. X.G. and T.H.G designed the collective coordinates with the help of J.H.,T.K.,C.Y. X.G. and T.H.G. wrote the paper. All authors discussed the results and made comments and edits to the manuscript.

4.9 REFERENCES

- [1] Brandon Anderson, Truong Son Hy, and Risi Kondor. “Cormorant: Covariant molecular neural networks”. In: *Adv. Neural Inf. Process. Syst.* 32.NeurIPS (2019). ISSN: 10495258. arXiv: 1906.04015.
- [2] Shi Jun Ang, Wujie Wang, Daniel Schwalbe-Koda, Simon Axelrod, and Rafael Gómez-Bombarelli. “Active learning accelerates ab initio molecular dynamics on reactive energy surfaces”. In: *Chem* 7.3 (2021), pp. 738–751. ISSN: 2451-9294. DOI: <https://doi.org/10.1016/j.chempr.2020.12.009>. URL: <https://www.sciencedirect.com/science/article/pii/S2451929420306410>.
- [3] A. Barducci, G. Bussi, and M. Parrinello. “Well-tempered metadynamics: A smoothly converging and tunable free-energy method”. In: *Physical Review Letters* 100 (2008), p. 020603. DOI: 10.1103/PhysRevLett.100.020603.
- [4] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. “E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials”. en. In: *Nature Communications* 13.11 (May 2022), p. 2453. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29939-5.
- [5] Jörg Behler and Michele Parrinello. “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces”. In: *Phys. Rev. Lett.* 98 (14 Apr. 2007), p. 146401. DOI: <https://doi.org/10.1103/PhysRevLett.98.146401>.
- [6] Luke W. Bertels, Lucas B. Newcomb, Mohammad Alaghemandi, Jason R. Green, and Martin Head-Gordon. “Benchmarking the Performance of the ReaxFF Reactive Force Field on Hydrogen Combustion Systems”. In: *The Journal of Physical Chemistry A* 124.27 (July 2020), pp. 5631–5645. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.0c02734.

- [7] Lennard Bösel, Moritz Thürlmann, and Sereina Riniker. “Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-Phase Systems”. In: *Journal of Chemical Theory and Computation* 17.5 (May 2021), pp. 2641–2658. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.0c01112.
- [8] Stefan Chmiela, Huziel E. Saucedo, Klaus Robert Müller, and Alexandre Tkatchenko. “Towards exact molecular dynamics simulations with machine-learned force fields”. In: *Nat. Commun.* 9.1 (2018). ISSN: 20411723. DOI: 10.1038/s41467-018-06169-2. eprint: 1802.09238.
- [9] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Saucedo, Igor Poltavsky, Kristof T Schütt, Klaus-robert Müller, Igor Poltavsky, and Kristof T Sch. “Machine learning of accurate energy-conserving molecular force fields”. In: *Sci. Adv.* 3.5 (2017). DOI: 10.1126/sciadv.1603015.
- [10] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Saucedo, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. “Machine learning of accurate energy-conserving molecular force fields”. In: *Science Advances* 3.5 (2017), e1603015. DOI: 10.1126/sciadv.1603015. URL: <http://advances.sciencemag.org/content/3/5/e1603015.abstract>.
- [11] Anders S. Christensen, Lars A. Bratholm, Felix A. Faber, and O. Anatole von Lilienfeld. “FCHL revisited: Faster and more accurate quantum machine learning”. In: *The Journal of Chemical Physics* 152.4 (2020), p. 044107. ISSN: 0021-9606. DOI: 10.1063/1.5126701.
- [12] Ralf Drautz. “Atomic cluster expansion for accurate and transferable interatomic potentials”. In: *Phys. Rev. B* 99 (1 Jan. 2019), p. 014104. DOI: 10.1103/PhysRevB.99.014104. URL: <https://link.aps.org/doi/10.1103/PhysRevB.99.014104>.
- [13] Stefan Elfving, Eiji Uchibe, and Kenji Doya. “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning”. In: *Neural Networks* 107.2015 (2018), pp. 3–11. ISSN: 18792782. DOI: 10.1016/j.neunet.2017.12.012. arXiv: 1702.03118.
- [14] Evgeny Epifanovskiy, Andrew TB Gilbert, Xintian Feng, Joonho Lee, Yuezhi Mao, Narbe Mardirossian, Pavel Pokhilko, Alec F White, Marc P Coons, Adrian L Dempwolff, et al. “Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package”. In: *The Journal of Chemical Physics* 155.8 (2021), p. 084801.

- [15] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. “Neural message passing for quantum chemistry”. In: *arXiv preprint arXiv:1704.01212* (2017).
- [16] Zachary L. Glick, Alexios Koutsoukas, Daniel L. Cheney, and C. David Sherrill. “Cartesian message passing neural networks for directional properties: Fast and transferable atomic multipoles”. In: *The Journal of Chemical Physics* 154.22 (2021), p. 224103. DOI: <https://doi.org/10.1063/5.0050444>. eprint: <https://doi.org/10.1063/5.0050444>.
- [17] Xingyi Guan. *THGLab/H2Combustion_AL: v1.0.0*. 2023. DOI: 10.5281/ZENODO.8378075. URL: <https://zenodo.org/record/8378075>.
- [18] Xingyi Guan, Joseph Heindel, Taehee Ko, Chao Yang, and Teresa Head-Gordon. *Hydrogen Combustion supplementary data from an active learning study*. 2023. DOI: 10.6084/M9.FIGSHARE.23290115. URL: https://figshare.com/articles/dataset/Hydrogen_Combustion_supplementary_data_from_an_active_learning_study/23290115.
- [19] Xingyi Guan et al. “A benchmark dataset for Hydrogen Combustion”. en. In: *Scientific Data* 9.11 (May 2022), p. 215. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01330-5.
- [20] Mojtaba Haghghatlari et al. “NewtonNet: a Newtonian message passing network for deep learning of interatomic potentials and forces”. en. In: *Digital Discovery* 1.3 (2022), pp. 333–343. DOI: 10.1039/D2DD00008C.
- [21] Or Herman-Saffar. *An approach for choosing number of clusters for K-means*. June 2021. URL: <https://towardsdatascience.com/an-approach-for-choosing-number-of-clusters-for-k-means-c28e614ecb2c>.
- [22] Yuriy Khalak, Gary Tresadern, David F. Hahn, Bert L. de Groot, and Vytautas Gapsys. “Chemical Space Exploration with Active Learning and Alchemical Free Energies”. In: *Journal of Chemical Theory and Computation* 18.10 (2022), pp. 6259–6270. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.2c00752. URL: <https://doi.org/10.1021/acs.jctc.2c00752>.
- [23] Rustam Z. Khaliullin, Erika A. Cobar, Rohini C. Lochan, Alexis T. Bell, and Martin Head-Gordon. “Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals”. In: *The Journal of Physical Chemistry A* 111.36 (2007), pp. 8753–8765. DOI: 10.1021/jp073685z.
- [24] Johannes Klicpera, Janek Groß, and Stephan Günnemann. “Directional Message Passing for Molecular Graphs”. In: *arXiv preprint arXiv:2003.03123v1* (2020), pp. 1–13. arXiv: arXiv:2003.03123v1.

- [25] Taehee Ko, Joseph P. Heindel, Xingyi Guan, Teresa Head-Gordon, David B. Williams-Young, and Chao Yang. “Using Diffusion Maps to Analyze Reaction Dynamics for a Hydrogen Combustion Benchmark Dataset”. In: *Journal of Chemical Theory and Computation* 19.17 (Sept. 2023), pp. 5872–5885. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.3c00426.
- [26] Maksim Kulichenko, Kipton Barros, Nicholas Lubbers, Ying Wai Li, Richard Messerly, Sergei Tretiak, Justin S. Smith, and Benjamin Nebgen. “Uncertainty-driven dynamics for active learning of interatomic potentials”. en. In: *Nature Computational Science* 3.33 (Mar. 2023), pp. 230–239. ISSN: 2662-8457. DOI: 10.1038/s43588-023-00406-5.
- [27] A. Laio and M. Parrinello. “Escaping free-energy minima”. In: *Proceedings of the National Academy of Sciences* 99 (2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399.
- [28] Ask Hjorth Larsen et al. “The atomic simulation environment—a Python library for working with atoms”. en. In: *Journal of Physics: Condensed Matter* 29.27 (June 2017), p. 273002. ISSN: 0953-8984. DOI: 10.1088/1361-648X/aa680e.
- [29] Juan Li, Zhenwei Zhao, Andrei Kazakov, and Frederick L. Dryer. “An updated comprehensive kinetic model of hydrogen combustion”. In: *Int. J. Chem. Kinet.* 36.10 (2004), pp. 566–575. ISSN: 05388066. DOI: <https://doi.org/10.1002/kin.20026>.
- [30] N. Mardirossian and M. Head-Gordon. “ ω B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy”. In: *Phys. Chem. Chem. Phys.* 16 (2014), pp. 9904–9924. DOI: <https://doi.org/10.1039/c3cp54374a>.
- [31] Tim Mueller, Alberto Hernandez, and Chuhong Wang. “Machine learning for interatomic potential models”. In: *The Journal of Chemical Physics* 152.5 (Feb. 2020), p. 050902. ISSN: 0021-9606. DOI: 10.1063/1.5126336.
- [32] Cas van der Oord, Matthias Sachs, Dávid Péter Kovács, Christoph Ortner, and Gábor Csányi. “Hyperactive Learning (HAL) for Data-Driven Interatomic Potentials”. In: arXiv:2210.04225 (Nov. 2022). arXiv:2210.04225 [physics, stat]. DOI: 10.48550/arXiv.2210.04225. URL: <http://arxiv.org/abs/2210.04225>.

- [33] Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R. Manby, and III Miller Thomas F. “OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features”. In: *The Journal of Chemical Physics* 153.12 (2020), p. 124111. ISSN: 0021-9606. DOI: 10.1063/5.0021955. URL: <https://doi.org/10.1063/5.0021955>.
- [34] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. “Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach”. In: *Journal of Chemical Theory and Computation* 11.5 (May 2015), pp. 2087–2096. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00099.
- [35] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. “Fast and accurate modeling of molecular atomization energies with machine learning”. In: *Physical Review Letters* 108.5 (2012). DOI: 10.1103/physrevlett.108.058301.
- [36] Christoph Schran, Fabian L. Thiemann, Patrick Rowe, Erich A. Müller, Ondrej Marsalek, and Angelos Michaelides. “Machine learning potentials for complex aqueous systems made simple”. In: *Proceedings of the National Academy of Sciences* 118.38 (2021), e2110077118. DOI: 10.1073/pnas.2110077118. URL: <https://doi.org/10.1073/pnas.2110077118>.
- [37] K. T. Schutt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller. “SchNet - A deep learning architecture for molecules and materials”. In: *Journal of Chemical Physics* 148.24 (Mar. 2018), p. 241722. ISSN: 0021-9606. DOI: 10.1063/1.5019779.
- [38] Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. “Equivariant message passing for the prediction of tensorial properties and molecular spectra”. In: *arXiv preprint arXiv:2102.03150* (2021). arXiv: 2102.03150.
- [39] H. S. Seung, M. Opper, and H. Sompolinsky. “Query by Committee”. In: COLT ’92. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1992, pp. 287–294. ISBN: 089791497X. DOI: 10.1145/130385.130417. URL: <https://doi.org/10.1145/130385.130417>.
- [40] Yihan Shao et al. “Advances in molecular quantum chemistry contained in the Q-Chem 4 program package”. In: *Molecular Physics* 113.2 (Sept. 2014), pp. 184–215. DOI: 10.1080/00268976.2014.952696.

- [41] Alexander Shapeev, Konstantin Gubaev, Evgenii Tsymbalov, and Evgeny Podryabinkin. *Active learning and uncertainty estimation*. Jan. 1970. URL: https://link.springer.com/chapter/10.1007/978-3-030-40245-7_15#Sec8.
- [42] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. “ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost”. In: *Chemical Science* 8.4 (2017), pp. 3192–3203.
- [43] Justin S. Smith, Olexandr Isayev, and Adrian E. Roitberg. “ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules”. In: *Scientific Data* 4.1 (2017), p. 170193. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.193. URL: <https://doi.org/10.1038/sdata.2017.193>.
- [44] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. “Less is more: Sampling chemical space with active learning”. In: *The Journal of Chemical Physics* 148.24 (2018), p. 241733. ISSN: 0021-9606. DOI: 10.1063/1.5023802. URL: <https://doi.org/10.1063/1.5023802>.
- [45] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. “Less is more: Sampling chemical space with active learning”. In: *The Journal of Chemical Physics* 148.24 (2018), p. 241733. DOI: 10.1063/1.5023802.
- [46] Justin S. Smith, Benjamin T. Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E. Roitberg. “Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning”. In: *Nature Communications* 10.1 (2019), p. 2903. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10827-4. URL: <https://doi.org/10.1038/s41467-019-10827-4>.
- [47] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. “Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds”. In: *arXiv preprint arXiv:1802.08219* (2018).
- [48] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. “Plumed 2: New feathers for an old bird”. In: *Computer Physics Communications* 185.2 (2014), pp. 604–613. DOI: 10.1016/j.cpc.2013.09.018.

- [49] Oliver T. Unke, Stefan Chmiela, Michael Gastegger, Kristof T. Schütt, Huziel E. Saucedo, and Klaus-Robert Müller. “SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects”. In: *Nature Communications* 12.1 (2021), p. 7273. ISSN: 2041-1723. DOI: 10.1038/s41467-021-27504-0. URL: <https://doi.org/10.1038/s41467-021-27504-0>.
- [50] Troy Van Voorhis and Martin Head-Gordon. “A geometric approach to direct minimization”. In: *Molecular Physics* 100.11 (2002), pp. 1713–1721. ISSN: 0026-8976. DOI: 10.1080/00268970110103642. URL: <https://doi.org/10.1080/00268970110103642>.
- [51] Manyi Yang, Luigi Bonati, Daniela Polino, and Michele Parrinello. “Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water”. In: *Catalysis Today* 387 (2022), pp. 143–149. ISSN: 0920-5861. DOI: <https://doi.org/10.1016/j.cattod.2021.03.018>. URL: <https://www.sciencedirect.com/science/article/pii/S092058612100136X>.
- [52] Shuhao Zhang, Małgorzata Makoś, Ryan Jadrich, Elfi Kraka, Kipton Barros, Benjamin Nebgen, Sergei Tretiak, Olexandr Isayev, Nicholas Lubbers, and Richard and Messerly. “Exploring the frontiers of chemistry with a general reactive machine learning potential”. In: *ChemRxiv* (2023). DOI: 10.26434/chemrxiv-2022-15ct6-v3.

Appendix

4.A Proof of Equivariance and Invariance

We prove that our model is rotationally equivariant on the atomic positions $\mathbf{R}_i \in \mathbb{R}^3$ and atomic numbers Z_i for a rotation matrix $T \in \mathbb{R}^{3 \times 3}$. In the equation 1, the euclidean distance is invariant to the rotation, as it can be shown that

$$\begin{aligned}
 \|T\mathbf{r}_{ij}\|^2 &= \\
 \|T\mathbf{R}_j - T\mathbf{R}_i\|^2 &= \\
 (\mathbf{R}_j - \mathbf{R}_i)^\top T^\top T (\mathbf{R}_j - \mathbf{R}_i) &= \\
 (\mathbf{R}_j - \mathbf{R}_i)^\top \mathbf{I} (\mathbf{R}_j - \mathbf{R}_i) &= \\
 \|\mathbf{R}_j - \mathbf{R}_i\|^2 &= \\
 \|\mathbf{r}_{ij}\|^2, &
 \end{aligned} \tag{4.21}$$

which means that the euclidean distance is indifferent to the rotation of the positions as it is quite well-known for this feature. Consequently, feature arrays m_{ij} , a_i , and all the linear or non-linear functions acting on them will result in invariant outputs. The only assumptions for this proof is that a linear combination of vectors or their product with invariant features will remain rotationally equivariant. Base on this assumption we claim that equation 5 to 11 will remain equivariant to the rotations. For instance, the same rotation matrix T propagates to equation 5 such that,

$$\phi_F(Tm_{ij}) T\hat{r}_{ij} = \phi_F(m_{ij}) T\hat{r}_{ij} = T \phi_F(m_{ij}) \hat{r}_{ij} = T\vec{F}_{ij}^l. \tag{4.22}$$

The last operator, equation 12, will remain invariant to the rotations due to the use of dot product. The proof for the invariant atomic energy changes is that,

$$\begin{aligned}
& -\phi_u(a_i^{l+1}) (T \mathbf{f}_i^{l+1} \cdot T d\mathbf{r}_i^{l+1}) = \\
& -\phi_u(a_i^{l+1}) (\mathbf{f}_i^{l+1} T^\top T d\mathbf{r}_i^{l+1}) = \\
& -\phi_u(a_i^{l+1}) (\mathbf{f}_i^{l+1} \mathbf{I} d\mathbf{r}_i^{l+1}) = \\
& -\phi_u(a_i^{l+1}) (\mathbf{f}_i^{l+1} \cdot d\mathbf{r}_i^{l+1}) = \\
& \delta U_i.
\end{aligned} \tag{4.23}$$

This is how we contract equivariant features to invariant arrays. The addition of these arrays to atomic features preserves the invariance for the final prediction of atomic contributions to the total potential energy.

4.B Supplementary Figures

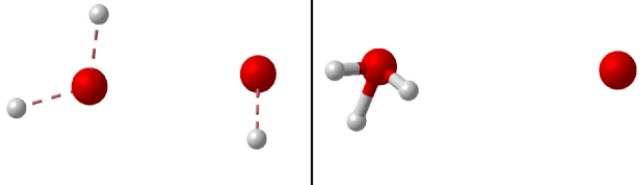
structure	
ML energy prediction (kcal/mol)	-324.08
DFT energy (kcal/mol)	-129.16

Figure 4.B.1: *Two representative structures that the original ML model predicts with large error.* Structures are shown as ball-and-stick with oxygen in red and hydrogen in silver. Energy predictions from the original ML model and DFT reference are provided to show that the original ML model predicts substantially lower energy values.

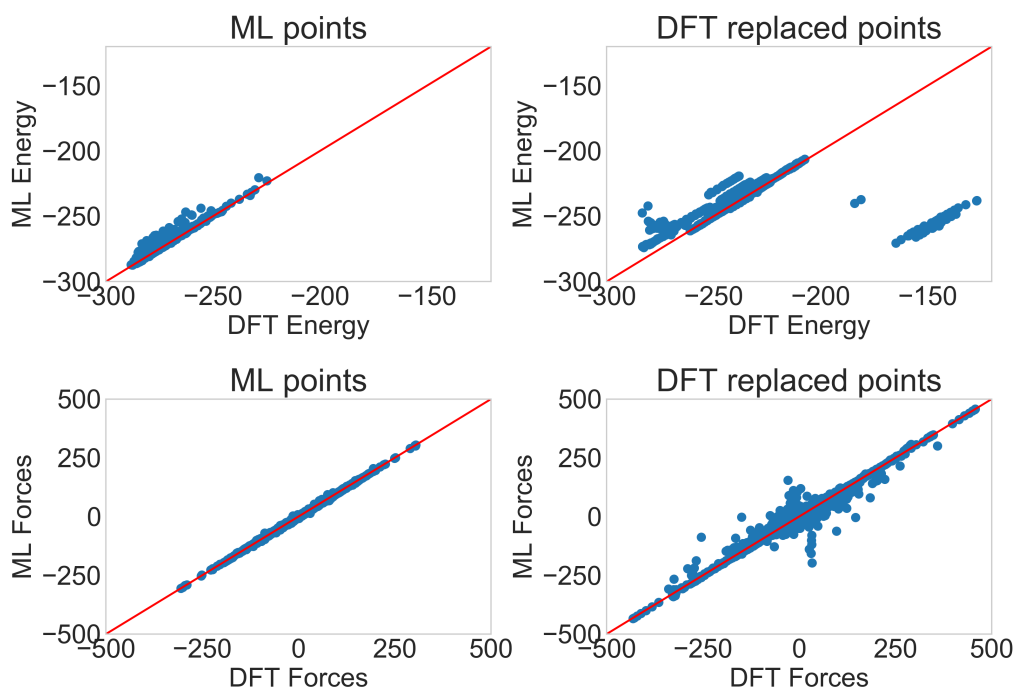


Figure 4.B.2: *Spot checking the hybrid mode model for Rxn18 for energies and forces.* On the left are ML points with standard deviations under 10 kcal/mol in energy or 10 kcal/mol/Å for forces selected from the trajectory. We can see that the ML model is in good agreement with the DFT reference. On the right are ML model standard deviations that are larger than 10 kcal/mol/(Å) and requiring DFT replacement. We see that poor predictions can be high energy states that are predicted to be low energy, and vice versa. Using the hybrid model, it is necessary to consider standard deviations regardless of their origin to flag high ML error, and to replace forces by directly calling the DFT calculation.

4.C Supplementary Tables

No. Reaction	Atoms	DoF	DoF _{int}
Association/Dissociation			
5. $\text{H}_2 \longrightarrow 2\text{H}$	2	6	1
6. $\text{O}_2 \longrightarrow 2\text{O}$	2	6	1
7. $\text{OH} \longrightarrow \text{O}+\text{H}$	2	6	1
8. $\text{H}+\text{OH} \longrightarrow \text{H}_2\text{O}$	3	9	3
9. $\text{H}+\text{O}_2 \longrightarrow \text{HO}_2$	3	9	3
15. $\text{H}_2\text{O}_2 \longrightarrow 2\text{OH}$	4	12	6
Substitution			
16. $\text{H}_2\text{O}_2+\text{H} \longrightarrow \text{H}_2\text{O}+\text{OH}$	5	15	9
O-transfer			
1. $\text{OH}+\text{O} \longrightarrow \text{H}+\text{O}_2$	3	9	3
11. $\text{HO}_2+\text{H} \longrightarrow 2\text{OH}$	4	12	6
12. $\text{HO}_2+\text{O} \longrightarrow \text{OH}+\text{O}_2$	4	12	6
H-transfer			
2. $\text{O}+\text{H}_2 \longrightarrow \text{OH}+\text{H}$	3	9	3
3. $\text{H}_2+\text{OH} \longrightarrow \text{H}_2\text{O}+\text{H}$	4	12	6
4. $\text{H}_2\text{O} \longrightarrow 2\text{OH}$	4	12	6
10. $\text{HO}_2+\text{H} \longrightarrow \text{H}_2+\text{O}_2$	4	12	6
13. $\text{HO}_2+\text{OH} \longrightarrow \text{H}_2\text{O}+\text{O}_2$	5	12	9
14. $2\text{HO}_2 \longrightarrow \text{H}_2\text{O}_2+\text{O}_2$	6	18	12
17. $\text{H}_2\text{O}_2+\text{H} \longrightarrow \text{HO}_2+\text{H}_2$	5	15	9
18. $\text{H}_2\text{O}_2+\text{O} \longrightarrow \text{HO}_2+\text{OH}$	5	15	9
19. $\text{H}_2\text{O}_2+\text{OH} \longrightarrow \text{H}_2\text{O}+\text{HO}_2$	6	18	12

Table 4.C.1: *The 19 reactions contained in the hydrogen combustion benchmark dataset.* The number of atoms involved in each reaction, the total number of degrees of freedom (DoF) in Cartesian coordinates, and total number of degrees of freedom in ICs (DoF_{int}.)

rxn	Old CV1	Old CV2	Final CV1	Final CV2
09	CN(O1-H3)	Angle(O2-O1-H3)	CN(O1-H3)	CN(O2-H3)
10	CN(O2-H3)	CN(H3-H4)	CN(O2-H3)	CN(O2-H4)
13	CN(O1-H3)	CN(O5-H3)	CN(O1-O5)	CN(O5-H3)
16	CN(O2-H3)	CN(O1-O2)	CN(O3-H5)	CN(O2-H5)
17	CN(O2-H4)	CN(H4-H5)	CN(H4-H5)	CN(O2-H5)
18	CN(O2-H4)	CN(O5-H4)	CN(O2-O5)	CN(O5-H4)

Table 4.C.2: *Metadynamics collective variables used in the active learning and free energy reconstruction.* As stated in the main text of the paper, in the sampling stage the method is not very sensitive to CV selection. Thus, we have used some different set of CVs throughout the active learning process and later for free energies. Here we list the CVs we have used in the sampling stage of the active learning. Old CVs are at the start of active learning and final CVs are for the later rounds of active learning as well as free energy surface reconstruction.

rxn	n data added
09	6686
10	6777
13	8175
16	8234
17	8230
18	8480

Table 4.C.3: *Total number of data points added in active learning for each reaction.* Throughout the 50 active learning cycles, a total of 46,182 data points each with DFT energy and forces are added into the dataset that trained the final model.

rxn	Reactant (%)	product(%)
01	44	56
09	53	47
10(TS1)	49	51
10(TS2)	50	50
14	44	56

Table 4.C.4: *AIMD Committer Analysis on identified Free Energy Transition State from the hybrid model at 500K.* We further validated the free energy transition states we identified using the hybrid model by running committer analysis with AIMD. All of these gives a close to 50-50 committer statistics for reactant and product, suggesting the ML predicted free energy transition states are accurate.