Accurate and transferable molecular-orbital-based machine learning for molecular modeling

Thesis by

Lixue Cheng

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

Caltech

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended March, 29, 2022

ACKNOWLEDGEMENTS

I want to thank my advisor, Professor Thomas F. Miller III, for his help and mentoring during my graduate studies, and my official advisor, Professor William Goddard, for his kindness and willingness to help my graduation progress over the past few months. I would also like to thank my thesis committee, Professor Garnet Chan, Professor Animashree Anandkumar, and Professor Niles Pierce, for their suggestions for my research and future career and your patience and help in scheduling all the meetings.

I feel lucky to be a graduate student at Caltech and in the Miller group, where I met so many talented young scientists who always provided valuable suggestions and supported me throughout all the difficult periods. I appreciate all the suggestions and efforts from my collaborator, Professor Anne McCoy in the University of Washington-Seattle; our postdocs, Dr. Matthew Welborn, Dr. Feizhi Ding, Dr. Tamara Husch, and Dr. J. Emiliano Deustua; and other graduate students, Dr. Sebastian J.R. Lee and Jiace Sun in this MOB-ML project. Without your help and and our collaborations, we not have developed such a fantastic tool. I also appreciate the nice discussions and suggestions from Zhuoran Qiao, which have greatly inspired us to improve MOB-ML.

Finally, I want to thank my parents, who are always supporting my dream and helping me to realize it. Special thanks to Jiace for being my partner in both work and life.

ABSTRACT

Quantum simulation is a powerful tool for chemists to understand the chemical processes and discover their nature accurately by expensive wavefunction theory or approximately by cheap density function theory (DFT). However, the cost-accuracy trade-offs in electronic structure methods limit the application of quantum simulation to large chemical and biological systems. In this thesis, an accurate, transferable, and physical-driven molecular modelling framework, i.e., molecular-orbital-based machine learning (MOB-ML), is introduced to provide accurate wavefunction-quality molecular descriptions with at most mean-field level computational cost. Instead of directly predicting the total molecular energies, MOB-ML describes the post-Hartree-Fock correlation energy from molecular orbital information at the cost of Hartree-Fock computations. Preserving all the physical constraints, molecular orbital based (MOB) features represent the chemical space faithfully in both supervised clustering and unsupervised learning for chemical space explorations. The development of local regressions with scalable exact Gaussian processes within clusters further allows MOB-ML to provide the most accurate approach in both low and big data regimes. As exciting and general new tool to tackle various problems in chemistry, MOB-ML offers great accuracies of predicting total energies and serves as a universal density functional for organic molecules and non-covalent interactions in various chemical systems. With the availability of analytical nuclear gradients, MOB-ML is also capable of generating accurate PESs with few reference high-level electronic structure computations in the diffusion Monte Carlo accurately and efficiently for computational spectroscopy.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. DiRisio, R. J., Cheng, L., Boyer, M. A., Lu, F., Sun, J., Lee, S. J. R., Deustua, J. E., Miller III, T. F. & McCoy, A. B. Near *ab Initio* potential energy surfaces for diffusion Monte Carlo using machine learning[1]. *In preparation.*
   L.C. contributed to the design of the research, wrote the computer codes, performed electronic structure computations, constructed ML models, and participated in the manuscript preparation.

2. Cheng, L., Kovachki, N. B., Welborn, M. & Miller III, T. F. Regression clustering for improved accuracy and training costs with molecular-orbital-based machine learning. *J. Chem. Theory Comput.* **15,** 6668–6677. `10.1021/acs.jctc.9b00884` (2019).
   L.C. generated part of the electronic structure data, wrote the computer codes, trained the ML models, made figures, performed data analysis, and participated in the writing of the manuscript.

3. Cheng, L., Welborn, M., Christensen, A. S. & Miller III, T. F. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *J. Chem. Phys.* **150,** 131103. `10.1063/1.5088393` (2019).
   L.C. generated part of the electronic structure data, trained the ML models, made figures, performed data analysis, and wrote the manuscript.

4. Welborn, M., Cheng, L. & Miller III, T. F. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **14,** 4772–4779. `10.1021/acs.jctc.8b00636` (2018).
   L.C. participated in the conception of the project, wrote the machine learning (ML) codes, trained the ML models, and wrote part of the manuscript.

[1] RJD and LC contributed equally to this work

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Number *Page*

NOMENCLATURE

AIMD. *Ab Initio* Molecular Dynamics.

AltBBMM. Alternative Blackbox Matrix-Matrix Multiplication.

BBMM. Blackbox Matrix-Matrix Multiplication.

CC. Coupled Cluster Theory.

CCSD. Coupled Cluster with Singles and Doubles.

CCSD(T). Coupled Cluster with Singles and Doubles with Perturbative Triples.

DF. Density Fitting.

DFT. Density Functional Theory.

DMC. Diffusion Monte Carlo.

FCI. Full Configuration Interaction.

GDB-13-T. A Thermalized Version of 1000 Selected GDB-13 Molecules with Thirteen C, O, N, S, and Cl Heavy Atoms.

GMM. Gaussian Mixture Model.

GP. Gaussian Process.

GPR. Gaussian Process Regression.

HF. Hartree-Fock.

IBO. Intrinsic Bond Orbital.

LMO. Localized Molecular Orbitals.

MAE. Mean Absolute Error.

MD. Molecular Dynamics.

ML. Machine Learning.

MO. Molecular Orbital.

MOB-ML. Molecular-Orbital-Based Machine Learning.

MP. Møller-Plesset Perturbation Theory.

MP2. Second-order Møller-Plesset Perturbation Theory.

NN. Neural Network.

NN + MOB-ML. Neural Network Refitted Molecular-Orbital-Based Machine Learning.

PES. Potential Energy Surface.

Post-HF. Post-Hartree-Fock.

QM7b-T. A Thermalized Version of the QM7b Set of 7211 Molecules with up to Seven C, O, N, S, and Cl Heavy Atoms.

RFC. Random Forest Classifier.

SCF. Self-Consistent Field.

SE. Schrödinger Equation.

SOTA. State-of-the-Art.

Chapter 1

INTRODUCTION

Quantum simulations have been shown to be powerful and widely-used tools to enhance our understanding of chemical and biological processes and facilitate the discovery of new drugs and materials. The ultimate goal of quantum simulations is to find the accurate numerical solutions to the Schrödinger equation (SE) with a reasonable computational cost. The Born-Oppenheimer approximation [1, 2] allows us to solve the SE by treating the nuclear and electronic wavefunctions separately since nuclei are much heavier than the electrons. The heavy nuclei can then be well-approximated as classical particles, while the proper treatment of electrons requires quantum mechanics. Electronic structure is the area that focuses on solving the wavefunctions of electrons after this wavefunction separation. This thesis introduces a physically-informed machine learning (ML) approach for electronic structure, known as molecular-orbital-based machine learning (MOB-ML), to solve the wavefunctions accurately and transferably in different chemical applications. To provide readers with more motivations for our work, in this chapter, we first introduce the commonly-used electronic structure theories and illustrate the high computational costs of the accurate theories. Then, we discuss the recent ML approaches to speed up and scale up the quantum simulations and the contributions of our MOB-ML approach to the field of ML for electronic structure. Finally, the overview of the contents of this thesis on the development and applications of MOB-ML is also provided.

Brief review of the theoretical developments in electronic structure During the past few decades, many traditional electronic structure methods in quantum chemistry have been developed to solve the electronic SEs accurately and approximately to obtain the ground and excited state energies and other molecular properties (Fig. 1.1). The exact quantum mechanical treatment can be realized by the Full Configuration Interaction (FCI) [3] method. It allows the many-body wavefunction to be represented as a linear combination of all possible electronic configurations, and thus its solution approaches the exact solution to the SE in the complete basis limit. However, the number of such possible configurations grows exponentially with the number of electrons (N), limiting it to systems with only a few electrons.

Wavefunction theories and density functional theory (DFT) [4] have been proposed in theoretical chemistry to solve the SE approximately. As the commonly used wavefunction theories, HF and methods like Møller-Plesset perturbation theory (MP) [5] and coupled cluster (CC) [6] are briefly introduced here.

In HF theory, the wavefunction can be represented as the Slater determinant of single-electron orbitals $\{\chi\}$.

$$|\Phi\rangle = \begin{vmatrix} \chi_1(1) & \chi_2(1) & \dots & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & \dots & \chi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(N) & \chi_2(N) & \dots & \chi_N(N) \end{vmatrix} \tag{1.1}$$

The orbitals $\{\chi\}$ are chosen to minimize the total energy of the system iteratively. Practically, the electron orbitals relax under the mean field created by other electrons self-consistently during the iterations, and this entire process is known as the self-consistent field (SCF) method. HF is an SCF method that captures most molecular energies in a mean-field way but does not include the crucial correlations between electrons for most chemical applications. In the MOB-ML approach, HF theory will provide input information to make predictions.

Based on HF theory, post-HF methods are designed to systematically correct the HF wavefunctions and energies by capturing the electron correlations. The perturbation theory treatment divides the full Hamiltonian $\hat{H}$ to an unperturbed Hamiltonian $\hat{H}_0$ and a perturbed Hamiltonian $\hat{H}_1$ ($\hat{H} = \hat{H}_0 + \lambda \hat{H}_1$) and then expands the wavefunction into the power series of the perturbation strength.

$$\begin{aligned} |\Psi\rangle &= \sum_{i=0} \lambda^i |\Psi\rangle^{(i)} \\ E &= \sum_{i=0} \lambda^i E^{(i)}, \end{aligned} \tag{1.2}$$

where λ is taken to be one after the expansion. As the most commonly used case of the perturbation theory, the MP theory takes the unperturbed Hamiltonian $\hat{H}_0$ as the Fock operator, which is defined as the effective one-electron Hamiltonian in HF, and leaves the rest part as the perturbed Hamiltonian $\hat{H}_1$. In this setup, the sum of the zeroth-order and the first-order energy equals the HF energy. Therefore, the second-order perturbation energy is the first non-vanishing correction in the MP theory. The method computing this correction is thus known as the second-order Møller–Plesset perturbation theory (MP2), which will be one of the reference theories learned by MOB-ML.

Another common treatment for electron correlation is CC, where the wavefunction is expressed as an exponential ansatz

$$|\Psi\rangle = e^{\hat{T}}|\Phi\rangle, \tag{1.3}$$

where $|\Phi\rangle$ is the HF ground state and the cluster operator $\hat{T}$ is the sum of series of excitation operators $\hat{T} = \hat{T}_1 + \hat{T}_2 \dots$, where

$$\begin{aligned}
\hat{T}_1|\Phi\rangle &= \sum_{i,a} t_i^a \Phi_i^a, \\
\hat{T}_2|\Phi\rangle &= \sum_{i>j,a>b} t_{ij}^{ab} \Phi_{ij}^{ab}, \\
&\dots
\end{aligned} \tag{1.4}$$

In practice, $\hat{T}$ is truncated to a specific order to reduce the computational cost, and the highest order term in truncated form can be treated perturbatively. The most common truncations are coupled cluster singles and doubles (CCSD) [7] and coupled-cluster singles, doubles, and perturbative triples (CCSD(T)) [8]. CCSD(T) is usually viewed as the gold-standard method in quantum chemistry for accurate energies and molecular properties.[9, 10]

Although the post-HF methods introduced above have good approximations to FCI, they are generally computationally expensive and thus are prohibitive for large chemical systems. For instance, MP2, CCSD, and CCSD(T) have a complexity of $\mathscr{O}(N^5)$, $\mathscr{O}(N^6)$ and $\mathscr{O}(N^7)$, respectively. The complexity of HF is as low as $\mathscr{O}(N^4)$, but the electron correlations are not included. To calculate larger systems with reasonable computational costs, theoretical chemists also put much effort into developing DFT to capture electron correlations. DFT is built on the Hohenberg-Kohn theorem [4], which states that for any system with a non-degenerate ground state, its ground state energy is a unique functional of the ground state electron density $n(r)$, and such ground state can be obtained variationally. Kohn-Sham DFT [11] further writes the electron density as the sum of the density of Kohn-Sham orbitals $\{\varphi_i(r)\}$ and expresses the energy functional as the sum of the Kohn-Sham kinetic energy, external potential energy, Coulomb interaction energy, and the exchange-correlation energy:

$$n(r) = \sum_i |\varphi_i(r)|^2, \tag{1.5}$$

$$\begin{aligned} E[n] = \sum_{i=1}^{N} \int dr \varphi_i^*(r)(-\frac{\hbar^2}{2m}\nabla^2)\varphi_i(r) + \int dr v_{\text{ext}}(r)n(r) \\ + \frac{e^2}{2}\int dr \int dr' \frac{n(r)n(r')}{|r-r'|} + E_{xc}[n]. \end{aligned} \tag{1.6}$$

The ground state energy can then be obtained by minimizing the functional $E[n]$. Since the exact exchange-correlation functional $E_{xc}[n]$ is unknown, the accuracies of DFT energies depend on how good the estimations of the exact exchange-correlation functionals are. The simplest but less accurate DFT is the local-density approximation (LDA) [12]. More accurate but more expensive functionals include generalized gradient approximations (GGA) and meta-GGA and hybrid-GGA. [13–15]

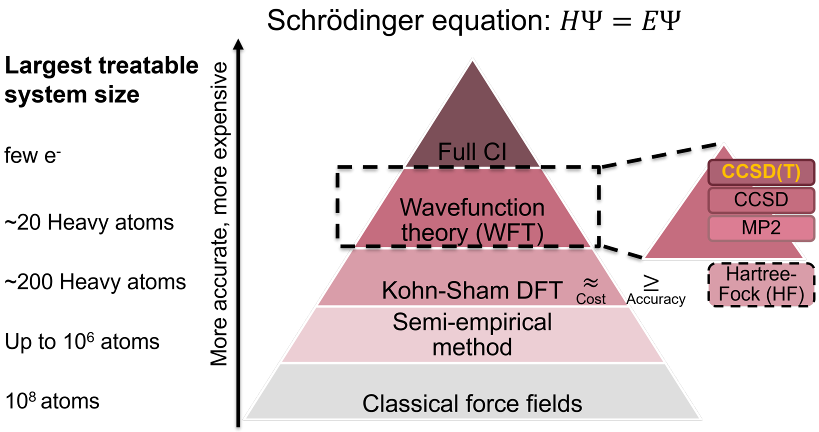

Figure 1.1: The pyramid for common methods in computational chemistry to solve Schrödinger equation. The computational costs and complexity scaling generally decrease, but the largest treatable system sizes increase from the top to the bottom. The accuracy of the computational methods is also considered decreasing from top to bottom. The three post-HF theories introduced in this thesis are also highlighted.

Besides the quantum mechanical treatment methods mentioned above, some classical force fields methods express the potential energies as functions of a set of atomic parameters, such as masses, coordinates, charges, and Lennard-Jones parameters. Such parametrizations could enable the computations of large systems but are not always accurate and transferable in different chemical applications due to the heavily numerical fitting towards some applications.

Brief review of the machine learning developments for electronic structure
The application of machine learning (ML) to electronic structure theory has been developing rapidly with an increasing number of studies in various chemical systems and applications [16, 17], such as directly predicting the molecular properties, developing force fields and interatomic potentials, and designing novel and efficient catalysts [18, 19], drugs [20–23] and materials [24, 25]. The major applications of ML in quantum simulations include predicting chemical properties to reduce computational costs by supervised learning [26, 27], detecting the patterns of chemical spaces by unsupervised learning [28, 29], and proposing more suitable chemical systems by reinforcement learning [22, 30] and generative models [31, 32]. This thesis will briefly review the different representations in supervised learning approaches to predict the molecular energies to reach chemical accuracy (1 kcal/mol) at different levels of electronic structure theories.

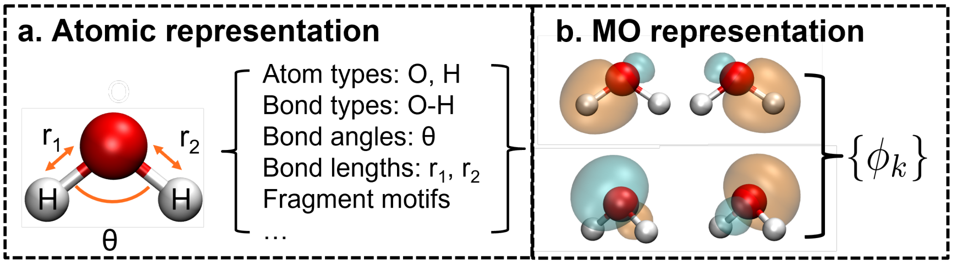

Figure 1.2: Atomic and physically informed molecular-orbital-based (MOB) representations to describe chemical systems in ML for electronic structure. The example system in this figure is H_2O. a. Common atomic representations in literature. b. Novel MOB representations introduced in this thesis.

There are two main categories of ML approaches to facilitate the electronic structure computations of energies in practice. The first category focuses on reaching excellent accuracy at the level of DFT with a computational cost of classical force fields [33–51]. These ML approaches usually describe the chemical systems using atomic representations (Fig. 1.2a) and have shown great advantages to replace the more expensive electronic structure potential energy surfaces [33–36] and to facilitate the molecular dynamics (MD) simulations in a large chemical system more than 100,000 atoms with an accuracy of DFT [33]. However, there are two notable disadvantages of these atomic representations. First, the complication of building an ML model to describe a diverse set of elements and chemistries leads to the rapid growth of features with the increasing number of atom and bond types. In addition, there are significant accuracy losses in the predictions of the untrained types of chemical environments due to the lack of information. Both issues result in the unavoidable need for vast amounts of reference data in training (usually more than 50,000) to achieve the desired accuracy using atomic representations and hinder the degree of chemical transferability of existing ML models to new systems.

The second category of ML methods targets achieving an accuracy at wavefunction level using information from lower-level electronic structure theory, which usually describes the chemical systems using physically informed representations computed from quantum simulations (or known as quantum representations). The quantum information used in ML includes atomic orbitals [52–56], molecular orbitals[57–62], and slate determinants [63] obtained from HF or DFT. To reach the same accuracy, the approaches using quantum representations require much fewer data points (usually less than 5,000) than the ones using atomic representations and can also achieve better model transferability.

One of these approaches is MOB-ML that uses the information of the set of molecular orbitals from HF computations to describe the chemical systems (Fig. 1.2b). MOB-ML uses information from HF to create a simpler and more direct mapping from the input features to the molecular energies. A molecular orbital (MO) describes the spatial distribution of an electron in a molecule and represents the single-electron wavefunctions faithfully in all the system. Thanks to the generality of MO, the MO theory has also become a valuable tool to predict and interpret molecular properties and understand chemical processes in chemistry. We note that the types of MOs, including σ and π MOs, are much fewer than the atomic connectives in chemistry. Therefore, the usage of molecular-orbital-based (MOB) representation in ML improves the transferability of MOB-ML across chemical systems that are not trained in the model with very few training data. MOB-ML has achieved significantly better accuracy with training data smaller than 200 training molecules and slightly better accuracy with 6500 than all other literature results for datasets composed of drug-like organic molecules. MOB-ML is the first generation of ML approach that shows the benefits of incorporating the properties of electronic structure theories (e.g., symmetries and size-consistency) and quantum information into data generation or ML framework to improve the learning efficiency and the transferability for highly accurate predicted molecular energies. As the first ML approach using MO representations, MOB-ML inspires many benchmark ML approaches using quantum information as input, such as NeuralXC [52], DeePHF [53], DeePKS [64], PauliNet [63], and OrbNet [54–56], and provides promising directions for the field of ML for electronic structure.

Overview of theories and applications of MOB-ML in this thesis In this thesis, we will introduce the problem setup, feature designs, specialized ML algorithms, and the applications of the MOB-ML approach for molecular energy learning in various chemical systems. In Chapter 2, we introduce this novel approach of molecular-orbital-based machine learning (MOB-ML) to accurately and transferability predict molecular energies at the levels of MP2, CCSD, or CCSD(T) using MOB features and information from HF . Instead of directly predicting the total energies, MOB-ML learns the pairwise contributions of a wavefunction correlation energy as a function of MOs via Gaussian process regression (GPR) , and thus provides accurate energy predictions at a wavefunction level with a mean-field computation cost. Its accuracy and transferability are illustrated by training and testing on chemical systems of various sizes and chemical nature. Although MOB-ML shows the great potential of being a universal density functional, the cubic scaling of the training cost for GPR becomes a bottleneck to include more training examples for better prediction accuracy. A clustering/regression/classification framework is thus introduced in Chapter 3 to improve the learning efficiency and scale up the training in MOB-ML. The training data are first clustered into subsets by a supervised clustering approach known as regression-clustering, and then independently regressed using GPR or linear regressions. The test data are classified by a classifier and predicted by their corresponding local regressors. The resulting MOB-ML models significantly reduce the training costs by over 4500-fold while preserving prediction accuracy and transferability for the thermalized drug-like organic molecule datasets with different molecular sizes, i.e., QM7b-T and GDB-13-T. The key accuracy loss in this clustering/regression/classification framework is attributed to the classification step, and we also discuss the difficulty of unsupervised clustering due to the design of MOB features.

Husch et al. [60] addresses the issues of original MOB feature design and proposes an improved MOB feature (or size-consistent feature) design by consistently ordering and numerically adjusting the features. The introduction of this improved MOB feature design not only enhances the prediction accuracy and transferability of MOB-ML but also enables the unsupervised clustering on the chemical space. In addition, Sun et al. [62] further applies a scalable GPR with exact GP inference but a lower scaling ($\mathscr{O}(N^2)$), i.e. blackbox matrix-matrix multiplication (BBMM) algorithm, to scale up the Gaussian Process (GP) training of molecular energies. An alternative implementation (AltBBMM) is also proposed to further improve the performance of MOB-ML to predict molecular energies. In Chapter 4, by applying the improved MOB feature design, an enhanced clustering algorithm is proposed to unsupervisedly cluster the chemical space via the Gaussian mixture model (GMM) and then regress molecular energies via alternative blackbox matrix-matrix multiplication (AltBBMM) for MOB-ML and to eliminate the training of an additional classifier. This improved clustering accurately reproduces chemically intuitive groupings of frontier molecular orbitals, and regression on top of the resulting clusters provides the most accurate molecular energy predictions for QM7b-T and GDB13-T compared with other state-of-the-art (SOTA) approaches.

The availability of analytical gradient of MOB-ML in Ref. 61 opens an avenue of applying MOB-ML to provide accurate potential energy surfaces (PESs). We thus explore the efficiency of MOB-ML as PESs in the diffusion Monte Carlo (DMC) simulations for H_2O, CH_5^+ and $C_2H_5^+$ in Chapter 5. The most popular ML-assisted PES approaches usually require over 10,000 to 100,000 high-level reference electronic structure computations (e.g., CCSD(T)/aug-cc-pVTZ level), while MOB-ML only needs under 3000 reference data to achieve the same level of accuracy. To further facilitate the larger DMC simulations and reduce the simulation costs comparable to mechanical force fields, neural networks (NNs) referred as NN $+$ MOB-ML are trained to refit the MOB-ML potential energy surfaces. As a result, the PESs from MOB-ML and NN $+$ MOB-ML achieve comparable accuracy as the standard literature PESs and provide insights of the experimental spectroscopy results for H_2O and CH_5^+.

As an ML approach, it is of great significance for MOB-ML to achieve SOTA results on the benchmark systems and push the accuracy limits higher and higher. More importantly, MOB-ML has become a valuable tool for computational chemists to perform accurate simulations for larger systems without unaffordable costs. This thesis not only shows the power of ML to study molecular properties with a much lower scaling compared with traditional quantum simulations, but also emphasizes the importance of incorporating physical and chemical knowledge, such as consistent ordering and size-consistence, to improve the design of features and ML algorithms. As a universal density matrix functional and general PES generator, it is promising to extend MOB-ML to more complicated systems and theories and apply it to simulate large chemical systems accurately.

Chapter 2

A UNIVERSAL DENSITY MATRIX FUNCTIONAL FROM MOLECULAR ORBITAL-BASED MACHINE LEARNING: TRANSFERABILITY ACROSS ORGANIC MOLECULES.

1. Welborn, M., Cheng, L. & Miller III, T. F. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **14,** 4772–4779 (2018).
2. Cheng, L., Welborn, M., Christensen, A. S. & Miller III, T. F. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *J. Chem. Phys.* **150,** 131103 (2019).

In this chapter, we present a machine learning (ML) method for predicting electronic structure correlation energies using Hartree-Fock input, i.e., molecular orbital based machine learning (MOB-ML). The total correlation energy is expressed in terms of individual and pair contributions from occupied molecular orbitals, and Gaussian process regression is used to predict these pairwise contributions from a feature set that is based on molecular orbital properties, such as Fock, Coulomb, and exchange matrix elements. With the aim of maximizing transferability across chemical systems and compactness of the feature set, we avoid the usual specification of ML features in terms of atom- or geometry-specific information, such atom/element-types, bond-types, or local molecular structure. Refined strategies for feature design and orbital localizations are shown to provide better accuracy. This method maintains accuracy while providing transferability both within and across chemical families; this includes predictions for molecules with atom-types and elements that are not included in the training set. To explore the breadth of chemical diversity that can be described, MOB-ML is also applied to new datasets of thermalized geometries of 7211 organic molecules with up to seven heavy atoms and 1000 organic molecules with up to thirteen heavy atoms. MOB-ML holds promise both in its current form and as a proof-of-principle for the use of ML in the design of generalized density-matrix functionals.

2.1 Introduction

Recent interest in the use of machine learning (ML) for electronic structure has focused on models that are formulated in terms of atom- and geometry-specific features, such as atom-types and bonding connectivities. The advantage of this approach is that it can yield excellent accuracy with computational cost that is comparable to classical force fields.[27, 34, 35, 37–51, 65–67] However, a disadvantage of this approach is that building an ML model to describe a diverse set of elements and chemistries requires training with respect to a number of features that grows quickly with the number of atom- and bond-types, and also requires vast amounts of reference data for the selection and training of those features; these issues have hindered the degree of chemical transferability of existing ML models for electronic structure. For example, previous methods have not demonstrated predictions for molecules with chemical elements that are not included in the training data.

In this work, we focus on the more modest goal of using ML to describe the post-Hartree-Fock correlation energy. Assuming willingness to incur the cost of a Hartree-Fock (HF) self-consistent field (SCF) calculation, we aim to describe the correlation energy associated with MP [5], CC [68], or other post-HF methods. Our approach focuses on training not with respect to atom-based features, but instead using features based on the HF molecular orbitals (MOs), which have no explicit dependence on the underlying atom-types and may thus be expected to provide greater chemical transferability. We then demonstrate the performance of MOB-ML across a broad swath of chemical space, as represented by the QM7b [39] and GDB-13 [69] test sets of organic molecules.

2.2 Theory

The current work aims to predict post-HF correlated wavefunction energies using features of the HF molecular orbitals (MOs). The starting point for the MOB-ML method[57] is that the correlation energy can be decomposed into pairwise occupied MO contributions[7, 70]

$$E_{\mathrm{c}} = \sum_{ij}^{\mathrm{occ}} \varepsilon_{ij}, \tag{2.1}$$

where the pair correlation energy ε_{ij} can be written as a functional of the full set of MOs, $\{\phi_p\}$, appropriately indexed by i and j

$$\varepsilon_{ij} = \varepsilon\left[\{\phi_p\}^{ij}\right]. \tag{2.2}$$

The functional ε is universal across all chemical systems; for a given level of correlated wavefunction theory, there is a corresponding ε that maps the HF MOs to the pair correlation energy, regardless of the molecular composition or geometry. Furthermore, ε simultaneously describes the pair correlation energy for all pairs of occupied MOs (i.e., the functional form of ε does not depend on i and j). For example, the pair correlation energies in MP2 [5] are

$$\varepsilon_{ij}^{\mathrm{MP2}} = \frac{1}{4}\sum_{ab}^{\mathrm{virt}} \frac{|\langle ij||ab\rangle|^2}{e_a + e_b - e_i - e_j}, \tag{2.3}$$

where a and b index virtual MOs, e_p is the HF orbital energy corresponding to MO ϕ_p, and $\langle ij||ab\rangle$ are antisymmetrized electron repulsion integrals.[7] A corresponding expression for the pair correlation energy exists for any post-HF method, but it is typically costly to evaluate in closed form.

In MOB-ML, a machine learning model is constructed for the pair energy functional

$$\varepsilon_{ij} \approx \varepsilon^{\mathrm{ML}}\left[\mathbf{f}_{ij}\right], \tag{2.4}$$

where $\mathbf{f}_{ij}$ denotes a vector of features associated with MOs i and j. Eq. 2.4 thus presents the opportunity for the machine learning of a universal density matrix functional for correlated wavefunction energies, which can be evaluated at the cost of the MO calculation.

Following our previous work [57], the features $\mathbf{f}_{ij}$ correspond to unique elements of the Fock ($\mathbf{F}$), Coulomb ($\mathbf{J}$), and exchange ($\mathbf{K}$) matrices between ϕ_i, ϕ_j, and the set of virtual orbitals. In the current work, we additionally include features associated with matrix elements between pairs of occupied orbitals for which one member of the pair differs from ϕ_i or ϕ_j (i.e., non-i,j occupied MO pairs). The feature vector takes the form

$$\begin{aligned}\mathbf{f}_{ij} =&(F_{ii}, F_{ij}, F_{jj}, \mathbf{F}_i^{\mathrm{o}}, \mathbf{F}_j^{\mathrm{o}}, \mathbf{F}_{ij}^{\mathrm{vv}},\\ &J_{ii}, J_{ij}, J_{jj}, \mathbf{J}_i^{\mathrm{o}}, \mathbf{J}_j^{\mathrm{o}}, \mathbf{J}_i^{\mathrm{v}}, \mathbf{J}_j^{\mathrm{v}}, \mathbf{J}_{ij}^{\mathrm{vv}},\\ &K_{ij}, \mathbf{K}_i^{\mathrm{o}}, \mathbf{K}_j^{\mathrm{o}}, \mathbf{K}_i^{\mathrm{v}}, \mathbf{K}_j^{\mathrm{v}}, \mathbf{K}_{ij}^{\mathrm{vv}}),\end{aligned} \tag{2.5}$$

where for a given matrix ($\mathbf{F}$, $\mathbf{J}$, or $\mathbf{K}$) the superscript o denotes a row of its occupied–occupied block, the superscript v denotes a row of its occupied–virtual block, and the superscript vv denotes its virtual–virtual block. Redundant elements are removed, such that the virtual–virtual block is represented by its upper triangle and the diagonal elements of $\mathbf{K}$ (which are identical to those of $\mathbf{J}$) are omitted. To increase transferability and accuracy, we choose ϕ_i and ϕ_j to be localized molecular orbitals (LMOs) rather than canonical MOs and employ valence virtual LMOs[71] in place of the set of all virtual MOs (as detailed in Ref. 57). We separate Eq. 2.4 to independently machine learn the cases of $i = j$ and $i \neq j$,

$$\varepsilon_{ij} \approx \begin{cases} \varepsilon_{\mathrm{d}}^{\mathrm{ML}}\left[\mathbf{f}_i\right] & \text{if } i = j \\ \varepsilon_{\mathrm{o}}^{\mathrm{ML}}\left[\mathbf{f}_{ij}\right] & \text{if } i \neq j, \end{cases} \tag{2.6}$$

where $\mathbf{f}_i$ denotes $\mathbf{f}_{ii}$ (Eq. 2.5) with redundant elements removed; by separating the pair energies in this way, we avoid the situation where a single ML model is required to distinguish between the cases of $i = j$ and ϕ_i being nearly degenerate to ϕ_j, a distinction which can represent a sharp variation in the function to be learned.

In the current work, several technical refinements are introduced to improve training efficiency (i.e., the accuracy and transferability of the model as a function of the number of training examples). These are now described.

Occupied LMO symmetrization. The feature vector is preprocessed to specify a canonical ordering of the occupied and virtual LMO pairs. This reduces permutation of elements in the feature vector, resulting in greater ML training efficiency. Matrix elements M_{ij} ($\mathbf{M} = \mathbf{F}, \mathbf{J}, \mathbf{K}$) associated with ϕ_i and ϕ_j are rotated into gerade and ungerade combinations

$$\begin{aligned} M_{ii} &\leftarrow \frac{1}{2}M_{ii} + \frac{1}{2}M_{jj} + M_{ij} \\ M_{jj} &\leftarrow \frac{1}{2}M_{ii} + \frac{1}{2}M_{jj} - M_{ij} \\ M_{ij} &\leftarrow \frac{1}{2}M_{ii} - \frac{1}{2}M_{jj} \\ M_{ip} &\leftarrow \frac{1}{\sqrt{2}}M_{ip} + \frac{1}{\sqrt{2}}M_{jp} \\ M_{jp} &\leftarrow \frac{1}{\sqrt{2}}M_{ip} - \frac{1}{\sqrt{2}}M_{jp} \end{aligned} \tag{2.7}$$

with the sign convention that F_{ij} is negative. Here, p indexes any LMO other than i or j (i.e. an occupied LMO k, such that $i \neq k \neq j$, or a valence virtual LMO).

LMO sorting. The virtual LMO pairs are sorted by increasing distance from occupied orbitals ϕ_i and ϕ_j. Sorting in this way ensures that features corresponding to valence virtual LMOs are listed in decreasing order of heuristic importance, and that the mapping between valence virtual LMOs and their associated features is roughly preserved. We recognize this issue could also potentially be addressed through the use of symmetry functions,[72] but these are not employed in the current work.

For purposes of sorting, distance is defined as

$$R_a^{ij} = \left\| \langle \phi_i | \hat{R} | \phi_i \rangle - \langle \phi_a | \hat{R} | \phi_a \rangle \right\| + \left\| \langle \phi_j | \hat{R} | \phi_j \rangle - \langle \phi_a | \hat{R} | \phi_a \rangle \right\|, \qquad (2.8)$$

where ϕ_a is a virtual LMO, $\hat{R}$ is the Cartesian position operator, and $||.||$ denotes the 2-norm. $\left\| \langle \phi_i | \hat{R} | \phi_i \rangle - \langle \phi_a | \hat{R} | \phi_a \rangle \right\|$ represents the Euclidean distance between the centroids of orbital i and orbital a. Previously,[57] distances were defined based on Coulomb repulsion, which was found to sometimes lead to inconsistent sorting in systems with strongly polarized bonds. The non-i,j occupied LMO pairs are sorted in the same manner as the virtual LMO pairs.

Orbital localization. We employ Boys localization[73] to obtain the occupied LMOs, rather than intrinsic bond orbital (IBO) localization[71] which was employed in our previous work.[57] Particularly for molecules that include triple bonds or multiple lone pairs, it is found that Boys localization provides more consistent localization as a function of small geometry changes than IBO localization; and the chemically unintuitive mixing of σ and π bonds in Boys localization ("banana bonds")[74] does not present a problem for the MOB-ML method.

Feature selection. Prior to training, automatic feature selection is performed using random forest regression [75] with the mean decrease of accuracy criterion (sometimes called permutation importance).[76] This technique was found to be more effective than our previous use[57] of the Gini importance score[75] which led to worse accuracy and failed to select any features for the case of methane.

The reason for using feature selection in this way is twofold. First, GPR performance is known to degrade for high-dimensional datasets (in practice 50-100 features);[77] and second, the use of the full feature set with small molecules can lead to overfitting as features can become correlated.

2.3 Computational details

Results are presented for a single water molecule; a series of alkane molecules; a thermalized version of the QM7b set of 7211 molecules with up to seven C, O, N, S, and Cl heavy atoms; and a thermalized version of the GDB-13 set of molecules with thirteen C, O, N, S, and Cl heavy atoms. All datasets employed in this work are provided in Ref. 78.

Training and test geometries are sampled at 50 fs intervals from *ab initio* molecular dynamics trajectories performed with the Q-CHEM 5.0 software package,[79] using the B3LYP[80–83]/6-31g*[84] level of theory and a Langevin thermostat[85] at 350 K.

The features and training pair energies associated with these geometries are computed using the MOLPRO 2018.0 software package[86] in a cc-pVTZ basis set unless otherwise noted.[87] Valence virtual orbitals used in feature construction are determined with the Intrinsic Bond Orbital method.[71] Reference pair correlation energies are computed with MP2[5, 88] CCSD[68, 89] as well as CCSD(T).[8, 90] Density fitting (DF) for both Coulomb and exchange integrals [91] is employed for all results below except those corresponding to the water molecule. The frozen core approximation is used in all cases.

Gaussian process regression (GPR)[92] is employed to machine learn $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$ (Eq. 2.6) using the GPY 1.9.6 software package. [93] The GPR kernel is Matérn 5/2 with white noise regularization[92]. Kernel hyperparameters are optimized with respect to the log marginal likelihood objective for the water and alkane series results, as well as for $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ of the QM7b results. We use the Matérn 3/2 kernel instead of the Matérn 5/2 kernel for the case of $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$ for QM7b results, as it was empirically found to yield slightly better accuracy. [1] Feature selection is performed using the random forest regression implementation in the SCIKIT-LEARN v0.20.0 package. [94]

2.4 Results

The ML model of Eq. 2.6 is a universal functional for any molecular Hamiltonian. In principle, with an adequate feature list and unlimited training data (and time), it should accurately and simultaneously describe all molecular systems. In practice, we must train the ML model using a truncated feature list and finite data. These choices determine the accuracy of the model.

Below, we examine the performance of the MOB-ML method in three increasingly broad regions of chemical space: (i) training on randomized water molecule geometries and predicting the energies of other water molecule geometries; (ii) training on geometries of short alkanes and predicting the energies of longer alkanes; and (iii) training on a small set of organic molecules and predicting the energies of a broader set of organic molecules. In all cases, we report the ML prediction accuracy as a function of the number of training examples and all the results are trained using

[1] In principle, the smoothness of the Matérn kernel could be taken as a kernel hyperparameter; however, this possibility was not explored in this work.

different selected feature sets obtained from feature selections.

Transferability across different geometries

As a first example, we consider the performance of MOB-ML for a single water molecule. A separate model is trained to predict the correlation energy at the MP2, CCSD, and CCSD(T) levels of theory, using reference calculations on a subset of 1000 randomized water geometries to predict the correlation energy for the remainder. Feature selection with an importance threshold of 1×10^{-3} results in 12, 11, and 10 features for $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$ for MP2, CCSD, and CCSD(T), respectively; ten features are selected for $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ for all three post-HF methods.

Figure 2.1 presents the test set prediction accuracy of each MOB-ML model as a function of the number of training geometries (i.e., the "learning curve"). MOB-ML predictions are shown for MP2, CCSD, and CCSD(T), and the model shows the same level of accuracy for all three methods. Remarkably, all three models achieve a prediction mean absolute error (MAE) of 1 mH when trained on only a single water geometry, indicating that only a single reference calculation is needed to provide chemical accuracy for the remaining 999 geometries at each level of theory. Since it contains 10 distinct LMO pairs, this single geometry provides enough information to yield a chemically accurate MOB-ML model for the global thermally accessible potential energy surface.

For all three methods (Fig. 2.1), the learning curve exhibits the expected[95] power-law behavior as a function of training data, and the total error reaches microhartree accuracy with tens of water training geometries. As compared to our previous results, where training on 200 geometries resulted in a prediction MAE of 0.027 mH for the case of CCSD,[57] the current implementation of the MOB-ML model is substantially improved; the improvement for this case stems primarily from the use of Boys localization,[73] which specifies unique and consistent LMOs corresponding to the oxygen lone pairs.

To further illustrate the excellent performance of MOB-ML in predicting different geometries in single molecule, a set of small molecules are tested with all three theory levels at cc-pVTZ basis sets if not specified. Table 2.1[2] summarizes the corresponding results for these small molecules, with $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$ trained on randomly selected 10 geometries and used to predict correlation energy for other 90

[2] Results listed here are regenerated using the feature design introduced in this chapter to make the content consistent, and are different from the ones shown in Ref. 57

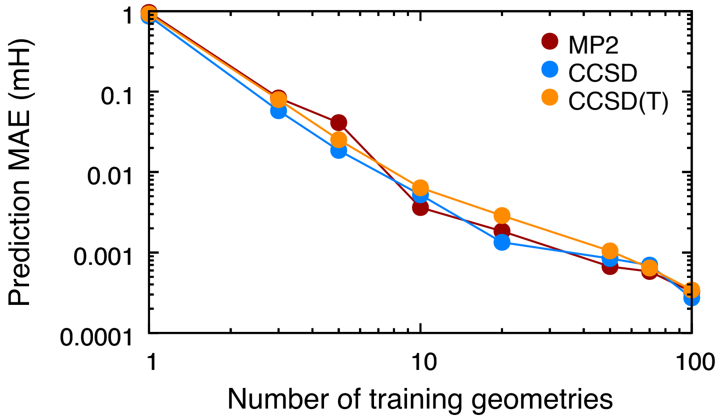

Figure 2.1: Learning curves for MOB-ML models trained on the water molecule and used to predict the correlation energy of different water molecule geometries at three levels of post-Hartree-Fock theory. Prediction errors are summarized in terms of mean absolute error (MAE).

geometries. The MAEs are smaller than 1 milliHartree for all the small molecules at all three theory levels, and the relative error (Rel. Error) is also very small (ranging from 0.00003% to 0.1717%). As pointed in Fig. 2.1 and shown in Table 2.1, MOB-ML is insensitive to the choice of theory level. In addition, the water results for basis sets of double-zeta and triple-zeta suggest that the choice of basis set will not affect the MOB-ML performance as well.

Transferability across within chemical families

Next, we explore the transferability of MOB-ML predictions for a model that is trained on thermalized geometries of short alkanes and then used for predictions on thermalized geometries of larger and more branched alkanes (n-butane and isobutane). For these predictions, the absolute zero of energy is shifted for each molecule to compare relative energies on its potential energy surface (i.e., parallelity errors are removed). These shifts are reported in the figure caption; for no other results reported in the paper are parallelity errors removed.

In our previous work,[57] this test was performed using training data that combined of 100 geometries of methane, 300 of ethane, and 50 or propane; the resulting predictions are reproduced here in Fig. 2.2a. This earlier implementation of MOB-ML led to predictions for n-butane and isobutane with substantial errors (0.59 mH for n-butane and 0.93 mH for isobutane) and noticable skew with respect to the true correlation energy.

Table 2.1: MOB-ML predictions of MP2, CCSD and CCSD(T) correlation energies for a collection of small molecules with 10 training and 90 test geometries.

	MAE (milliHartree)			Rel. Error (%)		
Molecule	MP2	CCSD	CCSD(T)	MP2	CCSD	CCSD(T)
H_2	1.00e-5	1.05e-5	1.05e-5	0.00003	0.00003	0.00003
N_2	9.53e-3	4.00e-3	4.92e-3	0.0023	0.0010	0.0012
F_2	2.13e-2	4.02e-2	1.27e-2	0.0041	0.0076	0.0023
HF	8.73e-5	1.67e-4	1.53e-4	0.00003	0.00006	0.0001
NH_3	1.43e-2	1.33e-2	1.24e-2	0.0062	0.0053	0.0048
CH_4	2.23e-2	1.35e-2	1.72e-2	0.0112	0.0062	0.0076
CO	1.46e-4	1.46e-4	3.96e-4	0.00004	0.00004	0.0001
CO_2	1.20e-3	1.46e-3	1.75e-3	0.0005	0.0005	0.0006
HCN	1.46e-2	1.05e-2	1.17e-2	0.0042	0.0030	0.0032
HNC	2.75e-2	3.52e-2	3.77e-2	0.0082	0.0103	0.0105
C_2H_2	1.02e-1	1.03e-1	1.11e-1	0.0327	0.0319	0.0327
C_2H_4	4.08e-1	6.20e-1	5.34e-1	0.1212	0.1717	0.1421
C_2H_6[†]	1.42e-1	1.41e-1	1.48e-1	0.0381	0.0351	0.0356
CH_2O	1.75e-2	2.59e-2	5.01e-2	0.0044	0.0064	0.0119
HCO_2H	4.34e-1	3.22e-1	4.56e-1	0.0686	0.0505	0.0689
CH_3OH	2.98e-1	2.11e-1	3.70e-1	0.0693	0.0472	0.0801
CH_2F_2	7.87e-1	5.60e-1	7.68e-1	0.1156	0.0811	0.1078
H_2O[† ‡]						
cc-pVDZ	3.38e-3	2.65e-3	2.61e-3	0.0017	0.0012	0.0012
cc-pVTZ	3.62e-3	5.23e-3	6.37e-3	0.0014	0.0019	0.0023

[†] Training and test sets contain 50 and 950 geometries, respectively.

[‡] Results for two basis sets.

The predictions of MOB-ML in the current work (Fig. 2.2b) are markedly improved. First, the overall prediction accuracy is improved for all four summary statistics (inset in Fig. 2.2) despite substantial reduction in the number of training examples used. (The current work uses only 50 geometries of ethane, 20 geometries of propane, and no methane data.) Second, n-butane and isobutane are predicted with nearly identical accuracy. Finally, the prediction errors are no longer skewed as a function of true correlation energy. The primary methodological sources of these improvements are found to be symmetrization of occupied orbitals (Eq. 2.7) and the improved feature selection methodology. The MOB-ML features in the current work are selected with an importance threshold of 1×10^{-4}, resulting in 27 features for $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and 12 features for $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$; results presented in Fig 2.2b for CCSD(T) are qualitatively identical to those obtained for CCSD (not shown).

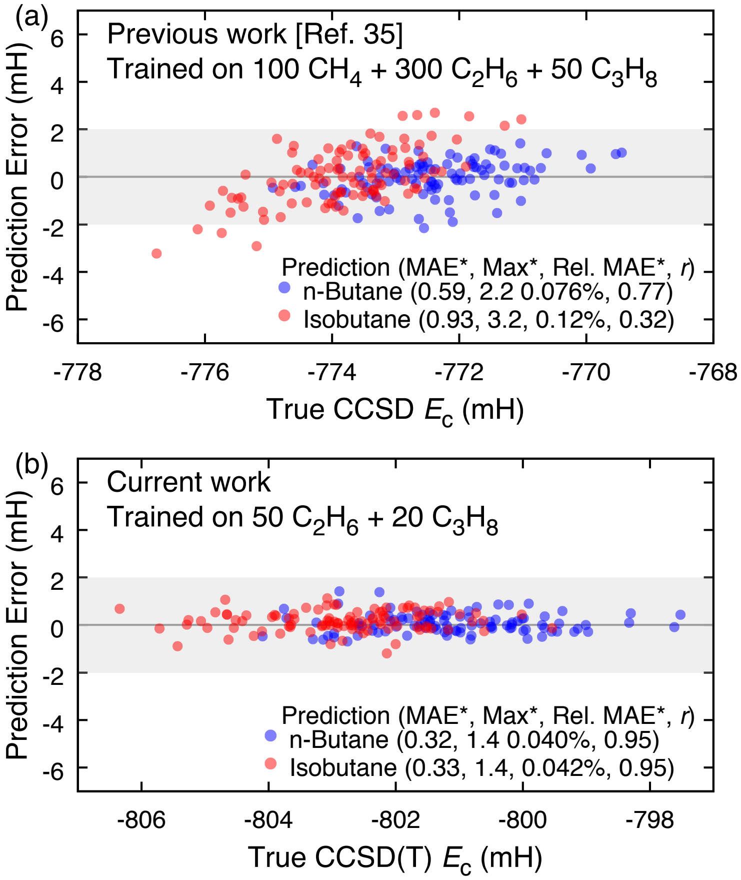

Figure 2.2: MOB-ML predictions of the correlation energy for 100 n-butane and isobutane geometries, using MOB-ML features described in the (b) current work, compared to (a) the previous MOB-ML features of Ref. 57. Training sets are indicated in each panel of the figure. MOB-ML prediction errors are plotted versus the (a) true CCSD and (b) true CCSD(T) correlation energy. To remove parallelity error, a global shift is applied to the predictions of n-butane and isobutane by (a) 3.3 and 0.73 mH and (b) 0.90 and 0.17 mH, respectively. Summary statistics that include this shift (indicated by an asterisk) are presented, consisting of mean absolute error (MAE*), maximum absolute error (Max*), MAE* as a percentage of E_c (Rel. MAE*), and Pearson correlation coefficient (r)[96]. The gray shaded region corresponds to errors of ±2 mH.

Transferability across drug-like molecules

We now examine the transferability of the MOB-ML method across a broad swath of chemical space. Specifically, we consider the QM7b dataset,[97] which is comprised of 7,211 plausible organic molecules with 7 or fewer heavy atoms. The chemical elements in QM7b are limited to those likely to be found in drug-like compounds: C, H, O, N, S, and Cl. We refer to the dataset used herein as QM7b-T to reflect the fact that it contains geometries sampled at a temperature of 350 K (as described in Sec. 2.3), as opposed to DFT-optimized geometries. The MOB-ML model is trained on a randomly chosen subset of QM7b-T molecules and used to predict the correlation energy of the remainder. Active learning was also tested as a training data selection strategy, but was not found to improve the predictions in the regime of chemical accuracy, and in fact led to slightly worse transferability.

For comparison, a Δ-ML model[42] was trained on the same molecules using kernel-ridge regression using the FCHL representation[98] with a Gaussian kernel function (FCHL/Δ-ML), as implemented in the QML package.[99] All hyperparameters of the model were set to those obtained in Ref. 98, which have previously been demonstrated to work well for datasets containing structures similar to those in QM7b-T.[99]

A possible source of concern for MOB-ML is that the number of selected features would grow with the chemical complexity of the training data. For example, 27 features for $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and 12 features for $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$ were selected in the alkane test case using ethane + propane training data (Fig. 2.2b), whereas only 10 features for $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and 10 features for $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$ were selected for the water test case at the CCSD(T) level of theory (Fig. 2.1). To examine this, we perform feature selection on increasing numbers of randomly selected molecules from the QM7b-T dataset. Table 2.2 presents two statistics on the feature importance as a function of the number of training molecules: (i) the number of "important features" (i.e., those whose permutation importance[76] exceeds a set threshold of 2×10^{-4} and 5×10^{-5} for $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$, respectively) and (ii) the inverse participation ratio[100] of the feature importance scores. The latter is a threshold-less measure of the number of important features; it takes a value of 1 when only 1 feature has nonzero importance and N when all N features have equal importance. Although the QM7b-T dataset contains many different chemical elements and bonding motifs, Table 2.2 reveals that the selected features remain compact and do not grow with the number of training molecules. Indeed, for a large number of training molecules, the number of selected features slightly decreases, reaching 42 and 24 selected features for $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$, respectively, for the largest training sizes considered.

Table 2.2: Number of features selected as a function of the number randomly chosen training molecules for the QM7b-T dataset at the CCSD(T)/cc-pVDZ level. The number of features that exceed an importance threshold as well as the inverse participation ratio (IPR) of the feature importance scores are reported (see text).

	# of important features		feature weight IPR	
Training size	$\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$	$\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$	$\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$	$\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$
20	50	28	4.720	1.116
50	46	28	3.718	1.097
100	46	26	3.450	1.115
200	42	24	3.430	1.120

The learning curves for MOB-ML models trained at MP2/cc-pVTZ and CCSD(T) /cc-pVDZ levels of theory are shown in Fig. 2.3a, as well as the FCHL/Δ-ML learning curve for MP2/cc-pVTZ. At the MP2 level of theory, the MOB-ML model achieves an accuracy of 2 mH with 110 training calculations (representing 1.5% of the molecules in the QM7b-T dataset), whereas the FCHL/Δ-ML requires over 300 training geometries to reach the same accuracy threshold. Fig. 2.3a also illustrates the relative insensitivity of MOB-ML to the level of electronic structure theory, with the learning curve for CCSD(T)/cc-pVDZ reaching 2 mH accuracy with 140 training calculations. An analysis of the sensitivity of the MOB-ML predictions to the number of selected features is presented in **Appendix** Fig. 2.4, which indicates that the reported results are robust with respect to the number of selected features.

As a final test of transferability of the MOB-ML and FCHL/Δ-ML methods across chemical space, Figs. 2.3b and 2.3c show results in which the ML methods are trained on QM7b-T molecules and then used to predict results for a dataset of 13-heavy-atom organic molecules at thermalized geometries, GDB-13-T, which includes six thermally sampled geometries each of 1,000 13-heavy-atom organic molecules chosen randomly from the GDB-13 dataset.[69] Like QM7b, the members of GDB-13 contain C, H, N, O, S, and Cl. The size of these molecules precludes the use of CC to generate reference data; we therefore make comparison at the MP2/cc-pVTZ level of theory, noting that MOB-ML has consistently been shown to be insensitive to the employed post-Hartree–Fock method (as in Fig. 2.3a). Transfer learning results as a function of the number of training molecules are presented in Figs. 2.3b (on a linear-linear scale) and 2.3c (on a log-log scale).

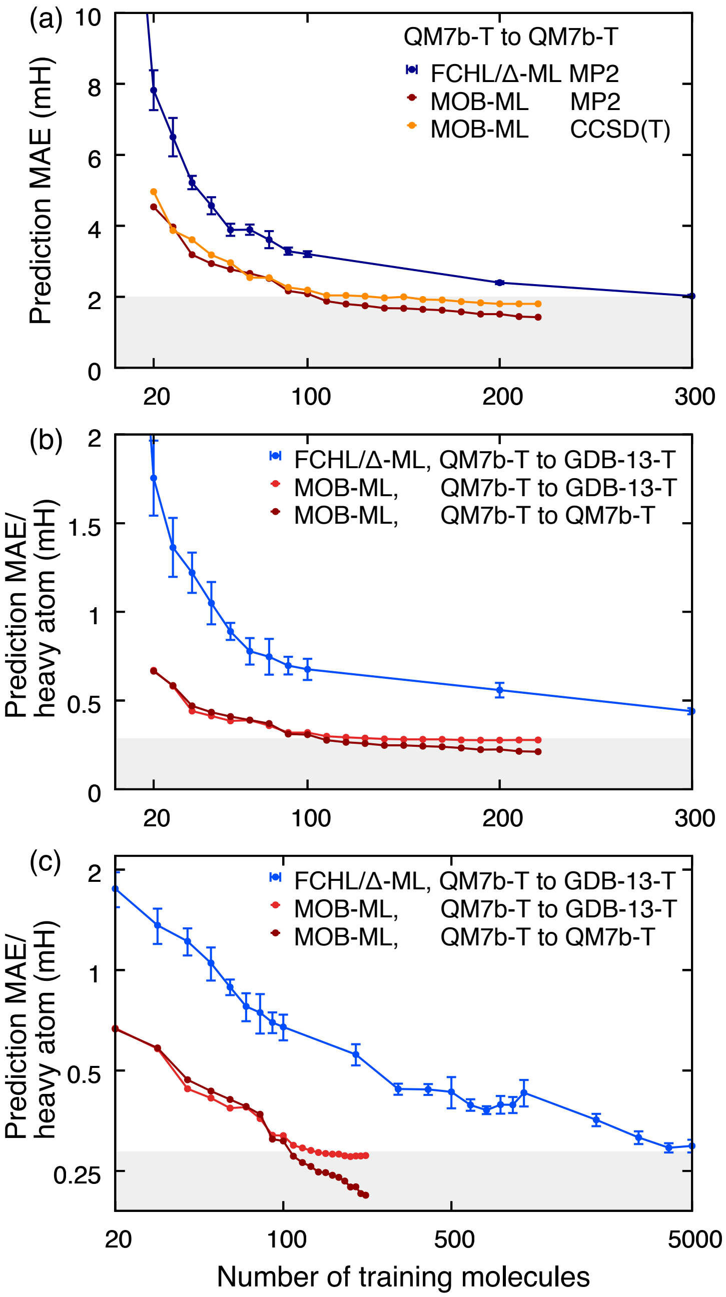

Figure 2.3: Learning curves for MOB-ML trained on QM7b-T and applied to QM7b-T and GDB-13-T (see text for definition of these datasets). FCHL/Δ-ML [98] results are provided for comparison. (a) Predictions are made for QM7b-T at the MP2/cc-pVTZ (red) and CCSD(T)/cc-pVDZ (orange) levels of theory. (b) Using the same models trained on QM7b-T, predictions are made for GDB-13-T, and reported in terms of MAE per heavy atom. (MOB-ML predictions for QM7b-T are included for reference.) (c) As in the previous panel, but plotted on a logarithmic scale and extended to show the full range of FCHL/Δ-ML predictions. Error bars for FCHL/Δ-ML represent prediction standard errors of the mean as measured over 10 models. The gray shaded area corresponds to errors of 2 mH per 7 heavy atoms.

Using the MOB-ML model that is trained on 110 seven-heavy-atom molecules (corresponding to a prediction MAE of 1.89 mH for QM7b-T), we observe a prediction MAE of 3.88 mH for GDB-13-T. Expressed in terms of size-intensive quantities, the prediction MAE per heavy atom is 0.277 mH and 0.298 mH for QM7b-T and GDB-13-T, respectively, indicating that the accuracy of the MOB-ML results are only slightly worse when the model is transferred to the dataset of larger molecules. On a per-heavy-atom basis, MOB-ML reaches chemical accuracy with the same number of QM7b-T training calculations (approximately 100), regardless of whether it is tested on QM7b-T or GDB-13-T.

In contrast with MOB-ML, the FCHL/Δ-ML method is found to be significantly less transferable from QM7b-T to GDB-13-T. For models trained using 100 seven-heavy-atom molecules, the MAE per heavy atom of FCHL/Δ-ML is over twice that of MOB-ML (Fig. 2.3b). Moreover, whereas MOB-ML reaches the per-heavy-atom chemical accuracy threshold with 140 training calculations, the FCHL/Δ-ML method only reaches that threshold with 5000 training calculations.

2.5 Conclusions

Molecular-orbital-based machine learning (MOB-ML) has been shown to be a simple and strikingly accurate strategy for predicting correlated wavefunction energies at the cost of a Hartree-Fock calculation, benefiting from the intrinsic transferability of the localized molecular orbital representation. The starting point for the MOB-ML method is a rigorous mapping from the Hartree-Fock molecular orbitals to the total correlation energy, which ensures that the use of sufficient training data and molecular orbital features will produce a model that matches the corresponding correlated wavefunction method across the entirety of chemical space. The current work explores this possibility within the subspace of organic molecules. It is shown that MOB-ML predicts energies of the QM7b-T dataset to within a 2 mH accuracy using only 110 training calculations at the MP2/cc-pVTZ level of theory and using 140 training calculations at the CCSD(T)/cc-pVDZ level of theory. Direct comparison with FCHL/Δ-ML reveals that MOB-ML is threefold more efficient in reaching chemical accuracy for describing QM7b-T. Further, a transferability test of a MOB-ML model trained on QM7b-T to GDB-13-T reveals that MOB-ML exhibits negligible degradation in accuracy; as a result, chemical accuracy is achieved with 36-times fewer training calculations using MOB-ML versus FCHL/Δ-ML. With the similar level of costs, MOB-ML provides much more reliable atomization energies compared with several commonly used DFTs using CCSD/cc-pVTZ as a reference theory. These results suggest that MOB-ML provides a promising approach toward the development of density matrix functionals that are applicable across broad swathes of chemical space.

2.6 Appendix

The datasets used in this work are available for download[78] and they include MOB-ML features, HF energies, pair correlation energies, and geometries. MOB-ML and FCHL/Δ-ML predictions corresponding to Fig. 2.3 are shown in Table 2.3 and 2.4. An analysis of the sensitivity of the results of Fig. 2.3 to the number of selected features are available in Fig. 2.4.

Table 2.3: MAE and MAE/heavy atom (MAE/HA) of MOB-ML on predicting QM7b-T and GDB-13-T using a model trained on QM7b-T (energies in mH).

	QM7b-T				GDB-13-T	
Training	MP2/cc-pVTZ		CCSD(T)/cc-pVDZ		MP2/cc-pVTZ	
size	MAE	MAE/HA	MAE	MAE/HA	MAE	MAE/HA
20	4.536	0.6664	4.962	0.7314	8.711	0.6701
30	3.966	0.5844	3.865	0.5690	7.554	0.5811
40	3.183	0.4696	3.605	0.5309	5.731	0.4408
50	2.938	0.4338	3.180	0.4678	5.375	0.4135
60	2.774	0.4094	2.960	0.4371	5.020	0.3862
70	2.660	0.3906	2.540	0.3751	5.055	0.3888
80	2.519	0.3701	2.538	0.3755	4.669	0.3591
90	2.165	0.3116	2.266	0.3354	4.161	0.3201
100	2.085	0.3076	2.187	0.3235	4.150	0.3192
110	1.878	0.2768	2.037	0.3017	3.880	0.2985
120	1.797	0.2650	2.040	0.3023	3.809	0.2930
130	1.747	0.2582	2.013	0.2987	3.746	0.2882
140	1.681	0.2484	1.967	0.2921	3.692	0.2840
150	1.674	0.2475	1.998	0.2962	3.665	0.2820
160	1.645	0.2429	1.921	0.2855	3.654	0.2810
170	1.620	0.2394	1.911	0.2834	3.652	0.2809
180	1.577	0.2333	1.865	0.2778	3.611	0.2778
190	1.511	0.2240	1.827	0.2728	3.592	0.2763
200	1.511	0.2244	1.802	0.2696	3.605	0.2773
210	1.443	0.2140	1.801	0.2696	3.607	0.2774
220	1.427	0.2115	1.802	0.2698	3.617	0.2782

Table 2.4: MAE of FCHL/Δ-ML on predicting QM7b-T and GDB-13-T using a model trained on QM7b-T (energies in mH). The standard error of the mean (SEM) over 10 trials is also reported.

	QM7b-T, MP2/cc-pVTZ		GDB-13-T, MP2/cc-pVTZ			
Training	MAE	SEM	MAE	SEM	MAE	SEM
size					/HA	/HA
1	227.7	16.94	444.4	44.37	34.18	3.413
2	120.5	16.38	212.3	35.36	16.33	2.720
3	94.65	24.05	169.4	32.20	13.03	2.477
4	51.88	9.660	115.1	20.51	8.857	1.578
5	34.99	4.574	78.56	11.20	6.043	0.8618
6	20.37	1.943	56.29	5.873	4.330	0.4518
7	23.07	3.799	51.16	8.810	3.935	0.6777
8	19.04	1.639	42.21	5.878	3.247	0.4521
9	19.23	1.975	43.06	8.492	3.313	0.6532
10	14.22	1.671	43.05	6.783	3.312	0.5217
20	7.823	0.5624	22.80	2.744	1.754	0.2111
30	6.501	0.5400	17.72	2.161	1.363	0.1663
40	5.219	0.1874	15.87	1.477	1.221	0.1136
50	4.567	0.2395	13.64	1.549	1.049	0.1192
60	3.887	0.1713	11.57	0.6267	0.8897	0.04821
70	3.889	0.1453	10.11	0.9725	0.7780	0.07480
80	3.608	0.2412	9.704	1.311	0.7465	0.1008
90	3.283	0.1016	9.062	0.6463	0.6971	0.04971
100	3.205	0.08087	8.787	0.7807	0.6759	0.06006
200	2.396	0.03973	7.265	0.5289	0.5588	0.04068
300	2.022	0.03468	5.722	0.2212	0.4401	0.01701
400	1.870	0.01906	5.706	0.2140	0.4389	0.01646
500	1.760	0.02530	5.615	0.6035	0.4319	0.04642
600	1.648	0.01538	5.128	0.2007	0.3945	0.01544
700	1.581	0.02471	4.946	0.1344	0.3805	0.01034
800	1.503	0.02184	5.140	0.3127	0.3954	0.02405
900	1.445	0.01963	5.134	0.2843	0.3949	0.02187
1000	1.408	0.02135	5.584	0.5120	0.4295	0.03938
2000	1.135	0.01120	4.626	0.1944	0.3559	0.01495
3000	0.9837	0.003951	4.094	0.1812	0.3149	0.01394
4000	0.8995	0.006155	3.816	0.1211	0.2935	0.00931
5000	0.8618	0.005251	3.865	0.1691	0.2973	0.01301

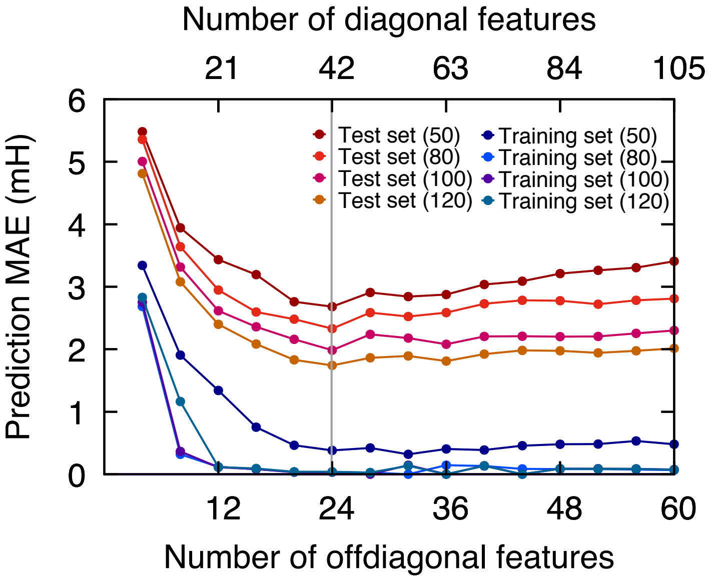

Figure 2.4: Prediction MAE for MOB-ML models trained on the QM7b-T dataset as a function of the number of MOB-ML features selected. Predictions are made for the training set and for a test set comprised of the remainder of QM7b-T, with the number of molecules included in the training set indicated in parentheses. Features are included in order of decreasing RFR-MDA importance. The gray line indicates the number of features employed for training on the QM7b-T dataset in the main text (Fig. 2.3); here, the ratio of the number of diagonal features to off-diagonal features is fixed at 42:24. Regardless of whether the MOB-ML models are trained using either 50, 80, 100 and 120 molecules, the accuracy of the test-set prediction is relatively insensitive to the number of selected MOB-ML features.

Chapter 3

REGRESSION-CLUSTERING FOR IMPROVED ACCURACY AND TRAINING COST WITH MOLECULAR-ORBITAL-BASED MACHINE LEARNING

1. Cheng, L., Kovachki, N. B., Welborn, M. & Miller III, T. F. Regression clustering for improved accuracy and training costs with molecular-orbital-based machine learning. *J. Chem. Theory Comput.* **15,** 6668–6677 (2019).

In previous chapter, MOB-ML employed Gaussian Process Regression (GPR) has been show to provides good prediction accuracy with small training sets; however, the cost of GPR training scales cubically with the amount of data and becomes a computational bottleneck for large training sets. In this chapter, we address this problem by introducing a clustering/regression/classification implementation of MOB-ML. By independently regressing these subsets of the data, we obtain MOB-ML models with greatly reduced training costs while preserving prediction accuracy and transferability. For a dataset of thermalized (350 K) geometries of 7211 organic molecules of up to seven heavy atoms (QM7b-T), the chemical accuracy (1 kcal/mol prediction error) can be reached with only 300 training molecules, while providing 35000-fold and 4500-fold reductions in the wall-clock training time, respectively, compared to MOB-ML without clustering. The resulting models are also demonstrated to retain transferability for the prediction of large-molecule energies with only small-molecule training data. Finally, it is shown that capping the number of training datapoints per cluster leads to further improvements in prediction accuracy with negligible increases in wall-clock training time.

3.1 Introduction

We has introduced a rigorous factorization of the post-HF correlation energy into contributions from pairs of occupied molecular orbitals and showed that these pair contributions could be compactly represented in the space of molecular-orbital-based (MOB) features to allow for straightforward ML regression.[57, 58] This MOB-ML method was demonstrated to accurately predict MP2[5, 88] and CCSD(T)[8, 90] energies of different benchmark systems, including the QM7b-T and GDB-13-T datasets of thermalized drug-like organic molecules. While providing good accuracy with a modest amount of training data, the accuracy of MOB-ML in these initial studies was limited by the high computational cost ($\mathcal{O}(N^3)$) of applying Gaussian Process Regression (GPR) to the full set of training data.[58]

In this chapter, we combine MOB-ML with regression clustering (RC) to overcome this bottleneck in computational cost and accuracy. The training data are clustered via RC to discover locally linear structures. In a first step, regression clustering (RC) is used to partition the training data to best fit an ensemble of linear regression (LR) models; in a second step, each cluster is regressed independently, using either LR or GPR; and in a third step, a random forest classifier (RFC) is trained for the prediction of cluster assignments based on MOB feature values.

Upon inspection, RC is found to recapitulate chemically intuitive groupings of the frontier molecular orbitals, and the combined RC/LR/RFC and RC/GPR/RFC implementations of MOB-ML are found to provide good prediction accuracy with greatly reduced wall-clock training times. The resulting models are also demonstrated to retain transferability for the prediction of large-molecule energies with only small-molecule training data. It is also shown that capping the number of training datapoints per cluster leads to further improvements in prediction accuracy with negligible increases in wall-clock training time.

3.2 Theory

Local linearity of MOB feature space

It has been previously emphasized that MOB-ML facilitates transferability across chemical systems, even allowing for predictions involving molecules with elements that do not appear in the training set,[57] due to the fact that MOB features provide a compact and highly abstracted representation of the electronic structure. However, it is worth additionally emphasizing that this transferability benefits from the smooth variation and local linearity of the pair correlation energies as a function of MOB feature values associated with different molecular geometries and even different molecules.

Figure 3.1 illustrates these latter properties for a σ-bonding orbital in a series of simple molecules. On the y-axis, we plot the diagonal contribution to the correlation energy associated with this orbital (ε_{ii}), computed at the MP2/cc-pvTZ level of theory. On the x-axis, we plot the value of a particular MOB feature, the Fock matrix element for the that localized orbital, F_{ii}. For each molecule, a range of geometries is sampled from the Boltzmann distribution at 350 K, with each plotted point corresponding to a different sampled geometry.

It is immediately clear from the figure that the pair correlation energy varies smoothly and linearly as a function of the MOB feature value. Moreover, the slope of the linear curve is remarkably consistent across molecules. This illustration suggests that MOB features may lead to accurate regression of correlation energies using simple ML models (even linear models), and it also indicates the basis for the robust transferability of MOB-ML across diverse chemical systems, including those with elements that do not appear in the training set.

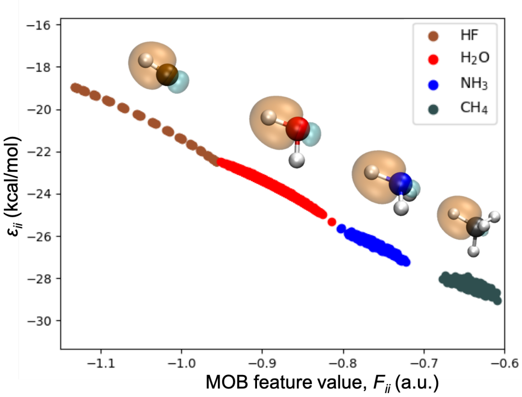

Figure 3.1: The diagonal pair correlation energy (ε_{ii}) for a localized σ-bond in four different molecules at thermally sampled geometries (at 350 K), computed at the MP2/cc-pvTZ level of theory. The diagonal pair correlation energies for HF, NH_3, and CH_4 are shifted vertically downward relative to those of HF by 3.407, 6.289, and 7.772 kcal/mol for H_2O, NH_3, and CH_4. Illustrative σ-bond LMOs are shown for each molecule.

Regression clustering with a greedy algorithm

To take advantage of the local linearity of pair correlation energies as a function of MOB features, we propose a strategy to discover optimally linear clusters using regression clustering (RC).[101] Consider the set of M datapoints $\{\mathbf{f}_t, \varepsilon_t\} \subset \mathbb{R}^d \times \mathbb{R}$, where d is the length of the MOB feature vector and where each datapoint is indexed by t and corresponds to a MOB feature vector and the associated reference value (i.e., label) for the pair correlation energy. To separate these datapoints into locally linear clusters, $(S_1, ..., S_N)$, we seek a solution to the optimization problem

$$\min_{S_1,...,S_N} \sum_{k=1}^{N} \sum_{t \in S_k} |A(S_k) \cdot \mathbf{f}_t + b(S_k) - \varepsilon_t|^2, \tag{3.1}$$

where $(A(S_k) \in \mathbb{R}^d)$ and $(b(S_k) \in \mathbb{R})$ are obtained via ordinary least squares (OLS) solution,

$$\begin{bmatrix} \mathbf{f}_{t_1}^T & 1 \\ \vdots & \vdots \\ \mathbf{f}_{t_{|S_k|}}^T & 1 \end{bmatrix} \begin{bmatrix} A(S_k) \\ b(S_k) \end{bmatrix} = \begin{bmatrix} \varepsilon_{t_1} \\ \vdots \\ \varepsilon_{t_{|S_k|}} \end{bmatrix}. \tag{3.2}$$

Each resulting S_k is the set of indices t assigned to cluster k composed of $|S_k|$ datapoints. To perform the optimization in Eq. 3.1, we employ a modified version of the greedy algorithm proposed in Ref. 102 (Algorithm 1). In general, solutions to Eq. 3.1 may overlap, such that $S_k \cap S_l \neq \emptyset$ for $k \neq l$; however, the proposed algorithm enforces that clusters remain pairwise-disjoint.

Algorithm 1 Greedy algorithm for the solution of Eq. 3.1.

```
Input: Initial clusters: $S_1, \ldots, S_N$
Output: Data clusters $S_1, \ldots, S_N$
 1: for $k \leftarrow 1$ to $N$ do
 2:     $A(S_k), b(S_k) \leftarrow$ OLS solution of Eq. 3.2
 3: end for
 4: while not converged do
 5:     for $k \leftarrow 1$ to $N$ do
 6:         $S_k \leftarrow \{t \in \{1, \ldots, M\} : \arg\min_{n \in \{1,...,N\}} |A(S_n) \cdot \mathbf{f}_t + b(S_n) - \varepsilon_t|^2 = k\}$
 7:     end for
 8:     for $k \leftarrow 1$ to $N$ do
 9:         $A(S_k), b(S_k) \leftarrow$ OLS solution of Eq. 3.2
10:     end for
11: end while
```

Algorithm 1 has a per-iteration runtime of $\mathcal{O}(Md^2)$, since we compute N OLS solutions each with runtime $\mathcal{O}(|S_k|d^2)$ and since $\sum_{k=1}^{N} |S_k| = M$. However, the algorithm can be trivially parallelized to reach a runtime of $\mathcal{O}(\max(|S_k|)d^2)$. A key operational step in this algorithm is line 6, which can be explained in simple terms as follows: we assign each datapoint, indexed by t, to the cluster to which it is closest, as measured by the squared linear regression distance metric,

$$|D_{n,t}|^2 = |A(S_n) \cdot \mathbf{f}_t + b(S_n) - \varepsilon_t|^2, \tag{3.3}$$

where $D_{n,t}$ is the distance of this point to cluster n. In principle, a datapoint could be equidistant to two or more different clusters by this metric; in such cases, we randomly assign the datapoint to only one of those equidistant clusters to enforce the pairwise-disjointness of the resulting clusters. Convergence of the greedy algorithm is measured by the decrease in the objective function of Eq. 3.1.

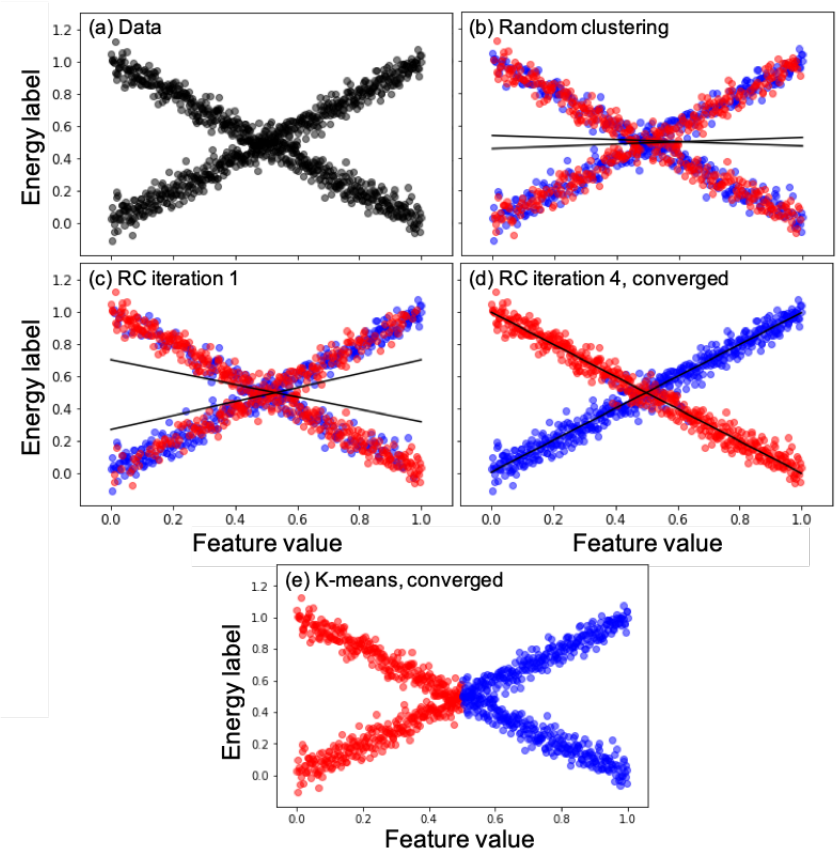

Figure 3.2: Comparison of clustering algorithms for (a) a dataset composed of two cluster of nearly linear data that overlap in feature space, using (b-d) RC and (e) standard K-means clustering. (b) Random initialization of the clusters for the greedy algorithm, with datapoint color indicating cluster assignment. (c) Cluster assignments after one iteration of the greedy algorithm. (d) Converged cluster assignments after four iterations of the greedy algorithm. For panels (b-d), two linear regression lines at each iteration are shown in black. (e) Converged cluster assignments obtained using K-means clustering, which fails to reveal the underlying linear structure of the clusters.

Figure 3.2 illustrates RC in a simple one-dimensional example for which unsupervised clustering approaches will fail to reveal the underlying linear structure. To create two clusters of nearly linear data that overlap in feature space, the interval of feature values on $[0,1]$ is uniformly discretized, such that $\mathbf{f}_t = (t-1)/(M-1)$ for $t = 1, \ldots, M$. Then, $M/2$ of the feature values are randomly chosen without replacement for cluster S_1 while the remainder are placed in S_2; the energy labels associated with each feature value are then generated using

$$\varepsilon_t = \mathbf{f}_t + \xi_{t,1}, \quad t \in S_1$$

and

$$\varepsilon_t = -\mathbf{f}_t + 1 + \xi_{t,2}, \quad t \in S_2$$

, where $\xi_{t,k} \sim \mathscr{N}(0, 0.1^2)$ is an i.d.d. sequence. The resulting dataset is shown in Fig. 3.2a.

Application of the RC method to this example is illustrated in Figs. 3.2(b-d). The greedy algorithm is initialized by randomly assigning each datapoint to either S_1 or S_2 (Fig. 3.2b). Then, with only a small number of iterations (Figs. 3.2c and d), the algorithm converges to clusters that reflect the underlying linear character. For comparison, Fig. 3.2e shows the clustering that is obtained upon convergence of the standard K-means algorithm,[103] initialized with random cluster assignments. Unlike RC, the K-means algorithm prioritizes the compactness of clusters, resulting in a final clustering that is far less amenable to simple regression. While we recognize that the correct clustering could potentially be obtained using K-means when the dimensions of $\mathbf{f}_t$ and ε_t are comparable, this is not the case for MOB-ML applications since $\mathbf{f}_t$ is typically at least 10-dimensional and ε_t is a scalar; the RC approach does not suffer from this issue. Finally, we have confirmed that initialization of RC from the clustering in Fig. 3.2e rapidly returns to the results in Fig. 3.2d, requiring only a couple of iterations of the greedy algorithm.

3.3 Computational Details

Results are presented for QM7b-T,[58] a thermalized version of the QM7b set [97] of 7211 molecules with up to seven C, O, N, S, and Cl heavy atoms, as well as for GDB-13-T, [58] a thermalized version of the GDB-13 set [69] of molecules with thirteen C, O, N, S, and Cl heavy atoms. The MOB-ML features employed in the current study are identical to those previously provided. [58] Reference pair correlation energies are computed using MP2 [5] and using CCSD(T). [8, 90] The MP2 reference data were obtained with the cc-pVTZ basis set, [87] whereas the

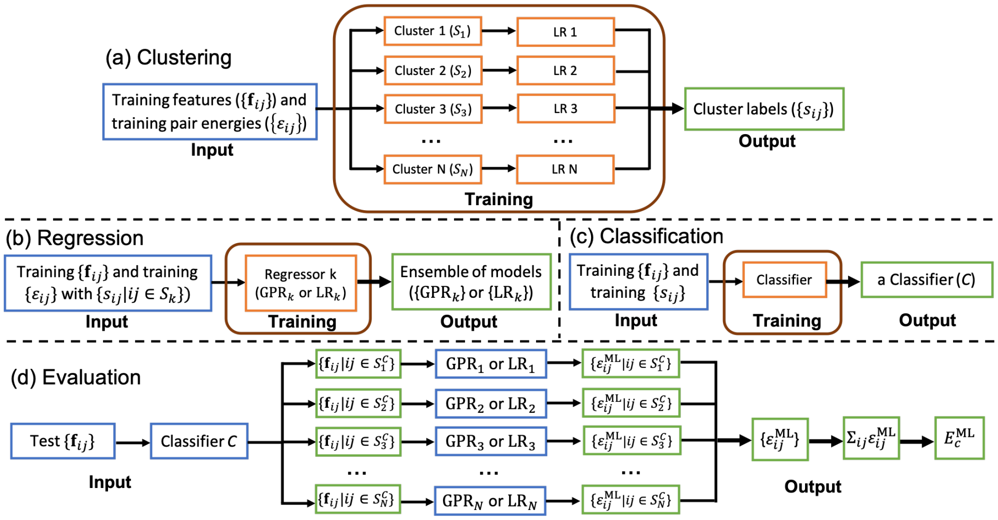
(a) Clustering
Training features ({$\mathbf{f}_{ij}$}) and training pair energies ({ε_{ij}})
Input
Cluster 1 (S_1)
Cluster 2 (S_2)
Cluster 3 (S_3)
Cluster N (S_N)
LR 1
LR 2
LR 3
LR N
Training
Cluster labels ({s_{ij}})
Output
(b) Regression
Training {$\mathbf{f}_{ij}$} and training {ε_{ij}} with {$s_{ij}|ij \in S_k$})
Input
Regressor k (GPR$_k$ or LR$_k$)
Training
Ensemble of models ({GPR$_k$} or {LR$_k$})
Output
(c) Classification
Training {$\mathbf{f}_{ij}$} and training {s_{ij}}
Input
Classifier
Training
a Classifier (C)
Output
(d) Evaluation
Test {$\mathbf{f}_{ij}$}
Classifier C
Input
{$\mathbf{f}_{ij}|ij \in S_1^C$}
{$\mathbf{f}_{ij}|ij \in S_2^C$}
{$\mathbf{f}_{ij}|ij \in S_3^C$}
{$\mathbf{f}_{ij}|ij \in S_N^C$}
GPR$_1$ or LR$_1$
GPR$_2$ or LR$_2$
GPR$_3$ or LR$_3$
GPR$_N$ or LR$_N$
{$\varepsilon_{ij}^{\mathrm{ML}}|ij \in S_1^C$}
{$\varepsilon_{ij}^{\mathrm{ML}}|ij \in S_2^C$}
{$\varepsilon_{ij}^{\mathrm{ML}}|ij \in S_3^C$}
{$\varepsilon_{ij}^{\mathrm{ML}}|ij \in S_N^C$}
{$\varepsilon_{ij}^{\mathrm{ML}}$}
$\Sigma_{ij}\varepsilon_{ij}^{\mathrm{ML}}$
E_c^{ML}
Output

Figure 3.3: General clustering/regression/classification workflow for MOB-ML. (a) Clustering of the training dataset of MOB-ML feature vectors and energy labels using RC to obtain optimized linear clusters and to provide the cluster labels for the feature vectors. (b) Regression of each cluster of training data (using LR or GPR), to obtain the ensemble of cluster-specific regression models. (c) Training a classifier (RFC) from the MOB-ML feature vectors and cluster labels for the training data. (d) Evaluating the predicted MOB-ML pair correlation energy from a test feature vector is performed by first classifying the feature vector into one of the clusters, then evaluating the cluster-specific regression model. In each panel, blue boxes indicate input quantities, orange boxes indicate training intermediates, and green boxes indicate the resulting labels, models, and pair correlation energy predictions.

CCSD(T) data were obtained using the cc-pVDZ basis set. [87] All employed training and test datasets are provided in Ref. 58.

Regression Clustering (RC)

RC is performed using the ordinary least square linear regression implementation in the SCIKIT-LEARN package [94]. Unless otherwise specified, we initialize the greedy algorithm from the results of K-means clustering, also implemented in SCIKIT-LEARN; K-means initialization was found to improve the subsequent training of the random forest classifier (RFC) in comparison to random initialization. It is found that neither L1 nor L2 regularization had significant effect on the rate of convergence of the greedy algorithm, so neither is employed in the results presented here. It is found that a convergence threshold of 1×10^{-8} kcal^2/mol^2 for the loss function of the greedy algorithm (Eq. 3.1) leads to no degradation in the final MOB-ML regression accuracy (Fig. S2); this value is employed throughout.

Regression

Two different regression models are employed in the current work. The first is ordinary least-squares linear regression (LR), as implemented in SCIKIT-LEARN. The second is Gaussian Process Regression, as implemented in the GPY 1.9.6 software package [93]. Regression is independently performed for the training data associated with each cluster, yielding a local regression model for each cluster. Also, as in our previous work,[57, 58] regression is independently performed for the diagonal and off-diagonal pair correlation energies ($\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ and $\varepsilon_{\mathrm{o}}^{\mathrm{ML}}$) yielding independent regression models for each.

GPR is performed using a negative log marginal likelihood objective. As in our previous work,[58] the Matérn 5/2 kernel is used for regression of the diagonal pair correlation energies and the Matérn 3/2 kernel is used for the off-diagonal pair correlation energies; in both cases, white noise regularization[92] is employed, and the GPR is initialized with unit lengthscale and variance.

Classification

An RFC is trained on MOB-ML features and cluster labels for a training set and then used to predict the cluster assignment of test datapoints in MOB-ML feature space. We employ the RFC implementation in SCIKIT-LEARN, using with 200 trees, the entropy split criteria,[104] and balanced class weights.[104] Alternative classifiers were also tested in this work, including K-means, Linear SVM,[105] and AdaBoost;[106] however, these schemes were generally found to yield less accurate MOB-ML energy predictions than RFC.

For comparison, a “perfect" classifier is obtained by simply including the test data within the RC training set. While useful for the analysis of prediction errors due to classification, this scheme is not generally practical because it assumes prior knowledge of the reference energy labels for the test molecules. Since the perfect classifier avoids mis-classification of the test data by construction, it should be regarded as a best case scenario for the performance of the clustering/regression/classification approach.

The clustering/regression/classification workflow

Fig. 3.3 summarizes the combined work flow for training and evaluating a MOB-ML model with clustering. The training involves three steps: First, the training dataset of MOB-ML feature vectors and energy labels are assigned to clusters using the RC method (panel a). Second, for each cluster of training data, the regression model (LR or GPR) is trained, to enable the prediction of pair correlation energies from the MOB-ML vector. Third, a classifier is trained from the MOB-ML feature vectors and cluster labels for the training data, to enable the prediction of the cluster assignment from a MOB-ML feature vector.

The resulting MOB-ML model is specified in terms of the method of clustering (RC, for all results presented here), the method of regression (either LR or GPR), and the method of classification (either RFC or the perfect classifier). In referring to a given MOB-ML model, we employ a notation that specifies these options (e.g., RC/LR/RFC or RC/GPR/perfect).

Evaluation of the trained MOB-ML model is explained in Fig. 3.3d. A given molecule is first decomposed into a set of test feature vectors associated with the pairs of occupied MOs. The classifier is then used to assign each feature vector to an associated cluster. The cluster-specific regression model is then used to predict the pair correlation energy from the MOB feature vector. And finally, the pair correlation energies are summed to yield the total correlation energy for the molecule.

To improve the accuracy and reduce the uncertainty in the MOB-ML predictions, ensembles of 10 independent models using the clustering/regression/classification workflow are trained, and the predictive mean and the corresponding standard error of the mean (SEM) are computed by averaging over the 10 models; a comparison between the learning curves[95] from a single run and from averaging over the 10 independent models is included in **Appendix** Fig. 3.11. As described here, the predicted correlation energies may exhibit discontinuities as a function of nuclear position, due to changes in the assignment of feature vectors among the clusters; moving forward, this may be avoided with the use of soft (or fuzzy) clustering algorithms.[107]

3.4 Results

Clustering and classification in MOB feature space

We begin by showing that the situation explored in Fig. 3.2, in which locally linear clusters overlap, also arises in realistic chemical applications of MOB-ML. We consider the QM7b-T set of drug-like molecules with thermalized geometries, using the diagonal pair correlation energies $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ computed at the MP2/cc-pVTZ level. Randomly selecting 1000 molecules for training, we perform RC on the dataset composed of these energy labels and feature vectors, using $N = 20$ optimized clusters; the sensitivity of RC to the choice of N is examined later.

In many cases, the resulting clusters are well separated, such that the datapoints for one cluster have small distances (as measured by the linear regression distance metric, Eq. 3.3) to the cluster which it belongs to and large distances to all other clusters. However, the clusters can also overlap. Fig. 3.4a illustrates this overlap for two particular clusters (labeled 1 and 2) obtained from the QM7b-T diagonal-pair training data.

Each datapoint assigned to cluster 1 (blue) is plotted according to its distance to both cluster 1 and cluster 2; likewise for the datapoints in cluster 2 (red). The datapoints for which the distances to both clusters approach zero correspond to regions of overlap between the clusters in the high-dimensional space of MOB-ML features, akin to the case shown in Fig. 3.2.

Finally, in Fig. 3.4b, we illustrate the classification of the feature vectors into clusters. An RFC is trained on the feature vectors and cluster labels for the diagonal pairs of 1000 QM7b-T molecules in the training set, and the classifier is then used to predict the cluster assignment for the feature vectors associated with the remaining diagonal pairs of 6211 molecules in QM7b-T. For clusters 1 and 2, we then analyze the accuracy of the RFC by plotting the linear regression distance for each datapoint to the two clusters, as well as indicating the RFC classification of the feature vector. Each red datapoint in Fig. 3.4b that lies above the diagonal line of reflection is mis-classified into cluster 2, and similarly, each blue datapoint that lies below the line of reflection is mis-classified into cluster 1. The figure illustrates that while RFC is not a perfect means of classification, it is at least qualitatively correct. Later, in the results section, we will analyze the sources of MOB-ML prediction errors due to mis-classification by comparing energy predictions obtained with perfect classification versus RFC.

Chemically intuitive clusters

To address this, we employ a training set of 500 randomly selected molecules from QM7b-T, and we perform regression clustering for the diagonal pair correlation energies $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$ with a range of total cluster numbers, up to $N = 20$. For each clustering, we then train an RFC. Finally, each trained RFC is independently applied to a set of test molecules with easily characterized valence molecular orbitals (listed in the caption of Fig. 3.5), to see how the feature vectors associated with valence occupied LMOs are classified among the optimized clusters.

Figure 3.5 presents the results of this exercise, clearly indicating the agreement be-

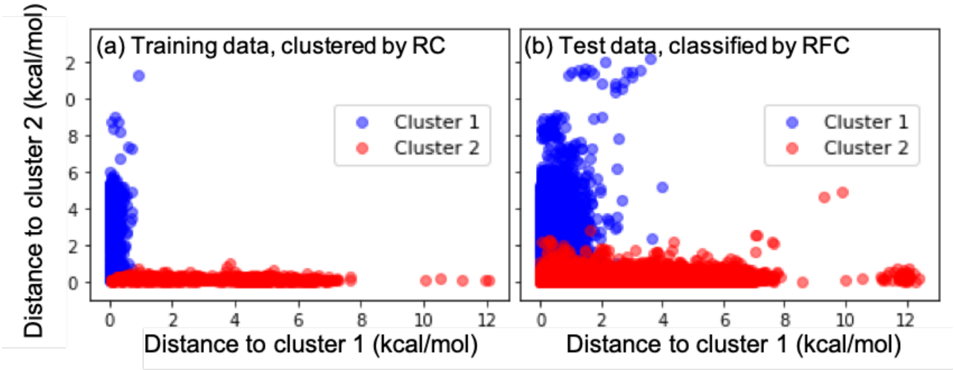

Figure 3.4: (a) Illustration of the overlap of clusters obtained via RC for the training set molecules from QM7b-T. (b) Classification of the datapoints for the remaining test molecules from QM7b-T, using RFC. Distances correspond to the linear regression metric defined in Eq. 3.3.

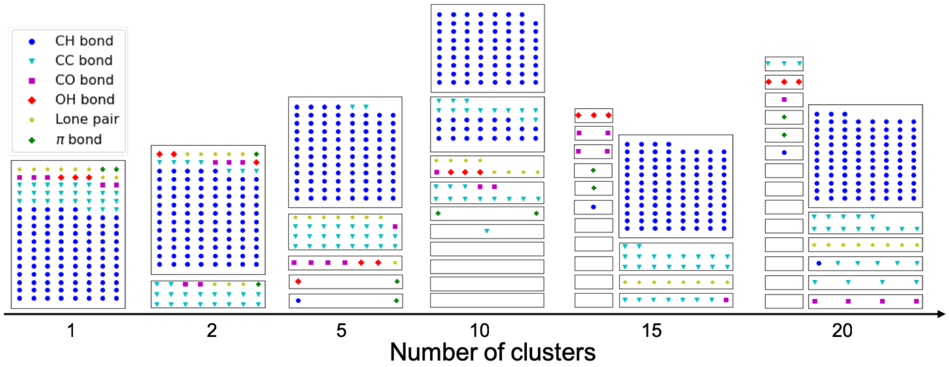

Figure 3.5: Analyzing the results of clustering/classification in terms of chemical intuition. Using a a training set of 500 randomly selected molecules from QM7b-T, RC is performed for the diagonal pair correlation energies, $\varepsilon_{\mathrm{d}}^{\mathrm{ML}}$, with a range of cluster numbers, N, and for each clustering, an RFC is trained. Then, the trained classifier is applied to a set of test molecules (CH_4, C_2H_6, C_2H_4, C_3H_8, CH_3CH_2OH, CH_3OCH_3, $CH_3CH_2CH_2CH_3$, $CH_3CH(CH_3)CH_3$, $CH_3CH_2CH_2CH_2CH_2CH_2CH_3$, $(CH_3)_3CCH_2OH$, and $CH_3CH_2CH_2CH_2CH_2CH_2OH$) which have chemically intuitive LMO types, as indicated in the legend. The LMOs are successfully resolved according to type by the classifier as N increases. Empty boxes correspond to clusters into which none of the LMOs from the test set is classified; these are expected since the training set is more diverse than the test set.

tween chemical intuition and the predictions of the RFC. As the number of clusters increases, the feature vectors associated with different valence LMO types are resolved into different clusters; and with a sufficiently large number of clusters (15 or 20), each cluster is dominated by a single type of LMO while each LMO type is assigned to a small number of different clusters. The empty boxes in Fig. 3.5 reflect that the training set contains a larger diversity of LMO types than the 11 test molecules, which is expected. The observed consistency of the clustering/classification method presented here with chemical intuition is of course promising for the accurate local regression of pair correlation energies, which is the focus of the current work; however, the results of Fig. 3.5 also suggest that the clustering/classification of chemical systems in MOB-ML feature space provides a powerful and highly general way of mapping the structure of chemical space for other applications, including explorative or active ML applications.[108]

Sensitivity to the number of clusters

We now explore the sensitivity of the MOB-ML clustering/regression/classification implementation to the number of employed clusters. In particular, we investigate the mean absolute error (MAE) of the MOB-ML predictions for the diagonal ($\sum_i \varepsilon_{ii}$) and off-diagonal ($\sum_{i\neq j} \varepsilon_{ij}$) contributions to the total correlation energy, as a function of the number of clusters, N, used in the RC. The MOB-ML models employ linear regression and RFC classification (i.e., the RC/LR/RFC protocol); the training set is composed of 1000 randomly chosen molecules from QM7b-T, and the test set contains the remaining molecules in QM7b-T.

Figure 3.6 presents the result of this calibration study, plotting the prediction MAE as a function of the number of clusters. Not surprisingly, the prediction accuracy for both the diagonal and off-diagonal contributions improves with N, although it eventually plateaus in both cases. For the diagonal contributions, the accuracy improves most rapidly up to approximately 20 clusters, in accord with the observations in Fig. 3.5; and for the off-diagonal contributions, a larger number of clusters is useful for reducing the MAE error, which is sensible given the greater variety of feature vectors that can be created from pairs of LMOs rather than only individual LMOs. Appealingly, there does not seem to be a strong indication of MAE increases due to "over-clustering". While recognizing that the optimal number of clusters will, in general, depend somewhat on the application and the regression method (i.e., LR versus GPR), the results in Fig. 3.6 nonetheless provide useful guidance with regard to the appropriate values of N. Throughout the remainder of the study, we employ a value of $N = 20$ for the MOB-ML prediction of diagonal contributions to the correlation energy and a value of $N = 70$ for the off-diagonal contributions; however, we recognize that these choices could be further optimized.

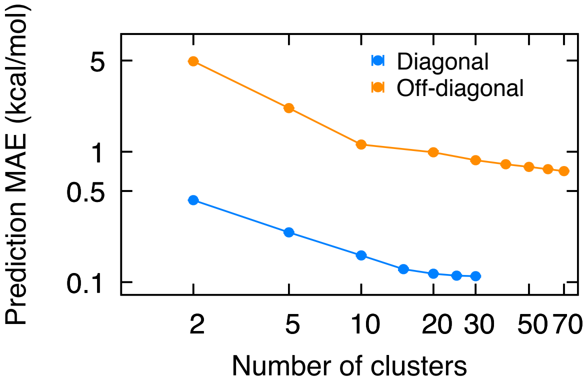

Figure 3.6: Illustration of the sensitivity of MOB-ML predictions for the diagonal and off-diagonal contributions to the correlation energy for the QM7b-T set of molecules, using a subset of 1000 molecules for training and the RC/LR/RFC protocol. The standard error of the mean (SEM) for the predictions is smaller than the size of the plotted points.

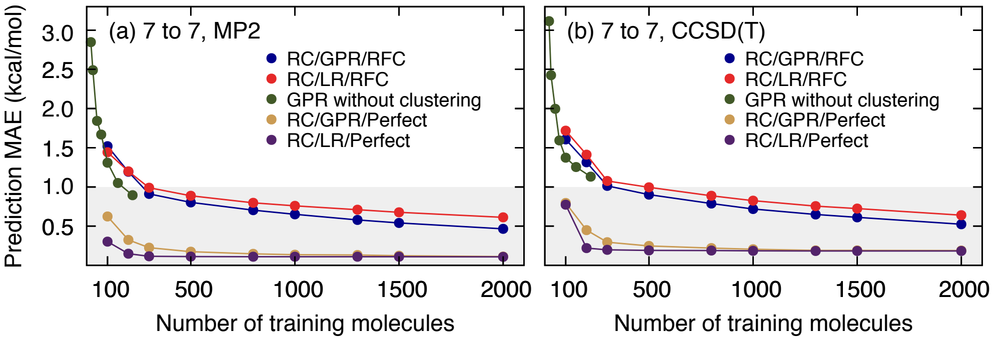

Figure 3.7: Learning curves for various implementations of MOB-ML applied to (a) MP2/cc-pVTZ and (b) CCSD(T)/cc-pVDZ correlation energies, with the training and test sets corresponding to non-overlapping subsets of the QM7b-T set of drug-like molecules with up to heavy seven atoms. Results obtained using GPR without clustering (green) are reproduced from Ref. 58. The gray shaded area corresponds to a MAE of 1 kcal/mol per seven heavy atoms. The prediction SEM is smaller than the plotted points. The log-log version of this plot is provided in Fig. S3.

Performance and training cost of MOB-ML with RC

We now investigate the effect of clustering on the accuracy and training costs of MOB-ML for applications to sets of drug-like molecules. Figure 3.7a presents learning curves (on a linear-linear scale) for various implementations of MOB-ML

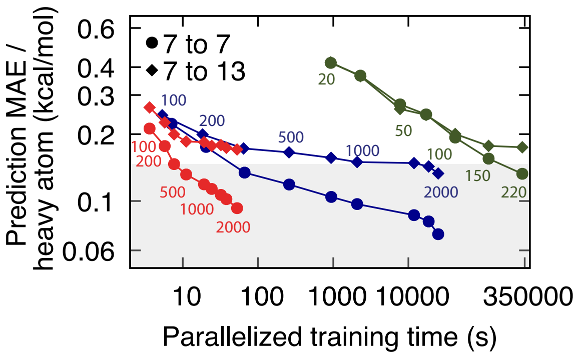

Figure 3.8: Training costs and transferability of MOB-ML with clustering (RC/LR/RFC, red; RC/GPR/RFC, blue) and without clustering (green, Ref. 58), applied to correlation energies at the MP2/cc-pVTZ level. Prediction errors are plotted as a function of wall-clock training time. Training sets are composed of subsets of the QM7b-T dataset, with the number of training molecules indicated via datapoint labels. Correlation energy predictions are made for test sets composed of the remaining seven-heavy-atom molecules from QM7b-T (circles) and the thirteen-heavy-atom molecules from GDB-13-T (diamonds). Both MAE prediction errors and parallelized wall-clock training times are plotted on a log scale. The gray shaded area corresponds to a MAE of 1 kcal/mol per seven heavy atoms. The prediction SEM is smaller than the plotted points. Details of the parallelization and employed computer hardware are described in the text.

applied to MP2/cc-pVTZ correlation energies, with the training and test sets corresponding to non-overlapping subsets of QM7b-T. In addition to the new results obtained using RC, we include the MOB-ML results from our previous work (GPR without clustering).[58]

Figure 3.7a yields three clear observations. The first is that the use of RC with RFC (i.e., RC/GRP/RFC and RC/LR/RFC) leads to slightly less efficient learning curves than our previous implementation without clustering, at least when efficiency is measured in terms of the number of training molecules. Both the RC/GPR/RFC and RC/LR/RFC protocols require approximately 300 training molecules to reach the 1 kcal/mol per seven heavy atoms threshold for chemical accuracy employed here, whereas MOB-ML without clustering requires approximately half as many training molecules. The second observation is that the classifier is the dominant source of prediction error in these results. Comparison of results using RFC versus the perfect classifier (which utilizes prior knowledge of the energy labels and thus is not generally practical), reveals a dramatic reduction in the prediction error, regardless of the regression method. This result indicates that there is potentially much to be gained from the development of improved classifiers for MOB-ML applications. A third observation is that with a perfect classifier, the LR slightly outperforms GPR, given that the clusters are optimized to be locally linear; however, GPR slightly outperforms LR in combination with the RFC, indicating that GPR is less sensitive to classification error that LR.

Figure 3.7b presents the corresponding results at the CCSD(T)/cc-pVDZ level of theory. The same trends emerge as the ones at the MP2/cc-pVTZ level of theory. As seen in previous work, the training efficiency of MOB-ML with respect to the size reference dataset is found to be largely insensitive to the level of electronic structure theory.[57, 58]

Figure 3.8 explores the training costs and transferability of MOB-ML models that employ RC. In all cases, the models are trained on random subsets of molecules from QM7b-T with up to seven heavy atoms, and predictions are made either on the remaining molecules of QM7b-T (circles) or on the GDB-13-T set (diamonds); it has previously been shown than that MOB-ML substantially outperforms the FCHL atom-based-feature method in terms of transferability from small to large molecules.[58] The parallelization of the training steps are implemented as follows. Within the RC step, the LR for each cluster is performed independently on a different core of a 16-core Intel Skylake (2.1 GHz) CPU processor. Within the regression step, the LR or GPR for each cluster is likewise performed independently on a different core. For RFC training, we apply parallel 200 cores using the parallel implementation of SCIKIT-LEARN, since there are 200 trees. The regression and RFC training are independent of each other and are thus also trivially parallelizable.

Focusing first on the predictions for seven-heavy-atom molecules (circles), it is clear from Fig. 3.8 that RC leads to large improvements in the efficiency of the MOB-ML wall-clock training costs. Although it requires somewhat more training molecules than MOB-ML without clustering, MOB-ML with clustering enables chemical accuracy to be reached with the training cost reduced by a factor of approximately 4500 for RC/GPR/RFC and of 35000 for RC/LR/RFC. Remarkably, for predictions within the QM7b-T set, chemical accuracy can be achieved using RC/LR/RFC with a wall-clock training time of only 7.7 s.

Figure 3.8 also demonstrates the transferability of the MOB-ML models for predictions on the GDB-13-T set of thirteen-heavy-atom molecules (diamonds). In general, it is seen that the degradation in the MAE per atom is greater for the RC/LR/RFC than for RC/GPR/RFC, due to the previously mentioned sensitivity of LR to classification error. However, we note that the RC/GPR/RFC enables predictions on GDB-13-T (blue, diamonds) that meet the per-atom threshold of chemical accuracy used here, whereas that threshold was not achievable without clustering (green, diamonds) due to the prohibitive training costs involved.

The improved efficiency of MOB-ML training with the use of clustering arises from the cubic scaling of standard GPR in terms of training time ($\mathscr{O}(M^3)$, where M is number of training pairs).[92] Trivial parallelization over the independent regression of the clusters reduces this training time cost to the cube of largest cluster. We note that other kernel-based ML methods with high complexity in training time, like Kernel Ridge Regression,[109] would similarly benefit from clustering. For the RC/LR/RFC and RC/GPR/RFC results presented in Fig. 3.8, a breakdown of the training time contributions for each step of the clustering/regression/classification workflow as a function of the size of the training dataset is shown in Fig. S4; this supporting information figure confirms that the GPR regression dominates the total training (and prediction) costs for the RC/GPR/RFC implementation, whereas training the RFC dominates the training costs for RC/LR/RFC. In addition to improved efficiency in terms of training time, clustering also bring benefits in terms of the memory costs for MOB-ML training, due to the quadratic scaling of GPR memory costs in terms of the size of the dataset.

Finally, returning to the learning curves, we compare the results for MOB-ML both with and without clustering to recent work[110] using Faber-Christensen-Huang-Lilienfeld (FCHL) features. Fig. 3.9 shows these various learning curves for the MP2/cc-pVTZ correlation energies. For Fig. 3.9a, the training and test sets correspond to non-overlapping subsets of QM7b-T, and Fig. 3.9b shows the transferability of the same models trained using QM7b-T to predict the energies for GDB-13-T. Fig. 3.9a again shows that MOB-ML RC/GPR/RFC requires slightly more training geometries than MOB-ML without clustering, yet both MOB-ML protocols are more efficient in terms of training data than either the FCHL18[98] or FCHL19 implementations[110]. Like MOB-ML with clustering, the FCHL19 implementation was developed to reduce training times.

Capping the cluster size

Since the parallelized training time for RC/GPR/RFC is dominated by the GPR regression of the largest cluster (Fig. S4), a natural question is whether additional computational savings and adequate prediction accuracy could achieved by simply capping the number of datapoints in the largest cluster. In doing so, we define

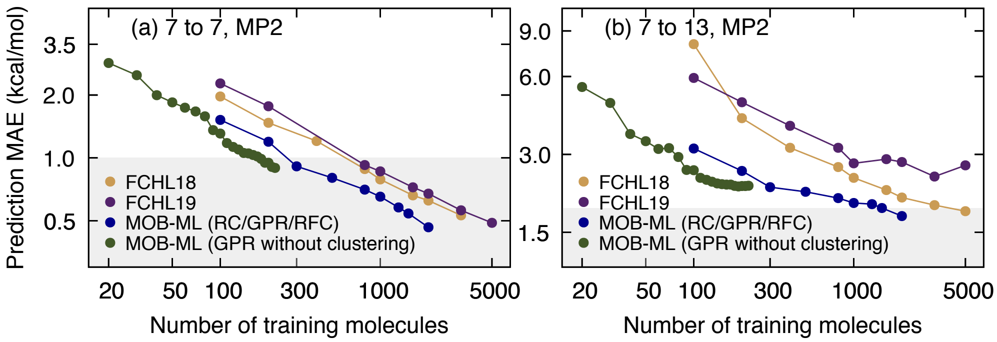

Figure 3.9: Comparison of learning curves for MP2/cc-pVTZ correlation energies obtained using MOB-ML (with and without clustering) versus FCHL18 and FCHL19. Part (a) presents results for which both the training and test sets include molecules from QM7b-T, and part (b) presents results for which the training set includes molecules from QM7b-T and the test set includes molecules from GDB-13-T. The MAE are plotted on a log-log scale as a function of number of training molecules. The gray shaded area corresponds to a MAE of 1 kcal/mol per seven heavy atoms. Results for FCHL18 and FCHL19 were digitally captured from Ref. 110.

$S_{\max}^{N_{\text{cap}}}$ to be the number of datapoints in the largest cluster obtained when the RC with the greedy algorithm is applied to a training dataset of N_{cap} molecules from QM7b-T. Upon specifying N_{cap} (and thus $S_{\max}^{N_{\text{cap}}}$), the RC/GPR/RFC implementation is modified as follows. For a given number of training molecules (which will typically exceed N_{cap}), the RC step is performed as normal. However, at the end of the RC step, datapoints for clusters whose size exceeds $S_{\max}^{N_{\text{cap}}}$ are discarded at random until all clusters contain $S_{\max}^{N_{\text{cap}}}$ or fewer datapoints. The GPR and RFC training steps are performed as before, except using this set of clusters that are capped in size. The precise value of $S_{\max}^{N_{\text{cap}}}$ will vary slightly depending on which training molecules are randomly selected for training and the convergence of the greedy algorithm, but typical values for $S_{\max}^{N_{\text{cap}}}$ are $672, 1218, 1863, 3005$, and 4896 for $N_{\text{cap}} = 100, 200, 300, 500$, and 800, respectively, and those values will be used for the numerical tests presented here.

Figure 3.10a demonstrates that capping the maximum cluster size allows for substantial improvements in accuracy when the number of training molecules exceeds N_{cap}. Specifically, the figure shows the effect of capping on RC/GPR/RFC learning curves for MP2/cc-pVTZ correlation energies, with the training and test sets corresponding to non-overlapping subsets of QM7b-T. As a baseline, note that with 100 training molecules, the RC/GPR/RFC implementation yields a prediction MAE of approximately 1.5 kcal/mol. However, if the maximum cluster size is capped at $N_{\text{cap}} = 100$ and 300 training molecules are employed, then the prediction MAE drops to approximately 1.0 kcal/mol while the parallelized training cost for RC/GPR/RFC will be unchanged so long as it remains dominated by the size of the largest cluster. As expected, Fig. 3.10a shows that the learning curves saturate at higher prediction MAE values when smaller values of N_{cap} are employed. Nonetheless, the figure demonstrates that if additional training data is available, then the prediction accuracy for MOB-ML with RC can be substantially improved while capping the size of the largest cluster.

Figure 3.10b demonstrates the actual effect of capping on the parallelized training time, plotting the prediction MAE versus parallelized training time as a function of the number of training molecules. For reference, the results obtained using RC/LR/RFC and RC/GPR/RFC without capping are reproduced from Fig. 3.8. As is necessary, the RC/GPR/RFC results obtained with capping exactly overlap those obtained without capping when the number of training molecules is not greater than N_{cap}. However, for each value of N_{cap}, a sharp drop in the prediction MAE is seen when the number of training molecules begins to exceed N_{cap}, demonstrating that prediction accuracy can be greatly improved with minimal increase in parallelized training time. For example, it is seen that for RC/GPR/RFC with $N_{\text{cap}} = 100$, chemical accuracy can be reached with only 7.4 s of parallelized training, slightly less than even RC/LR/RFC. For small values of N_{cap}, this prediction MAE eventually levels-off versus the training time, since the RFC training step becomes the dominant contribution to the training time.

3.5 Conclusions

Molecular-orbital-based (MOB) features offer a complete representation for mapping chemical space and a compact representation for evaluating correlation energies. In the current work, we take advantage of the intrinsic structure of MOB feature space, which cluster according to types of localized molecular orbitals, as well as the fact that orbital-pair contributions to the correlation energy contributions vary linearly with the MOB features, to overcome a fundamental bottleneck in the efficiency of ML correlation energies. Specifically, we introduce a regression clustering (RC) approach in which MOB features and pair correlation energies are clustered according to their local linearity; we then individually regress these clusters and train a classifier for the prediction of cluster assignments on the basis of MOB features. This combined clustering/regression/classification approach is found to

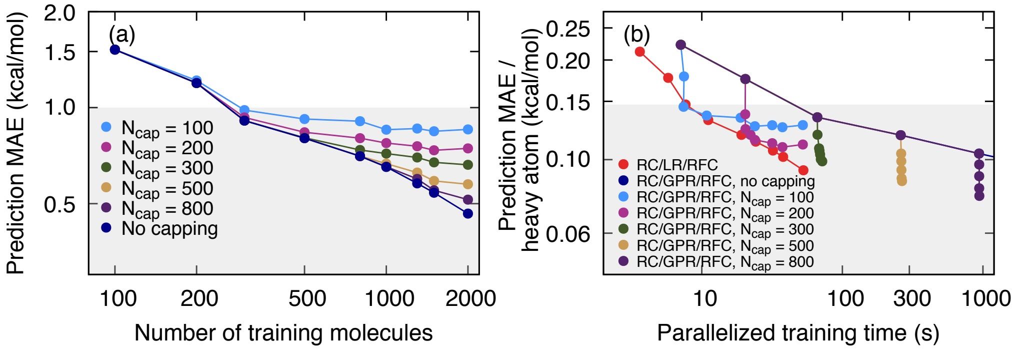

Figure 3.10: The effect of cluster-size capping on the prediction accuracy and training costs for MOB-ML with RC. Results reported for correlation energies at the MP2/cc-pVTZ level, with the training and test sets corresponding to non-overlapping subsets of the QM7b-T set of drug-like molecules with up to heavy seven atoms. (a) Prediction MAE versus the number of training molecules, with the clusters capped at various maximum sizes. The RC/GPR/RFC curve without capping is reproduced from Fig. 3.7a. (b) Prediction MAE per heavy atom versus parallelized training time as a function of the number of training molecules, as in Fig. 3.8. The results for MOB-ML with clustering and without capping cluster size (RC/LR/RFC, red; RC/GPR/RFC, blue) are reproduced from Fig. 3.8. Also, the results for RC/GPR/RFC with various capping sizes N_{cap} are shown. For part (a), the gray shaded area corresponds to a MAE of 1 kcal/mol, and for part (b), it corresponds to 1 kcal/mol per seven heavy atoms, to provide consistency with preceding figures. The prediction SEM is smaller than the plotted points.

reduce MOB-ML training times by 3-4 orders of magnitude, while enabling prediction accuracies that are substantially improved over that which is possible using MOB-ML without clustering. The use of a random forest classifier for the cluster assignments, while better than alternatives that were explored, is found to be the limiting factor in terms of MOB-ML accuracy within this new approach, motivating future work on improved classifiers. This work provides a useful step towards that development of accurate, transferable, and scalable quantum ML methods to describe ever-broader swathes of chemical space.

3.6 Appendix

Figures in this **Appendix** show the effect of averaging over independently trained MOB-ML-models (Fig. 3.11), the sensitivity of the prediction accuracy to the RC convergence threshold (Fig. 3.12), learning curves for various implementations of MOB-ML plotted on a log-log scale (Fig. 3.13), and a detailed breakdown of the parallelized wall-clock timings (Fig. 3.14). Tables 3.1, 3.2, 3.3, 3.4, and 3.5 provide the numerical data for the plots appearing in the main text.

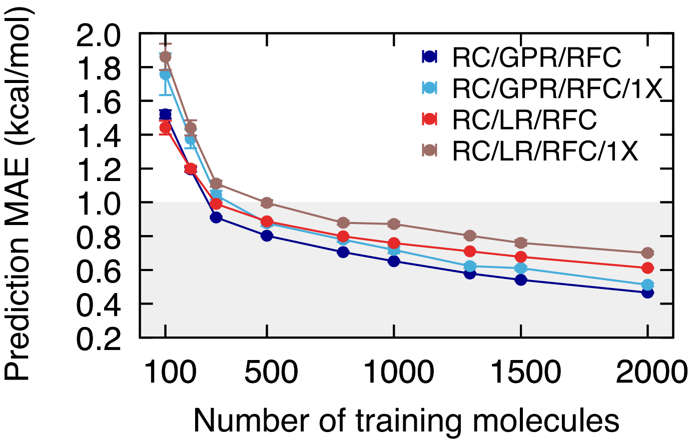

Figure 3.11: Learning curves for various implementations of MOB-ML applied to MP2/cc-pVTZ correlation energies, with the training and testing sets corresponding to non-overlapping subsets of QM7b-T. Results obtained from averaging over 10 independent models are compared to results from a single model (/1X) without averaging. For both the RC/GPR/RFC and RC/LR/RFC implementations, averaging over independent models reduces the prediction MAE.

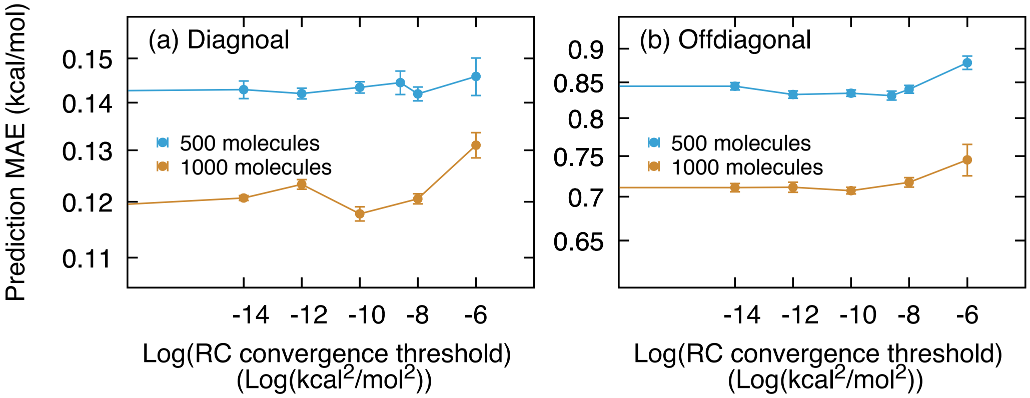

Figure 3.12: Sensitivity of MOB-ML predictions to the RC convergence threshold. Results are obtained using the RC/LR/RFC implementation of MOB-ML applied to MP2/cc-pVTZ correlation energies, with the training and testing sets corresponding to non-overlapping subsets of QM7b-T. The prediction MAEs for the contributions from the (a) diagonal and (b) off-diagonal pair energies are shown for two different training set sizes. For both the diagonal and off-diagonal pair contributions, a threshold value of 1×10^{-8} kcal^2/mol^2 for the RC loss function (Eq. 4 in the main text) provides similar results as tighter convergence thresholds.

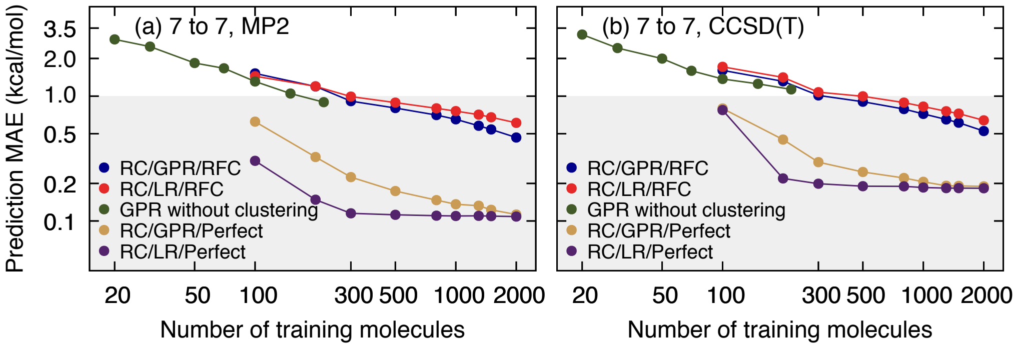

Figure 3.13: Learning curves for various implementations of MOB-ML applied to (a) MP2/cc-pVTZ and (b) CCSD(T)/cc-pVDZ correlation energies, with the training and test sets corresponding to non-overlapping subsets of the QM7b-T set of drug-like molecules with up to heavy seven atoms. These results are identical to those of Fig. 7 in the main text, except plotted on a log-log scale.

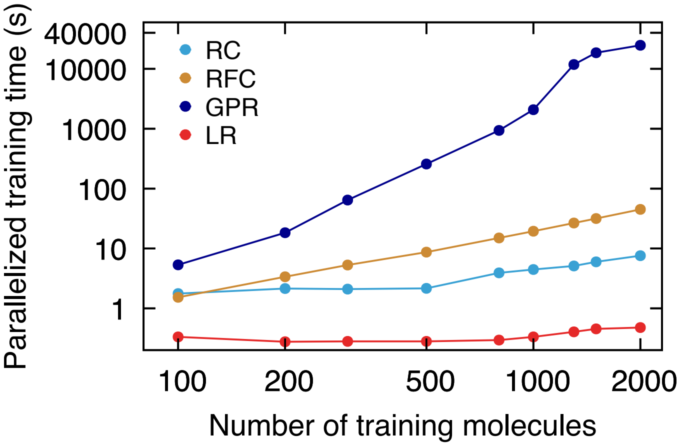

Figure 3.14: Breakdown of the wall-clock timings for RC, RFC, GPR and LR for different number of training molecules from the QM7b-T set at the MP2/cc-pvTZ level of theory. The parallelization is implemented as follows. Within the RC step, the LR regression of each cluster is performed independently on a different core of a 16-core Intel Skylake (2.1 GHz) CPU processor. With in the regression step, the LR and GPR regression of each cluster is likewise performed independently on a different core. For RFC training, we apply employ parallel 200 cores using the parallel implementation of SCIKIT-LEARN, since there are 200 trees.

Table 3.1: MOB-ML prediction accuracy for the RC/GPR/RFC implementation, applied to correlation energies at the MP2/cc-pVTZ level. Training sets are composed of subsets of the QM7b-T dataset, with the number of training molecules indicated. Correlation energy predictions are made for test sets composed of the remaining seven-heavy-atom molecules from QM7b-T and the thirteen-heavy-atom molecules from GDB-13-T. Energies in kcal/mol.

QM7b-T	QM7b-T prediction		GDB-13-T prediction			
Training size	MAE	SEM	MAE	SEM	MAE/HA	SEM/HA
100	1.520	0.025	3.415	0.046	0.2431	0.0035
200	1.194	0.014	2.785	0.024	0.1992	0.0018
300	0.9109	0.0056	2.366	0.027	0.1724	0.0020
500	0.8028	0.0048	2.278	0.023	0.1655	0.0018
800	0.7048	0.0036	2.161	0.020	0.1564	0.0015
1000	0.6517	0.0043	2.088	0.021	0.1497	0.0016
1300	0.5791	0.0032	2.062	0.019	0.1482	0.0015
1500	0.5414	0.0052	1.993	0.012	0.1432	0.0009
2000	0.4654	0.0027	1.913	0.016	0.1332	0.0012

Table 3.2: MOB-ML prediction accuracy for the RC/LR/RFC implementation, applied to correlation energies at the MP2/cc-pVTZ level. Training sets are composed of subsets of the QM7b-T dataset, with the number of training molecules indicated. Correlation energy predictions are made for test sets composed of the remaining seven-heavy-atom molecules from QM7b-T and the thirteen-heavy-atom molecules from GDB-13-T. Energies in kcal/mol.

QM7b-T	QM7b-T prediction		GDB-13-T prediction			
Training size	MAE	SEM	MAE	SEM	MAE/HA	SEM/HA
100	1.442	0.041	3.427	0.086	0.2636	0.0066
200	1.199	0.018	2.935	0.035	0.2258	0.0027
300	0.9909	0.0084	2.596	0.029	0.1997	0.0023
500	0.8869	0.0051	2.412	0.016	0.1855	0.0013
800	0.7984	0.0042	2.394	0.020	0.1842	0.0015
1000	0.7586	0.0062	2.301	0.026	0.1770	0.0020
1300	0.7100	0.0038	2.321	0.021	0.1786	0.0017
1500	0.6769	0.0037	2.257	0.014	0.1736	0.0011
2000	0.6115	0.0028	2.218	0.022	0.1706	0.0017

Table 3.3: MOB-ML prediction accuracy for the RC/GPR/Perfect and RC/LR/Perfect implementations, applied to correlation energies at the MP2/cc-pVTZ level, with the training and testing sets corresponding to non-overlapping subsets of QM7b-T. Energies in kcal/mol.

	RC/GPR/Perfect		RC/LR/Perfect	
Training size	MAE	SEM	MAE	SEM
100	0.6235	0.0331	0.3031	0.0574
200	0.3254	0.0113	0.1481	0.0231
300	0.2246	0.0075	0.1153	0.0031
500	0.1734	0.0052	0.1120	0.0029
800	0.1470	0.0031	0.1104	0.0031
1000	0.1361	0.0026	0.1096	0.0032
1300	0.1324	0.0014	0.1099	0.0033
1500	0.1230	0.0019	0.1095	0.0034
2000	0.1127	0.0010	0.1085	0.0035

Table 3.4: MOB-ML prediction accuracy for the RC/GPR/RFC and RC/LR/RFC implementations, applied to correlation energies at the CCSD(T)/cc-pVDZ level, with the training and testing sets corresponding to non-overlapping subsets of QM7b-T. Energies in kcal/mol.

	RC/GPR/RFC		RC/LR/RFC	
Training size	MAE	SEM	MAE	SEM
100	1.607	0.041	1.718	0.065
200	1.314	0.016	1.412	0.025
300	1.013	0.006	1.075	0.012
500	0.9026	0.0036	0.9951	0.0063
800	0.7880	0.0031	0.8876	0.0051
1000	0.7194	0.0053	0.8253	0.0062
1300	0.6495	0.0047	0.7559	0.0035
1500	0.6116	0.0034	0.7251	0.0034
2000	0.5243	0.0026	0.6402	0.0028

Table 3.5: MOB-ML prediction accuracy for the RC/GPR/RFC implementation with cluster-size capping, applied to correlation energies at the MP2/cc-pVTZ level, with the training and testing sets corresponding to non-overlapping subsets of QM7b-T. Energies in kcal/mol.

Training size	$N_{\text{cap}} = 100$	$N_{\text{cap}} = 200$	$N_{\text{cap}} = 300$	$N_{\text{cap}} = 500$	$N_{\text{cap}} = 800$
100	1.520	1.520	1.520	1.520	1.520
200	1.217	1.194	1.194	1.194	1.194
300	0.9827	0.9318	0.9109	0.9109	0.9109
500	0.9211	0.8370	0.8049	0.8028	0.8028
800	0.9066	0.8028	0.7368	0.7054	0.7048
1000	0.8532	0.7745	0.7178	0.6676	0.6534
1300	0.8602	0.7568	0.6983	0.6261	0.5980
1500	0.8432	0.7353	0.6740	0.5892	0.5511
2000	0.8549	0.7456	0.6620	0.5753	0.5148

Chapter 4

IMPROVED ACCURACY OF MOLECULAR ENERGY LEARNING VIA UNSUPERVISED CLUSTERING FOR THE ORGANIC CHEMICAL SPACE WITH MOLECULAR-ORBITAL-BASED MACHINE LEARNING

Adapted from:

1. Cheng, L., Sun, J. & Miller III, T. F. Improved accuracy of molecular energy learning via unsupervised clustering for organic chemical space with molecular-orbital-based machine learning. *In preparation.*

In this chapter, we consider improved clustering and regression algorithms to unsupervisedly cluster the chemical space via the Gaussian mixture model (GMM) and then regress molecular energies via alternative blackbox matrix-matrix multiplication (AltBBMM) for MOB-ML. Although regression clustering (RC) combined with local Gaussian process regression (GPR) or linear regression (LR) provides a useful framework for accurately constructing local regression models for MOB-ML, the accuracy loss associated with classification limits this approach's ability to provide the most accurate energy predictions. The improved feature design allows unsupervised clustering, and the introduction of an exact scalable GPR algorithm, i.e., AltBBMM, further scales the training of MOB-ML up. Without any additional information from label space, the resulting clusters from GMM agree with the chemically intuitive groupings of MO types. This unsupervised clustering with AltBBMM local regression provides superior accuracy over all other learning protocols and the state-of-the-art learning efficiency on the QM7b-T and GDB-13-T benchmark systems by training on 6500 QM7b-T molecules.

4.1 Introduction

Machine learning (ML) approaches have attracted considerable interest in the chemical sciences for a variety of applications, including molecular and material design, [22, 31, 44, 111–115], protein property prediction [116–118], reaction mechanism discovery [112, 119–123], and analysis and classification tasks for new insight [124–126]. As an alternative to physics-based computations, ML has also shown promise for the prediction of molecular energies [46–49, 51, 52, 127–135], intermolecular interactions [133, 136], electron densities [48, 127, 137–139], and linear response properties [50, 99, 140–143]. ML applications in the chemical sciences have relied on atom- or geometry-specific representations in the majority of cases, yet wavefunction-specific and deep-learning representations are becoming increasingly available [43, 52, 57, 58, 60, 134, 144–146]. Among these recent approaches, MOB-ML [57–61] has been shown to exhibit excellent learning efficiency and transferability for the prediction of energies of post-Hartree-Fock (post-HF) wavefunction methods.

MOB-ML initially had a limited training size due to the high computational cost of Gaussian process regression (GPR) training. A local regression with supervised clustering algorithm and a scalable exact GPR algorithm, termed as Regression-clustering (RC) algorithm (RC/GPR) [59] and alternative blackbox matrix-matrix multiplication algorithm (AltBBMM) [62], respectively, are introduced to reduce training costs and enable training on large datasets.

In this work, we use an unsupervised clustering method, namely the Gaussian mixture model (GMM), to accurately cluster the organic chemical space using MOB features. The current work determines clusters via GMM in an entirely blackbox manner and simplifies an earlier supervised clustering approach in Chap.3 by eliminating the necessity for user-specified parameters and the training of an additional classifier. Unsupervised clustering produces clusters that are consistent with chemically intuitive groupings of chemical space and exhibit linear relationships between pair energies and MOB features. All the regression (GPR or linear regression) with clustering (supervised or unsupervised) methods offer exceptional efficiency and transferability for molecular energy learning. GPR with GMM clustering is the most efficient training protocol for MOB-ML and delivers the best accuracy on QM7b-T and transferability on GDB-13-T out of all the available literature results.

4.2 Theory

Size consistency and feature ordering consistency in MOB-ML and the improved MOB feature design

Feature ordering consistency in MOB-ML MOB features have to be properly sorted to ensure consistent feature ordering according to an appropriate importance criterion, which is one of the challenges in this vector representation of MOB features. [57–60] In Ref. 60, the sorting strategy is to determine the importance of each element of f_{ij} by its contribution to ε_{ij} estimated from the MP2 (Eq. 2.3) or third-order MP (MP3) expression.

Size consistency in MOB-ML Size consistency is one of the most significant properties of an electronic structure theory since it provides a correct scaling with the system size. It states that for two infinitely separated subsystems A and B, the energy should satisfy $E(A+B) = E(A) + E(B)$. This principle requires the independence of the energies of two subsystems A and B and zero interactions at infinite distances. In the MOB-ML framework, the requirement for energy reduces to the requirement for predicted pair energies, and furthermore reduces to the following four feature conditions:

- The dependence of f_{ij} on k decays as $1/r_{ik}^3$ when $r_{ik} \to \infty$.
- f_{ij} is independent of k when $r_{ik} = \infty$.
- f_{ij} decays as $1/r_{ij}^6$ when $r_{ij} \to \infty$.
- $f_{ij} = 0$ when $r_{ij} = 0$.

Unfortunately, the MOB feature design introduced in Chapter 2 satisfies none of these criteria, which hinders the prediction accuracies of MOB-ML in many-body systems and large molecules[60]. The Coulomb interaction elements are changed from J_{pq} to J_{pq}^3 according to the first condition. To satisfy the second condition, each matrix element is also multiplied by its corresponding feature ordering importance. A global factor of $(1+\frac{1}{6}(\frac{r_{ij}}{r_0})^6)^{-1}$ is also multiplied to each feature vector according to the third requirement. However, all the manipulations above cannot guarantee the condition 4. To ensure the prediction of the zero feature vector 0 is always 0, we manually add a small number of zero training points $(0,0)$ in the GPR instead of changing feature elements directly.

We have made all the above changes in an improved MOB feature design in Ref. 60 and proved its enhanced abilities to predict the energies of organic molecules, closed-shell transition metal complexes, non-covalent interactions, and the transition states. More detailed discussions about this improved feature design and changes for each feature are shown in the Supporting Information of Ref. 60. In this chapter and Chapter 5, we employ the same feature generation and sorting approach described in Ref. 60.

Supervised and unsupervised clustering schemes for chemical spaces

A straightforward application of Gaussian progress regression (GPR) with MOB features encounters a bottleneck due to computational demands since GPR introduces the complexity of $\mathcal{O}(N^2)$ in memory and $\mathcal{O}(N^3)$ in training cost. The property of local linearity for MOB features has been investigated, which allows pair energies to be fitted as a linear function of MOB features within local clusters. [59] Thanks to this property, we proposed a comprehensive framework for local regression with clusters to further scale MOB-ML to the large data regime with lower training costs in Ref. 59. Our previous work applied supervised clustering, i.e., regression-clustering (RC), to cluster the training set and then performed GPR or linear regression (LR) as local regressors. A random forest classifier (RFC) is also trained to classify the test data.

However, supervised clustering has its limitations. RC requires a predetermined number of clusters and an additional classifier [59]. The performance of the supervised clustering scheme is also hindered by the classifier, which struggles to classify the results from RC due to the fact that the pair energy label information is only provided to RC but not RFC [59]. Therefore, a more precise and efficient clustering and classification strategy is needed to enhance the performance of the entire framework.

Improved MOB feature engineering results in a continuous MOB feature space [60], and consequently enables the unsupervised clustering scheme for MOB-ML. The points close in the feature space have similar chemical groupings, cluster identities, and label values, and therefore distance is an appropriate measure to cluster the MOB feature space. As a result, any distance-based clustering approach should perform well using the improved MOB features. K-means is the simplest and fastest distance-based unsupervised clustering method, which can effectively cluster the MOB feature space and produce reliable regression results when used in conjunction with GPR (details are shown in **Appendix**). Unfortunately, the lack of intrinsic probability measure and the assumption of isotropy makes it not as accurate as other distance-based clustering methods, such as DBSCAN [147], OPTICS [148] and Gaussian mixture model (GMM).

GMM can be treated as a generalized k-means method and is chosen for further investigation in this study. It assumes that all the N data points belong to a mixture of a certain number of multivariate Gaussian distributions in the feature space with means and covariance to be determined, and each distribution can represent a cluster. For a number of K clusters (or Gaussian distributions) with D feature dimensions, the cluster centers (or means of the distributions) $\{\mu_i \in \mathbb{R}^D, i = 1,2,...,K\}$ and their corresponding covariance matrices $\{\Sigma_i \in \mathbb{R}^{D\times D}, i = 1,2,...,K\}$ are solved by maximizing the likelihood L using Expectation-Maximization (EM) algorithm. The expectation, parameters, and clusters identities are computed and reassigned in the Expectation (E) stage, and the parameters to maximize likelihood are updated in the Maximization (M) stage. The two stages are repeated until reaching convergence. For a test point, GMM can not only provide hard cluster assignments with the maximum posterior probability, but also enable soft clustering by computing the normalized posterior probability of a test point belonging to each cluster [149]. To make the GMM training completely blackboxed, we also perform the model selection using the Bayesian information criterion (BIC) to determine the number of clusters used in GMM via scanning a reasonable series of candidate cluster sizes based on the training size N. BIC penalizes the likelihood increase due to including more clusters and more fitting parameters to avoid overfitting with respective to the number of clusters (Eq. 4.1),

$$\mathrm{BIC} = q\ln(N) - 2\ln(L), \tag{4.1}$$

where q is the number of parameters in the GMM model.

Local regression by alternative blackbox matrix-matrix multiplication algorithm

While the general framework of regression with clustering considerably improves the efficiency of MOB-ML [59], local regression with full GPR with a cubic time complexity remains the computational bottleneck for MOB-ML. The computational bottleneck is the calculation of GP inference, $\omega = \hat{K}^{-1}y$, when we only care about the predicted mean, where $\hat{K} = K(X,X) + \sigma_n^2 I$ is the regularized kernel, K is the kernel, and σ_n^2 is the Gaussian noise variance. The standard Cholesky decomposition treatment has a computational complexity of $O(N^3)$ in time and $O(N^2)$ in memory. Blackbox matrix-matrix multiplication (BBMM) is a SOTA exact GP algorithm [150, 151] that applies the pivoted Cholesky decomposition (pCD) preconditioner to precondition the kernel and conjugate gradient (CG) approach to solve the GP inference. A low Gaussian noise ($10^{-8} \sim 10^{-6}$) is preferred to obtain an accurate prediction in MOB-ML but less efficient by slowing down the convergence speed of CG. To improve the learning efficiency in MOB-ML, an alternative implementation of BBMM, known as AltBBMM, has been proposed to improve the performance by adapting a symmetric preconditioner to increase stability and the block conjugate gradient (BCG) method to speedup the CG convergence [62]. The derivation and implementation details for AltBBMM are discussed in Ref. 62. AltBBMM has been shown to be able to speed up and scale the GP training in MOB-ML for molecular energies with exact inferences. It reduces the training time complexity to $\mathscr{O}(N^2)$, enables the training on 1 million pair energies, or equivalently, 6500 QM7b-T molecules without sacrificing transferability across chemical systems of different molecular sizes[62]. By applying AltBBMM as the local regressor within each cluster, the efficiency of MOB-ML training can be further increased compared with regressing with full GPR.

4.3 Computational details

The performance of clustering and subsequent local regression approaches are evaluated on QM7b-T and GDB-13-T benchmark systems, which comprise molecules with at most seven and only thirteen C, O, N, S, and Cl heavy atoms, respectively. Each molecule in QM7b-T and GDB-13-T has seven and six conformers, respectively, and only one conformer of each randomly selected QM7b-T molecule is picked in the training set. The features are computed at HF/cc-pVTZ level with the Boys–Foster localization scheme [73, 152] using ENTOS QCORE [153], and reference MP2 [5, 88] pair energy labels with cc-pVTZ basis set [87] are generated from Molpro 2018.0 [86]. All the features, selected features and reference pair energies employed in the current work are identical to those reported in Ref. 60.

Supervised Clustering with MOB features

RC can cluster the organic chemical space represented by QM7b-T and GDB-13-T [59] by maximizing the local linearity of the MOB feature space. On the MOB feature space, we apply the same standard RC protocol introduced in Ref. 59 using k-means cluster initialization [59] and ordinary least square linear regressions (LR) implemented using CuPy [154]. The RC step is fully converged (zero training MAE change between two iterations) to obtain the training clusters. Random forest classification (RFC) with 200 balanced-tress implemented in Scikit-learn [59] is performed to classify the test data. To reduce the cost of training RFC and local regressors for the off-diagonal clusters with a large number of pairs, we adapted the same capping strategy illustrated in Ref. 59 with a capping size of 10,000 for each training off-diagonal cluster during the training over 2000 QM7b-T molecules. There is no capping applied to all diagonal and off-diagonal pairs with training sizes of less than 2000 molecules.

Unsupervised clustering with MOB features

Following the implementation in Scikit-learn, we reimplemented GMM to enable multi-GPU usage using CuPy, which is initialized by k-means clustering and constructed with a full covariance matrix. The objective function of GMM is to maximize the likelihood, which is solved iteratively by the EM algorithm. A regularization of 1e-6 is added to its diagonal terms to ensure the positive definiteness of the covariance matrix of GMM.

The number of clusters K_{best} used in GMM is automatically detected by scanning a series of reasonable cluster sizes and finding the GMM model with the most negative BIC score. According to the previous study in Ref. 59, the optimal numbers of clusters for 1000 training molecules in RC are 20 and 70, respectively. The scanning series of possible K are $\{5i|i = 1,2,...,10\}$ and $\{5i|i = 7,8,...,32\}$ for diagonal and off-diagonal pairs, receptively. Empirical equations for estimating the scanning range of the number of clusters are also presented. We note that this auto-determination procedure is completely unsupervised and does not require any cross-validation from regression.

A hard clustering from GMM assigns the test point to the cluster with the highest probability, and a soft clustering from GMM provides probabilities of the point belonging to each possible cluster. Only a few pairs (under 10%) in QM7b-T can have a second most probable cluster with a probability over 1e-4 (Table 4.3). More details about soft clustering are described in **Appendix**. The current work presents and analyzes the results from hard clustering without specifying any parameters. To demonstrate the smoothness and continuity of GMM clusters constructed using MOB features on the chemical space, the Euclidean distance between the feature vector of each diagonal pair and the corresponding hard cluster center μ_i is computed and analyzed.

MO type determinations

To compare the cluster compositions and the MO compositions in QM7b-T and GDB-13-T, we also apply an algorithm to determine MO types represented by atomic connectivity and bond order for closed-shell molecules following the octet rule. This procedure requires the coordinates of atoms and the centroids of MOs computed using HF information. More details and the pseudo-code of the MO detection algorithm are included in the **Appendix** (Algorithm 2).

Regression within local clusters

Regressions by GPR or LR on top of RC or GMM clustering are used to predict molecular energies. For LR, we use the ordinary least square linear regression with no regularization for diagonal and off-diagonal pairs. To reduce the training cost of local GPRs, as a scalable exact GP algorithm, AltBBMM [62] is performed with Matérn 5/2 kernel with white noise regularization of 1e-5 for both diagonal and off-diagonal pair energies. For the clusters with training points fewer than 10,000, GPR models are directly obtained by minimizing the negative log marginal likelihood objective with the BFGS algorithm until full convergence. For the clusters with more than 10,000 training points, the variance and lengthscale are first optimized using randomly selected 10,000 training points within the cluster, and the Woodbury vector [92] is further solved by the block conjugate gradient method with preconditioner sizes of 10,000 and block sizes of 50.

In order to improve the accuracy and reduce uncertainty, without specifications, the predicted energies are reported as the averages of ten independent runs for all MOB-ML with clustering protocols. We abbreviate the RC then RFC classification and GPR regression as RC/GPR since no other classifier is used with RC. Similarly, RC then RFC classification and LR regression, GMM clustering with GPR regression, and GMM clustering with LR regression are abbreviated as RC/LR, GMM/GPR, and GMM/LR, respectively. The entire workflow on the general framework of MOB-ML with clustering is also introduced in Ref. 59.

4.4 Results and discussions

Number of clusters detected in GMM

Rather than predetermining the number of clusters through pilot experiments [59], GMM automatically selects the most suitable model among the ones with different cluster numbers by finding the lowest BIC score, which prevents overfitting due to a large number of clusters and is faithful to the intrinsic feature space structure in the training set [155]. Figure 4.1 depicts the optimal number of clusters determined by BIC scores as a function of the number of training QM7b-T molecules. The numbers of diagonal and off-diagonal clusters are roughly proportional to the training sizes in a logarithm scale if the training set is larger than 250 molecules, and the best number of clusters can be estimated as functions of the number of training molecules N_{mol} from this set of results as K_d and K_o for diagonal and off-diagonal pairs, respectively.

$$K_d = 0.296 N_{mol}^{0.579}, \tag{4.2}$$

$$K_o = 2.117 N_{mol}^{0.502}. \tag{4.3}$$

These two empirical equations serve as estimation functions to avoid searching an excessive amount of candidate clustering numbers. For the future multi-molecule dataset training, it is sufficient to construct the scanning region of possible K values as $[K_{est} - 10,\ K_{est} - 5,\ K_{est},\ K_{est} + 5,\ K_{est} + 10]$, where K_{est} is the five multiple closest to the estimated value computed by the above empirical equations.

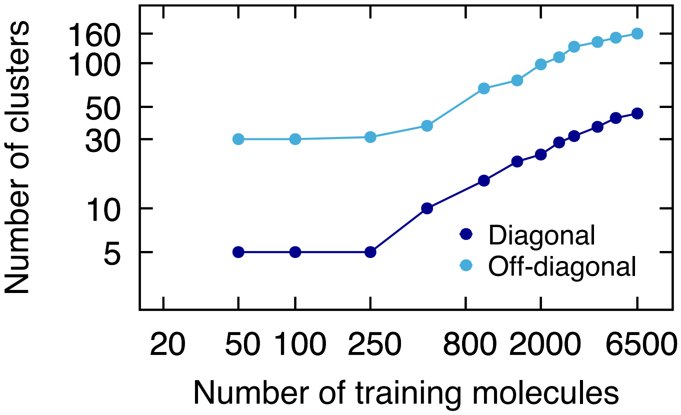

Figure 4.1: Numbers of clusters in GMMs for diagonal and off-diagonal pairs detected by BIC scores. The average number of clusters over ten runs is plotted versus the number of training molecules in QM7b-T on a logarithm scale.

Unsupervised clustering organic chemical space

Chemically intuitive clusters from unsupervised clustering

A specific MO can be one-to-one represented by its diagonal feature space, and thus all the MO analyses are conducted with the clustering results from GMMs trained

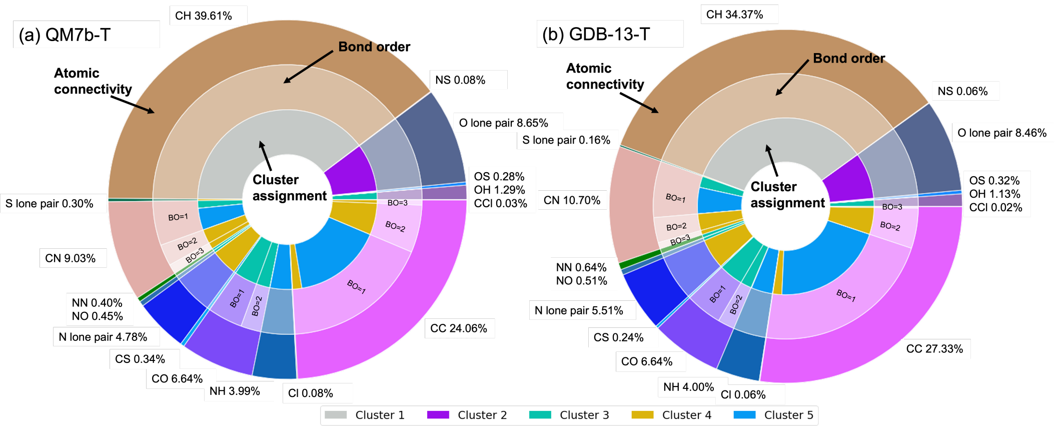

Figure 4.2: MO types and cluster compositions of (a) QM7b-T and (b) GDB-13-T predicted by GMM model trained on diagonal features of 250 QM7b-T molecules (250 GMM model). The layers from outer to inner are the atomic connectivity of MOs, the bond orders (BO) of MOs, and the GMM classification results, respectively. The abundance of each type of atomic connectivity in each dataset is labeled. The BOs are only marked for cases where one type of atomic connection has more than one possible bond order. If the two atoms of a MO can only form a single bond or the MO is a lone pair, the bond order is not listed in the figure.

on diagonal features using different numbers of QM7b-T molecules. The GMM clustering results and MO types are categorized by multiple pie charts layer by layer for QM7b-T and GDB-13-T datasets in Fig. 4.2. The first (outermost) layer depicts the atomic connectivity of a MO, which is further classified by the bond order in the second (intermediate) layer. The third (innermost) layer illustrates the classification results for each type of MO obtained from the diagonal GMM model trained on 250 QM7b-T molecules.

MOB-ML has been shown to be transferable in supervised clustering and regression tasks by creating an interpolation to weak extrapolation tasks between different chemical systems using the MOB representation. [57–60] When the first and second layers of two sets of pie charts in Fig. 4.2a and b are compared, it becomes clear that QM7b-T and GDB-13-T share the same categories of MOs with slightly different abundances. CH MO is the most prevalent MO type in QM7b-T, and its popularity declines as the popularity of other less-trained MOs increases in GDB-13-T. This discovery implies that QM7b-T has the majority of the information necessary to predict the properties of molecules in GDB-13-T with any MO-based representation and any transferable ML approach. The almost identical grouping patterns in the third layers of Fig. 4.2a and b suggest that unsupervised clustering via GMM is transferable as expected. In addition, the cluster assignments match the chemically intuitive groupings for both QM7b-T and GDB-13-T (Fig.4.2a inner layer). Each type of MO is clustered into one type of cluster except CC single bond, while most clusters contain more than one type of MOs. For example, all CC double bonds are clustered into Cluster 2, but Cluster 2 contains CC single, double, and triple bonds, CN double and triple bonds, and lone pairs on the N atom. More training points are required to capture finer clustering patterns in the chemical space using GMM.

We note that while clustering based on intuitive groupings or MO types is theoretically feasible, but not practical as a general approach in MOB-ML to predict molecular properties. It is hard to intuitively define the types of MOs within chemical systems with complicated electronic structures, such as transition states, while MOB features can still represent these MOs. As demonstrated in the first layers in both Fig. 4.2a and b, the organic chemical space is biased heavily towards CH and CC MOs significantly. To avoid tiny clusters and achieve accurate local regression models, a careful design of training sets is also required by including various MO types for clustering based on MO types. In addition, the local regression models with clusters based on MO types cannot predict the properties of a new type of MO without explicitly including it in the training set.

Resolutions of GMM clustering with different training sizes

As the number of training pairs increases, the number of clusters recognized by GMM for diagonal feature space increases from 5 at 250 training molecules to 15 at 1000 training molecules (Fig. 4.1). Figure 4.3a and b compares the clustering patterns predicted by GMMs trained on different training sizes for QM7b-T and GDB-13-T, respectively. In both panels, the layers show the cluster compositions determined by the GMM trained on 250 molecules (250 GMM model) in the outer layer and 1000 molecules (1000 GMM model) in the inner layer. Training on more molecules not only provides more diverse chemical environments for the same type of MOs, but also aids in the resolution of the local structures in the MOB feature space. The MO types with high abundance in QM7b-T and GDB-13-T could be split into multiple clusters. For instance, the one cluster for CH single bond trained on 250 molecules is split into two clusters trained on 1000 molecules. In addition, the MO types with low abundances in QM7b-T and GDB-13-T could be resolved with more training data, rather than mixed into one cluster. For example, CO single bonds and CO double bonds are classified into two clusters by the 1000 GMM model instead of one cluster by the 250 GMM model.

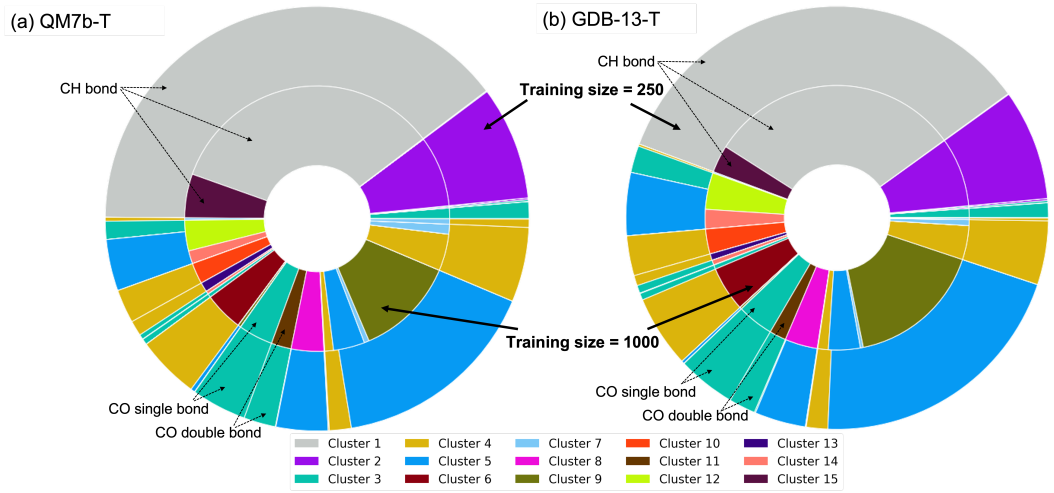

Figure 4.3: Cluster assignments of (a) QM7b-T and (b) GDB-13-T predicted by the diagonal GMM models trained on 250 (250 GMM model) and 1000 (1000 GMM model) QM7b-T molecules. The outer layers display the same clustering results as the most inner layers in Fig. 4.2 predicted by the GMM model trained on 250 molecules with five detected clusters. The inner layers show the clustering results predicted by the GMM model trained on 1000 molecules with 15 detected clusters. In both panels, the clusters in the inner layers further split up the ones in the outer layers. The MO identities of example clusters analyzed in the main text are labeled in the figure as well.

Molecular energy learning by regression with clustering

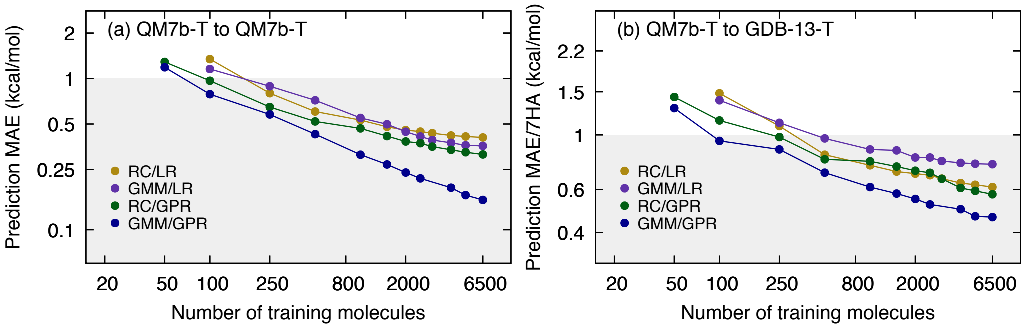

Figure 4.4: Learning curves for MP2/cc-pVTZ energy predictions with different clustering methods trained on QM7b-T and applied to (a) QM7b-T and (b) GDB-13-T. The models are the same ones trained on QM7b-T for (a) and (b). The prediction performance is reported in terms of MAEs and MAE per 7 heavy atoms (7HA) for (a) and (b), respectively, by averaging over ten runs. All the data are plotted on a logarithm scale, and the shaded areas correspond to an MAE/7HA of 1 kcal/mol.

We now present the results of predicting molecular energies utilizing GPR or LR on top of supervised or unsupervised clustering methods in MOB-ML. The RC/LR and GMM/LR training results on 50 molecules are omitted due to the instability of local

LR models with 50 training molecules. The prediction accuracy is assessed by mean absolute error (MAE) of total energies predicted by each MOB-ML model on test sets and plotted as a function of the number of training molecules on a logarithm scale ("learning curve" [95]) in Fig. 4.4. The test sets consist of all remaining QM7b-T thermalized geometries not included in the training sets in Fig.4.4a and all the GDB-13-T thermalized geometries in Fig.4.4b. All the test errors for different training protocols with different training sizes are reported in **Appendix** Table 4.1 and Table 4.2.

Among all four training protocols, GMM/GPR provides the best learning accuracy on QM7b-T and transferability on GDB-13-T. By training on 6500 molecules, GMM/GPR can achieve an MAE of 0.157 kcal/mol and an MAE/7HA of 0.462 kcal/mol for QM7b-T and GDB-13-T, respectively. The performances of all three other approaches are similar on QM7b-T, but RC/GPR and RC/LR have slightly better performance on GDB-13-T than GMM/LR. For the models clustered by RC, LR provides similar accuracy and transferability compared with GPR since RC maximizes the local linearity for each local cluster. While the accuracy loss due to non-linearity of local regression is more significant with GMM clustering with an MAE of 0.202 kcal/mol and an MAE/7HA of 0.298 kcal/mol for QM7b-T and GDB-13-T, respectively, training on 6500 molecules. Although GMM/LR is not as accurate as GMM/GPR, the reasonably accurate predictions from GMM/LR for both QM7b-T and GDB-13-T infer that GMM still can capture local linearity to some degree, despite the fact that GMM is not trained to maximize local linearity. In comparison to GMM/GPR, the learning efficiency of RC/GPR is harmed by the classification errors from RFC for test points[59], and hence RC/GPR provides twice as large errors for QM7b-T and 0.111 kcal/mol worse MAE/7HA for GDB-13-T.

With the same clustering method, GPR is a more accurate local regressor compared with LR and generally offers superior accuracy across all the training sizes. GMM/LR has 1.47 to 2.28 times higher MAEs than GMM/GPR, and RC/GPR also marginally outperforms RC/LR. The chemical accuracy of 1 kcal/mol for test QM7b-T molecules can be reached by training on 100 and 250 training molecules using GPR and LR local regressors, respectively in Fig. 4.4a. In addition, GMM/GPR only requires 100 training molecules to reach the chemical accuracy for GDB-13-T.

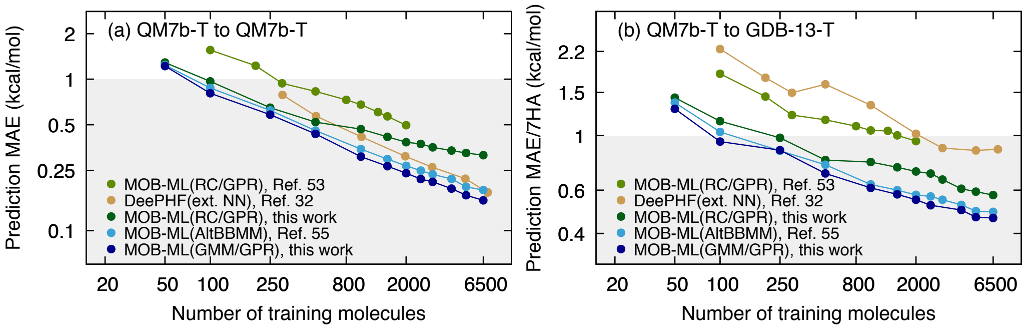

Figure 4.5: Accuracy comparison between different ML methods trained on QM7b-T and tested on (a) QM7b-T and (b) GDB-13-T. The learning curves of RC/GPR and GMM/GPR are the same ones shown in Fig. 4.4. Results from RC/GPR in Ref. 59 were trained on non-size consistent features and therefore different from the ones obtained from RC/GPR in this work. In addition, MOB-ML regressed with AltBBMM (MOB-ML (AltBBMM)) [62] and DeePHF (ext. NN) [134] are also plotted for comparison. All the data are digitally extracted from the corresponding studies and plotted on a logarithm scale. The shaded area corresponds to the chemical accuracy of 1 kcal/mol.

Comparison with molecular energy learning results from literature

In Fig.4.5, the learning curves of RC/GPR and GMM/GPR in this study are further compared to those of state-of-the-art methods in literature trained on randomly selected QM7b-T molecules, including MOB-ML regressed with RC/GPR using outdated MOB features from Ref. 59(MOB-ML (RC/GPR)), DeePHF trained with an NN regressor [134] (DeePHF (ext. NN)), and MOB-ML regressed with AltBBMM (MOB-ML (AltBBMM)) [62].

The introduction of the most recent improved MOB features [60] considerably enhances the accuracy of MOB-ML, and therefore RC/GPR from this work is more accurate than the literature RC/GPR with outdated features. Training on the best available MOB features leads to over around 30 % accuracy improvements with RC/GPR on both QM7b-T and GDB-13-T test molecules. This observation suggests that better feature engineering not only can improve the accuracy for GPR without clustering [60], but also can enhance the efficiency of regression with clusters. GMM/GPR achieves slightly higher prediction accuracy than AltBBMM without clustering when using the same MOB feature design, indicating that an additional GMM clustering step prior to regression benefits the entire training process by replacing the global regression model with more accurate local regression models.

As another ML framework to predict molecular energies at HF cost, DeePHF [134] also achieves accurate predictions for QM7b-T, while its transferability on GDB-13-T is less than MOB-ML [60]. GMM/GPR training on 6500 molecules (MAE=0.157 kcal/mol) outperforms the best DeePHF model training on 7000 molecules (MAE=0.159 kcal/mol) on the total energies of QM7b-T test molecules. Without sacrificing transferability on GDB-13-T, the best model from GMM/GPR can achieve half of the error from DeePHF on GDB-13-T and become the most accurate model for molecular energies in GDB-13-T.

Efficient learning by local AltBBMM with GMM clustering

Due to the limited resources of GPUs, we report the results and timing of single-run regression with clustering in this section. Figure 4.6 plots the test MAEs of QM7b-T and GDB-13-T from single-run models as a function of parallelized training time on 8 NVIDIA Tesla V100-SXM2-32GB GPUs for three most accurate MOB-ML training protocols. GMM/GPR provides slightly improved accuracy and transferability compared to direct regression by AltBBMM without clustering and significantly reduces the training time of MOB-ML by 10.4 folds with 6500 training molecules. As the most cost-efficient and accurate training protocol for MOB-ML, a single run of GMM/GPR only requires 2170.4s wall-clock time to train the best model with 6500 molecules.

We note that the computational costs of GMM and local AltBBMM in GMM/GPR are comparable and lower than AltBBMM without clustering. The complexity analysis is as follows. The training complexity of GMM of each EM iteration is $\mathcal{O}(NK)$ with a fixed number of features[156], where N is the number of training points and scales linearly with N_{mol}, and K is the number of clusters. Local AltBBMM has a training complexity of $\mathcal{O}(KN_{loc}^2)$, where N_{loc} is the number of training points in each local cluster[92]. Since N_{loc} roughly scales as $\mathcal{O}(N/K)$, the complexity of local AltBBMM in GMM/GPR can be approximated as $\mathcal{O}(N^2/K)$. Therefore, GMM becomes the computational bottleneck in GMM/GPR when K grows faster than $N^{0.5}$; otherwise, local AltBBMM is more expensive than GMM. As discussed in Sec. 4.4, the optimal K_d and K_o for QM7b-T are fitted as functions of N with an approximate scaling of $\mathcal{O}(N^{0.579})$ and $\mathcal{O}(N^{0.502})$, respectively. GMM and local AltBBMM share similar computational costs in this case, and the overall complexity of GMM/GPR using local AltBBMM is around $\mathcal{O}(N^{1.58})$, which is lower than AltBBMM without clustering. Therefore, GMM/GPR is no longer the computational bottleneck in training and is able to scale MOB-ML to train more than 1 million data.

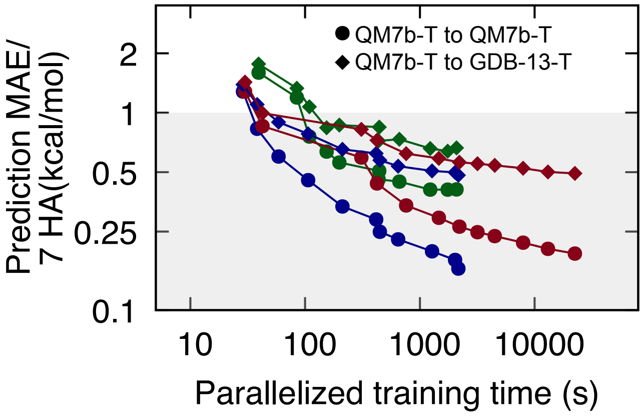

Figure 4.6: Accuracy and training costs of MP2/cc-pVTZ energy using single-run GPR with RC and GMM clustering (RC/GPR/single-run, green; GMM/GPR/single-run, blue) and AltBBMM without clustering (red, Ref. 62). Prediction MAEs of test QM7b-T (circles) and GDB-13-T (diamonds) from single runs are plotted as a function of wall-clock training time with parallelization on 8 NVIDIA Tesla V100 GPUs on a log scale. The models are the same as the ones reported in Fig. 4.4 and the corresponding training sizes of QM7b-T are labeled in the figure. The shaded areas correspond to an MAE/7HA of 1 kcal/mol.

4.5 Conclusion

We extend our previous work on supervised clustering to unsupervised clustering the organic chemical space with the improved MOB features, and introduce an accurate, efficient, and transferable regression with clustering scheme to learn molecular energies of QM7b-T and GDB-13-T. Without specifying the number of clusters ahead, unsupervised clustering via Gaussian mixture model (GMM) is fully blackboxed and able to cluster the organic chemical space represented by QM7b-T and GDB-13-T in ways consistent with the chemically intuitive groupings of MO types. As the amount of training data increases, the finer grouping patterns of MOB feature space are captured, and the resulting clusters are gradually separated following the chemical intuition. As the most efficient training protocol for MOB-ML, GMM/GPR surpasses RC/GPR and AltBBMM without clustering in prediction accuracy and transferability with a training cost at a tenth of the one of AltBBM without clustering. GMM/GPR not only reaches the chemical accuracy for QM7b-T and GDB-13-T by only training on 100 QM7b-T molecules, but also offers superior performance to all other state-of-the-art ML methods in literature with an MAE of 0.157 kcal/mol for QM7b-T and an MAE/7HA of 0.462 kcal/mol for GDB-13-T. We finally illustrate that the overall complexity of GMM/GPR is lower than AltBBMM without clustering and local AltBBMM regression is no longer the computational bottleneck in GMM/GPR. As a future direction, it is promising to apply GMM/GPR to even larger datasets with more diverse chemistry due to its low complexity. The unsupervised nature of GMM also opens an avenue to regress other molecular properties with MOB features by GMM/GPR.

4.6 Appendix

MO type determination

The raw atomic connectivity of each MO is identified by searching the two atoms which have the smallest euclidean distances to the centroid of the corresponding MO. For each MO, we assume that its final atomic connectivity can only be two cases. If this MO is a bond, then two selected atoms are connected; and if this MO is a lone pair, it only belongs to the atom with smaller distance to its centroid. To judge the MO identity (a bond or a lone pair), we define "atom-bond angle", i.e., $\angle ACB$, where C is the centroid position of this bond, and A, B are its two nearest atoms. Ideally, the center of the bond between two atoms should be collinear with these two atoms, i.e. the atom-bond angle is 180°. The final atomic connectivity is determined by iteratively classifying the MOs with small atom-bond angles as lone pairs until all atoms satisfying the octet rule in chemistry (details see **Appendix**). The bond order of each MO is computed by the number of bonds between the two corresponding detected atoms.

Algorithm 2 states the details of determination process of MO types of a closed-shell molecule using the MO centroid coordinates $\{M_1, ..., M_N\} \in \mathbb{R}^3$ and the atom coordinates $\{A_1, ..., A_n\} \in \mathbb{R}^3$. Additionally, for each atom a_i with certain number of connected bonds, we define $\hat{S}_i$ as the expected number of bonds connected to each atom i (i.e. 1 for H, 2 for O, 3 for N, 4 for C, 1 for Cl, and not defined for S) also as part of the algorithm input. The output of this algorithm is the atomic connectivity represented as tuple $(I_{k,1}, I_{k,2}, BO_k)$ for each MO k, where $I_{k,1}$, $I_{k,2}$ are the two connected atoms of the MO k, and BO_k is its bond order.

For each MO k, a boolean variable T_k is introduced to determine if the MO k is a bond (B) or a lone pair (L). We initialize the atomic connectivity of the MO k as the atoms α_k and β_k, which equal to the indices of the first and second smallest elements in $\{D_k^i | i = 1, ..., n\}$, where D_k^i is the euclidean distance between M_k and A_i. We define the atom-bond angle of the MO k as $\theta_k = \angle \alpha_k M_k \beta_k$, which tends to be large for bond because it is 180° in the ideal case. For the MO k, we initialize $T_k = L$ if $\theta_k < 72°$, and $T_k = B$ if $\theta_k > 72°$, because 72° is small enough that, for any MO k, $\theta_k < 72^\circ$ guarantees $T_k = L$. The number of the bonds connected to each non-sulfur atom i, i.e., S_i, is computed. We iteratively converge $\{S_i\}$ by decreasing the values until $S_i = \hat{S}_i$ for each non-sulfur i (which leads to "success"), or there is at least one $S_i < \hat{S}_i$ so that $\{S_i\}$ is no longer possible to agree with $\{\hat{S}_i\}$ (which leads to "failure"). We note that sulfur is not checked because it is more complicated than the rest types of atoms. In each iteration, a atom u satisfying $S_u > \hat{S}_u$ is selected randomly, and then we find the set of bonds connected to u. We change T_p with the smallest θ_p in this set to lone pair, i.e., $T_p = L$, and update S_{α_p} and S_{β_p} by decreasing one. After the iteration finishes, each T_k has been successfully determined, so we can now determine $(I_{k,1}, I_{k,2})$ as (α_k, β_k) if $T_k = B$, or as (α_k, None) if $T_k = L$. Finally, for each MO k that $T_k = B$, the bond order BO_k can be determined by the number of bonds having the same unordered pair (α_k, β_k) with it, which finishes the algorithm.

Since randomness is introduced in the algorithm, we repeat the algorithm several times until success or it fails more then 10 times so that we believe a solution cannot be found. In this work, all the MO types of 99.9% of QM7b-T and 98.7% of GDB-13-T molecules have been successfully recognized without any contradictions to the octet rule. The molecules with at least one atom violating the octet rule are excluded only in the analysis of unsupervised clustering on organic chemical space but included in the energy predictions.

Molecular energy learning with k-means and GMM clusters with single run

In the main text, learning curves of molecular energies obtained from four regression with clustering protocols are plotted and investigated. In fact, other unsupervised clustering method like k-means can also work well in this general framework for energy learning. Here we include learning curves of single run with k-means and GMM unsupervised clusters using GPR as regressors in Fig. 4.7. The number of clusters in k-means is auto-detected by Davies-Bouldin index, which compares distance between clusters with the size of the clusters themselves[94]. Although averaging over 10 independent runs offers slightly better prediction accuracy for GMM/GPR, the results from single run GMM/GPR (GMM/GPR/1X) only have little accuracy loss. GMM is considered as a general and better clustering and classifier method than k-means. In this application, k-means is found to be a reasonably good choice for clustering and classification. The prediction accuracies of K-means/GPR/1X is only at most 36.57% and 10.30% worse for QM7b-T and GDB-13-T than the ones of GMM/GPR/1X, respectively.

Algorithm 2 MO type determination algorithm

```
Input: MO coordinates $M_1, \dots, M_N$, atoms coordinates $A_1, \dots, A_n$, expected number of bonds for atoms $\hat{S}_1, ..., \hat{S}_n$
Output: Success or Failure of the process, a set of MO type descriptors $\{(I_{k,1}, I_{k,2}, BO_k)\}_{k=1,...,N}$ if the process is successful
for $k \leftarrow 1$ to $N$ do
    Compute $\{D_k^i\}_{i=1,\ ...,\ n}$
    $\alpha_k, \beta_k \leftarrow$ First two $i$ sorted by increasing order of $\{D_k^i\}_{i=1,\ ...,\ n}$
    if $\theta_k < 72°$ then
        $T_k \leftarrow$ L
    else
        $T_k \leftarrow$ B    ▷ Temporarily classify them to be bonds, may be changed to lone pair later
    end if
end for
Compute $S_1,\ ...,\ S_n$ from $T_1,\ ...,\ T_N$    ▷ Initialization of $\{S_i\}$
while $(\exists S_i > \hat{S}_i)$ do
    if $(S_i > \hat{S}_i)$ then
        return Failure
    end if
    randomly pick $u \in \{i | S_i > \hat{S}_i\}$
    $p \leftarrow \arg\max_{k} \theta_k$, subject to $T_k = B, u \in \{\alpha_k,\ \beta_k\}$
    $(T_p,\ S_{\alpha_p},\ S_{\beta_p}) \leftarrow (L, S_{\alpha_p} - 1,\ S_{\beta_p} - 1)$
end while
for $k \leftarrow 1$ to $N$ do
    if $T_k = B$ then
        $I_{k,1} \leftarrow \alpha_k,\ I_{k,2} \leftarrow \beta_k,\ BO_k = \#\{j | \{\alpha_j, \beta_j\} = \{\alpha_k, \beta_k\},\ T_j = B\}$    ▷ Unordered pair, i.e. $\{a, b\} = \{b, a\}$
    else
        $I_{k,1} \leftarrow \alpha_k,\ I_{k,2} \leftarrow$ None, $BO_k = 0$
    end if
end for
return Success, $\{(I_{k,1}, I_{k,2}, BO_k)\}_{k=1,...,N}$
```

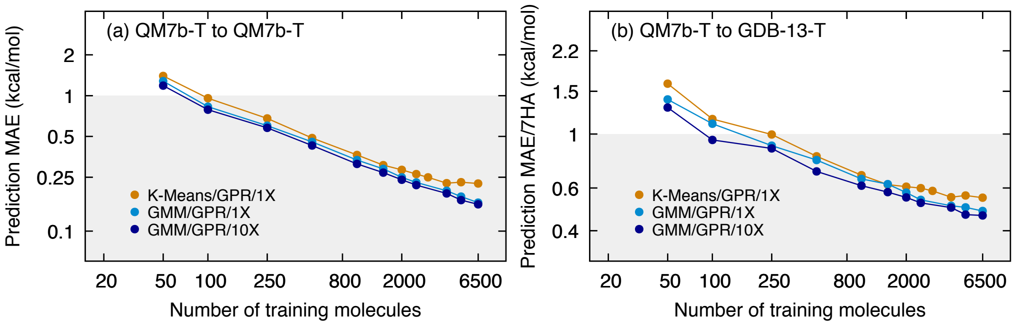

Figure 4.7: Learning curves for MP2/cc-pVTZ energy predictions with different clustering methods trained on QM7b-T and applied to (a) QM7b-T and (b) GDB-13-T. The MAEs of GMM/GPR are also plotted for comparison. All the data are plotted on a logarithmic scale, and the shaded areas correspond to an MAE/7HA of 1 kcal/mol.

Molecular energy learning with unsupervised clustering

In Table 4.1 and 4.2, we summarize the MAEs of molecular energies in kcal/mol using different clustering-then-regression protocols plotted in Fig. 4.4 in the main text for QM7b-T and GDB-13-T, respectively.

Table 4.1: MOB-ML prediction accuracy (kcal/mol) for four regression with clustering protocols applied to MPC/cc-pVTZ energies of QM7b-T. The training and testing sets corresponding to non-overlapping subsets of QM7b-T.

Training sizes	GMM/GPR/10X	RC/GPR/10X	GMM/LR/10X	RC/LR/10X
50	1.187	1.289	–	–
100	0.788	0.968	1.156	1.344
250	0.579	0.648	0.889	0.800
500	0.429	0.520	0.718	0.605
1000	0.313	0.468	0.549	0.531
1500	0.270	0.416	0.499	0.479
2000	0.239	0.383	0.445	0.456
2500	0.219	0.373	0.414	0.446
4000	0.189	0.338	0.375	0.420
5000	0.169	0.325	0.362	0.413
6500	0.157	0.315	0.359	0.407

Comparison between training costs of supervised clustering and unsupervised clustering

Figure 4.8 plots the wall-clock timing of RC and GMM clustering on 8 NVIDIA Tesla V100-SXM2-32GB GPUs as functions of number of training sizes. The costs of RC and GMM are similar across all the training sizes.

Table 4.2: MOB-ML prediction accuracy (kcal/mol) for four regression with clustering protocols applied to MPC/cc-pVTZ energies of GDB-13-T. Models are the same as the ones in Table 4.1

Training sizes	GMM/GPR/10X	RC/GPR/10X	GMM/LR/10X	RC/LR/10X
50	1.286	1.428	–	–
100	0.945	1.145	1.383	1.477
250	0.873	0.980	1.120	1.086
500	0.702	0.795	0.968	0.830
1000	0.613	0.781	0.873	0.752
1500	0.577	0.743	0.865	0.709
2000	0.549	0.715	0.809	0.696
2500	0.521	0.701	0.809	0.685
4000	0.498	0.608	0.770	0.637
5000	0.466	0.591	0.763	0.626
6500	0.462	0.573	0.760	0.613

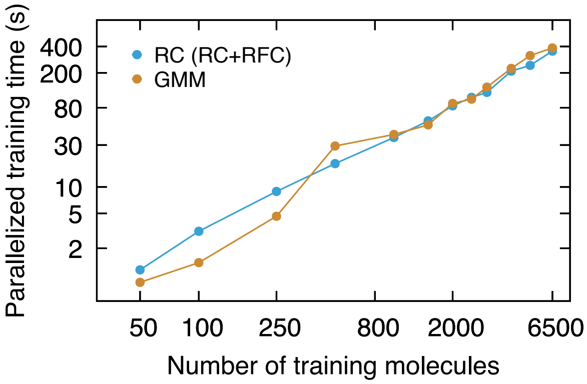

Figure 4.8: Wall-clock timings for RC+RFC and GMM for different number of training molecules from the QM7b-T set with 8 NVIDIA Tesla V100 GPUs. The LR regression of each cluster is performed independently on a different core in the RC step and RFC is trained using parallel implementation of SCIKIT-LEARN. GMM is trained using parallel implementation of EM algorithm.

Soft clustering from GMM

GMM provides probabilities of every possible cluster for a test point, the predictions of molecular energies with soft clustering are possible. The prediction $\varepsilon^{ML}_{ij,soft}$ of each pair energy is a weighted average of the predictions $\varepsilon^{ML}_{ij,n}[\mathbf{f}_{ij}]$ from all K possible clusters evaluated as Eq. 4.4.

$$\varepsilon^{ML}_{ij,soft} = \sum_{n=1}^{K} P_n[\mathbf{f}_{ij}] * \varepsilon^{ML}_{ij,n}[\mathbf{f}_{ij}], \tag{4.4}$$

where $\mathbf{f}_{ij}$ is the set of MOB features for pair ij, $n = 1, 2, ..., K$ is the cluster ID; and P_n is the corresponding probability of the pair ij being classified into cluster n satisfying $\sum_{n=1}^{K} P_n = 1$.

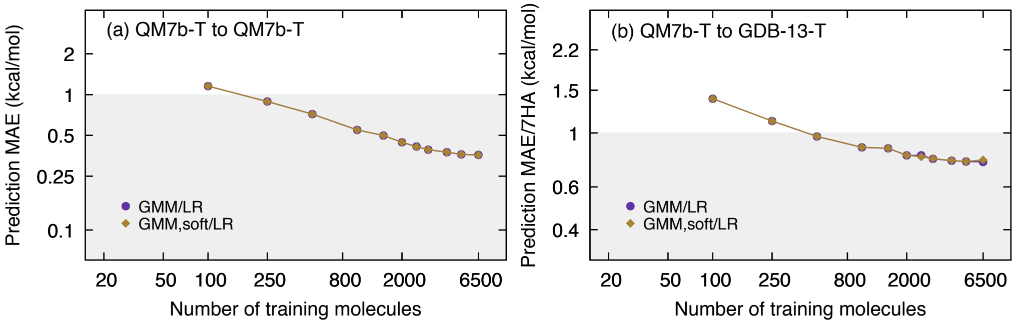

Figure 4.9: Comparison between hard clustering and soft clustering on molecular energy learning regressed by LR for (a) QM7b-T and (b) GDB-13-T. The results from hard clustering (purple circles) and the ones from soft clustering (dark gold diamond) overlap with each other well. All the data are plotted on a logarithmic scale, and the shaded areas correspond to an MAE/7HA of 1 kcal/mol.

Figure 4.9 displays the comparison of the prediction accuracies of molecular energies regressed by LR on top of hard and soft clustering from the same set of GMMs for QM7b-T and GDB-13-T. In both panels, the results from soft clustering method (GMM,soft/LR) overlap with the ones from hard clustering (GMM/LR) with accuracy differences smaller than 0.003 kcal/mol, which suggests that soft clustering does not provide any extra benefits in this application. Since we create an interpolation to weak extrapolation problem, the cluster identities of the tests points are therefore unambiguous. Table 4.3 shows the percentages of pairs that have more than one clusters with probability higher than 0.0001, namely, how many pairs can be influenced by adapting soft clustering method during the predictions of energies. For QM7b-T, under 10% of pairs in both diagonal and off-diagonal feature spaces have more than one possible clusters. The numbers of pairs with more than one possible cluster identities increase for GDB-13-T, but are still not significant enough to change the predicted energies in Fig. 4.9. We note that soft clustering from GMM might provide some accuracy improvements in some future applications.

Table 4.3: Percentages of pairs in QM7b-T and GDB-13-T having n number of clusters that their predicted probability over 0.0001 by GMMs

Pair type	GMM training size	QM7b-T		GDB-13-T	
		n =2	n≥3	n =2	n≥3
Diagonal	250	0.37%	0	0.69%	0
	1000	6.29%	0	8.83%	0
Off-diagonal	250	8.74%	0.83%	20.42%	1.85%
	1000	6.40%	1.21%	10.63%	3.99%

Chapter 5

NEAR AB INITIO POTENTIAL ENERGY SURFACES FOR DIFFUSION MONTE CARLO USING MOB-ML AND NEURAL NETWORK REFITTED MOB-ML

Adapted from:

1. DiRisio, R. J., Cheng, L., Boyer, M. A., Lu, F., Sun, J., Lee, S. J. R., Deustua, J. E., Miller III, T. F. & McCoy, A. B. Near *ab Initio* potential energy surfaces for diffusion Monte Carlo using machine learning[1]. *In preparation.*

Reprinted with permission from all other authors.

The molecular-orbital-based machine learning (MOB-ML) method provides an accurate, efficient, and general approach for obtaining high-level electronic energies at the same cost as a Hartree-Fock calculation for an arbitrary molecular system. In this chapter, we explore the applicability of the MOB-ML method for the generation of potential energy surfaces suitable for the computation of ground- and vibrational excited-state wavefunctions and energies. Specifically, we use small-scale diffusion Monte Carlo (DMC) simulations to evaluate the zero-point energies of H_2O and CH_5^+. In the case of H_2O, we also calculate vibrationally excited state energies. To facilitate the larger DMC simulations, a Neural Network (NN) is trained on MOB-ML energies. The resulting MOB-ML-based NN-DMC method allows us to take advantage of GPU accelerated energy evaluations, with which we can perform large-scale DMC simulations on both systems as well as the five deuterated isotopologues of CH_5^+. For CH_5^+, we achieve excellent agreement with the the ground state energies and probability amplitudes obtained using a potential surface that had been fit to more than two orders of magnitude more CCSD(T) energies than were used to train the MOB-ML potential.

[1] RJD and LC contributed equally to this work

5.1 Introduction

Quantum descriptions of molecular vibrations require an accurate representation of the potential energy surface (PES) for the system of interest. For systems where the vibrational ground state is localized near the potential minimum, and which undergo small amplitude vibrational motions, harmonic treatments of the potential may be sufficient.[157] Such a description can be readily achieved at a broad range of levels of electronic structure theory and bases using electronic structure packages, as all that is required is the optimized geometry and Hessian. Significant insights may also be obtained from quartic expansions of the potential about the minimum, as this forms the basis for second-order perturbation theory calculations.[158] Unfortunately, there are many problems for which such low-order expansions of the potential are insufficient, and it becomes desirable to be able to evaluate the potential at arbitrary molecular configurations.

A common strategy for developing potentials for molecular spectroscopy, quantum dynamics or other non-local quantum applications is to evaluate the electronic energies over a broad range of geometries, and fit this data to a potential function. This has often involved fitting the electronic energies to functional forms that reflect the expected physics.[159–163] Consider, for example, the two systems explored in the present study. Partridge and Schwenke fit a potential surface for H_2O based on 1056 ICMRCI energies, and adjusted the parameters to include Born-Oppenheimer corrections and match experimental data[159]. This surface will be referred to as the PS potential in the remainder of this paper. Jin, Braams and Bowman (**JBB**) fit a potential surface for CH_5^+ to more than 35 000 energies, which were evaluated at the CCSD(T)/aug-cc-pVTZ level of theory/basis. To allow the ion to properly dissociate to CH_3^+ + H_2, they extrapolated their fit surface to long range by splicing a long-range CH_3^+ + H_2 potential onto the fit surface using a switching function in the CH_3^+-H_2 distance.[164]

More recently, ML approaches have been successfully used to fit potential surfaces to calculated electronic energies[165–167]. Many studies provide potential energy surfaces at the level of DFT with classical force field costs atom- and geometry-specific representations [168–172] have received much attention. Several alternative ML approaches adapting quantum level representations are proposed to reproduce energies from highly accurate electronic structure methods beyond DFT[52, 53, 110], for instance MP[5] and CC[173]. As one of the representative approaches to provide highly accurate electronic structure energies with HF computations, the

MOB-ML approach[57–60] allows for the prediction of high-level post-HF correlation energies, such as those provided by CC calculations. By exploiting localized molecular orbitals obtained from HF calculations, MOB-ML is able to reproduce highly accurate potential energy surfaces at a fraction of the cost incurred by the target high-level methods, even when small data sets are employed during training of the MOB-ML models [57, 58, 60]; meanwhile, other ML methods require hundreds of thousands of molecules to reach the same level of accuracy [61].

In the present study, we explore the accuracy and utility of the MOB-ML approach to generate potential energy surfaces for use in diffusion Monte Carlo (DMC)[174–177] calculations. Even though MOB-ML replaces the N^7 scaling of CCSD(T) with the N^3 scaling of HF, which already represents substantial computational savings, a typical DMC calculation requires on the order of 10^8 single-point energy evaluations. This becomes a computationally demanding task even with the N^3 scaling of HF calculations. To make this study feasible, we have developed an implementation of DMC for use in high-performance computing (HPC) environments, as well as a GPU-based NN regression scheme to generate potential energy surfaces for use in the DMC calculations of ground-state wavefunctions and zero-point energies.[178] In the remainder of the discussion, we will refer to this approach as NN-DMC. To explore the efficacy of the combined MOB-ML/NN-DMC approach results will be compared to those obtained using the well-established PS and JBB potential surfaces for H_2O and CH_5^+, respectively. These are two systems that we previously studied using the NN-DMC approach,[178] however the computational cost of the MOB-ML approach leads to challenges in the collection of training data compared to previous studies. As such, we will examine the possibility of using the previously outlined neural network training procedure in the context of the MOB-ML approach. The water monomer provides an example of a molecule that is straightforward to study by a variety of approaches, but where the relatively large amplitude OH stretching vibrations sample into the dissociative part of the potential even in the ground vibrational state. CH_5^+, on the other hand, is a ion that undergoes large amplitude vibrational motions in its ground vibrational state. In fact, the ground state wavefunction has been shown to have comparable amplitude at the 120 equivalent minima on the potential and the 180 low-energy saddle points that connect these minima.[179]

5.2 Theory and methods

Molecular-Orbital-Based Machine Learning (MOB-ML)

MOB-ML is a method for accurately predicting high-level molecular energies, such as those provided by CC, MP, and other wave-function-based electronic structure theories, by using only molecular orbital information obtained from HF computations with much reduced costs. The main idea behind MOB-ML is rooted in Nesbet's theorem (Eq. 5.1),[7, 70] which ensures that the correlation energy of an N-electron system, E^{corr}, can always be expressed as the sum over energy contributions comprising pairs of occupied orbitals

$$E^{\mathrm{corr}} = \sum_{ij}^{\mathrm{occ}} \varepsilon_{ij}. \tag{5.1}$$

The pair energies, ε_{ij}, take various functional forms, which can be readily defined for the specific electronic structure theory of choice, such as CCSD(T) or MP2.[60] Indeed, computing pair energies is of ten times computationally intractable since the high-order polynomial costs associated with CC, MP, and other theories far exceed HF CPU time steps. MOB-ML is designed to alleviate this issue by approximating the pair energy contributions via the general ML mapping

$$\varepsilon_{ij} \approx \varepsilon\left[\{\phi_p\}^{ij}\right], \tag{5.2}$$

which associates pair energies to MOs directly, bypassing high-level calculations altogether.

This general MOB-ML approach can be imbued with any particular ML methodology to define the mapping and trained to approximate energies of virtually any wave-function-based electronic structure method. We employ GPR to fit pair energies computed at the CCSD(T) level of theory. We do this by first subdividing Eq. 5.2 into diagonal and off-diagonal contributions

$$\varepsilon_{ij} \approx \begin{cases} \varepsilon_{\mathrm{d}}^{\mathrm{ML}}\left[\mathbf{f}_i\right] & \text{if } i = j \\ \varepsilon_{\mathrm{o}}^{\mathrm{ML}}\left[\mathbf{f}_{ij}\right] & \text{if } i \neq j, \end{cases} \tag{5.3}$$

which separates the different character of both types of pair energies and improves on the accuracy of the machine-learned models. The feature vectors $\mathbf{f}_i$ and $\mathbf{f}_{ij}$ are constructed from consistently ordered Fock, Coulomb, and exchange interaction matrix elements using localized HF molecular orbitals. We employ the IBO or Boys localization procedures to guarantee the transferability of MOB-ML models across different chemical systems and conformations [57, 58, 60]. We note there are several key invariances and physical properties that MOB-ML satisfies:

- Rotational invariance: ensured by HF.
- Translational invariance: ensured by HF.
- Atom index permutation invariance ($\mathbf{f}_{ij}(\{A_\mu\}) = \mathbf{f}_{ij}(P\{A_\mu\})$: ensured by HF.
- Pair energy index symmetry ($\varepsilon_{ij} = \varepsilon_{ji}$): ensured by LMO symmetrization [58].
- Orbital index permutation invariance ($\mathbf{f}_{ij}(\{\phi_p\}) = \mathbf{f}_{ij}(P\{\phi_k\})$, ensured by appropriate ordering of the feature vector elements. [60]
- Size-consistency ($E(A+B) = E(A) + E(B)$ for non-interacting subsystems A and B): ensured by the careful design of MOB-ML features to satisfy the corresponding properties in the long-distance limit. [60]

In this study, we use the same feature generation protocol described in Ref. 60 to ensure these invariances, the correct physical limit of the features and consistent feature ordering. The details of MOB feature designs have been fully described in our previous studies. [57, 58, 60]

Diffusion Monte Carlo (DMC)

DMC and the details of our implementation have been described elsewhere.[175–177, 180–182] In this study, we use both guided and unguided DMC simulations to obtain the ground state energy and wavefunction for the systems of interest. We also use our recently developed NN-DMC approach,[178] in which we replace the potential energy surface with a neural network potential energy surface for the DMC simulation. This results in a significant savings in the computational resources required for the DMC calculations.

In an unguided DMC simulation, the ground state wavefunction, Φ_0, is represented by an ensemble of N_w localized functions, which we will refer to as walkers. The density of walkers in a particular region of configuration space provides the amplitude of the ground state wavefunction at that geometry. The ensemble of walkers explores the potential energy surface of the system of interest through a propagation in imaginary time, $\tau = it/\hbar$, based on the imaginary-time time-dependent Schrödinger equation,

$$\begin{aligned} \Phi_0(\tau+\Delta\tau) &= \exp\left[-(H - V_\mathrm{ref}(\tau))\Delta\tau\right]\Phi_0(\tau) \\ &\approx \exp\left[-\{V(\mathbf{x}_i(\tau)) - V_\mathrm{ref}(\tau)\}\Delta\tau\right]\exp\left[-T\Delta\tau\right]\Phi_0(\tau). \end{aligned} \quad (5.4)$$

At each time step, $\Delta\tau$, the position, $\mathbf{x}_i(\tau)$, and weight, $w_i(\tau)$, of each of the walkers are updated. Specifically, the coordinates of each of the atoms that are described by the walkers are displaced according to a Gaussian distribution, with a standard deviation of $\sqrt{\Delta\tau/m_j}$, where m_j is the mass of the atom that is displaced. The weight of the ith walker is updated based on

$$w_i(\tau+\Delta\tau) = \exp\left[-\{V(\mathbf{x}_i(\tau)) - V_{\text{ref}}(\tau)\}\Delta\tau\right] w_i(\tau) \tag{5.5}$$

To ensure that a small fraction of walkers do not carry most of the weight, a branching step is introduced. In this step, the weights of the walkers are compared to upper and lower bound thresholds. All walkers with weights that are smaller than the lower bound threshold are removed from the ensemble. To keep the ensemble size and sum of the weights constant, an equal number of walkers with the highest weight are duplicated, and each of the walkers and their copies are given a weight that is half the original weight of the duplicated walker. After all the low-weight walkers have been removed from the ensemble, walkers that have a weight larger than the upper bound threshold are also duplicated, as described above, and an equal number of walkers with the lowest weights are removed from the simulation. For the NN-DMC simulations, the weights of all walkers in the simulation are constrained to 1, and the duplication or removal of walkers is achieved by an additional Monte Carlo step.[183] In this case, the ensemble size will fluctuate as the simulation progresses. This technique is referred to as discrete weighting, and the algorithm that allows the weights of the walkers to evolve with τ is called continuous weighting.

Next,

$$V_{\text{ref}}(\tau) = \frac{\sum_{i=1}^{N_{\text{w}}} w_i(\tau)V(\mathbf{x}_i(\tau))}{\sum_{i=1}^{N_{\text{W}}} w_i(\tau)} - \alpha\left[\frac{\sum_{i=1}^{N_{\text{w}}} w_i(\tau) - w_i(\tau=0)}{\sum_{i=1}^{N_{\text{w}}} w_i(\tau=0)}\right] \tag{5.6}$$

is evaluated, where $\alpha = 0.5/\Delta\tau$. The introduction of the second term in Eq.5.6 ensures the sum of the weights of the walkers is roughly constant throughout the simulation. The time averaged value of V_{ref} provides the zero-point energy of the system once the simulation has equilibrated.

The main difference between guided DMC simulations and unguided DMC simulations is that in the guided simulations, $f = \Phi_0\Psi_T$ is represented by the ensemble of walkers, where Ψ_T is the guiding function. This change leads to the potential energy evaluations being replaced by evaluations of the local energy,

$$E_\mathrm{L} = \frac{H\Psi_T}{\Psi_T}. \tag{5.7}$$

When Ψ_T provides a good approximation to Φ_0, the local energy is approximately constant. Using a guiding function also introduces a drift term that moves the walkers away from regions where the amplitude of Ψ_T is small and towards regions with large amplitude.

In several recent studies, we showed that using guiding functions that are direct products of one-dimensional wavefunctions in the high frequency stretches and, in the case of H_2O, the HOH bend, provide effective guiding functions for H_2O and CH_5^+ [181, 182]. Finally, descendant weighting is used to obtain projections of the probability amplitude onto coordinates of interest.[177, 180, 184] The unguided NN-DMC simulations were performed using PyVibDMC, a general-purpose, open source simulation package.[185]

Diffusion Monte Carlo in High-Performance Computing Environments

The MOB-ML surfaces provide CCSD(T)-quality potential energy evaluations in a computationally efficient way by reducing the cost of a single point energy calculation to effectively that of a HF calculation[57, 58, 60]. Unfortunately, DMC can require tens of millions of potential energy evaluations per simulation, which means that even for relatively simple systems, performing that many HF calculations on the fly still requires a considerable amount of computational resources. To make these calculations computationally tractable, we adapted the DMC procedure for high-performance computing (HPC) environments through a hybrid MPI/threading parallelism paradigm. In this approach, the DMC calculation is run in parallel over a restricted number of MPI jobs, usually equal to the number of compute nodes available to the calculations. Because a constant simulation size simplifies the MPI communication, these DMC calculations are run using the continuous weighting scheme described above. Each MPI job then has access to a number of threads to parallelize potential calls. The threading is handled either through Intel's Threaded Build Blocks (TBB) library or OpenMP. The MOB-ML surfaces used in the study were accessed through the ENTOS QCORE software package.[153]

To minimize the effects of process-to-process communication latency in the potential evaluations and to improve load balancing, we introduced a small variation to the continuous weighting DMC algorithm described above. In this modified approach, we propagate the coordinates and evaluate the potential energy of each of the walkers for N_τ steps before considering branching. Once the N_τ potential evaluations are complete, we update V_{ref}, the weights, and perform branching as necessary at each time step. This introduces an approximation into the DMC algorithm. Although we check for branching after each time step in the simulation, the branching is only applied every N_τ steps. For the purposes of this study, such an approach does not impact the overall accuracy of the DMC simulation, as typically fewer than 0.5% of the walkers undergo branching at each time step. By performing a smaller number of total MPI calls, we are able to cut down on the latency overhead involved in node-to-node communication. Additionally, both TBB and OpenMP have schedulers that can start a new potential evaluation the moment a previous one finishes. This improves load balancing, because some geometries require more time to complete the HF calculations. With threading, multiple faster evaluations can be completed while a more expensive one is calculated. Therefore, less time is spent waiting for all potential evaluations to complete. This package has previously been used to obtain zero-point energies for the water hexamer[186], running simulations with up to 10^6 walkers for 15 000 time steps.[187]

5.3 Computational details

Numerical details of training the MOB-ML potential energy surfaces

The 3000 training and test configurations for H_2O and CH_5^+, and 1000 training and test configurations for validation molecules are sampled at 50 fs intervals from *ab initio* molecular dynamics (AIMD) trajectories performed with the Q-CHEM 5.0 software package,[79] using the B3LYP[80–83]/6-31G*[84] level of theory. Following the same configuration generation protocol in Ref. 57 and 58, for eight validation molecules, single AIMD trajectories are performed by staring the corresponding optimized geometries at B3LYP/6-31G* level of theory with a Langevin thermostat at 350 K. In order to show the ability of MOB-ML to regress global PESs, we perform AIMDs starting from the optimized geometries at B3LYP/6-31G* level of theory at a Langevin thermostat[85] of 3000 K for the validation molecules. To ensure the full coverage of the two potential energy surfaces, a Langevin thermostat[85] at 6003 K is applied for the H_2O AIMD trajectory starting from the optimized H_2O geometry at B3LYP/6-31G* level of theory. For CH_5^+, three AIMD trajectories are performed by staring from the three literature local minima of CH_5^+[188] with a Langevin thermostat at 350 K, and each trajectory provides 1000 sampled configurations.

We follow the same feature generation protocol described in Husch et al.[60] to compute the associated features at density-fitted HF with aug-cc-pVTZ[189] basis set and aug-cc-pVTZ-JKFIT density fitting basis set [190] using ENTOS QCORE [153]. In this study, valence virtual orbitals are all localized by Intrinsic Bond Orbital method [71]. Valence occupied orbitals are localized by Boys–Foster localization for H_2O, and by Intrinsic Bond Orbital localizations [73, 152] for all the rest of molecules, including 8 small validation molecules and CH_5^+. Reference pair correlation energies are computed at the level of density-fitted CCSD(T)[8, 90] with the aug-cc-pVTZ-JKFIT density fitting basis sets. All these correlation computations are performed with frozen core approximation and full iterative triples treatments using the same LMOs computed by ENTOS QCORE.

For all the training, we employ GPRs [92] with white noise regularized Matérn 5/2 kernel to model the diagonal and offdiagonal pair energies separately using GPY 1.9.6 software package [93]. The generated reference dataset for each molecule is divided into training and test without data overlaps. The learning curves for H_2O and CH_5^+ are generated by training on the listed sizes and tested on the remainder. The results for 8 small validation molecules are collected by training on 500 configurations and testing on the 500 rest configurations. We note that since the numbers of valid features are under 80 and will not cause overfitting due to the small sizes of two molecules, all the valid features are used in the training without feature selection. The negative log marginal likelihood objective of GPR is optimized with respect to the kernel hyperparameters with a scaled conjugate gradient scheme for 100 steps and then apply the BFGS algorithm until full convergence [57, 58, 60].

DMC details

The guiding functions used in the guided DMC simulations are products of one-dimensional wavefunctions of the high frequency vibrations as described in our previous work.[182] The HOH bend is described by a harmonic oscillator with a frequency of 1668 cm^{-1} and a $G-$matrix element[157] of 2.338 amu^{-1} Å^{-2}. One-dimensional discrete variable representation (DVR) calculations were used to obtain the r_{OH} and r_{CH} wavefunctions.[191] The r_{OH} wavefunction was obtained via a potential scan along the r_{OH} coordinate with 900 r_{OH} bond lengths ranging from 0.27 Å to 1.59 Å. These potential values were then used as the potential function in the one-dimensional DVR calculation. A similar scan was done along the r_{CH} coordinate with 900 r_{CH} bond lengths ranging from 0.53 Å to 2.12 Å, keeping all other r_{CH} bond lengths and HCH angles constant.

All DMC simulations in this study were performed using a time step ($\Delta\tau$) of 1 a.u. The zero-point energies reported in Table 5.5 are calculated by averaging V_{ref} over the last two-thirds of the simulation time. For the guided **MOB-ML** DMC simulations, the minimum weight threshold was set at 0.01, and the maximum weight threshold is 1% of the ensemble size (e.g. 50 for a 5000 walker simulation). Each DMC simulation is run independently five times. The uncertainty of the reported zero-point energy is the standard deviation of these five simulations. In the H_2O **MOB-ML** guided DMC calculations, we propagated 2304 walkers for 10 000 a.u. and for CH_5^+ we propagated 5120 walkers for 5000 a.u. For the H_2O **NN+(MOB-ML)** unguided DMC simulations, we propagated 60 000 walkers for 50 000 a.u. and for CH_5^+, we propagated 60 000 walkers for 20 000 a.u. We ran analogous **PS** H_2O and **JBB** CH_5^+ calculations to compare energies and wavefunctions with the **NN+(MOB-ML)** unguided simulations.

Training the NN+(MOB-ML) potential energy surfaces

We used the Keras API implemented in the TensorFlow library[192] to construct, train, and evaluate the **NN+(MOB-ML)** surface. The neural network structure, hyperparameters and training procedure are identical to previous work[178]. To collect training data for the **NN+(MOB-ML)** surfaces, we performed two unguided DMC calculations for each system using the **MOB-ML** surface. For one of the DMC simulations, we multiplied all of the masses of each of the walkers by 0.5, and for the other we use standard masses. We propagated 7168 walkers for each DMC simulation. For all simulations used to collect training data, starting at the second time step, we collected all of the walkers and energies every 5 time steps until time step 50. Then, we collected all walkers every 50 time steps. The resultant training data consisted of approximately 8.6×10^5 configurations and energies for H_2O and 1.5×10^6 configurations and energies for CH_5^+, since for H_2O we propagated the walkers for 2500 a.u. and for CH_5^+ we propagated the walkers for 5000 a.u.

This procedure differs from previous work, where the masses were decreased during the second half of the simulation until the masses reached one-tenth of the original value. This is because the underlying HF calculation that is performed when calling the **MOB-ML** surface occasionally does not converge when unphysical geometries are used as input, resulting in simulation crashes. Based on our analysis, the training data collected from the DMC simulation in which the mass is multiplied by 0.5 sufficiently covers the high energy region, while the training data collected from the DMC simulation where the masses are multiplied by 1 adequately samples the ground state.

We also collected two types of test data sets to examine the error of the **NN+(MOB-ML)** surfaces. The **NN+(MOB-ML)** H_2O and CH_5^+ modified DMC test error reported in table 5.3 is evaluated using 10 000 geometries and energies collected from the same simulations as the training data sets where the masses are multiplied by 0.5. The **NN+(MOB-ML)** ground state DMC test set, also reported in Table 5.3 for each system, consists of three snapshots of walkers collected during the guided **MOB-ML** DMC simulations used to calculate the zero-point energy reported in the first column of Table 5.5. The H_2O ground state test set consists of 6912 configurations and the CH_5^+ ground state test set consists of 15 360 configurations.

Variational calculation [see also Supporting Information of Ref. 178]

The calculations of the vibrational levels of water were performed in Jacobi coordinates. While these are not the most efficient coordinates for describing low-lying vibrational levels of water, they have the advantage of a simple kinetic energy operator,

$$\hat{H} = \frac{\hat{p}_r^2}{2\mu_r} + \frac{\hat{P}_R^2}{2\mu_R} + \left(\frac{1}{2\mu_R R^2} + \frac{1}{2\mu_r r^2}\right)\hat{j}^2 + V(R,r,\theta), \tag{5.8}$$

where r represents one of the OH bond lengths, with reduced mass μ_r, R provides the distance between the second hydrogen atom and the center of mass of the OH bond described by r, and θ is the angle between $\vec{r}$ and $\vec{R}$. The reduced mass associated with R is

$$\mu_R = \left(\frac{1}{m_{\mathrm{H}}} + \frac{1}{m_{\mathrm{H}} + m_{\mathrm{O}}}\right) \tag{5.9}$$

To start, three cuts through the potential were taken, one along each of the three coordinates with the other two coordinates set to their equilibrium values. Each cut was used in a 1D Discrete Variable Representation (DVR) calculation[193], where a DVR based on the Hermite polynomials was used for R and r and the DVR in θ was based on Legendre Polynomials. For each DVR calculation, 250 DVR points were used. The resulting wavefunctions were used to obtain potential-optimized DVR points, with 35 in R and r and 30 in θ. These DVR points and the associated kinetic energy terms were used to set up the full Hamiltonian along with a potential cutoff of 35 000 cm^{-1}. With these parameters, we were able to converge the energies of the vibrational states of interest to within 1 cm^{-1}.

High performance computing DMC load balancing

Diffusion Monte Carlo (DMC), as an algorithm, only requires all-to-all communication between cores for the updating of weights, the calculation of V_{ref}, and the application of branching. Moreover, in the absence of branching, every walker would propagate independently of the others in the simulation, which would allow for a so-called "embarrassingly parallel" implementation in which multiple independent simulations are run and the data is brought back together only at the end. Obviously, branching is necessary. However, in the case that it occurs relatively rarely, we are able to get closer to the situation of truly independent simulations. To that effect, we have written our implementation of HPC DMC so that it can not only evaluate the potential in a distributed manner, but can also perform the diffusion of walkers in a distributed manner. When coupled with taking N_τ steps per propagation, this can significantly improve the performance of the simulation by minimizing latency.

5.4 Results and Discussion

Since we will be using several approaches for evaluating energies, before discussing the results we define the notation used to indicate how energies are evaluated. As noted in the Introduction, we will refer to the H_2O potential energy surface generated by Partridge and Schwenke as the **PS** surface and the global CH_5^+ potential energy surface generated by Jin, Braams, and Bowman as the **JBB** surface. We will refer to surfaces generated using the MOB-ML technique, as the **MOB-ML** surfaces, and we will refer to the neural network generated potential energy surface that was trained using the **MOB-ML** energies as the **NN+(MOB-ML)** surface. The energies obtained by using these surfaces will be denoted as $\mathrm{E}^{\text{potential}}_{\text{system}}$, where potential is replaced by **JBB**, **PS**, **MOB-ML**, or **NN+(MOB-ML)**, while system is replaced with either H_2O or CH_5^+.

Validation and comparison of the MOB-ML potential energy surfaces to previous work

The quality of the MOB-ML approach is commonly assessed by the mean absolute error (MAE) of the predicted CCSD(T)/aug-cc-pVTZ energies of a test set consisting of thermalized geometries. In our previous study [57], MOB-ML models that were based on small numbers of training configurations were shown to accurately predict the energies of geometries sampled at room temperature. The MAE for these models are provided in the right column of Table 5.1. Before applying DMC to studies of H_2O and CH_5^+, we explore the accuracy obtainable by MOB-ML protocols described in Ref. 60 when applied to eight small molecules considered in the earlier study, which we will refer to as the validation molecules. In this application, the training and test configurations are sampled from thermalized geometries at 3000 K. The results of this analysis are provided in Table 5.1.

Table 5.1: Predicted errors of the MOB-ML models relative to CCSD(T)/aug-cc-pVTZ energies. The models are trained on 500 configurations and tested on the rest 500 configurations

System	MAE[a]	MSE[b]	RMSE[c]	Max[d]	MAE (Ref. 57)
CH_4	1.800	-0.035	4.159	87.208	6.58
NH_3	1.046	0.106	3.169	45.969	35.12
HF	0.014	-0.008	0.188	3.436	6.58
CO	0.006	-0.004	0.041	0.228	6.58
N_2	0.028	0.026	0.845	13.119	13.17
F_2	0.544	-0.521	10.058	224.126	6.58
HCN	1.924	-0.809	16.467	303.119	8.78
HNC	2.388	1.034	23.560	191.068	19.75

[a] Mean Absolute Error in cm^{-1}.
[b] Mean Signed Error in cm^{-1}.
[c] Root Mean Square Error in cm^{-1}.
[d] Maximum Error in cm^{-1}.

As can be seen, with the revised protocols, the MOB-ML model achieves accuracies of better than 2.5 cm^{-1} for all eight molecules by training only on 500 configurations. This reflects a substantial improvement over the previously described approaches.[57]. Of equal importance for the DMC calculations is the fact that the errors are uniformly distributed, as indicated by the sub cm^{-1} mean signed errors reported in Table 5.1. It is notable that when there are the same number of electrons in the molecular system, the accuracy decreases with increased numbers of vibrational degrees of freedom. This can be seen by comparing the MAE for HF, NH_3, and CH_4. All three of these molecules have the same number of electrons as H_2O and CH_5^+, which are the focus of the remainder of this study.

Based on this analysis, for the development of the potentials for CH_5^+ and H_2O, we employ slightly larger training sets, and, in the case of H_2O sample geometries based on a thermalized trajectory at 6000 K. We are able to obtain MAE's of 1.8 and 2.0 cm^{-1} for H_2O and CH_5^+ respectively. To further explore the accuracy of these potentials, in Figure 5.1, we show a set of MAEs of predicted energies as functions of number of training configurations on a log-log scale, which are commonly re-

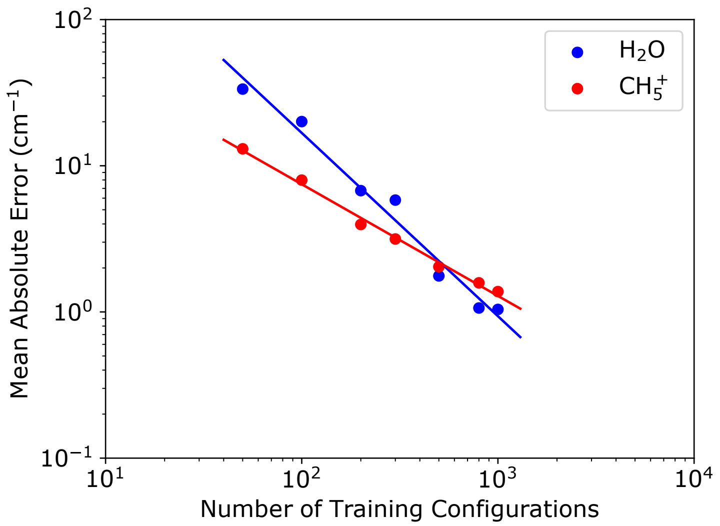

Figure 5.1: Prediction mean absolute errors (MAE) for total energies as a function of the number of training configurations (learning curves) of H_2O (blue circles, solid line) and CH_5^+ (green squares, dotted line) on a logarithm scale. The slopes of learning curves represent the learnability of the MOB-ML model for H_2O and CH_5^+ energies, and steeper learning curve suggests a higher learning efficiency.

ferred to as learning curves[95], for both H_2O and CH_5^+. The test geometries are the subset of 3000 randomized H_2O or CH_5^+ configurations that are not included in the training sets. By comparing the slopes of learning curves, we find that the MOB-ML approach has a slightly better learning efficiency for H_2O than for CH_5^+. This observation is consistent with the expectation that the larger number of vibrational degrees of freedom associated in CH_5^+ should make its potential surface a more difficult learning problem compared to H_2O. In both cases, high accuracies comprising MAEs below 5cm^{-1} are attainable by only including 200 configurations in the training set. Compared with the number of configurations used in the traditional parametric PESs, for instance, 1056 configurations in **PS**, and 36 173 configurations in **JBB**, **MOB-ML** requires significantly smaller number of configurations to provide high quality energies, closely resembling the ones provided by CCSD(T) calculations. Even when considering the most accurate **MOB-ML** models for the prediction of H_2O and CH_5^+ energies, which were trained on 1000 configurations each, **MOB-ML** achieves high accuracies with MAEs of only 1.04 and 1.38cm^{-1}, respectively. Throughout this work, we utilize these high-accuracy models to predict H_2O and CH_5^+ energies. Further details on the development of these models can be found in the Appendix.

The high accuracy of our **MOB-ML** model in describing the H_2O PES is further

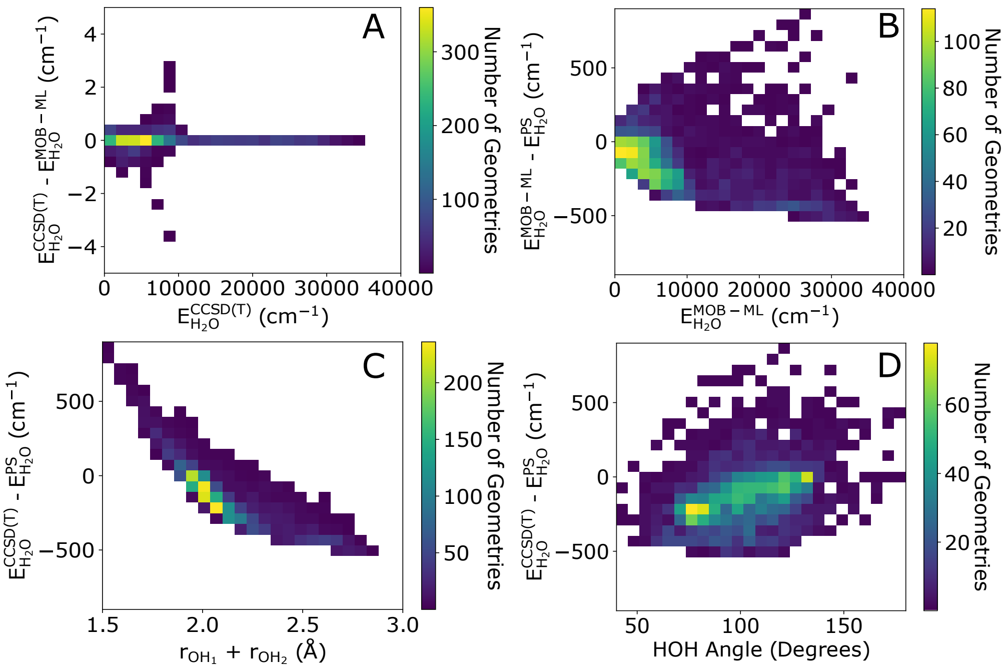

Figure 5.2: Comparison of the calculated energies of MOB-ML training and test set data. (A) The number of geometries plotted as a function of the CCSD(T) energy and the difference between the calculated **MOB-ML** and CCSD(T) energies. (B) The number of geometries plotted as a function of the **MOB-ML** energies and the difference between the **PS**[159] and the **MOB-ML** energies. (C) The number of geometries plotted as a function of the difference between the **MOB-ML** and **PS** energies and the sum of r_{OH} distances and (D) the HOH angle.

supported by comparing our **MOB-ML** predictions to a set of 2000 data points calculated at the CCSD(T)/aug-cc-pVTZ level of theory. As can be seen in panel (A) of Figure 5.2, 96% of the **MOB-ML**-predicted energies lie within 0.5 cm^{-1} of the corresponding CCSD(T) energies, while including all points only increases this value to 4 cm^{-1}. Nevertheless, **MOB-ML** accuracy is only as good as the underlying CCSD(T) level of theory. By comparing single-point energies obtained from **MOB-ML** predictions, $E_{H_2O}^{MOB\text{-}ML}$, and the **PS** PES, $E_{H_2O}^{PS}$, as shown in panel (B), (C), and (D), we immediately notice a large discrepancy over an order of magnitude larger than errors between **MOB-ML** and CCSD(T). These differences are also non-uniform, with a mean signed error (MSE) of -130 cm^{-1}. The large errors can be attributed to the failures of CCSD(T), and other CC methodologies based on perturbative energy corrections, in describing non-dynamical correlation effects, such as those dominating the symmetrically stretched geometries of the water molecule. [194, 195]

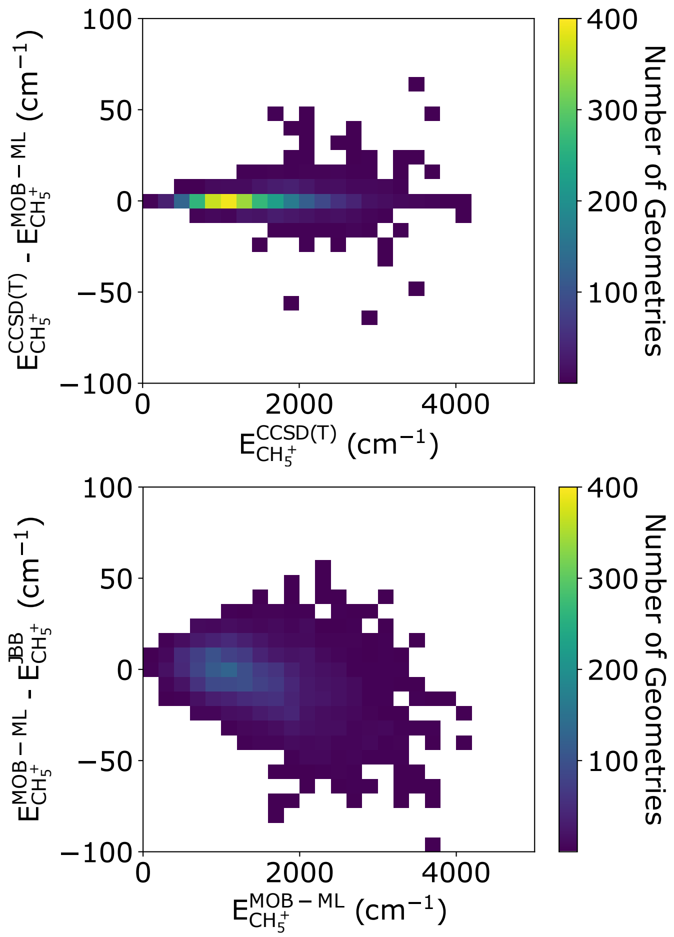

Figure 5.3: The comparison of the training and test geometries used to generate the CH_5^+ **MOB-ML** surface. The number of geometries plotted a function of the CCSD(T) energy and the difference between the **MOB-ML** and CCSD(T) energies (top), and the number of geometries plotted as a function of the **MOB-ML** energies and the difference between the **MOB-ML** and **JBB**[164] energies (bottom).

In Figure 5.3, we make an analogous comparison for CH_5^+. By comparing the **MOB-ML** and CCSD(T) energies computed for a combined selection of 3000 molecular geometries, containing both training and test set configurations, 99.5% of the **MOB-ML** predictions show energy errors smaller than 25 cm^{-1}, and 97 % are within 10 cm^{-1} of $E^{\mathrm{CCSD(T)}}_{\mathrm{CH_5^+}}$. Similarly, when we compare **MOB-ML** energies to those coming from the **JBB** PES, we find that 88% of the energy differences are smaller than 25 cm^{-1}, with the remaining higher energy configurations showing slightly larger errors. The MSE for configurations with calculated energies below 1500 cm^{-1} is 0.1 cm^{-1}, while for geometries with energies above 1500 cm^{-1} the MSE increases to 12.6 cm^{-1}. These differences mirror the root-mean squared fitting error (RMSE) for the **JBB** surface, which the authors report as approximately 10 cm^{-1} for energies below 1500 cm^{-1} and approximately 17 cm^{-1} for energies between 1500 and 4500 cm^{-1}.[164]

Calculating vibrational wavefunctions and energies using MOB-ML surfaces

While comparing single-point energies between **MOB-ML**, CCSD(T), and other previously reported sources provides a strong sense of the accuracy attainable by **MOB-ML** energy predictions, a more demanding task is to compute accurate molecular properties, such as vibrational energies and wavefunctions. To this end, we employ two different approaches combining **MOB-ML**-generated PESs and DMC simulations. In the first, the energies are evaluated using the **MOB-ML** surface directly. Even with the parallel implementation of DMC described above, these calculations are expensive. Therefore, to make this approach tractable, we performed the smallest calculations that are expected to provide reliable results. The parameters for these calculations were based on a previous DMC study performed using the **PS** PES for water[181], and the **JBB** surface for CH_5^+,[182] and are provided in the Supporting Information. While the parameters for these calculations were chosen to be as small as possible, while still providing accurate results, they are still expensive. In order to perform larger DMC calculations, we used the NN-DMC approach.[178] Finally, variational calculations were performed to obtain excited state energies for H_2O. We start by considering the ground state of H_2O. As shown in Table 5.5, the calculation based on the **MOB-ML** energies gives a zero-point energy of 4616(2) cm^{-1}, which is roughly 20 cm^{-1} lower than the corresponding zero-point energy obtained by performing a variational calculation using the **PS** potential. The smaller zero-point energy is consistent with the results plotted in Figure 5.2B, which show that the energies obtained from the **MOB-ML** surface are generally smaller than those obtained from the **PS** surface. It is also consistent with the 24 cm^{-1} lower harmonic zero-point energy obtained at the CCSD(T) level

Table 5.2: Harmonic frequencies for H_2O from underlying electronic structure calculations (cm^{-1})

Mode	$\omega_{H_2O}^{\text{MRCI},a}$	$\omega_{H_2O}^{\text{CCSD(T)}}$
1	1653.1	1646.0
2	3830.7	3810.7
3	3940.5	3919.8

[a] Ref. 159

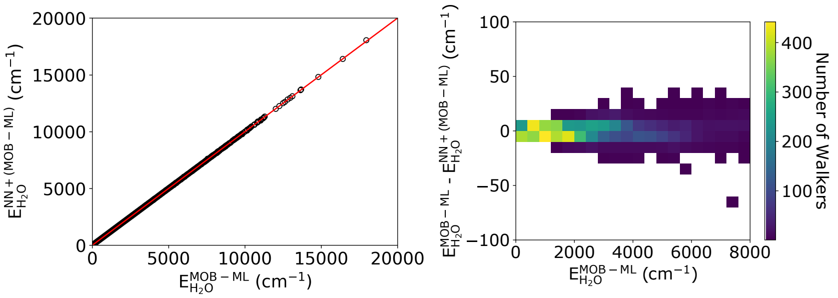

Figure 5.4: Comparison of the **NN+(MOB-ML)** and **MOB-ML** energies of the **NN+(MOB-ML)** ground state test data set for H_2O. This data is also used to calculate the ground state MAE in Table 5.3. The predicted **NN+(MOB-ML)** energy plotted as a function of the **MOB-ML** energy (left), and the number of geometries in the test set plotted as a function of the difference between the energies evaluated using the two surfaces and the **MOB-ML** energy (right).

compared to the MRCI calculations used to generate the **PS** surface (see Table 5.2). On the other hand, this result is based on a small DMC calculation. To verify this zero-point energy, we have performed a larger NN-DMC calculation, which gives a zero-point energy of 4615(1) cm^{-1}. This energy agrees with the results of the smaller calculation. While these results are promising, to further ensure that the **NN+(MOB-ML)** technique is adequately learning the **MOB-ML** surface for the purposes of DMC, we provide comparisons of the single point **NN+(MOB-ML)** energies to **MOB-ML** energies for both H_2O and CH_5^+ in Figures 5.4 and 5.5 the Supporting Information. Based on these comparisons, the **NN+(MOB-ML)** surface provides a similar level of accuracy when compared to our previous work, where the same neural network structure was used to learn the **PS** surface[178]. This gives us confidence in applying the neural network method to the **MOB-ML** surface beyond H_2O.

To this end, we performed a variational calculation of the vibrational energies of water. The details of this calculation are reported in a previous study[178] and reproduced in the Supporting Information. As can be seen in the results reported in Table 5.4, the energies obtained from the variational calculation using the **MOB-ML** surface and the **NN+(MOB-ML)** surface are in very good agreement, further validating the NN-DMC approach on the **MOB-ML** surface. When we compare the energies based on the **MOB-ML** and **PS** potentials, larger differences are observed. The anharmonic zero-point energy on the **MOB-ML** surface is approximately 20 cm^{-1} lower than the **PS** surface, and the energies of the levels with one quantum of excitation in the OH stretches each deviate by an additional 20 cm^{-1}. As mentioned

Table 5.3: MAE of the NN+(MOB-ML) training and test sets of H_2O and CH_5^+(cm^{-1}).

System	Training Error	Test Error Modified DMC[a]	Test Error Ground State DMC[b]
H_2O	18	24	4
CH_5^+	115	153	68

[a] Calculated based on the energies of 10 000 configurations collected from the training **MOB-ML** DMC simulation in which the masses of each atom are multiplied by 0.5.

[b] Calculated based on the energies of three sets of walkers collected from the small-scale **MOB-ML** DMC simulation whose zero-point energy is reported in the first column of Table 5.5.

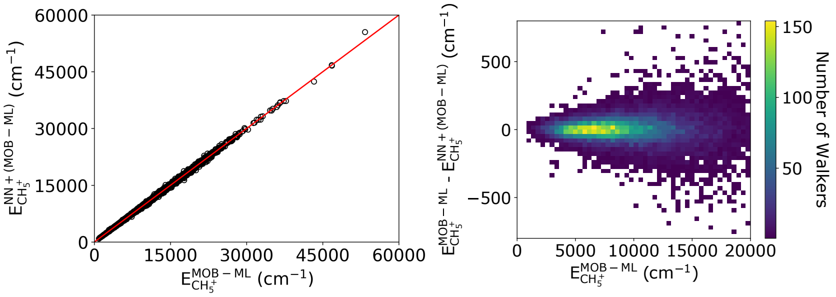

Figure 5.5: Comparison of the **NN+(MOB-ML)** and **MOB-ML** energies of the **NN+(MOB-ML)** ground state test data set for CH_5^+. This data is also used to calculate the ground state MAE in Table 5.3. The predicted **NN+(MOB-ML)** energy plotted as a function of the **MOB-ML** energy (left), and the number of geometries in the test set plotted as a function of the difference between the energies evaluated using the two surfaces and the **MOB-ML** energy (right).

above, the harmonic zero-point energy between the two surfaces differ by around 24 cm^{-1}, and the deviation can be traced to a 20 cm^{-1} discrepancy in each of the OH stretch frequencies. Finally, the difference between the energies of the bend states, calculated using these two potentials, differ by 1 to 4 cm^{-1}.

We also calculated the ground state energy and wavefunction for CH_5^+ based on the **MOB-ML** potential. Due to its larger number of vibrational degrees of freedom, two of which are large-amplitude vibrations, we have only performed ground state DMC calculations for this ion. Additionally, the increased dimensionality makes the evaluation of the **MOB-ML** potential approximately twice as expensive, and the minimum number of walkers needed to obtain a reliable ground state wavefunction and energy are roughly twice as large as for H_2O. This makes DMC calculations based on the **MOB-ML** potential barely feasible. Using this approach, we obtain a

Table 5.4: Calculated ground and excited state vibrational energies[a] for H_2O (cm^{-1})

v_s^b	v_b	v_a	**MOB-ML**	**MOB-ML** – **NN+(MOB-ML)**	**PS**[c]
0	0	0	4614.6	−0.01	4636.8
0	1	0	1594.1	−0.01	1594.4
0	2	0	3151.8	0.8	3150.8
1	0	0	3638.8	−0.3	3656.2
0	0	1	3734.5	−0.2	3755.1
0	3	0	4669.5	−0.3	4665.7
1	1	0	5216.3	0.2	5233.8
0	1	1	5308.9	−0.5	5330.0

[a] The first row corresponds to the calculated zero-point energy E_0, and all subsequent rows correspond to E-E_0.
[b] v_s, v_b, and v_a correspond to the number of quanta in the symmetric OH stretch, HOH bend, and antisymmetric OH stretch, respectively.
[c] Ref. 178.

zero-point energy of 10 912(15) cm^{-1}. When we use the NN-DMC approach, the zero-point energy becomes 10 909(2) cm^{-1}. While both values are slightly lower than the energies reported based on the global $E^{JBB}_{CH_5^+}$ surface, they are in excellent agreement with the DMC zero-point energy of 10 908(5) reported by Johnson and McCoy using the CCSD(T)-based surface (**JBB:CC**) from which the global surface was developed[188]. These results are summarized in Table 5.5.

Table 5.5: Calculated zero-point energies obtained using DMC (cm^{-1})

System	**MOB-ML**	**NN+(MOB-ML)**	**PS**[a]/**JBB**[b]	**JBB:CC**[c]
H_2O	4616 (2)	4615 (1)	4637 (2)	–
CH_5^+	10 912 (15)	10 908 (2)	10 917 (5)	10 908 (5)
CH_4D^+	–	10 301 (2)	10 303 (4)	10 298 (5)
$CH_3D_2^+$	–	9689 (4)	9698 (7)	9690 (5)
$CH_2D_3^+$	–	9086 (3)	9010 (3)	9090 (5)
CHD_4^+	–	8553 (2)	8565 (3)	8559 (5)
CD_5^+	–	8040 (3)	8044 (2)	8039 (5)

[a] Results of DMC simulations using the Partridge-Schwenke surface[159].
[b] Results of DMC simulations using the Jin, Braams, and Bowman surface[164].
[c] Results of DMC simulations on the CCSD(T) surface on which the **JBB** potential is based.[188]

CH_5^+ is an unusual ion in that it exhibits two large amplitude motions, which result in low barriers for permutation of the hydrogen atoms. Specifically, there are 120 equivalent minima on the potential surface that describes CH_5^+. Based on the

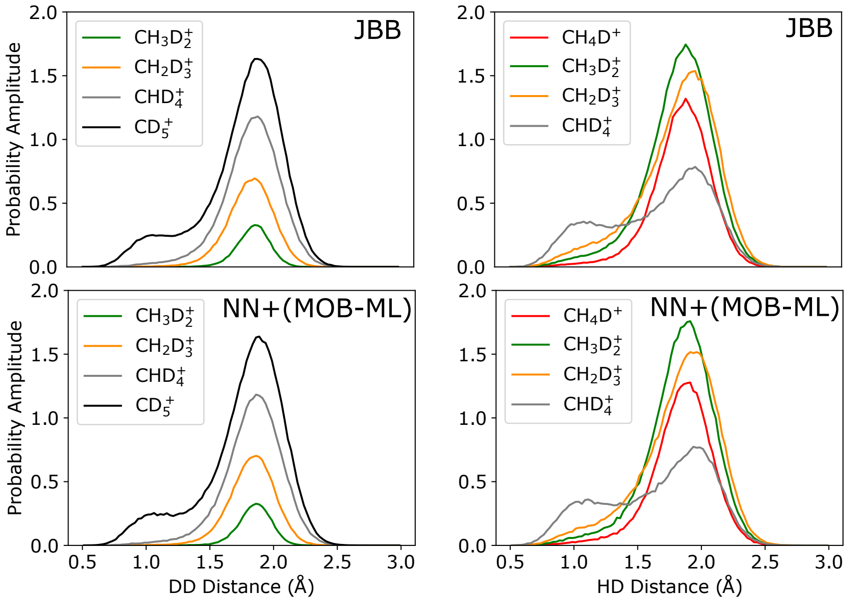

Figure 5.6: The DMC probability amplitude projected onto DD (left) and HD (right) distances using wavefunctions from the **JBB** surface (upper panels) and the **NN+(MOB-ML)** surface (lower panels).

CCSD(T) calculations used to develop the **MOB-ML** surface, these minima are separated by a series of transition states none of which is higher than 332 cm^{-1}. As a result, the ground state wavefunction for CH_5^+ has roughly equal amplitude at all 120 equivalent minima as well as the 180 low-energy transition states that connect these minima.[196] While isomerization is facile, the five CH bonds are not equivalent at any of the low-energy stationary points. This is illustrated by the harmonic frequencies for the CH stretches, which range from 2400 to 3250 cm^{-1}[164, 188]. As a result, when one or more of the hydrogen atoms is replaced by a deuterium atom, the ground state probability amplitude is no longer equally distributed among the 120 minima on the potential surface. This can be seen in the plots of the projection of the probability amplitude onto the HH distances, shown in Figure 5.7. In this figure, we compare the distributions obtained using NN-DMC calculations based on the **NN+(MOB-ML)** potential to results obtained running the analogous unguided DMC calculations on the **JBB** potential. The distributions change as hydrogen atoms are replaced with deuterium atoms, and the evolution of the distributions with deuteration reflects the localization described above. This effect has been discussed previously,[197, 198] and the important observation for the current study is that calculations of the ground state probability amplitude based on both the **NN+(MOB-ML)** potential and the **JBB** potential yield nearly identical distributions. Analogous distributions, which show similar agreement for the HD and DD distance distributions, are provided in Figure 5.6 in the Supporting Information. For all isotopomers, the difference in the zero-point energies calculated using the **JBB** and the **NN+(MOB-ML)** potentials remain smaller than 15 cm^{-1}. The deviations in the energies among isotopomers reflect a sensitivity of this quantity to small differences among the potentials. As mentioned above, the primary source of these differences is most likely from the introduction of a switching function that allows the **JBB** surface to dissociate properly. When that correction is not included, the differences between the zero-point energies reported in Ref. 188, and reproduced in Table 5.5, and those obtained using the **NN+(MOB-ML)** surface are less than 6 cm^{-1}. This difference is within the uncertainties of the previously reported values.

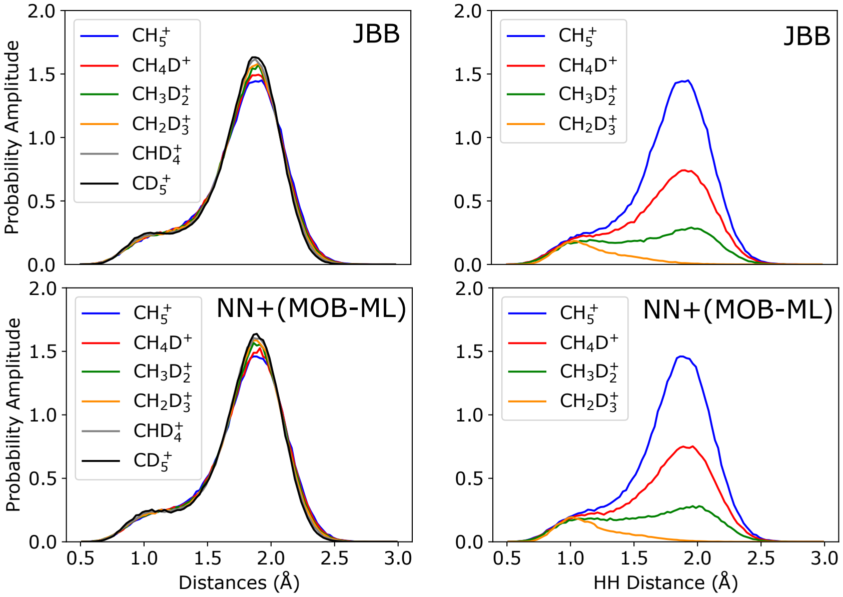

Figure 5.7: The calculated DMC probability amplitude projected onto all H/D distances (left) and HH distances (right) for the appropriate isotopologues of CH_5^+. The top two panels show the DMC probability amplitude using the **JBB** potential energy surface[164], where the bottom two are using the **NN+(MOB-ML)** surface.

The above agreement between the results of these two sets of calculations should not be surprising, as both the **MOB-ML** and the **JBB** surface are based on the same levels of electronic structure theory. On the other hand, whereas the earlier surface is based on fitting more than 35 000 electronic energies with energies up to 150 000 cm^{-1} to a potential function with 2300 coefficients[197], the **MOB-ML** potential is based on 1000 electronic energies with energies below 4500 cm^{-1}. The similarity between the calculated properties based on these two surfaces provides an illustration of the power of the MOB-ML approach.

5.5 Conclusion

In this work, we introduce a general approach to generate efficient and highly accurate potential energy surfaces for their use in large-scale molecular simulations. Specifically, we take advantage of the MOB-ML approach to generate CCSD(T)-quality potential energy surfaces for H_2O and CH_5^+ systems, at a small fraction of the computational cost relative to CCSD(T). We show that by relying on a training set of only 1000 molecular configurations and CCSD(T) energies, we can construct accurate MOB-ML models suitable for demanding DMC simulations. Furthermore, we demonstrate that by employing a NN approach to refit the MOB-ML energies, we can increase the computational efficiency of the MOB-ML approach by exploiting GPU technology, and achieve large scale DMC simulations while maintaining high accuracy.

BIBLIOGRAPHY

1. Born, M. & Oppenheimer, R. Zur quantentheorie der molekeln. *Annalen der physik* **389,** 457–484 (1927).

2. Tully, J. C. Perspective on “zur quantentheorie der molekeln”. *Theor. Chem. Acc.* **103,** 173–176 (2000).

3. Ross, I. G. Calculations of the energy levels of acetylene by the method of antisymmetric molecular orbitals, including σ-π interaction. *Trans. Faraday Soc.* **48,** 973–991 (1952).

4. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136,** B864 (1964).

5. Møller, C. & Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **46,** 618 (1934).

6. Bartlett, R. J. & Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **79,** 291 (2007).

7. Szabo, A. & Ostlund, N. S. *Modern quantum chemistry: introduction to advanced electronic structure theory* 231–239 (Courier Corporation, 2012).

8. Schütz, M. Low-order scaling local electron correlation methods. III. Linear scaling local perturbative triples correction (T). *J. Chem. Phys.* **113,** 9986–10001 (2000).

9. Ramabhadran, R. O. & Raghavachari, K. Extrapolation to the gold-standard in quantum chemistry: computationally efficient and accurate CCSD (T) energies for large molecules using an automated thermochemical hierarchy. *J. Chem. Theory Comput.* **9,** 3986–3994 (2013).

10. Scuseria, G. E. & Lee, T. J. Comparison of coupled-cluster methods which include the effects of connected triple excitations. *J. Chem. Phys.* **93,** 5851–5855 (1990).

11. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140,** A1133 (1965).

12. Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.* **87,** 897 (2015).

13. Tao, J., Perdew, J. P., Staroverov, V. N. & Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **91,** 146401 (2003).

14. Perdew, J. P., Chevary, J. A., Vosko, S. H., Jackson, K. A., Pederson, M. R., Singh, D. J. & Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **46,** 6671 (1992).

15. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38,** 3098 (1988).

16. Von Lilienfeld, O. A. & Burke, K. Retrospective on a decade of machine learning for chemical discovery. *Nat. Commun.* **11,** 1–4 (2020).

17. Westermayr, J., Gastegger, M., Schütt, K. T. & Maurer, R. J. Perspective on integrating machine learning into computational chemistry and materials science. *J. Chem. Phys.* **154,** 230903 (2021).

18. Freeze, J. G., Kelly, H. R. & Batista, V. S. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chem. Rev.* **119,** 6595–6612 (2019).

19. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **4,** 828–849 (2019).

20. Von Lilienfeld, O. A., Lins, R. D. & Rothlisberger, U. Variational particle number approach for rational compound design. *Phys. Rev. Lett.* **95,** 153002 (2005).

21. Von Lilienfeld, O. A. & Tuckerman, M. Alchemical variations of intermolecular energies according to molecular grand-canonical ensemble density functional theory. *J. Chem. Theory Comput.* **3,** 1083–1090 (2007).

22. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4,** eaap7885 (2018).

23. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* **119,** 10520–10594 (2019).

24. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **4,** 331–348 (2019).

25. Tran, A., Tranchida, J., Wildey, T. & Thompson, A. P. Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys. *J. Chem. Phys.* **153,** 074705 (2020).

26. Grisafi, A., Fabrizio, A., Meyer, B., Wilkins, D. M., Corminboeuf, C. & Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **5,** 57–64 (2019).

27. Manzhos, S. & Carrington Jr, T. Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **121,** 10187–10217 (2020).

28. Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* **150,** 150901 (2019).

29. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G. & Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571,** 95–98 (2019).

30. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9,** 1–10 (2019).

31. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361,** 360–365 (2018).

32. Schwalbe-Koda, D. & Gómez-Bombarelli, R. in *Machine Learning Meets Quantum Physics* 445–467 (Springer, 2020).

33. Deringer, V. L., Bernstein, N., Csányi, G., Ben Mahmoud, C., Ceriotti, M., Wilson, M., Drabold, D. A. & Elliott, S. R. Origins of structural and electronic transitions in disordered silicon. *Nature* **589,** 59–64 (2021).

34. Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A. & Müller, K.-R. Machine learning force fields. *Chem. Rev.* **121,** 10142–10186 (2021).

35. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145,** 170901 (2016).

36. Jiang, B., Li, J. & Guo, H. High-fidelity potential energy surfaces for gas-phase and gas–surface scattering processes from machine learning. *J. Phys. Chem. Lett.* **11,** 5120–5131 (2020).

37. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104,** 136403 (2010).

38. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108,** 58301 (2012).

39. Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15,** 95003 (2013).

40. Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, O. A., Tkatchenko, A. & Müller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9,** 3404 (2013).

41. Gasparotto, P. & Ceriotti, M. Recognizing molecular patterns by machine learning: an agnostic structural definition of the hydrogen bond. *J. Chem. Phys.* **141,** 174110 (2014).

42. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* **11,** 2087 (2015).

43. Brockherde, F., Vogt, L., Li, L., Tuckerman, M. E., Burke, K. & Müller, K.-R. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8,** 872 (2017).

44. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30,** 595 (2016).

45. Paesani, F. Getting the right answers for the right reasons: toward predictive molecular simulations of water with many-body potential energy functions. *Acc. Chem. Res.* **49,** 1844 (2016).

46. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K.-R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8,** 13890 (2017).

47. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8,** 3192–3203 (2017).

48. Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K. & Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **9,** 513 (2018).

49. Nguyen, T. T., Székely, E., Imbalzano, G., Behler, J., Csányi, G., Ceriotti, M., Götz, A. W. & Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **148,** 241725 (2018).

50. Yao, K., Herr, J. E., Toth, D. W., McKintyre, R. & Parkhill, J. The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics. *Chem. Sci.* **9,** 2261–2269 (2018).

51. Fujikake, S., Deringer, V. L., Lee, T. H., Krynski, M., Elliott, S. R. & Csányi, G. Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures. *J. Chem. Phys.* **148,** 241714 (2018).

52. Dick, S. & Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.* **11,** 1–10 (2020).

53. Chen, Y., Zhang, L., Wang, H. & E, W. Ground state energy functional with Hartree–Fock efficiency and chemical accuracy. *J. Phys. Chem. A* **124,** 7155–7165 (2020).

54. Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R. & Miller III, T. F. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **153,** 124111 (2020).

55. Qiao, Z., Ding, F., Welborn, M., Bygrave, P. J., Smith, D. G., Anandkumar, A., Manby, F. R. & Miller III, T. F. Multi-task learning for electronic structure to predict and explore molecular potential energy surfaces. *arXiv preprint arXiv:2011.02680* (2020).

56. Christensen, A. S., Sirumalla, S. K., Qiao, Z., O'Connor, M. B., Smith, D. G., Ding, F., Bygrave, P. J., Anandkumar, A., Welborn, M., Manby, F. R., *et al.* OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **155,** 204103 (2021).

57. Welborn, M., Cheng, L. & Miller III, T. F. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **14,** 4772–4779 (2018).

58. Cheng, L., Welborn, M., Christensen, A. S. & Miller III, T. F. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *J. Chem. Phys.* **150,** 131103 (2019).

59. Cheng, L., Kovachki, N. B., Welborn, M. & Miller III, T. F. Regression clustering for improved accuracy and training costs with molecular-orbital-based machine learning. *J. Chem. Theory Comput.* **15,** 6668–6677 (2019).

60. Husch, T., Sun, J., Cheng, L., Lee, S. J. R. & Miller, T. F. Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **154,** 064108 (2021).

61. Lee, S. J., Husch, T., Ding, F. & Miller III, T. F. Analytical gradients for molecular-orbital-based machine learning. *J. Chem. Phys.* **154,** 124120 (2021).

62. Sun, J., Cheng, L. & Miller III, T. F. *Molecular energy learning using alternative blackbox matrix-matrix multiplication algorithm for exact Gaussian process* in *NeurIPS 2021 AI for Science Workshop* (2021).

63. Hermann, J., Schätzle, Z. & Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.* **12,** 891–897 (2020).

64. Chen, Y., Zhang, L., Wang, H. & E, W. DeePKS: A comprehensive data-driven approach toward chemically accurate density functional theory. *J. Chem. Theory Comput.* (2020).

65. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.* **56,** 12828–12840 (2017).

66. Mueller, T., Hernandez, A. & Wang, C. Machine learning for interatomic potential models. *J. Chem. Phys.* **152,** 050902 (2020).

67. Gkeka, P., Stoltz, G., Barati Farimani, A., Belkacemi, Z., Ceriotti, M., Chodera, J. D., Dinner, A. R., Ferguson, A. L., Maillet, J.-B., Minoux, H., *et al.* Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems. *J. Chem. Theory Comput.* **16,** 4757–4775 (2020).

68. Čížek, J. On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell–type expansion using quantum–field theoretical methods. *J. Chem. Phys.* **45,** 4256 (1966).

69. Blum, L. C. & Reymond, J.-L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131,** 8732 (2009).

70. Nesbet, R. K. Brueckner's theory and the method of superposition of configurations. *Phys. Rev.* **109,** 1632 (1958).

71. Knizia, G. Intrinsic atomic orbitals: An unbiased bridge between quantum theory and chemical concepts. *J. Chem. Theory Comput.* **9,** 4834 (2013).

72. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98,** 146401 (2007).

73. Boys, S. F. Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another. *Rev. Mod. Phys.* **32,** 296–299 (1960).

74. Kaldor, U. Localized orbitals for NH_3, C_2H_4, and C_2H_2. *J. Chem. Phys.* **46,** 1981–1987 (1967).

75. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

76. Breiman, L. Statistical Modeling: The Two Cultures. *Stat. Sci.* **16,** 199–215 (2001).

77. Tripathy, R., Bilionis, I. & Gonzalez, M. Gaussian processes with built-in dimensionality reduction : Applications to high-dimensional uncertainty propagation. *J. Comput. Phys.* **321,** 191–223 (2016).

78. Cheng, L., Welborn, M., Christensen, A. S. & Miller III, T. F. *Thermalized (350K) QM7b, GDB-13, water, and short alkane quantum chemistry dataset including MOB-ML features* `https://data.caltech.edu/records/1177`. 2019.

79. Shao, Y. *et al.* Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **113,** 184 (2015).

80. Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **58,** 1200 (1980).

81. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37,** 785 (1988).

82. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648 (1993).

83. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98,** 11623 (1994).

84. Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **28,** 213 (1973).

85. Bussi, G. & Parrinello, M. Accurate sampling using Langevin dynamics. *Phys. Rev. E* **75,** 056707 (2007).

86. Werner, H.-J., Knowles, P. J., Knizia, G., Manby, F. R., Schütz, M., Celani, P., Györffy, W., Kats, D., Korona, T., Lindh, R., Mitrushenkov, A., Rauhut, G., Shamasundar, K. R., Adler, T. B., Amos, R. D., Bennie, S. J., Bernhardsson, A., Berning, A., Cooper, D. L., Deegan, M. J. O., Dobbyn, A. J., Eckert, F., Goll, E., Hampel, C., Hesselmann, A., Hetzer, G., Hrenar, T., Jansen, G., Köppl, C., Lee, S. J. R., Liu, Y., Lloyd, A. W., Ma, Q., Mata, R. A., May, A. J., McNicholas, S. J., Meyer, W., Miller III, T. F., Mura, M. E., Nicklass, A., O'Neill, D. P., Palmieri, P., Peng, D., Pflüger, K., Pitzer, R., Reiher, M., Shiozaki, T., Stoll, H., Stone, A. J., Tarroni, R., Thorsteinsson, T., Wang, M. & Welborn, M. *MOLPRO, version 2018.3, a package of ab initio programs* see http://www.molpro.net. Cardiff, UK, 2018.

87. Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90,** 1007 (1989).

88. Saebo, S. & Pulay, P. Local treatment of electron correlation. *Annu. Rev. Phys. Chem.* **44,** 213–236 (1993).

89. Hampel, C. & Werner, H. J. Local treatment of electron correlation in coupled cluster theory. *J. Chem. Phys.* **104,** 6286–6297 (1996).

90. Bartlett, R. J., Watts, J. D., Kucharski, S. A. & Noga, J. Non-iterative fifth-order triple and quadruple excitation energy corrections in correlated methods. *Chem. Phys. Lett.* **165,** 513–522 (1990).

91. Polly, R., Werner, H.-J., Manby, F. R. & Knowles, P. J. Fast Hartree-Fock theory using local density fitting approximations. *Mol. Phys.* **102,** 2311–2321 (2004).

92. Rasmussen, C. E. & Williams, C. K. I. *Gaussian processes for machine learning* (MIT Press, Cambridge, MA, 2006).

93. GPy. *GPy: A Gaussian process framework in python* `http://github.com/SheffieldML/GPy`.

94. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: machine learning in python (v0.21.2). *J. Mach. Learn. Res.* **12,** 2825 (2011).

95. Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V. & Denker, J. S. in *Advances in Neural Information Processing Systems 6* (eds Cowan, J. D., Tesauro, G. & Alspector, J.) 327–334 (Morgan-Kaufmann, 1994).

96. Pearson, K. Mathematical contributions to the theory of evolution. III. regression, heredity, and panmixia. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **187,** 253–318 (1896).

97. Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15,** 095003 (2013).

98. Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148,** 241717 (2018).

99. Christensen, A. S., Faber, F. A. & von Lilienfeld, O. A. Operators in quantum machine learning: Response properties in chemical space. *J. Chem. Phys.* **150,** 064105 (2019).

100. Kramer, B. & MacKinnon, A. Localization: theory and experiment. *Rep. Prog. Phys.* **56,** 1469 (1993).

101. Spath, H. Correction to Algorithm 39: Clusterwise Linear Regression. *Computing* **26,** 275 (1979).

102. Späth, H. Algorithm 39: Clusterwise linear regression. *Computing,* 367–373 (1979).

103. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28,** 129–137 (1982).

104. Criminisi, A., Shotton, J., Konukoglu, E., *et al.* Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision* **7,** 81–227 (2012).

105. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A library for large linear classification. *Journal of machine learning research* **9,** 1871–1874 (2008).

106. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class adaboost. *Stat. Interface.* **2,** 349–360 (2009).

107. Baraldi, A. & Blonda, P. A survey of fuzzy clustering algorithms for pattern recognition. I. *IEEE Trans. Syst. Man Cybern. B Cybern.* **29,** 778–785 (1999).

108. Browning, N. J., Ramakrishnan, R., von Lilienfeld, O. A. & Roethlisberger, U. Genetic optimization of training sets for improved machine learning models of molecular properties. *J. Phys. Chem. Lett.* **8,** 1351–1359 (2017).

109. Murphy, K. P. *Machine learning: a probabilistic perspective* 492–493 (MIT Press, Cambridge, MA, 2012).

110. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **152,** 044107 (2020).

111. Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. *Mol. Inform.* **35,** 3–14 (2016).

112. Mater, A. C. & Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **59,** 2545–2559 (2019).

113. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *Npj Comput. Mater.* **3,** 53 (2017).

114. Ren, F., Ward, L., Williams, T., Laws, K. J., Wolverton, C., Hattrick-Simpers, J. & Mehta, A. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4,** eaaq1566 (2018).

115. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559,** 547–555 (2018).

116. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16,** 687–694 (2019).

117. Casalino, L., Dommer, A. C., Gaieb, Z., Barros, E. P., Sztain, T., Ahn, S.-H., Trifan, A., Brace, A., Bogetti, A. T., Clyde, A., Ma, H., Lee, H., Turilli, M., Khalid, S., Chong, L. T., Simmerling, C., Hardy, D. J., Maia, J. D., Phillips, J. C., Kurth, T., Stern, A. C., Huang, L., McCalpin, J. D., Tatineni, M., Gibbs, T., Stone, J. E., Jha, S., Ramanathan, A. & Amaro, R. E. AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. *Int. J. High. Perform.* **35,** 432–451 (2021).

118. Gussow, A. B., Park, A. E., Borges, A. L., Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Bondy-Denomy, J. & Koonin, E. V. Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat. Commun.* **11,** 1–12 (2020).

119. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2,** 725–732 (2016).

120. Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J. & Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533,** 73–76 (2016).

121. Ulissi, Z. W., Medford, A. J., Bligaard, T. & Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **8,** 14621 (2017).

122. Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H. & Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4,** 1465–1476 (2018).

123. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555,** 604–610 (2018).

124. Aarva, A., Deringer, V. L., Sainio, S., Laurila, T. & Caro, M. A. Understanding X-ray spectroscopy of carbonaceous materials by combining experiments, density functional theory, and machine learning. Part I: Fingerprint spectra. *Chem. Mater.* **31,** 9243–9255 (2019).

125. Zhang, Y., Tang, Q., Zhang, Y., Wang, J., Stimming, U. & Lee, A. A. Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nat. Commun.* **11,** 1–6 (2020).

126. Magdau, I.-B. & Miller III, T. F. Machine Learning solvation environments in conductive polymers: Application to ProDOT-2Hex with solvent swelling. *ChemRxiv preprint* (2020).

127. Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O. & Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10,** 2903 (2019).

128. Lubbers, N., Smith, J. S. & Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **148,** 241715 (2018).

129. Li, H., Collins, C., Tanha, M., Gordon, G. J. & Yaron, D. J. A density functional tight binding layer for deep learning of chemical Hamiltonians. *J. Chem. Theory Comput.* **14,** 5764–5776 (2018).

130. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120,** 143001 (2018).

131. Nandy, A., Duan, C., Janet, J. P., Gugler, S. & Kulik, H. J. Strategies and software for machine learning accelerated discovery in transition metal chemistry. *Ind. Eng. Chem. Res.* **57,** 13973–13986 (2018).

132. Bogojeski, M., Vogt-Maranto, L., Tuckerman, M. E., Müller, K.-R. & Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **11,** 1–11 (2020).

133. Glick, Z. L., Metcalf, D. P., Koutsoukas, A., Spronk, S. A., Cheney, D. L. & Sherrill, C. D. AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials. *J. Chem. Phys.* **153,** 044112 (2020).

134. Chen, Y., Zhang, L., Wang, H. & E, W. Ground state energy functional with Hartree–Fock efficiency and chemical accuracy. *J. Phys. Chem. A* **124,** 7155–7165 (2020).

135. Christensen, A. S. & von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach. learn.: sci. technol.* **1,** 045018 (2020).

136. Mezei, P. D. & von Lilienfeld, O. A. Noncovalent Quantum Machine Learning Corrections to Density Functionals. *J. Chem. Theory Comput.* **16,** 2647–2653 (2020).

137. Grisafi, A., Wilkins, D. M., Willatt, M. J. & Ceriotti, M. in *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions* 1–21 (2019).

138. Pereira, F. & Aires-de-Sousa, J. Machine learning for the prediction of molecular dipole moments obtained by density functional theory. *J. Cheminform.* **10,** 43 (2018).

139. Fabrizio, A., Grisafi, A., Meyer, B., Ceriotti, M. & Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **10,** 9424–9432 (2019).

140. Ramakrishnan, R., Hartmann, M., Tapavicza, E. & von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **143,** 084111 (2015).

141. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8,** 6924–6935 (2017).

142. Ghosh, K., Stuke, A., Todorović, M., Jørgensen, P. B., Schmidt, M. N., Vehtari, A. & Rinke, P. Deep learning spectroscopy: Neural networks for molecular excitation spectra. *Adv. Sci.* **6,** 1801367 (2019).

143. Veit, M., Wilkins, D. M., Yang, Y., DiStasio, R. A. & Ceriotti, M. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *J. Chem. Phys.* **153,** 024113 (2020).

144. McGibbon, R. T., Taube, A. G., Donchev, A. G., Siva, K., Hernández, F., Hargus, C., Law, K.-H., Klepeis, J. L. & Shaw, D. E. Improving the accuracy of Møller-Plesset perturbation theory with neural networks. *J. Chem. Phys.* **147,** 161725 (2017).

145. Nudejima, T., Ikabata, Y., Seino, J., Yoshikawa, T. & Nakai, H. Machine-learned electron correlation model based on correlation energy density at complete basis set limit. *J. Chem. Phys.* **151,** 024104 (2019).

146. Townsend, J. & Vogiatzis, K. D. Data-driven acceleration of the coupled-cluster singles and doubles iterative solver. *J. Phys. Chem. Lett.* **10,** 4129–4135 (2019).

147. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al. A density-based algorithm for discovering clusters in large spatial databases with noise.* in *Kdd* **96** (1996), 226–231.

148. Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record* **28,** 49–60 (1999).

149. Bishop, C. M. *Pattern recognition and machine learning* (Springer, 2006).

150. Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q. & Wilson, A. G. *GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration* in *Advances in Neural Information Processing Systems* (2018), 7576–7586.

151. Wang, K. *et al. Exact Gaussian Processes on a Million Data Points* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019), 14648–14659.

152. Foster, J. M. & Boys, S. F. Canonical configurational interaction procedure. *Rev. Mod. Phys.* **32,** 300–302 (1960).

153. Manby, F. R., Miller III, T. F., Bygrave, P., Ding, F., Dresselhaus, T., Batista-Romero, F., Buccheri, A., Bungey, C., Lee, S. J. R., Meli, R., Miyamoto, K., Steinmann, C., Tsuchiya, T., Welborn, M., Wiles, T. & Williams, Z. entos: A Quantum Molecular Simulation Package (2019).

154. Okuta, R., Unno, Y., Nishino, D., Hido, S. & Loomis, C. *CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations* in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)* (2017).

155. Findley, D. F. Counterexamples to parsimony and BIC. *Ann. Inst. Stat. Math.* **43,** 505–514 (1991).

156. Pinto, R. C. & Engel, P. M. A fast incremental gaussian mixture model. *PloS one* **10,** e0139931 (2015).

157. Wilson, E. B., Decius, J. C. & Cross, P. C. *Molecular Vibrations* (Dover, New York, 1955).

158. Nielsen, H. H. The vibration-rotation energies of molecules. *Rev. Mod. Phys.* **23,** 90–136 (2 1951).

159. Partridge, H. & Schwenke, D. W. The determination of an accurate isotope dependent potential energy surface for water from extensive *ab initio* calculations and experimental data. *J. Chem. Phys.* **106,** 4618–4639 (1997).

160. Aziz, R. A. & Salaman, M. J. The Ne-Ne interatomic potential revisited. *Chem. Phys.* **130,** 187 (1989).

161. Babin, V., Leforestier, C. & Paesani, F. Development of a "first principles" water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient. *J. Chem. Theory Comput.* **9.** PMID: 26592277, 5395–5403 (2013).

162. Schatz, G. C. The analytical representation of electronic potential-energy surfaces. *Rev. Mod. Phys.* **61,** 669 (1989).

163. Huang, X., Schwenke, D. W. & Lee, T. J. Rovibrational spectra of ammonia. I. Unprecedented accuracy of a potential energy surface used with nonadiabatic corrections. *J. Chem. Phys.* **134,** 044320 (2011).

164. Jin, Z., Braams, B. J. & Bowman, J. M. An *ab initio* based global potential energy surface describing $CH_5^+ \rightarrow CH_3^+ + H_2$. *J. Phys. Chem. A* **110,** 1569–1574 (2006).

165. Nguyen, T. T., Székely, E., Imbalzano, G., Behler, J., Csányi, G., Ceriotti, M., Götz, A. W. & Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **148,** 241725 (2018).

166. Liu, Y., Li, J., Felker, P. M. & Bačić, Z. HCl–H 2 O dimer: an accurate full-dimensional potential energy surface and fully coupled quantum calculations of intra-and intermolecular vibrational states and frequency shifts. *Phys.Chem. Chem. Phys.* **23,** 7101–7114 (2021).

167. Kondati Natarajan, S., Morawietz, T. & Behler, J. Representing the potential-energy surface of protonated water clusters by high-dimensional neural network potentials. *Phys. Chem. Chem. Phys.* **17,** 8356–8371 (2015).

168. Wang, H., Zhang, L., Han, J. & Weinan, E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228,** 178–184 (2018).

169. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8,** 3192–3203 (2017).

170. Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., Chmiela, S., Tkatchenko, A. & Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. arXiv: `1706.08566` (2017).

171. Profitt, T. A. & Pearson, J. K. A shared-weight neural network architecture for predicting molecular properties. en. *Phys. Chem. Chem. Phys.* **21,** 26175–26183. ISSN: 1463-9084 (2019).

172. Park, C. W., Kornbluth, M., Vandermause, J., Wolverton, C., Kozinsky, B. & Mailoa, J. P. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *Npj Comput. Mater.* **7,** 1–9 (2021).

173. Bishop, R. An overview of coupled cluster theory and its applications in physics. *Theor. Chim. Acta* **80,** 95–148 (1991).

174. Metropolis, N. & Ulam, S. The Monte Carlo method. *J. Am. Stat. Assoc.* **44,** 334–341 (1949).

175. Anderson, J. B. A random-walk simulation of the Schrödinger equation: H_3^+. *J. Chem. Phys.* **63,** 1499–1503 (1975).

176. Anderson, J. B. Quantum chemistry by random walk. H ^2P, H_3^+ D_{3h} $^1A_1'$, H_2 $^3\Sigma_u^+$, H_4 $^1\Sigma_g^+$, Be ^1S. *J. Chem. Phys.* **65,** 4121–4127 (1976).

177. Suhm, M. A. & Watts, R. O. Quantum Monte Carlo studies of vibrational states in molecules and clusters. *Phys. Rep.* **204,** 293–329 (1991).

178. DiRisio, R. J., Lu, F. & McCoy, A. B. GPU-accelerated neural network potential energy surfaces for diffusion Monte Carlo. *J. Phys. Chem. A* **125,** 5849–5859 (2021).

179. Huang, X., Johnson, L. M., Bowman, J. M. & McCoy, A. B. Deuteration effects on the structure and infra-red spectrum of CH_5^+. *J. Am. Chem. Soc.* **128,** 3478–3479 (2006).

180. McCoy, A. B. Diffusion Monte Carlo approaches for investigating the structure and vibrational spectra of fluxional systems. *Int. Rev. Phys. Chem.* **25,** 77–107 (2006).

181. Lee, V. G. M. & McCoy, A. B. An efficient approach for studies of water clusters using diffusion Monte Carlo. *J. Phys. Chem. A* **123,** 8063–8070 (2019).

182. Finney, J. M., DiRisio, R. J. & McCoy, A. B. Guided diffusion Monte Carlo: A method for studying molecules and ions that display large amplitude vibrational motions. *J. Phys. Chem. A* **124,** 9567–9577 (2020).

183. McCoy, A. B., Dzugan, L. C., DiRisio, R. J. & Madison, L. R. Spectral signatures of proton delocalization in $H^+(H_2O)_{n=1-4}$ ions. *Faraday Discuss.* **212,** 443–466 (2018).

184. Barnett, R., Reynolds, P. & W.A Lester, J. Monte Carlo Algorithms for Expectation Values of Coordinate Operators. *J. Comput. Phys.* **96,** 258–276 (1991).

185. DiRisio, R. J. & McCoy, A. B. *rjdirisio/pyvibdmc:1.1.8* version 1.1.8. 2021.

186. Lee, V. G. M., Vetterli, N. J., Boyer, M. A. & McCoy, A. B. Diffusion Monte Carlo studies on the detection of structural changes in the water hexamer upon isotopic substitution. *J. Phys. Chem. A* **124,** 6903–6912 (Aug. 2020).

187. Boyer, M. A., DiRisio, R. J., Finney, J. M. & McCoy, A. B. *McCoyGroup/ PyHPCDMC* version v1.0.0. 2021.

188. Johnson, L. M. & McCoy, A. B. Evolution of structure in CH_5^+ and its deuterated analogs. *J. Phys. Chem. A* **110,** 8213–8220 (2006).

189. Kendall, R. A., Dunning Jr, T. H. & Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **96,** 6796–6806 (1992).

190. Weigend, F. A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Phys. Chem. Chem. Phys.* **4,** 4285–4291 (2002).

191. Colbert, D. T. & Miller, W. H. A novel discrete variable representation for quantum mechanical reactive scattering via the S-matrix Kohn method. *J. Chem. Phys.* **96,** 1982–1991 (1992).

192. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu & Xiaoqiang Zheng. *TensorFlow: Large-scale machine learning on heterogeneous systems* Software available from tensorflow.org. 2015. `https://www.tensorflow.org/`.

193. Lill, J., Parker, G. & Light, J. Discrete variable representations and sudden models in quantum scattering theory. *Chem. Phys. Lett* **89,** 483–489. ISSN: 0009-2614 (1982).

194. Bauman, N. P., Shen, J. & Piecuch, P. Combining active-space coupled-cluster approaches with moment energy corrections via the CC (P; Q) methodology: Connected quadruple excitations. *Mol. Phys.* **115,** 2860–2891 (2017).

195. Eriksen, J. J., Matthews, D. A., Jørgensen, P. & Gauss, J. Assessment of the accuracy of coupled cluster perturbation theory for open-shell systems. I. Triples expansions. *J. Chem. Phys.* **144,** 194102 (2016).

196. Huang, X., McCoy, A. B., Bowman, J. M., Johnson, L. M., Savage, C., Dong, F. & Nesbitt, D. J. Quantum deconstruction of the infrared spectrum of CH_5^+. *Science* **311,** 60–63 (2006).

197. McCoy, A. B., Braams, B. J., Brown, A., Huang, X., Jin, Z. & Bowman, J. M. *Ab initio* diffusion Monte Carlo calculations of the quantum behavior of CH_5^+ in full dimensionality. *J. Phys. Chem. A* **108,** 4991–4994 (2004).

198. Fore, M. E. & McCoy, A. B. Statistical analysis of the effect of deuteration on quantum delocalization in CH_5^+. *J. Phys. Chem. A* **123,** 4623–4631 (2019).