UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2022-027
The Faculty of Science, Technology and Medicine

DISSERTATION

Presented on 03/03/2022 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN PHYSIQUE

by

Valentin VASSILEV GALINDO
Born on 18 December 1991 in Mexico City (Mexico)

MACHINE LEARNING FORCE FIELDS: TOWARDS MODELLING FLEXIBLE MOLECULES

Dissertation defense committee

Dr. Alexandre Tkatchenko, Dissertation Supervisor
Professor, Université du Luxembourg

Dr. Massimiliano Esposito
Professor, Université du Luxembourg

Dr. Anatole von Lilienfeld
Professor, Universität Wien

Dr. Cecilia Clementi
Professor, Freie Universität Berlin

Dr. Etienne Fodor
Professor, Université du Luxembourg

University of Luxembourg
Campus Limpertsberg
162a, avenue de la Faïencerie
L-1511 Luxembourg
Luxembourg

Abstract

Accurate modelling of chemical and physical interactions is crucial for obtaining thermodynamic and dynamical properties of any chemical system, enabling a myriad of possible applications. Many of these applications are computationally prohibitive when using advanced Computational Chemistry (CompChem) methods even on modern supercomputers. Because of this, machine learning (ML) force fields (FFs), combining the accuracy of state-of-the-art *ab initio* methods and the efficiency of classical FFs, are being increasingly used to reconstruct potential-energy surfaces (PESs) of molecules and solids. It is precisely the synergy of ML and CompChem that has revolutionized the field in the last decade, rising the applications to a qualitatively new level. Despite this great success, there are still many unsolved challenges. In this context, my thesis aims to investigate the capability of the existing MLFFs to provide simultaneously accurate and efficient models offering unprecedented insights into the (thermo)dynamics of realistic molecular systems.

Using the examples of molecular interactions that are pervasive in (bio)chemical systems, we show a counterintuitive effect of strengthening of such interactions, as well as an unexpected prevalence of quantum nuclear fluctuations over thermal contributions at room temperature. We reveal that, when dealing with complex PESs, the predictions of state-of-the-art ML models (BPNN, SchNet, GAP, and sGDML) greatly depend on the descriptor used, and on the region of the PES being sampled. Given the varying performance of MLFFs, we present a descriptor optimization scheme improving simultaneously the accuracy and efficiency of ML models. Our results show that the commonly employed strategies followed to construct both local and global descriptors need to be improved because the optimal descriptors are a non-trivial combination of local and global features. Therefore, the work presented in this thesis highlights the potential of MLFFs to provide insights into chemical systems while, at the same time, discloses the current limitations preventing the construction of accurate MLFFs for more realistic systems. Also, I propose the optimization of the description of interactions within an ML model as a valuable step towards obtaining efficient and accurate MLFFs of large and flexible molecules.

Overall, this thesis suggests that the full workflow for building ML models still need significant elaboration. Despite this finding, the combination of CompChem and ML methods in atomistic modelling promises to enable us to solve multiple problems in different areas, such as medicine, materials design, pharmacology, energy production, environmental sciences, among others.

Preface

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Alexandre Tkatchenko, for providing me with the opportunity to work in this wonderful research topic, as well as for the guidance and numerous fruitful discussions. In the same vein, I would also like to thank Dr. Igor Poltavsky, who was always there to discuss new ideas and my research outputs, and to guide me on exploring different alternatives in my research. A special thanks goes to my friends and colleagues at the TCP group for enriching the work environment these past years in Luxembourg.

I would also like to acknowledge the financial support from the Fonds National de la Recherche Luxembourg through the DTU MASSENA (FNR PRIDE/15/10935404) that enabled this work and my attendance to various international workshops and conferences, not to mention the inspiring seminars and events held with other MASSENA PhD candidates.

Last but not least, I sincerely thank my family for all the support over these years. They have been the pillars of my achievements. Without them, nothing of this would be possible.

Note on Previous Publications

This thesis is based on and partly published in the following articles:

1) V. Vassilev-Galindo, G. Fonseca, I. Poltavsky, A. Tkatchenko. “Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules”. *J. Chem. Phys.* **154** (9), 094119 (2021).

2) H. E. Sauceda, V. Vassilev-Galindo, S. Chmiela, K.-R. Müller, A. Tkatchenko. “Dynamical strengthening of covalent and non-covalent molecular interactions by nuclear quantum effects at finite temperature”. *Nat. Commun.* **12** (1), 442 (2021).

3) J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko. “Combining machine learning and computational chemistry for predictive insights into chemical systems”. *Chem. Rev.* **121** (16), 9816-9872 (2021).

Table of Contents

List of Abbreviations

ACSFs	Atom-centered symmetry functions
BO	Born-Oppenheimer
BP	Back-propagation
BPNN	Behler-Parrinello neural network
CC	Coupled cluster
CCSD	Coupled cluster up to double excitations
CCSD(T)	Coupled cluster with single, double, and perturbative triple excitations
CompChem	Computational Chemistry
CV	Collective variable
DFT	Density functional theory
FF	Force field
GAP	Gaussian approximation potential
GGA	Generalized gradient approximation
GP	Gaussian process
HF	Hartree-Fock
KRR	Kernel ridge regression
KS-DFT	Kohn-Sham density functional theory
LJ	Lennard-Jones
MBD	Many-body dispersion
MD	Molecular dynamics
MEP	Minimum energy pathway
ML	Machine learning
NBO	Natural bond orbital
NEB	Nudged elastic band
NN	Neural Network
NQE	Nuclear quantum effects
PBE	Perdew-Burke-Ernzerhof
PES	Potential energy surface

PIMD	Path integral molecular dynamics
RMSE	Root mean squared error
SA	Sensitivity analysis
sGDML	Symmetric gradient domain machine learning
SOAP	Smooth overlap of atomic positions
TS	Tkatchenko-Scheffler
vdW	van der Waals
wACSFs	Weighted atom-centered symmetry functions
WBI	Wiberg bond indices
ZPE	Zero-point energy

List of Figures

List of Tables

CHAPTER 1

Introduction

The potential-energy surface (PES) is the basic element to study the dynamics, thermodynamics, and the reaction kinetics of chemical systems. A faithful description of PESs is crucial for understanding chemical reactions,[1,2] conformational changes in molecules,[3,4] nucleation events during phase transitions,[5,6] and folding and unfolding of proteins,[7–9] to name a few examples of a myriad of possible applications.[10–15] There are different approaches to reproducing PESs in Computational Chemistry (CompChem). Among them, we can find *ab initio* methods,[16] classical force fields (FFs),[17] and machine learning (ML) models.[15] *Ab initio* methods are based on solving the electronic Schrödinger equation with some approximations. In contrast, classical FFs and ML models do not explicitly solve the quantum mechanical problem. Classical FFs use manually predefined analytic functions to model PESs, while ML models are trained on precomputed CompChem datasets without explicitly assuming any functional forms. The selection of a given approach depends on the accuracy and computational cost requirements for a desired application (see Fig. 1.1 for a schematic representation). For instance, applications that only need a few calculations, such as single-point energy calculations and geometry optimizations, can be done using accurate *ab initio* methods. Conversely, the applications that require simulations of long trajectories consisting of millions of steps, like the computation of heat capacities and free energies or dealing with rare events, are limited to efficient FFs.

Overall, *ab initio* approaches, like wavefunction methods or density functional theory (DFT), yield the most accurate chemical/physical properties.[16] The state-of-the-art diffusion Monte Carlo[18] and the coupled cluster method that involves single, double, and perturbative triples excitations [CCSD(T)][19,20] calculations are often used as benchmarks in CompChem since they have been demonstrated to achieve “chemical accuracy” of 1 kcal mol^{-1} for atomization energies of many small molecular systems

(in comparison to experiment). Unfortunately, their excessive computational cost [e.g., diffusion Monte Carlo scales as $O(n^4)$ and CCSD(T) as $O(n^7)$ with respect to the number of electrons] limits their applicability to systems with only tens of atoms. When one needs to deal with larger systems, DFT is widely used due to its better scaling law [$O(n^3)$]. Despite having larger errors than wavefunction methods, DFT calculations with suitable exchange-correlation functionals and appropriate considerations for van der Waals (vdW) dispersion interactions are reliable for many applications, such as performing conformational searches for molecules and predicting physical/chemical properties of materials.[21–24] Their relatively high computational cost, however, still represents an obstacle for computationally demanding tasks, like modelling the dynamics and thermodynamics of realistic systems.

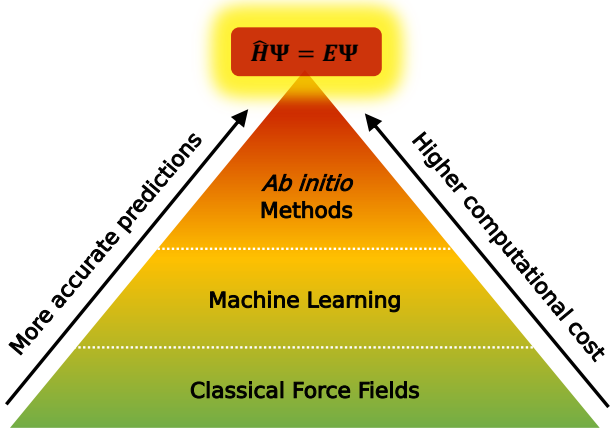

Figure 1.1: Hierarchies in Computational Chemistry methods. Only the methods relevant for this work are presented.

The interest in large scale applications fostered the development of classical FFs.[17,25–32] Such FFs are fitted empirically to experimental and/or CompChem data, and evaluated very efficiently by explicitly considering the interactions between atoms and treating electrons only in an implicit manner. The low computational cost of classical FFs makes them suitable for reconstructing the PESs of large biomolecular systems (e.g., proteins, biopolymers).[33–35] However, this advantage comes at the cost of accuracy because it is very hard to manually predefine the functional forms of all possible interactions within a studied system.[36]

The need for efficient, albeit accurate, PES models shifted the efforts of a broad part of the CompChem community to developing MLFFs.[37–67] MLFFs aim to keep the

accuracy of *ab initio* calculations while achieving an efficiency comparable to that of classical FFs. Contrary to other CompChem methods, MLFFs use CompChem data to estimate functional relationships (or surrogate models) that can recover mappings between molecular configurations and their corresponding energy and/or forces. Plenty of different MLFFs has been proposed in recent years: the Behler-Parrinello neural network (BPNN),[62] SchNet,[66] the symmetric gradient domain machine learning (sGDML),[45,46,48] Gaussian approximation potentials (GAPs),[56] PhysNet,[67] the embedded atom NN (EANN),[39] DeepMD,[40] the deep potential – smooth edition (DeepPot-SE),[41] the hierarchically interacting particle NN (HIP-NN),[42] ANI,[43] and TensorMol,[44] are a few examples. Most of these models allow reconstruction of PESs for a wide range of systems from small organic molecules to bulk condensed materials and interfaces with energy prediction errors under 1 kcal mol^{-1} with respect to the reference *ab initio* calculations.[11,36,44–47,64–86] Typical applications include understanding the origins of electronic and structural transitions in materials,[87] enhancing the sampling of rare events,[88] and computing molecular spectra.[65]

Despite the great success of MLFFs, there are still open challenges. In this context, this thesis aims to investigate in detail the capability of existing MLFFs to provide simultaneously accurate and efficient models offering unprecedented insights into the (thermo)dynamics of realistic molecular systems. To achieve this, we first study the consequences of nuclear quantum effects (NQE) at finite temperature on molecular interactions.[89] Long molecular dynamics (MD) simulations obtained using MLFFs revealed a counterintuitive strengthening of molecular interactions in the aspirin and toluene molecules, as well as a prevalence of quantum fluctuations over thermal contributions at a wide range of temperatures from 50 to 450 K. As a second important step, we assess the ability of state-of-the-art ML models, such as BPNN, SchNet, GAP, and sGDML, to reconstruct the complex PESs of large and/or flexible molecules, which might present several minima and many possible transition paths.[90] We found that ML models present an uneven accuracy across configurational space [up to an order of magnitude difference in root mean squared error (RMSE)], and their descriptors cannot equally resolve the states along different parts of the PES. This clearly demonstrates that typically employed strategies to construct both local (based on finite cutoff radius) and global (all degrees of freedom) descriptors are insufficient when one needs to deal with complex PESs. An optimal descriptor should keep only

some of the features included in local descriptors, together with the information about certain long distances (even beyond 10 Å) that are only present in global descriptors. Such descriptor optimization leads to simultaneously more efficient and accurate MLFFs, which is crucial for practical applications.

This thesis is organized as follows: Chapter 2 provides an introduction to the theoretical background that this work is based on. It contains brief discussions of CompChem and ML methods. Chapter 3 reports the use of MLFFs of the aspirin and toluene molecules to unveil and explain counterintuitive effects induced by NQE at finite temperature on molecular interactions that are ubiquitous in realistic systems.[89] In Chapter 4, we discuss the problems that state-of-the-art MLFFs, such as BPNN, SchNet, GAP, and sGDML, face when reconstructing the complex PESs of flexible molecules.[90] Chapter 5 addresses the need of finding optimal descriptors for obtaining accurate and efficient MLFFs of large and flexible molecules. Finally, Chapter 6 contains a summary of all the results and provides perspectives for future research.

CHAPTER 2

Theoretical Background

Parts of this chapter have been reproduced from
J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller & A. Tkatchenko, *Chem. Rev.* **121**, 9816-9872 (2021)[91]
https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00107
with the permission of ACS Publications. Further permissions related to the material excerpted should be directed to the ACS.

Sec. 2.1 summarizes CompChem methods relevant to the content of this thesis. In particular, wavefunction theory methods, DFT, and interatomic potentials are covered. For a deeper description of the methods, the reader is referred to one of the many books available on the topic. See Refs. 92–96, for example. The section concludes with an introduction to MD simulations, which includes a discussion of nuclear quantum effects (NQE).

Sec. 2.2 introduces the field of ML in general terms and provides an overview of the overall construction of ML models, starting from a dataset all the way to a trained and tested model. The section concludes with a brief introduction to MLFFs. Further details can be found in the original publications and reviews (e.g., Refs. 15,71,91).

2.1 Basics of Computational Chemistry Methods

Here, quantum mechanics is described by the nonrelativistic time-independent Schrödinger equation

$$\hat{H}\Psi = E\Psi \qquad (1)$$

where the Hamiltonian operator ($\hat{H}$) is the sum of the kinetic ($\hat{T}$) and potential ($\hat{V}$) operators, Ψ is the wavefunction (i.e., an eigenfunction) that represents the electrons

and nuclei in the system, and E is the energy (i.e., an eigenvalue). This is considered as the "standard model" because it accurately represents the physics of charged particles (electrons and nuclei) that make up almost all molecules and materials.

In a real system, however, the position and interactions of a single particle in the system with respect to all other particles will be correlated, and this makes the exact solution of the Schrödinger equation impossible for almost all systems of practical interest. Thus, different types of approximations can be exploited to make the problem more tractable. For instance, the Born-Oppenheimer (BO) approximation considers nuclei as fixed point charges at any part of the PES since they are expected to move much slower than electrons. Eq. 1, then, can be transformed into the electronic Schrödinger equation:

$$\begin{aligned}\hat{H}_{el}\Psi_{el}(\boldsymbol{r};\boldsymbol{R}) &= \left[\hat{T}_e(\boldsymbol{r}) + \hat{V}_{eN}(\boldsymbol{r};\boldsymbol{R}) + \hat{V}_{NN}(\boldsymbol{R}) + \hat{V}_{ee}(\boldsymbol{r})\right]\Psi_{el}(\boldsymbol{r};\boldsymbol{R}) \\ &= E_{el}\Psi_{el}(\boldsymbol{r};\ \boldsymbol{R})\end{aligned} \qquad (2)$$

where the Hamiltonian $\hat{H}_{el}$ and wavefunction $\Psi_{el}(\boldsymbol{r};\boldsymbol{R})$ only depend parametrically on the nuclear coordinates $\boldsymbol{R}$. The above expression is composed of single electron (*e*), electron-nuclear (*eN*), nuclear-nuclear (*NN*), and electron-electron (*ee*) terms.

Any theoretical method for predicting molecular or material phenomena must first be rooted in quantum-mechanical theory and then suitably coarse-grained and approximated so that it can be applied in a practical setting. CompChem, or more precisely, computational quantum chemistry defines computationally driven numerical analyses based on quantum mechanics.

This section provides a concise overview of some of the CompChem methods that are relevant in the context of this thesis.

2.1.1 Wavefunction Theory Methods

As mentioned before, the BO approximation can be applied to transform the time-independent Schrödinger equation (Eq. 1) into the electronic Schrödinger equation (Eq. 2), where nuclei are considered as fixed point charges. A second common approximation is to expand the total electronic wavefunction in terms of one-electron wave-functions (i.e., spin orbitals): $\phi(\boldsymbol{r}_i)$. Electrons are Fermions and therefore exhibit

antisymmetry, which in turn results in the Pauli exclusion principle. Antisymmetry means that the interchange of any two particles within the system should bring an overall sign change to the wavefunction (i.e., from + to −, or vice versa). This property is conveniently captured mathematically by combining one electron spin orbitals into the form of a Slater determinant:

$$\Psi(r_1, \cdots, r_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(\boldsymbol{r_1}) & \cdots & \phi_n(\boldsymbol{r}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{r}_n) & \cdots & \phi_n(\boldsymbol{r}_n) \end{vmatrix} \quad (3)$$

The spin orbitals can be treated as a mathematical expansion using a basis set of μ functions χ_μ, each having coefficients $c_{\mu i}$, which are generally Gaussian basis functions,[97–99] Slater-type hydrogenic orbitals,[100] or plane waves under periodic boundary conditions:[101–103]

$$\phi_i = \sum_\mu c_{\mu i} \chi_\mu \quad (4)$$

The different types of mathematical functions bring different strengths and weaknesses, but these will not be discussed further here. A universal point is that larger basis sets will have more basis functions and thus give a more flexible and physical representation of electrons within the system. On one hand this can be crucial for capturing subtle electronic structure effects due to electron correlation. On the other hand, larger basis sets also necessitate significantly higher computational effort.

Using the BO approximation (Eq. 2) together with a Slater determinant wavefunction (Eq. 3) expressed in a finite basis set (Eq. 4) brings about the simplest wavefunction based method, the Hartree−Fock (HF) approach (for historical context see Refs. 104–106). The HF method is a mean field approach, where each electron is treated as if it moves within the average field generated by all other electrons. It is generally considered inaccurate when describing many chemical systems, but it continues to serve as a critical pillar for CompChem electronic structure calculations since it either establishes the foundation for all other accurate methods or provides energy contributions (i.e., exact exchange) that is not provided in some CompChem methods.

CompChem methods that achieve accuracy higher than HF theory are said to contain electron correlation, a critical component for understanding molecules and materials. While correlation energy makes up a rather small contribution to the overall energy of a system (usually about 1 % of the total energy), because internal energies in molecular and material systems are so enormous, this contribution becomes rather significant. As an example, most molecular crystals would be unstable as solids if calculated using the HF level of theory. The missing component is attractive forces that are obtained from levels of theory that account for correlation energy. Correlation energies are obtained by calculating additional electron−electron interaction energies that arise from different arrangements of electron configurations (i.e., different possible excitations) that are not treated with the mean-field approach of HF theory. For instance, coupled cluster theory, a widely used post-HF method, includes additional electron configurations via cluster operators.[107] One coupled cluster method that involves single, double, and perturbative triples excitations, CCSD(T), is referred to as the "gold-standard" approach for CompChem electronic structure methods since it has been demonstrated to achieve "chemical accuracy" of 1 kcal mol^{-1} for many small molecular systems (in comparison to experiment).

A key point is that correlated wavefunction methods are founded on the HF theory, and so they are even more computationally demanding than HF calculations [$O(n^4)$], for example, CCSD(T) scales poorly with the number of electrons in the system as $O(n^7)$.

2.1.2 Density Functional Theory

DFT[96] is another method to calculate the quantum mechanical energy of a system using an energy expression that relies on functionals of electronic density $\rho = \left|\boldsymbol{\Psi}_{el(\boldsymbol{r};\boldsymbol{R})}\right|^2$:

$$E[\rho] = T[\rho] + V[\rho] \qquad (5)$$

Compared to wavefunction theory, DFT should be far more efficient since the dimensionality of a density representation for electrons will always be three rather than the 3*n* dimensions for any *n*-electron system described by a many-body wavefunction

method. DFT has an important drawback that the exact expression for the energy functional is currently unknown, all approximations bring some degree of uncontrollable error, and this has precipitated disagreeable opinions from purists in chemical physics, especially those who are developing correlated wavefunction methods. However, there is also substantial evidence that DFT approximations are reasonably reliable and accurate for many practical applications that bring information, knowledge, and sometimes insight.

The most used form of DFT (which is also one of the most widely used CompChem methods in use today) is called Kohn-Sham (KS-)DFT.[108] In KS-DFT, one assumes a fictitious system of noninteracting electrons with the same ground state density as the real system of interest. This makes it possible to split the energy functional in Eq. 5 into a new form that involves an exact expression of the kinetic energy for noninteracting electrons:

$$E[\rho] = T_{ni}[\rho] + V_{eN}[\rho] + V_{ee}[\rho] + \Delta T_{ee}[\rho] + \Delta V_{ee}[\rho] \qquad (6)$$

Here, $T_{ni}[\rho]$ is the kinetic energy of the noninteracting electrons, $V_{eN}[\rho]$ is the exact nuclear-electron potential, and $V_{ee}[\rho]$ is the Coulombic (classical) energy of the noninteracting electrons. The last two terms are corrections due to the interacting nature of electrons and nonclassical electron−electron repulsion, and these are combined into the so-called "exchange-correlation" term (E_{xc}), which uniquely defines which scheme of KS-DFT is being used. The formal scaling for KS-DFT is $O(n^3)$ with respect to the number of electrons. There are several modern exchange-correlation functionals that routinely achieve much higher accuracy than HF theory with less computational cost, and thus KS-DFT is a competitive alternative to many correlated wavefunction methods in for applications to molecules and materials.

A remaining problem is constructing a practical expression for the exchange-correlation functional, as its exact functional form remains unknown. This has spawned a wealth of approximations that have been founded with different degrees of first principles and/or empirical schemes. Classes of KS-DFT functionals are defined by whether the exchange-correlation functional is based on just the homogeneous electron gas (i.e., the "local density approximation", LDA), that and its derivative (i.e., the "generalized gradient approximation", GGA), as well as other additional terms that

should result in physically improved descriptions or error cancellations. For instance, hybrid DFT functionals (e.g., PBE0[109]) include a fraction of HF exact exchange repulsion, which reduces some of the shortcomings of pure GGA functionals.

2.1.3 Interatomic Potentials

Interatomic potentials, or force fields (FFs), introduce levels of abstraction compared to methods described above. Instead of using quantum mechanical expressions to create the PES for the system, analytic functions are used to model a presupposed PES that contains explicit interactions between atoms, while electrons are treated in an implicit manner (sometimes using partial charge schemes).[110–115] Interatomic potentials thus are (oftentimes drastically) more computationally efficient than correlated wavefunction and DFT. This efficiency makes it possible to study even larger systems of atoms (e.g., biomolecules, surfaces, and materials) than is possible with other computational methods. Note that different empirical potentials bring substantially different computational efficiencies. The degree of efficiency arises from the balance of using accurate or physically justified functional forms, approximations, and model parametrizations. While efficient and versatile, the accuracy of all interatomic potentials is inherently constrained by their functional forms. There are many different formulations, but a typical energy expression for classical FFs is

$$E_{tot} = \sum_{bonds} k_{ij}\left(r_{ij} - \bar{r}_{ij}\right)^2 + \sum_{angles} k_{ijk}\left(\theta_{ijk} - \bar{\theta}_{ijk}\right)^2 + \sum_{dihedrals} \sum_{\gamma} k_{ijkl}^{(\gamma)}\left[1 - cos\left(\gamma\phi_{ijkl} - \bar{\phi}_{ijkl}^{(\gamma)}\right)\right] + \sum_{i,j>i} \frac{q_i q_j}{r_{ij}} + E_{LJ} \quad (7)$$

where the first three terms are the energy contributions of the distances (r_{ij}), angles (θ_{ijk}) and dihedral angles (ϕ_{ijkl}) between bonded atoms. Because of this, they are also referred to as bonded contributions. Bond and angle energies are modeled via harmonic potentials, with the k_{ij} and k_{ijk} parameters modulating the potential strength and $\bar{r}_{ij}$ and $\bar{\theta}_{ijk}$ are the equilibrium distances and angles. The dihedral term is modeled with a Fourier series to capture the periodicity of dihedral angles, with k_{ijkl} and γ as free parameters. The last two terms account for nonbonded interactions. The long-

range electrostatics are modeled as the Coulomb energy between charges q_i and q_j, and the vdW energy is treated via a Lennard-Jones potential

$$\mathrm{E_{LJ}} = \sum_{\mathrm{i,j>i}} 4\varepsilon_{\mathrm{ij}} \left[\left(\frac{\sigma_{\mathrm{ij}}}{\mathrm{r_{ij}}} \right)^{12} - \left(\frac{\sigma_{\mathrm{ij}}}{\mathrm{r_{ij}}} \right)^{6} \right] \qquad (8)$$

where ε_{ij} modulates the strength of the interaction function and σ_{ij} defines where it reaches its minimum.

In FFs, empirical parameters are tabulated for a variety of elements and different chemical environments (for example Ref. 116). Parameters for any one system should not necessarily be assumed to transfer well to other systems, and reparametrizations may be needed depending on the application. Different sets of parametrization schemes give rise to different types of classical FFs, with CHARMM,[25] Amber,[26,27] GROMOS,[28–30] and OPLS[31,32] being a few of many examples.

For further discussions on these methods see the extensive review by Akimov and Prezhdo.[16] Also, an excellent review for interatomic potentials is provided by Harrison *et al.*,[17] and an excellent overview of modern methods can be found in a special issue of *J. Chem. Phys*.[117]

2.1.4 Molecular dynamics

So far, only the electronic structure problem has been addressed. However, the interactions between nuclei need to be considered with methods, such as MD, for properly studying the dynamical properties of molecules and materials.

In MD simulations, the motion of nuclei, despite being quantum mechanical in nature, is computed with the principles of classical mechanics. Namely, Newton's equations of motion are numerically solved to propagate the evolution of nuclei on the PES in time. MD averages computed over time provide insights into properties, such as average thermodynamic quantities, structure, and free energies along transition paths on the PES.

Nuclear Quantum Effects

A classical treatment of the nuclei is oftentimes sufficient to provide accurate simulations, but the quantum nature of lighter elements, such as H−Li, and even heavier elements that form strong chemical bonds (C−C bond in graphene for example[118]) gives rise to significant nuclear quantum effects (NQEs). Such effects are a consequence of the impact of zero-point energy (ZPE) and quantum tunneling, and are responsible, for instance, for large differences from the Dulong−Petit limit of the heat capacity of solids, isotope effects, and the deviations of the particle momentum distribution from the Maxwell-Boltzmann equation.[119]

One way to capture NQEs is the path-integral molecular dynamics (PIMD) method based on Feynman path integrals. PIMD generates the quantum-mechanical ensemble of a system by relating the statistics of a quantum particle to that of several replicas of a particle coupled by harmonic springs. The result of this approach is exact in the limit of infinite replicas, but in practice only a finite number of them is needed to achieve convergence.

For obtaining a reliable property prediction with (PI)MD simulations, one requires simulation times comparable to the timescale of the real process of interest. This requirement limits the applicability of accurate electronic-structure methods, since their use results in quite computationally demanding simulations that are normally restricted to small system sizes and short simulation times. Alternatives, such as ML methods constructed using accurate data, are appealing to overcome this issue.

2.2 Basics of Machine Learning

ML has had a drastic impact on many aspects of our daily lives and has arguably become one of the most far-reaching technologies of our era. It is hard to overstate its importance in solving long-standing computer science challenges, such as image classification[120–122] or natural language processing,[123–126] tasks that require knowledge that is hard to capture in a traditional computer program.[127,128] Over the past decade, advances in ML algorithms and computer technology made it possible to learn underlying regularities and relevant patterns from massive data sets that enable automatic constructions of powerful models that can sometimes even outperform

humans at those tasks. This development inspired researchers to approach challenges in science with the same tools, driven by the hope that ML would revolutionize their respective fields in a similar way.

In the most general sense, ML algorithms estimate functional relationships without being given any explicit instructions of how to analyze or draw conclusions from the data. Learning algorithms can recover mappings between a set of inputs and corresponding outputs. Specifically, ML techniques aim to identify functions to predict interpolations between data points and thus minimize the prediction error for new data points that might later appear.[129] Such functions, or models, $\hat{f}$ satisfy

$$\hat{f} = \underset{f \in \mathcal{F}}{arg\,min} \left[\sum_{i}^{M} \mathcal{L}(f(\boldsymbol{x}_i), y_i) + \left|\left|\Gamma\Theta\right|\right|^2 \right] \qquad (9)$$

where $\left|\left|\Gamma\Theta\right|\right|^2$ is a regularization term, which influences the selection of candidate models by introducing additional properties that promote generalization. Γ is a matrix that defines "simplicity" with regard to the model parameters Θ. Usually, $\Gamma = \lambda I$ (where I is the identity matrix and $\lambda > 0$) is chosen to simply favor a small L^2-norm on the parameters, such that the solution does not rely on individual input features too strongly.

A model that is heavily regularized (i.e., using a large λ) will eventually become biased in that it is too simplistic to fit the data well. In contrast, a lack of regularization might yield an overly complex model with high variance. Such an overfitted model will follow the data exactly to the point that it also models the noise components and consequently fails to generalize. Finding the appropriate amount of regularization λ to manage under- and overfitting is known as attaining a good bias-variance trade-off.[130]

In what follows, an overview of the different stages involved in finding a suitable model $\hat{f}$ is presented.

2.2.1 ML Workflow

The ML workflow typically includes the following stages:

A. Gathering and preparing the data

B. Choosing a representation
C. Training the model
    1. Train model candidates
    2. Evaluate model accuracy
    3. Tune hyperparameters
D. Testing the model out of sample

Note, that the progression to a good ML model is not necessarily linear and some steps (except the out of sample test) may require reiteration as one learns about the problem at hand.

Datasets

On a fundamental level, ML models could be simply regarded as sophisticated parametrized functions of datasets. While the architectural details of the model matter, the reference data set forms the backbone that ultimately determines the effectiveness of the model. If the data set is not representative of the problem at hand, the model will be incomplete and behave unpredictably in situations that have been improperly captured. The same applies to any other shortcomings of the dataset, such as biases or noise artifacts that will also be reflected in the model.

Robust models can generally only be constructed from comprehensive datasets, but it is possible to incorporate certain patterns into models to make them more data-efficient. Prior scientific knowledge or intuition about specific problems can be used to reduce the function space from which an ML algorithm has to select a solution. If some of the unphysical solutions are removed a priori, less data are necessary to identify a good model.

Descriptors

To apply ML, the dataset needs to be encoded into a numerical representation (i.e., features/descriptors) that allows the learning algorithm to extract meaningful patterns and regularities.[52,58,60,131–137] This is particularly challenging for unstructured data like molecular graphs that have well-defined invariable or equivariable characteristics that are hard to capture in a vectorial representation. For example, if one encodes a molecule using the Cartesian coordinates ($\mathcal{R}^{3N}$) and the chemical elements ($\mathcal{Z}^{N}$) of all

its N atoms, one would introduce substantial redundancies into the ML model (e.g., an ML model might consider two identical molecules that are rotated or translated as different molecules). To avoid these problems, atomic coordinates are transformed into an appropriate descriptor, tailored for a given task. This transformation allows the incorporation of physical invariances, such as rotations about an axis or the translation of the whole molecule in Cartesian space. The relevance of a descriptor in the accuracy of an ML model has motivated the design of multiple different descriptors,[52–58,60,131–135,138–148] and efforts to find suitable representations are still ongoing. Descriptors are designed to emphasize particular aspects of a system or to disambiguate similar chemical or physical principles across different molecules or materials. The set of features contained in the descriptor normally depends on the task at hand and on the chemical physical intuition.

Training

The training process is the key step that ties together the dataset and model architecture. Through the choice of the model architecture, we implicitly define a function space of possible solutions, which is then conditioned on the training data set by selecting suitable parameters. This optimization task is guided by a loss function that encodes the two somewhat opposing objectives: i) achieving a good fit to the data, while ii) keeping the parametrization general enough such that the trained model becomes applicable to data that is not covered in the training set (see the two terms in Eq. 9). Satisfying the latter objective involves a process called model selection in which a suitable model is chosen from a set of variants that have been trained with exclusive focus on the first objective. Depending on the model architecture, more or less sophisticated optimization algorithms can be applied to train the set of model candidates. In the context of CompChem, both NNs[37,38,49,60,62–66] and kernel-based methods[45–48,50–59,61] are the most used model architectures.

Kernel-based learning algorithms are typically linear in their parameters

$$f(\boldsymbol{x}) = \sum_{j=1}^{M} \alpha_j K(\boldsymbol{x}_j, \boldsymbol{x}) \qquad (10)$$

where α_j are the parameters adapted to the data, $\boldsymbol{x}$ denotes a query sample, $\boldsymbol{x}_j$ denotes the *j*-th training sample, M is the number of training points and K is the so-called kernel. Eq. 10 coupled with a quadratic loss function, $\mathcal{L}\big(\hat{f}(\boldsymbol{x}), y\big) = \big(\hat{f}(\boldsymbol{x}) - y\big)^2$, yield a convex optimization problem. Convex problems can be solved quickly and reliably due to only having a single solution that is guaranteed to be globally optimal. This solution can be found algebraically by taking the derivative of the loss function and setting it to zero. For example, kernel ridge regression (KRR) and Gaussian processes (GPs) then yield a linear system of the form

$$\nabla_{\boldsymbol{\alpha}} \mathcal{L}\big(\hat{f}(\boldsymbol{x}), y\big) = (\boldsymbol{K} + \lambda I)\boldsymbol{\alpha} - y = 0 \qquad (11)$$

which is typically solved in a robust way by factorizing the kernel matrix K. Factorization approaches are, however, only feasible if enough memory is available to store the matrix factors, and this can be a limitation for large-scale problems. In that case, numerical optimization algorithms provide an alternative. They take a multistep approach to solve the optimization problem iteratively by following the gradient:

$$\boldsymbol{\alpha}^t = \boldsymbol{\alpha}^{t-1} - \gamma \underbrace{\nabla_{\alpha} \mathcal{L}\big(\hat{f}(\boldsymbol{x}), y\big)}_{e.g.,(\boldsymbol{K}+\lambda I)\boldsymbol{\alpha}^{t-1} - y} \qquad (12)$$

where γ is the step size (or learning rate). Iterative solvers follow the gradient of the loss function until it vanishes at a minimum, which is much less computationally demanding per step, because it only requires the evaluation of the model $\hat{f}$.

On the other hand, NNs are constructed by nesting nonlinear functions in multiple layers.

$$f_k(x) = g\left(\sum_{j=1}^{H} w_{kj}\, g\left[\sum_{i=1}^{M} w_{ji} x_i - b\right]\right) \qquad (13)$$

Here, $g(\cdot)$ denotes an activation function that is the nonlinear transformation that allows complex mappings between input and output, and $\boldsymbol{x}$ denotes the M-dimensional

input to the node. The nodes have parameters consisting of w (so-called weights) and b (a so-called threshold) that are learned from data.

Eq. 13 yields nonconvex optimization problems and closed-form solutions like Eq. 11 do not exist. Hence, the parameters of multilayer NNs are learned using iterative algorithms that compute the gradient of the loss-function (analogous to Eq. 12) using the back-propagation (BP) algorithm.[149–151] Because of multiple local minima and saddle points on the loss surface, the global minimum is exponentially hard to obtain (since iterative algorithms usually converge to a local minimum). However, thanks to the strong modeling power of NNs, local solutions are usually good enough.[152]

In addition to the parameters that are determined when fitting an ML model to the data set (i.e., the node weights/biases in Eq. 13 or regression coefficients in Eq. 10), many models contain hyperparameters that need to be fixed before training. Two types of hyperparameters can be distinguished: ones that influence the model, such as the type of kernel or the NN architecture, and ones that affect the optimization algorithm, for example, the choice of regularization scheme or the aforementioned learning rate. Both tune a given model to the prior beliefs about the data set and thus play a significant role in model effectiveness. Hyperparameters can be used to gauge the generalization behavior of a model.

Hyperparameter spaces are often rather complex: certain parameters might need to be selected from unbounded value spaces, others could be restricted to integers or have interdependencies. This is why they are usually optimized using primitive exhaustive search schemes like grid or random searches in combination with educated guesses for suitable search ranges. Common gradient-based optimization methods typically cannot be applied for this task. Instead, the performance of a given set of hyperparameters is measured by evaluating the respective model on another training dataset called the validation dataset.

Out-of-Sample Testing

For a reasonably complex model, it is typically not challenging to generate the right responses for the data known from the training set. This is why the training error is not indicative of how the model will fulfill its ultimate purpose of predicting responses for new inputs. Alas, since the probability distribution of the data is typically unknown, it

is not possible to determine this generalization error exactly. Instead, this error is often estimated using an independent test subset that is held back and later passed through the trained model to compare its responses to the known test labels. If the model suffers from overfitting on the training data, this test will yield large errors. It is important to remember not to tweak any parameters in response to these test results, as this will skew this assessment of the model performance and will lead to overfitting on the test set.[140]

2.2.2 ML Potentials

In CompChem, one important type of ML models that are constructed following the workflow presented above are MLFFs.

Contrary to interatomic potentials introduced in Sec. 2.1 that require an explicit parametrization, MLFFs find suitable functional expressions for interactions and energy in a fully data-driven manner and are only limited by the amount and quality of available reference data. Because of this, MLFFs have emerged as way to achieve as high accuracy as KS-DFT or correlated wavefunction methods but with a fraction of the cost.

Training an MLFF to reproduce the PES of a system usually requires generating diverse and high quality CompChem data points that cover the relevant temperature and pressure conditions, reaction pathways, polymorphs, defects, compositions, etc.[153–159] After data points comprised of atomic configurations, system energies, and forces are obtained, different methods for constructing MLFFs employ either different descriptors or different ML architectures to perform interpolations of the sampled parts of the PES. Smoothness is an essential feature for any PES, so special considerations are needed to avoid numerical noise that would result in discontinuities.[160,161] Some of the most important MLFFs will be introduced in the following paragraphs.

Behler-Parrinello Neural Networks

The Behler-Parrinello NN (BPNN)[62] is a descriptor-based NN, which is the first developed NN architecture for training MLFFs. The method obtains molecular properties as the sum of individual atomic contributions. These contributions are assumed to depend on the local chemical environment of each atom. Because of such

assumption, the output does not change when the input is permuted, and models can be transferred between system sizes whenever a sufficient sampling of the local environments is performed and a suitable correction for long-range effects is considered.

In BPNNs, local environments are described by atom-centered symmetry functions (ACSFs)[60] consisting of two body terms

$$G_i^2 = \sum_{i \neq i}^{N} e^{-\eta\left(r_{ij} - r_s\right)^2} \, f_{cut}\left(r_{ij}\right) \qquad (14)$$

and three body terms

$$G_i^3 = 2^{1-\zeta} \sum_{j,k \neq i}^{N} \left(1 + \lambda cos\theta_{ijk}\right)^{\zeta} e^{-\eta\left(r_{ij}^2 + r_{ik}^2 + r_{jk}^2\right)} \times f_{cut}\left(r_{ij}\right) f_{cut}\left(r_{ik}\right) f_{cut}\left(r_{jk}\right) \qquad (15)$$

Where r_{ij} is the distance between atoms i and j, θ_{ijk} the angle spanned by atoms i, j and k centered around i, and N is the number of atoms. The parameters η, r_s, ζ, and λ determine the sensitiveness of ACSFs to the different distances, or combination of angles and distances. The cutoff function $f_{cut}(r)$ ensures the smoothness of G_i^2 and G_i^3, and its value depends on a predefined cutoff distance r_{cut}

$$f_{cut}(r) = \begin{cases} \frac{1}{2}\left[cos\left(\frac{\pi r}{r_{cut}}\right) + 1\right], & r \leq r_{cut} \\ 0, & r > r_{cut} \end{cases} \qquad (16)$$

ACSFs only use geometric information, so they work best for systems containing only atoms of one element. To improve its performance, a modification, labelled as weighted ACSFs (wACSFs),[162] accounts for the composition of the chemical environment in an implicit manner by introducing element-dependent weighting functions to Eqs. 14 and 15, instead of using separate functions to describe different combinations of elements.

SchNet

SchNet[66] is an end-to-end NN architecture. This means that the NN can learn the descriptors for molecules and materials, following fundamental symmetries of atomistic systems by construction. Such data-driven representations are constructed by iteratively refining a set of atom-wise features x_i based on interactions with neighboring atoms. Higher-order interactions can then be captured in a hierarchical fashion. For example, a first information pass would only capture radial information, but further interactions would recover angular relations and so on. In SchNet, continuous convolutions are used to describe the interactions, so the update function depending on the atomic environment takes the form

$$\boldsymbol{x}_i^{l+1} = \boldsymbol{x}_i^l + NN_{tr}^l\left(\sum_j \boldsymbol{x}_j^l NN_{rad}^l\left[\boldsymbol{g}^l\left(|\boldsymbol{r}_i - \boldsymbol{r}_j|\right)\right] f_{cut}\left(|\boldsymbol{r}_i - \boldsymbol{r}_j|\right)\right) \qquad (17)$$

Here, l indicates the number of overall update steps. The sum runs over all atoms j in the local environment, and a cutoff function f_{cut} ensures smoothness of the representation. Each feature is updated with information from all neighboring atoms with a feature transformation (NN_{tr}) and a radial dependence (NN_{rad}), which are both modeled as trainable NNs

Even though SchNet primarily learns local features, interaction blocks in its architecture embed such features into the local environments of different parts of the molecule to learn other interactions.

Gaussian Approximation Potentials

Gaussian approximation potentials (GAPs)[56] are an implementation of GPs regression to fit quantum mechanical data. Like BPNNs, GAPs consider the total energy as the sum of the energy of atom-centered environments. Such environments are defined by a cutoff radius that guarantees that atomic energy contributions are local, and no discontinuities are introduced.

Despite quantum mechanical calculations only provide total energies and its derivatives, atom-wise energy contributions can be approximated by GPs. Specifically,

one can construct a kernel function for two given systems $\boldsymbol{x}$ and $\boldsymbol{x}'$ with N and N' atoms, respectively, by summing over all kernels of the local environments

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{N} \sum_{j=1}^{N'} K_{local}(x_i, x_j') \quad (18)$$

For representing local environments within the GAP method, the smooth overlap of atomic positions (SOAP)[58] descriptor has been widely used. To construct SOAP descriptors, one first considers an atomic environment $\mathcal{X}$ and a Gaussian function of width σ is then placed on each atom i in $\mathcal{X}$ to make an atomic density function

$$\rho_{\mathcal{X}_i}(\boldsymbol{r}) = \sum_{i \in \mathcal{X}} exp\left[-\frac{|\boldsymbol{r} - \boldsymbol{r}_i|^2}{2\sigma^2}\right] f_{cut}(|\boldsymbol{r}|) \quad (19)$$

Here, $\boldsymbol{r}$ denotes a point in Cartesian space, $\boldsymbol{r}_i$ is the position of atom i relative to the central atom of $\mathcal{X}$, and the cutoff function f_{cut} smoothly decays to zero beyond the cutoff radius. This density representation ensures invariance with respect to translations and permutations of atoms of the same species but not rotations. Rotational invariance is obtained when constructing the SOAP kernel with an integral over rotations $\mathcal{R}$ of atomic densities

$$K(\rho, \rho') = \int d\mathcal{R} \left| \int \rho(\boldsymbol{r}) \rho'(\mathcal{R}\boldsymbol{r}) d\boldsymbol{r} \right|^n \quad (20)$$

In practice, the density is expanded in a basis of spherical harmonics and a set of orthogonal radial functions to evaluate the kernel in Eq. 20.

Gradient Domain Machine Learning

Gradient domain machine learning (GDML)[45] is a kernel-based ML method that allows the construction of complex multidimensional PESs by combining rigorous physical laws with data-driven ML techniques. Contrary to most other MLFFs, it learns directly from forces and its flexible nature can recover both local and non-local interactions.

For avoiding the violation of energy conservation, conservative forces are ensured by using a kernel that models the forces $\boldsymbol{F}$ as a transformation of an unknown PES E

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \nabla_{\boldsymbol{x}} K_E(\boldsymbol{x}, \boldsymbol{x}') \nabla_{\boldsymbol{x}'}^T \quad (21)$$

Then, GDML solves the normal equation of the ridge estimator, but it does it in the gradient domain since the kernel defined in Eq. 21 corresponds to the Hessian matrix of the original kernel K_E.

To disambiguate Cartesian coordinates, GDML encodes the chemical structure in a descriptor $\boldsymbol{x}$ consisting of the inverse of all pairwise distances. Such a representation is rotationally and translationally invariant but is not permutationally invariant. Because of this, an extension of the original method, symmetric GDML (sGDML),[46,48] exploits all the static and dynamic symmetries of the molecules to achieve permutational invariance. The imposition of relevant symmetries reduces the complexity of the learning process without incurring in additional computational costs.

In the following chapters, the use of MLFFs trained on accurate electronic-structure data will be addressed by highlighting their power to obtain insights into chemical systems, and by discussing the challenges and improvement opportunities that can extend their success to large and flexible molecules.

CHAPTER 3

Strengthening of Molecular Interactions by Nuclear Quantum Effects at Finite Temperature

---

The ML models and MD simulations, on which the results of Section 3.2 in this chapter are based, as well as the results shown in Fig. 3.2 and Fig. 3.4 were done by Dr. Huziel E. Sauceda.
Parts of this chapter have been published in this or similar form in
H. E. Sauceda, V. Vassilev-Galindo, S. Chmiela, K.-R. Müller & A. Tkatchenko, *Nat. Commun.* **12**, 442 (2021) (distributed under CC BY)[89]
and have been produced in a collaboration of the above authors.

Nuclear delocalization is a fundamental feature of quantum mechanics resulting from Heisenberg's uncertainty principle. In molecules, light elements such as protons and first row atoms are especially prone to delocalization. Even molecules or materials with heavier atoms and strong bonds can exhibit significant NQE.[118,163–167] Within the BO approximation, NQE tend to lower energy barriers and stimulate tunnelling. In addition, the inclusion of NQE promotes a delocalized sampling of the molecular configuration space, consequently exploring regions of the PES inaccessible by classical dynamics. As a result, this can enhance or inhibit certain molecular interactions.[168] A clear example is the hydrogen bond, where the NQE affect interactions in biological systems and molecular crystals by delocalizing protons. In the case of bulk water, NQE can even qualitatively change its fundamental physical and chemical properties.[169,170]

In general, the study of NQE in molecular and biological systems is a thriving research field covering from rigid and fluxional molecules[45,48,171–173] to liquids and DNA base pairs,[119,174–178] allowing the analysis of, for example, interactions between neighboring

molecules via hydrogen bonding,[168,169,179] spectroscopic properties[46,180] and proton transport.[181] Nevertheless, previous works have been mainly focused on the general implications of proton delocalization and much less is known about how NQE influence other types of interactions. Particularly, biological systems often use combinations of covalent and non-covalent interactions for carrying out a wide variety of different processes. Therefore, it is crucial to understand whether NQE can also play an important role beyond hydrogen bonding, where, for instance, they generate two competing phenomena in water: a weakening and a strengthening of interactions, leading to a reduction of the viscosity of liquid water and to an increment in the dipole moment in water molecules, respectively.[171]

In this chapter, counterintuitive effects induced by NQE are reported: nuclear delocalization can lead to a dynamical strengthening of different molecular interactions. The conclusions are shown to be valid for different molecules. In order to demonstrate these results, representative mechanisms ubiquitously occurring in biological systems have been selected: $n \rightarrow \pi^*$ interactions, and methyl rotors. The faithful description of such weak molecular interactions require high levels of theory (e.g., coupled cluster) which is not always computationally affordable when performing long *ab initio* PIMD simulations. In this study, PIMD simulations have been performed using MLFFs constructed using the sGDML framework[45–48,182–184] and trained on coupled cluster reference data (CCSD(T) or CCSD depending on the size of the molecule).

3.1 Reference Data Generation and Computational Details

(PI)MD simulations were performed using the i-PI package[185] coupled with sGDML FFs trained on 1000 configurations of the MD17 datasets[45] of aspirin and toluene at the CCSD/cc-PVDZ and CCSD(T)/cc-PVDZ levels of theory, respectively. All the simulations presented here were done for at least 500 ps with time steps of 0.2 fs. For an initial assessment of the impact of NQE on molecular interactions, (PI)MD simulations at room temperature (300 K) were run for aspirin and toluene.

Additional simulations for aspirin were run at temperatures from 50 to 450 K in intervals of 50 K to analyze the interactions and NQEs as a function of temperature.

To properly account for NQE at low temperatures in the PIMD formalism, the number of beads was set, for most of the temperatures, using the relation

$$P = \frac{4800}{T} \qquad (22)$$

where P is the number of beads, T is the temperature and the constant value of 4800 was calculated by considering that $P = 16$ when $T = 300$ K. Table 3.1 shows the values of P for all temperatures considered in this study. When T was greater than 300 K, P was kept fixed to 16. Since only even values of P can be used, the number of beads was set to 20 for $T = 250$ K, instead of 19, which is the result of Eq. 22. Finally, for $T = 50$ K, the number of beads was reduced from 96 (the result of Eq. 22) to 72 because convergence is already expected with this lower number of beads.

Table 3.1: Number of beads (P) used in the PIMD simulations at different temperatures (T).

T (K)	P
50	72
100	48
150	32
200	24
250	20
300	16
350	16
400	16
450	16

Additional post-processing of *ab initio* calculations was done using methods such as natural bond orbital (NBO)[186] to compute second order perturbative energies ($E_{n\rightarrow\pi^*}$) and Wiberg bond indices (WBI) with NBO 7.0[187] coupled with ORCA 4.1.2[188,189] at the CCSD/cc-pVDZ (for aspirin) and CCSD(T)/cc-pVDZ (for toluene) levels of theory. The $E_{n\rightarrow\pi^*}$ energies were taken as the stabilization energies due to $n \rightarrow \pi^*$ interactions. Such interactions, which play an important role in molecular reactivity and conformation (for instance, the Bürgi-Dunitz trajectory[190] preferred during nucleophilic

attacks at a carbonyl carbon), comprise delocalization of lone-pair electrons (n) of an electronegative atom into an empty π^*-antibonding orbital of an aromatic ring or a carbonyl group.[191–193]

3.2 Dynamical Molecular Stabilization Induced by NQE

Here, we analyze the role of NQE for two small molecules that serve as fundamental examples of mechanisms that are present in larger chemical and biological systems. First, $n \rightarrow \pi^*$ interactions in aspirin are analyzed as an example of local electronic orbital effects influenced by nuclear quantum delocalization. Then, the unexpected NQE-induced localization of methyl rotor in toluene is studied as a model for methyl groups in biomolecules. All results discussed in this section were obtained from simulations run at room temperature.

3.2.1 Enhanced $n \rightarrow \pi^*$ Interaction

A particularly important type of interaction often occurring between pairs of neighboring carbonyl groups is the so-called $n \rightarrow \pi^*$ interaction. It arises from the delocalization of lone–pair electrons on electronegative atoms (e.g., oxygen atom) into an antibonding π^* orbital of an aromatic ring or a carbonyl group (see Fig. 3.1A).[194] First discussed in early 1970, the $n \rightarrow \pi^*$ interaction has attracted significant attention in recent years and it is hypothesized to impart substantial structural stability to proteins[195–198] and molecules,[191,193,199,200] as well as define reactivity,[193] regulate isomerisation[194] and energy barriers,[201] and promote charge transfer.[202] Nevertheless, the actual dynamical implications at finite temperature of such interaction have not been explicitly studied.

To elucidate this matter, here the aspirin molecule is studied as a proof of concept. For this molecule the $n \rightarrow \pi^*$ interaction is the main contribution to the relative energy of the global minimum (Fig. 3.2a) and two other local minima (Fig. 3.2b, c), thereby defining their energetic ordering.[193,199] Fig. 3.2 shows the configuration space sampling obtained from classical MD and PIMD simulations at room temperature, where the dynamical implications of the NQE on aspirin's behavior are evident: NQE constrain the dynamics of the molecule to the global minimum in contrast to the results from classical MD. Hence, the NQE must be promoting a particular intramolecular

interaction and, given the evidence provided by Choudhary *et al.*,[193] the $n \rightarrow \pi^*$ interaction between the ester and carboxyl groups is the main candidate.

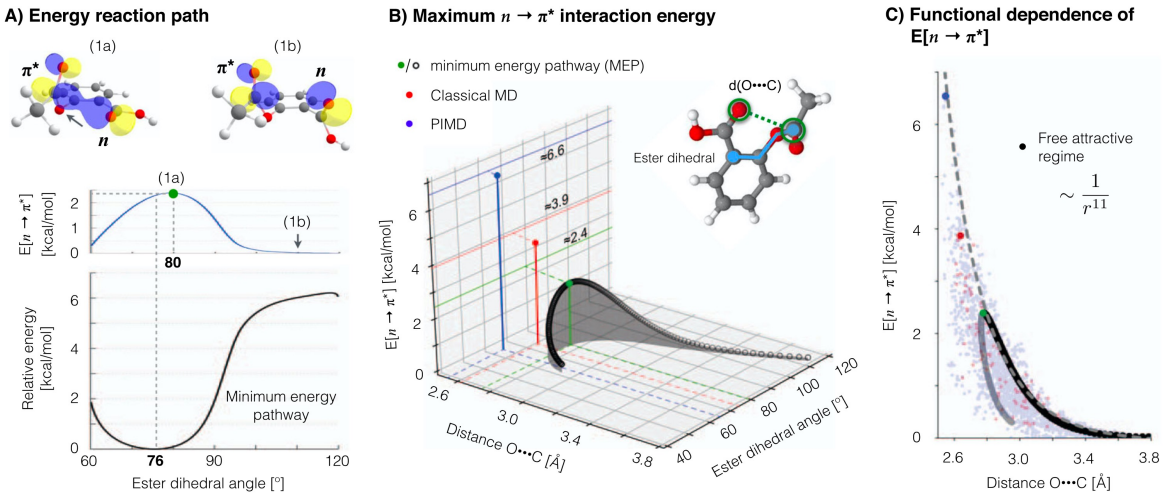

Figure 3.1: Enhancement of the $n \rightarrow \pi^*$ interaction by nuclear delocalization in aspirin. A)-Bottom aspirin's PES and **A)**-Middle $n \rightarrow \pi^*$ interaction energy $E_{n\rightarrow\pi^*}$ along the minimum energy path (MEP) of the ester's dihedral angle. The energy $E_{n\rightarrow\pi^*}$ was computed using NBO method[186,187] at CCSD/cc-pVDZ level of theory, whereby a positive energy value means stabilization and zero energy means absence of overlap between n and π^* orbitals. Hence, a positive value implies a stabilization of the molecule. Both plots are in the same energy scale. **A)**-Top The configuration (1a) (marked by the green circle) defines the maximum interaction (stabilization) energy $E_{n\rightarrow\pi^*}$ along the MEP, while (1b) represents a configuration where overlap between lone-pair electrons $\phi_{(n)}$ and the antibonding $\phi_{(\pi^*)}$ orbitals, and therefore the energy $E_{n\rightarrow\pi^*}$, has gone to zero. **B)** Estimations of the maximum $E_{n\rightarrow\pi^*}$ interaction energy values reached while running PIMD (blue circle, ~6.6 kcal mol^{-1}) and classical MD (red circle, ~3.9 kcal mol^{-1}) at 300 K using the sGDML@CCSD model. As a reference, the energy $E_{n\rightarrow\pi^*}$ curve is plotted along the MEP trajectory as a function of its two main degrees of freedom, the interatomic distance $d_{O\cdots C}$ and the ester's dihedral angle. The maximum energy $E_{n\rightarrow\pi^*}$ value along the MEP is ~2.4 kcal mol^{-1} (green circle). **C)** Approximate functional dependence of the interaction energy $E_{n\rightarrow\pi^*}$ on the oxygen (in hydroxyl) and carbon (in ester) interatomic distance $d_{O\cdots C}$. The $\frac{1}{r^n}$ function was fitted to the free-attractive-regime part of $E_{n\rightarrow\pi^*}$ (black circles) along the MEP starting from 3.8 Å, giving a value of n ~ 11. Reproduced from Ref. 89 (distributed under CC-BY).

To investigate further the contribution of the $n \rightarrow \pi^*$ interaction to the total energy, the $n \rightarrow \pi^*$ interaction energy ($E_{n\rightarrow\pi^*}$) was computed along the ester's minimum energy pathway (MEP) trajectory (Fig. 3.1A) using the NBO analysis.[186,187] In what follows, the NBO definition for the $n \rightarrow \pi^*$ energy is used,[186,187] whereby a positive energy value means stabilization and zero energy means absence of overlap between n and π^* orbitals. Hence, a positive value of the $n \rightarrow \pi^*$ energy in Fig. 3.1 implies stabilization

of the molecule. The visual representation of the n electron delocalization into π^* is presented in Fig. 3.1A-Top. These results show that the $E_{n\rightarrow\pi^*}$ is maximum near the global minimum of aspirin (green circle in Fig. 3.1A-Middle, 1a), such value quickly vanishes as the molecule moves away towards the transition state (at 180º). From these results it is seen that the $n \rightarrow \pi^*$ interaction contributes ~40 % of the global energy minimum relative to the transition state. Additionally, this result also shows that the two main degrees of freedom describing $E_{n\rightarrow\pi^*}$ in aspirin are the $d_{O\cdots C}$ distance, also known as Bürgi-Dunitz parameter,[193–195,197,200] and the ester's dihedral angle, as shown in Fig. 3.1B.

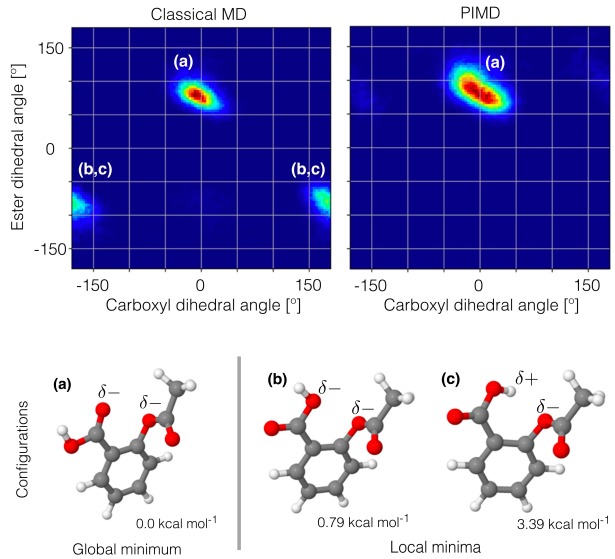

Figure 3.2: Classical MD and PIMD simulations at room temperature of aspirin described by the sGDML@CCSD molecular force field. The plots are projections of the dynamics to the two main degrees of freedom of aspirin: carboxyl and ester dihedral angles. Structures of three relevant molecular configurations are shown: a global minimum and b, c two of the lowest local minima. Reproduced from Ref. 89 (distributed under CC-BY).

Now, in order to understand the results from finite temperature simulations (Fig. 3.2) in the context of the $n \rightarrow \pi^*$ interaction (Fig. 3.1A-Middle), the $E_{n\rightarrow\pi^*}$ energy was computed for a set of configurations sampled from the classical MD and PIMD trajectories on the global minimum (Fig. 3.2a), and then we have plotted their maximum $E_{n\rightarrow\pi^*}$ respective values in Fig. 3.1B. All the computed samples are plotted against the Bürgi-Dunitz parameter $d_{O\cdots C}$ in Fig. 3.1C.

The results displayed in Fig. 3.1B already provide a clear picture of the behavior of the $n \rightarrow \pi^*$ interaction at finite temperature: The maximum $E_{n\rightarrow\pi^*}$ energy along the MEP (i.e., 0 K) is of 2.4 kcal mol^{-1} (green circle in Fig. 3.1), but this value can be enhanced by 160 % due to pure thermal fluctuations (red circle in Fig. 3.1) and up to 270 % by NQE at room temperature (blue circle in Fig. 3.1). This means that the NQE alone

could strengthen the attractive interaction energy between the carbonyl and the ester functional groups by up to ~2.7 kcal mol^{-1} at room temperature. Consequently, given the evidence of such a considerable increment of the $n \rightarrow \pi^*$ interaction energy and the configurational localization resulting from the MD simulations, both originated by the NQE, it has been found that nuclear quantum delocalization can stabilize intramolecular interactions and selected molecular conformations.

It is worth to analyze the underlying dynamics created by NQE that lead to such a prominent increase of the $E_{n\rightarrow\pi^*}$ energy, which could suggest ways to generalize the results found here to other systems. From the MEP trajectory in Fig. 3.1B it can be seen the approximate dependence of the $E_{n\rightarrow\pi^*}$ energy as a function of the $d_{O\cdots C}$ distance and the ester's dihedral angle. Given the nature of the $n \rightarrow \pi^*$ interaction, i.e., its increase with the orbital overlap,[198] variations of the $d_{O\cdots C}$ distance should generate the steepest changes of $E_{n\rightarrow\pi^*}$. This can be seen in Fig. 3.1B, where a small decrease of the $d_{O\cdots C}$ and dihedral values increase the interaction energy. More interestingly, if one only focus on the free-attractive-regime of the interaction energy as displayed in Fig. 3.1C, the fitting of a r^{-n} function to the attractive part of the MEP trajectory suggests that the $n \rightarrow \pi^*$ interaction energy can be approximated by $E_{n\rightarrow\pi^*} \sim r^{-11}$ (dashed line in Fig. 3.1C). This approximation serves as a upper limit envelope to the out-of-equilibrium configurations sampled from classical MD and PIMD simulations (red and blue circles in Fig. 3.1, respectively), and even extrapolates to the more extreme cases such as the maximum energy value reached by the quantum dynamics. Such a steep dependence on the distance between functional groups reveals that even a minor nuclear quantum delocalization leads to a substantial increase in stability.

From these results, and based on the fact that $n \rightarrow \pi^*$ interactions have been consistently reported to occur in different molecular and biological systems,[191,193,202,194–201] one can hypothesize that the strengthening of such interaction by the NQE at finite temperature could prompt similar localization effects in biological systems. Hence, it can be concluded that nuclear quantum fluctuations are not only the source of the enhanced sampling in atomic systems, but also they can promote molecular and intermolecular rigidity in systems with prominent $n \rightarrow \pi^*$ interactions such as polyproline helices in protein fragments which display a double carbonyl–carbonyl interaction.[200]

3.2.2 Methyl Rotor Hindering

The methyl (Me) functional group is a pervasive fragment in chemical and biological systems, playing a fundamental role in, for example, genetics[203] and protein synthesis.[204] The immediate chemical neighborhood of the Me group can drastically modify its energy landscape, going from a free rotor to a localized one with large energetic rotational barriers.

In general, NQE are known to play an important role in lowering energetic barriers when these are of the order of $k_B T$. In the particular case of the Me group, rotational barriers can be much lower than $k_B T$. For this reason the Me group is often considered to be a nearly free rotor at room temperature ($k_B T \approx 0.6$ kcal mol^{-1} at $T = 300$ K). The toluene molecule is one of the simplest representative examples of a molecule with a Me group. The Me rotor in toluene has a six-fold rotational PES whose best experimental estimates for the energetic barrier range from ~0.014 to 0.028 kcal mol^{-1} (~4.9–9.8 cm^{-1}),[205–207] while theoretical results at the CCSD(T) level of theory give ~0.024 kcal mol^{-1} (blue curve in Fig. 3.3B). After performing MD simulations at room temperature (classical MD and PIMD) at the sGDML@CCSD(T) level of theory and analyzing the Me rotor's dynamics, results show contrasting outcomes to what can be trivially assumed given the nature of the system. One would expect that NQE lower the rotational energy barriers even more, but Fig. 3.4C shows that NQE actually hinder the Me group rotations (red) contrary to the free rotation obtained from classical MD (blue). In fact, an incremental inclusion of the NQE via increasing the number of beads in PIMD simulations demonstrates that nuclear delocalization systematically localizes the Me rotor dynamics. Additionally, the PIMD results show that the Me group no longer stays in the plane defined by the benzene ring as in the classical case, instead higher amplitude out-of-plane oscillations are observed due to the NQE (Fig. 3.4D).

To understand the origin of this localization, it is important to first focus on the time evolution of the Me rotor shown in Fig. 3.4B. Here one can see that the classical description of the rotor is indeed a free rotor most of the time, nevertheless an interesting phenomenon emerges: The Me rotor can suddenly stop rotating for up to 4 ps. Still this is not apparent from the cumulative histograms in Figs. 3.4C, D due to the rare nature of this event. Contrasting with the classical model, PIMD results show

a qualitatively different picture. In this case the rotor localization is much more frequent but in general the lifetime of the localized state is shorter. From this, one can hypothesize that nuclear quantum delocalization promotes the localization of the Me rotor, but at the same time the NQE tunnel the system out of the localized state. In contrast, Me rotor localization is a rare event in classical dynamics, but when it occurs, it can take a longer time for purely thermal fluctuations to bring the system out of such state.

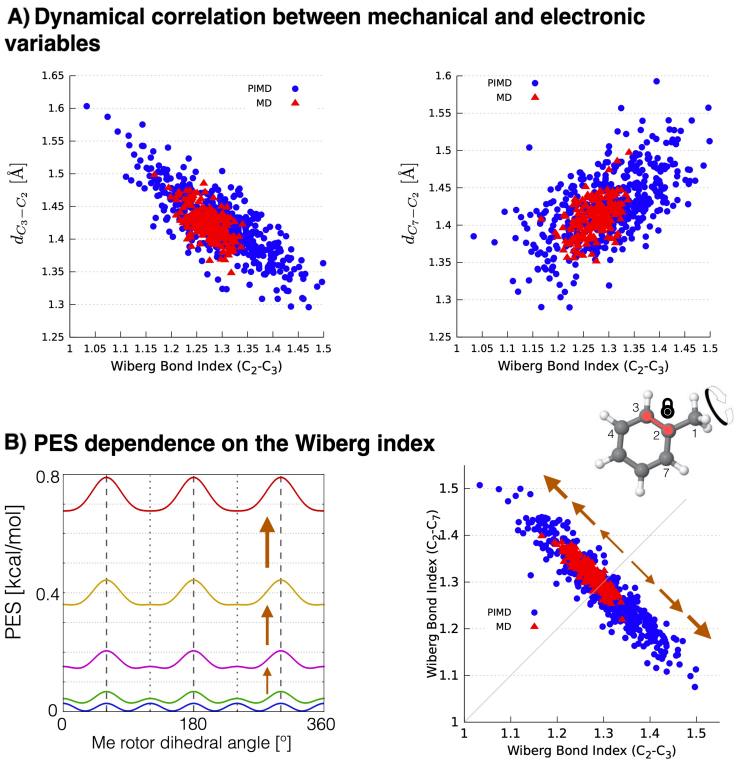

Figure 3.3: Relations between WBI, bond lengths, and the PES of toluene. A) Dynamical correlation between the WBI for the bond C_3−C_2 and bond length $d_{C_3-C_2}$ (left) and bond length $d_{C_7-C_2}$ (right). The classical MD (red) is contrasted with the PIMD (blue) simulations. **B)** Methyl rotor's PES for different fixed values of the C_2−C_3 distance (left) and its qualitative relationship with the WBI (right). Reproduced in altered form from Ref. 89 (distributed under CC-BY).

The hindering of the rotations of the Me group has a dynamical origin, and it is determined by the delocalization of the benzene carbon-carbon bonds generated by the NQE. Bond length delocalization is a well known implication of NQE, which, in this particular case, transforms the Me rotor's PES from a six-fold energy surface to a three-fold energy surface as shown in Fig. 3.3B. Furthermore, the magnitude of the transition state energy is determined by the two benzene ring bonds C_2−C_3 and C_2−C_7 near to the Me group (see Figs. 3.4A, 3.3B). To corroborate this and to assess its behaviour at finite temperatures, the WBI of these bonds were monitored during a Me localized state for classical MD and PIMD. The correlation $WBI_{C_2-C_3} \sim -WBI_{C_2-C_7}$ in the plot shown in Fig. 3.3B agrees with the linear correlations between $WBI_{C_2-C_3}$ and

the corresponding bond lengths ($d_{C_2\text{-}C_3}$ and $d_{C_2\text{-}C_7}$), giving the direct relationship $WBI_{C_2\text{-}C_3} \sim -d_{C_2\text{-}C_3} \sim +d_{C_2\text{-}C_7}$ (Fig. 3.3A). In this context, the electron gain/depletion (bond length decrease/increase) drastically changes the energy landscape. Computing the $d_{C_2\text{-}C_3}$-restricted Me rotor PES (Fig. 3.3B) gives a lower bound estimate for the increase of the Me rotor energy barrier, which can be up to ~400 % larger than the reference (~0.028 → 0.120 kcal mol^{-1}).

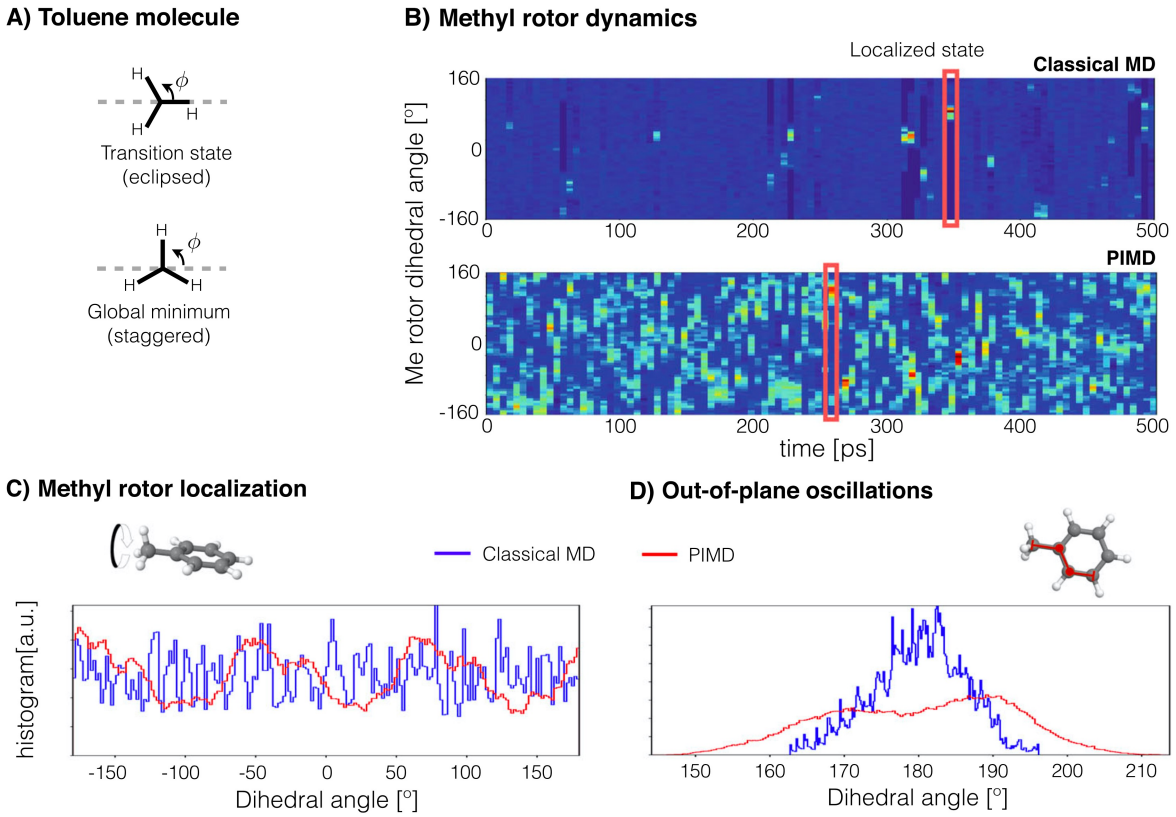

Figure 3.4: Hindering of methyl rotor dynamics by nuclear delocalization in toluene molecule. A) Global minima and transition state of Me rotor. **B)** Time evolution of the Me rotor for classical MD (top) and PIMD (bottom) simulations. The red rectangles indicate some of the localized states in the dynamics. The size of the bins are 5 ps x 6º. Histogram of the Me rotor's dihedral angle (**C**) and out-of-plane dihedral angle (**D**) computed from classical MD (blue) and PIMD (red). Reproduced in altered form from Ref. 89 (distributed under CC-BY).

Consequently, PIMD will tend to generate much higher rotation energy barriers given the extra dilation of such bond lengths induced by the NQE beyond the thermal dilation generated by classical MD. According to the results, the rotor described by PIMD experiences energetic barriers of up to 0.55 kcal mol^{-1}, energy comparable to k_BT, therefore hindering Me rotations. From here, one can conclude that the intricate

quantum dynamics exhibited by the Me rotor in toluene is due to two competing NQE: On one side the nuclear delocalization of the carbon atoms promotes higher rotational energetic barriers hindering the rotor, but the quantum fluctuation of the hydrogen atoms in the Me takes the rotor out of the localized state.

Even though the results shown here are for toluene, the electronic origin of the rotational energetic barrier of Me rotor is very similar in different molecular systems.[207–210] Hence, similar dynamical effects are to be expected in large biological systems given the ubiquity of Me groups in macro-molecules and protein fragments.

3.3 Prevalence of NQE over Thermal Contributions in Molecular Interactions at Temperatures from 50 to 450 K

Sec. 3.2 was dedicated to discussing the strengthening of molecular interactions due to NQE. Now, the present section focuses on the aspirin molecule for analyzing the $n \rightarrow \pi^*$ interaction and NQE as a function of the temperature. Simulations in this section were obtained at different temperatures ranging from 50 to 450 K, and only the regions of the PES corresponding to the global minimum of aspirin (see Fig. 3.2) are considered for the discussion because in these regions is where the $n \rightarrow \pi^*$ interaction is present.

To perform a reliable analysis of the interaction at different temperatures, one requires to sample the $E_{n\rightarrow\pi^*}$ for all relevant configurations (i.e., configurations in regions close to the global minimum) in each simulation. Although for computing this energy one only needs single-point calculations, at high levels of theory (like CCSD/cc-pVDZ) doing such calculations hundreds of thousand times becomes computationally prohibitive. To overcome this problem, an ML model to predict the $E_{n\rightarrow\pi^*}$ was constructed. First, 100 steps and 16 beads for each PIMD simulation were randomly selected (for those simulations where the number of beads was equal to 16, all beads were selected). Then, the $E_{n\rightarrow\pi^*}$ for all these configurations (1600 for each PIMD simulation) were computed using the NBO analysis.[186,187] In total, 14400 samples of $E_{n\rightarrow\pi^*}$ were obtained, from which 12000 were used to train a KRR model with an RBF kernel and all inverse pairwise distances as descriptor (RMSE of 0.2 kcal mol^{-1}). This model was used to predict the $E_{n\rightarrow\pi^*}$ of all configurations in all MD and PIMD simulations.

As it was mentioned in Sec. 3.2, the Bürgi-Dunitz parameter (the $d_{O\cdots C}$ distance) can be used to accurately describe the $E_{n\to\pi^*}$. Fig. 3.5 shows the functional behavior of the $E_{n\to\pi^*}$ with respect to the $d_{O\cdots C}$ distance for the PIMD simulations at all temperatures considered in this work. The curves were obtained by fitting the $E_{n\to\pi^*}$ values of the 100 configurations with the highest $E_{n\to\pi^*}$ at each interval of 0.05 Å of the $d_{O\cdots C}$ distance from 2.2 to 3.3 Å. Overall, the dependence of the $E_{n\to\pi^*}$ on the $d_{O\cdots C}$ distance is practically the same for all temperatures, meaning that $n \to \pi^*$ interaction does not explicitly depends on the temperature. This is especially true for configurations with a $d_{O\cdots C}$ distance between 2.5 and 2.9 Å, which present a small dispersion in their $E_{n\to\pi^*}$ at different temperatures (see top of Fig. 3.5). Such configurations are the ones closest to the global minimum ($d_{O\cdots C}$ distance of ~2.8 Å). The remaining configurations, featuring extreme $d_{O\cdots C}$ distances (closer to 2.2 and 3.3 Å) further away from the equilibrium value, present a larger dispersion of $E_{n\to\pi^*}$ values with respect the temperature. This is the consequence of the limitations of projecting the $E_{n\to\pi^*}$ on a single degree of freedom. This projection hides the fluctuations of other degrees of freedom (e.g., the ester and carboxylic dihedral angles in Fig. 3.2), which are more pronounced at higher temperatures. Thus, at high temperatures some configurations (with diverse $E_{n\to\pi^*}$ values) that are not accessible at lower temperatures are sampled.

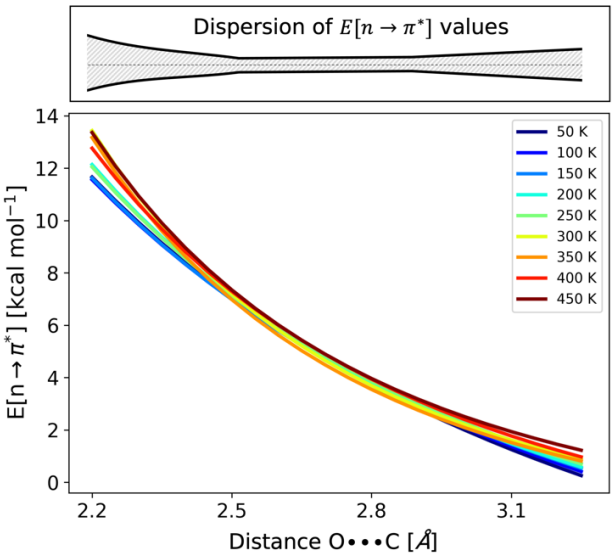

Figure 3.5: Functional behavior of $E_{n\to\pi^*}$ with respect $d_{O\cdots C}$ distances at different temperatures. Curves were obtained by fitting the $E_{n\to\pi^*}$ values of the 100 configurations with the highest $E_{n\to\pi^*}$ at each interval of 0.05 Å of the $d_{O\cdots C}$ distance from 2.2 to 3.3 Å. At the top, a schematic graph of the dispersion of $E_{n\to\pi^*}$ values at different temperatures is shown. Only PIMD simulations were considered for this plot.

The relevance of NQE for the $n \to \pi^*$ interaction can be further observed when comparing the evolution of maximum values of $E_{n\to\pi^*}$ with increasing temperature between classical MD and PIMD simulations. For doing this comparison, Fig. 3.6

presents the mean of the 100 highest values of $E_{n\to\pi^*}$ at each temperature. The $E_{n\to\pi^*}$ in classical MD simulations shows a steep linear increase with the temperature from 50 to 300 K and PIMD simulations have an almost constant value of $E_{n\to\pi^*}$ from 50 to 150 K. In classical MD simulations, only the thermal energy contributes to the $E_{n\to\pi^*}$, while in PIMD simulations the quantum fluctuations govern the dynamics of the molecule. Indeed, the constant $E_{n\to\pi^*}$ at low temperatures can be explained by the ZPE. The major contributions to the energy when temperatures are low are given by the vibrational energy, which remains fairly constant at these temperatures. If the temperature increases much, however, the contributions of the thermal energy become comparable to those of the ZPE. This leads to the almost equal behavior of $E_{n\to\pi^*}$ from 150 to 300 K in PIMD simulations and from 50 to 300 K in classical MD simulations. Finally, at high temperatures (greater than 350 K) the thermal energy contributions start to saturate in both classical MD and PIMD simulations, and the excess of energy starts to be distributed to other degrees of freedom in the molecule, thus promoting other processes, such as conformational changes.

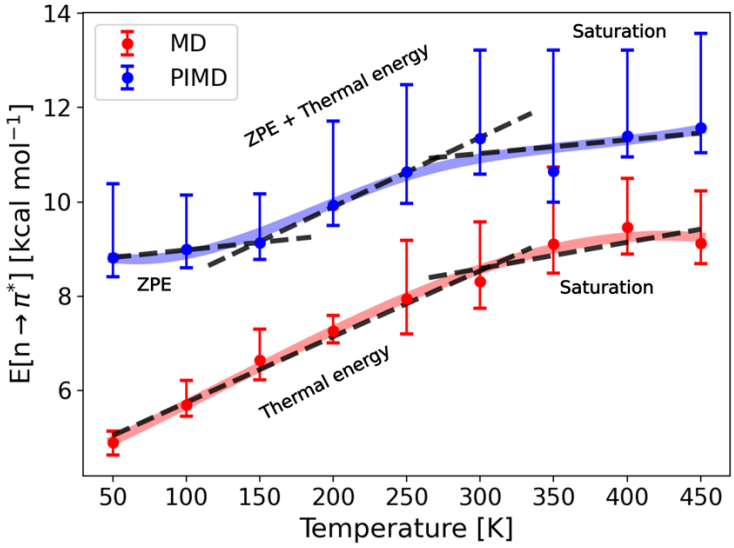

Figure 3.6: Change of maximum values of $E_{n\to\pi^*}$ as a function of the temperature. Circles mark the mean of the 100 highest $E_{n\to\pi^*}$ values at a given temperature. Data intervals on each circle indicate the maximum and minimum value of $E_{n\to\pi^*}$. Different stages for understanding the behavior of $E_{n\to\pi^*}$ are depicted with dashed lines.

Although thermal energy contributions do increase the maximum values of $E_{n\to\pi^*}$, NQE clearly dominate the interaction. For instance, in PIMD simulations the minimum value of $E_{n\to\pi^*}$ among the 100 highest at 450 K is less than 1.0 kcal mol^{-1} greater than the highest value of $E_{n\to\pi^*}$ at 50 K, whereas in classical MD simulations this difference exceeds 3.0 kcal mol^{-1} (Fig. 3.6). In addition, $E_{n\to\pi^*}$ values in classical MD simulations at 450 K are, in general, lower than those in PIMD simulation at 50 K. This prevalence

of quantum fluctuations over thermal contributions also explains the results observed in Fig. 3.5 and the “global” localization of aspirin in its global minimum due to local nuclear delocalization discussed in Sec. 3.2.

3.4 Conclusions

In this section, MLFFs were used to run long (PI)MD simulations of aspirin and toluene at CC level of theory. Such long simulations allowed the finding of counterintuitive consequences of NQE on the dynamics of the studied molecules. Namely, there is a strengthening of interactions that induces a localization of the aspirin molecule in its global minimum and the hindering of the methyl rotor in toluene. Furthermore, additional simulations at different temperatures for aspirin granted a deeper understanding on how NQE enhance the $n \rightarrow \pi^*$ interaction and the stability of the global minimum. Specifically, the predominance of quantum effects (e.g., ZPE) over thermal contributions determines the relative stability of different aspirin configurations even at high temperatures.

This section showcased the power of MLFFs to provide insights. The accuracy and efficiency of MLFFs allow obtaining of results that would be computationally prohibitive otherwise. Such accurate and insightful results are also desirable for flexible molecules. Nevertheless, there are still open challenges to solve before making MLFFs applicable to the complex PESs of such molecules. Because of this, the focus of the next sections will be the discussion of some of those challenges, as well as the possible solutions to further advance the applicability of MLFFs.

CHAPTER 4

Challenges for Machine Learning Force Fields in Reproducing Potential Energy Surfaces of Flexible Molecules

---

Parts of this chapter have been published in this or similar form in
V. Vassilev-Galindo, G. Fonseca, I. Poltavsky & A. Tkatchenko, *J. Chem. Phys.* **154**, 094119 (2021)[90]
and they are reprinted with the permission of AIP Publishing.

One of the main components of any PES are the transition paths that connect minima because they are related to chemical reactions and conformational changes. The state-of-the-art methods for finding transition pathways range from the optimization of a single direction on the PES[211,212] or a chain of states connecting both minima, e.g., the string[213] and nudged elastic band (NEB)[214,215] methods, to the more sophisticated transition path sampling techniques.[216,217] Most of them often provide only a single "optimal" transition path. The rate of success in finding the path highly depends on the dimensionality and complexity of the PES: flexible molecules containing a few tens of atoms, such as organic photoswitches and peptides, are already challenging to deal with. Moreover, due to the non-trivial interplay between covalent and non-covalent interactions, the transitions in such molecules may happen following several different pathways. In this case, one needs to consider the contribution of every path to the transition process, and the knowledge about just one optimal pathway is insufficient. Practical studies of such transitions require reliable FFs able to accurately reconstruct broad regions of the PES, including multiple local minima and all the relevant pathways connecting them.

MLFFs appear as a natural solution to solve the problem of accurately modelling transition processes. Indeed, some efforts have already been directed toward improving ML-aided search and sampling of transition states and pathways.[88,218,219] For instance, Noé *et al.*[88] showed a promising method to sample rare events between equilibrium states using Boltzmann generators. The method is by many orders of magnitude more efficient than "brute force" MD simulations. Other approaches[218,219] are built on the state-of-the-art methods for calculating transition states enhanced with ML techniques. ML-enhanced transition state search methods are more efficient than their precursors but present the same limitations. ML methods are often data demanding, making their application infeasible when computationally expensive *ab initio* methods are required. Hence, constructing robust ML models for flexible molecules is the necessary next step for practical applications of ML potentials in chemistry and biology.

There are two main challenges in building accurate ML models for complex PESs with many minima and transition paths: First, generating enough data around the transition regions of the PES. Second, building a highly accurate and data-efficient ML model that faithfully describes all relevant parts of the resulting complex PES. In this chapter, both of these challenges are addressed in an azobenzene ($C_{12}H_{10}N_2$) molecule, as example. While being small in size, azobenzene is flexible enough to feature a *cis* to *trans* thermal relaxation following at least three possible channels: a rotation, an inversion, and a rotation assisted by inversion mechanisms.[220–224] The chapter starts with a description of the computational details. Then, the problem of building reliable reference datasets for these transitions is discussed. Finally, the performance of the state-of-the-art ML methods is assessed in the prediction of forces and energies along the obtained transition paths. The methods include NNs, such as BPNNs[60,62] and SchNet[66,225,226] architectures, and kernel-based methods, such as sGDML[45–48] and GAPs[56,57] using the SOAP representation.[58] To highlight how the complexity in learning the PES increases with the flexibility of a molecule, the results of azobenzene are compared with those of a simpler glycine molecule. The training datasets were limited to 1000 geometries. The ML models unable to predict the PES of the considered small molecules correctly within this limit would face considerable problems for large flexible molecules where the cost of reference calculations increases very steeply.

4.1 Computational Details

4.1.1 Reference Energies and Forces

All reference data was obtained using FHI-aims software[227] with the DFT functional Perdew-Burke-Ernzerhof (PBE)[228] and tight settings. The Tkatchenko-Scheffler (TS)[229] scheme was employed to account for vdW interactions.

4.1.2 Transition Paths

The string[213] and NEB[214,215] methods were used coupled with a climbing image approach[230] as implemented in FHI-aims to find transition paths connecting the minima of glycine. The first two algorithms search for the MEP going from one metastable state to the other by building a chain of states. Convergence is achieved when the residual force of the curve connecting the states is optimized under a given threshold. Then, the climbing method looks for a saddle point by moving the points with highest energies toward a higher energy location along the path. Force thresholds were set at 0.2 eV $Å^{-1}$ for the MEP and at 0.02 eV $Å^{-1}$ for the saddle point.

4.1.3 Reference Datasets

MD was used to construct reference datasets for training ML models. All calculations were carried out with i-PI[185] wrapped with FHI-aims code to compute forces and energies. In all simulations a global Langevin thermostat was employed. Below it is described the set-up of the MD simulations performed for each dataset, separately.

Glycine

For building the dataset of glycine isomerization, two MD simulations of 5000 steps were run starting from each of the minima used in this work. Velocities were initialized at 500 K and the simulation temperature was also set to 500 K. A timestep of 1 fs and a friction coefficient of 2 fs were used.

Azobenzene

The datasets for the rotation and inversion mechanism of azobenzene were constructed combining the results of different MD runs:

1. Two long MD simulations (~120k steps) at the PBE+TS/light level of theory at 300 K starting from both the *cis* and *trans* isomers. The velocities were initialized also at 300 K. The timestep was set to 1 fs and a friction coefficient of 2 fs was used. Then, from these simulations, a configuration was selected every 25 steps (this interval was chosen to have a trade-off between enough decorrelated configurations and size of the datasets) and carried out single-point calculations with the PBE+TS/tight method for each of these geometries.
2. Short MD simulations of 300 steps at 100 K (initial temperature was set to 300 K) starting from each of the intermediate configurations on the rotation and inversion paths (see Sec. 4.2). A timestep of 0.5 fs and a friction coefficient equal to 1 fs were used. A low temperature and a strong friction coefficient were selected to allow the system to go from a configuration close to the transition state to an equilibrium state, thus avoiding high-energy states. A small timestep was chosen to collect enough points from the transition paths. All the low-energy *trans*- and *cis*-like configurations sampled in the simulations were removed because they do not give additional information.
3. Four MD simulations (of 2500 steps) with velocities initialized at 3000 K and a simulation temperature of 750 K, two starting close to the *cis* isomer and the other two close to the *trans* isomer. The timestep was set to 1 fs and, for keeping local geometry fluctuations under control, a friction coefficient of 0.2 fs was used.
4. MD simulations with velocities initialized at 300 K and a simulation temperature of 50 K with a timestep of 0.025 fs and a friction coefficient of 1 fs in order to have a progressive movement of the desired degree of freedom. The initial geometries were the steps 7 and 9 of the inversion path, and the steps 8 and 10 of the “optimized” rotation path (see Sec. 4.2, and Tables A.3 and A.5 in the Appendices for details). Each simulation is comprised of 2500 steps.

4.1.4 ML Models

The ML models were built with sGDML,[45–48] SchNet,[66,225,226] BPNN,[60,62] and GAPs[56,57] with the SOAP representation.[58] All models were trained on training sets containing from 100 to 1000 configurations in steps of 100.

sGDML models were validated using 1000 configurations and values for the hyperparameter σ from 2 to 200 with a step of 10 were explored.

SchNet models were trained using a cutoff radius of 5 Å, 128 features and 6 interaction layers.

BPNN models were trained using the wACSFs[162] descriptor. To match the number of features to those selected for SchNet (128 features) the amount of radial functions was set to 84 and the number of angular functions to 22. A cutoff radius of 5 Å was considered. The calculations were carried out with the implementation of BPNN in SchNetPack.[226]

GAPs were trained with the SOAP descriptor. 12 radial and 6 angular functions were used for the descriptor. The cutoff radius was set to 5 Å, although additional tests were done with larger cutoffs (up to 10 Å). δ was set to 0.25, the atom σ was set to 0.3, and the default σs for energy and forces were set to 0.001 and 0.2, respectively. All calculations were performed with the QUIP program package.[231]

4.1.5 Metadynamics Simulation

Metadynamics simulations for the azobenzene molecule were run using the dihedral ϕ and the angle θ as collective variables (see Fig. 4.1). The calculation was carried out with i-PI wrapped with FHI-aims code to compute forces and energies and PLUMED[232–234] to add the bias potential. The simulation temperature was set to 500 K, with a timestep of 1.5 fs to enhance the sampling, and a friction coefficient of 0.2 fs to keep local geometry fluctuations under control. The gaussian height was set to 2.5 kJ mol^{-1} and the gaussian widths were set to 0.5 rad for ϕ and 0.4 rad for θ. The bias was added every 50 steps.

4.1.6 Subset Selection Procedure

Before applying cross-validation for building the training and test sets, the training set selection process of sGDML was used to construct subsets of the datasets with a size equal to five times the number of training points (e.g., for 1000 training points a subset of 5000 configurations was constructed). The training set selection procedure of sGDML draws a sample from the dataset that preserves its original energy distribution. The distribution is estimated from a histogram were the bin size is determined using the Freedman-Diaconis rule.[235] This rule is designed to minimize the difference between the area under the empirical probability distribution and the area under the theoretical probability distribution. A reduced histogram is then constructed by sampling uniformly in each bin. It is intended to populate all bins with at least one sample in the reduced histogram, even for small training sizes.

4.1.7 MD Simulations with ML Potentials

The best GAP/SOAP, SchNet and sGDML models trained on 1000 training points were used to run with each method four MD simulations of 50k steps using a Langevin thermostat at 300 K with a timestep of 0.5 fs, and a friction coefficient of 2 fs. The simulations started from steps 7 and 9 of the inversion path, and steps 8 and 10 of the rotation path (see Sec. 4.2, and Tables A.3 and A.5 in the Appendices for details). All simulations were done with the atomic simulation environment (ASE).[236]

4.2 Constructing Reference Datasets for Isomerization

The starting point for building any MLFF is collecting reference data covering the relevant parts of the PES of interest. When modeling transition pathways, the reference data can be split into two parts: (i) data covering the vicinity of the equilibrium states between which the transition process happens and (ii) data of “far-from-equilibrium” parts of the PES defining the transition path(s). While equilibrium states are normally readily available, configurations describing transition paths connecting them are, in most cases, not trivial to find. Moreover, the complexity of this task rapidly grows with the increase in flexibility and size of the molecule. In view of this, we had to employ two different strategies for generating the datasets for glycine

and azobenzene isomerization. Below, we discuss in detail the process followed for each molecule separately.

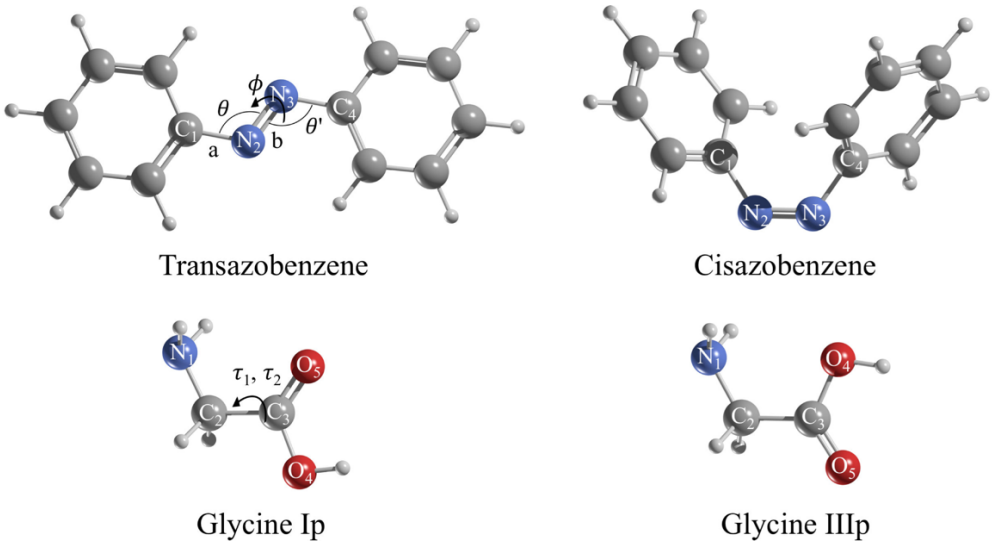

Figure 4.1: Optimized configurations of the minima considered in this study and labels of the main degrees of freedom: the bonds *a* (C_1−N_2 and N_3−C_4 in azobenzene) and *b* (N_2=N_3 in azobenzene), the bending angles θ and θ' (C_1−N_2=N_3 and N_2=N_3−C_4 in azobenzene), and the torsional angles ϕ (C_1−N_2=N_3−C_4 in azobenzene) and τ_1 and τ_2 (N_1−C_2−C_3−O_4 and N_1−C_2−C_3=O_5 in glycine).

4.2.1 Glycine

Glycine, being a rather small molecule, possesses numerous planar and non-planar conformers in the gas phase whose relative energies have been extensively studied.[237,238] Here, only the isomerization from the global minimum geometry, called Ip, to the IIIp conformer is considered because it is the closest "directly connected" minimum. Transitions to any other metastable state from Ip go through this conformer. The Ip–IIIp transformation can be characterized by a change in torsional angles τ_1 and τ_2 (see Fig. 4.1), both around the C bond. They go from 180.0º and 0.0º in isomer Ip to 0.0º and 180.0º in isomer IIIp, respectively.

Transition Path

To construct the transition path between the equilibrium states of glycine, the string[213] and NEB[214,215] methods were used. Both methods converge without any issues, providing similar pathways. The transition state obtained by employing the PBE+TS method lies only 2.4 kcal mol^{-1} above the Ip isomer (see Fig. 4.2). The mechanism is

defined by almost equal rotations of both τ_1 and τ_2 torsional angles (see Table A.1 in the Appendices for important geometric details).

Dataset

Since the relative energy between the Ip isomer and the highest-energy structure found on the MEP is less than 3.0 kcal mol^{-1}, the Ip–IIIp transformation is accessible via standard constant-temperature MD simulations. Therefore, two dynamics starting from both equilibrium geometries were ran to construct the dataset for glycine isomerization. A total of 5000 configurations at 500 K with a time step of 1 fs were obtained from each simulation. A transition was observed in the simulation starting from the IIIp isomer.

4.2.2 Azobenzene

Azobenzene is a photochemical compound; however, it also exhibits a *cis* to *trans* thermal relaxation, on which this work is focused. Hence, issues with electronic multireference states in azobenzene are avoided and one can use the generalized gradient approximation DFT-PBE functional and include vdW interactions with the TS-vdW method for generating the reference data. The *cis* and *trans* configurations of azobenzene (Fig. 4.1) differ mostly by a change in the torsional angle ϕ around the N=N double bond from close to 10.0º to 180.0º during the isomerization. Although the existence of the two forms has been known since the works of Hartley in the 1930s,[239,240] there is still an open debate regarding whether azobenzene primarily follows a rotation (changes around the dihedral angle ϕ), an inversion (changes in the angles θ and θ'), or a rotation assisted by an inversion (changes in both ϕ and θ and θ') mechanism. DFT, multi-reference methods, and ML approaches have been used in an attempt to unveil the actual mechanism of isomerization,[220–224,241–243] but conclusive evidence favoring a particular mechanism is yet to be found.

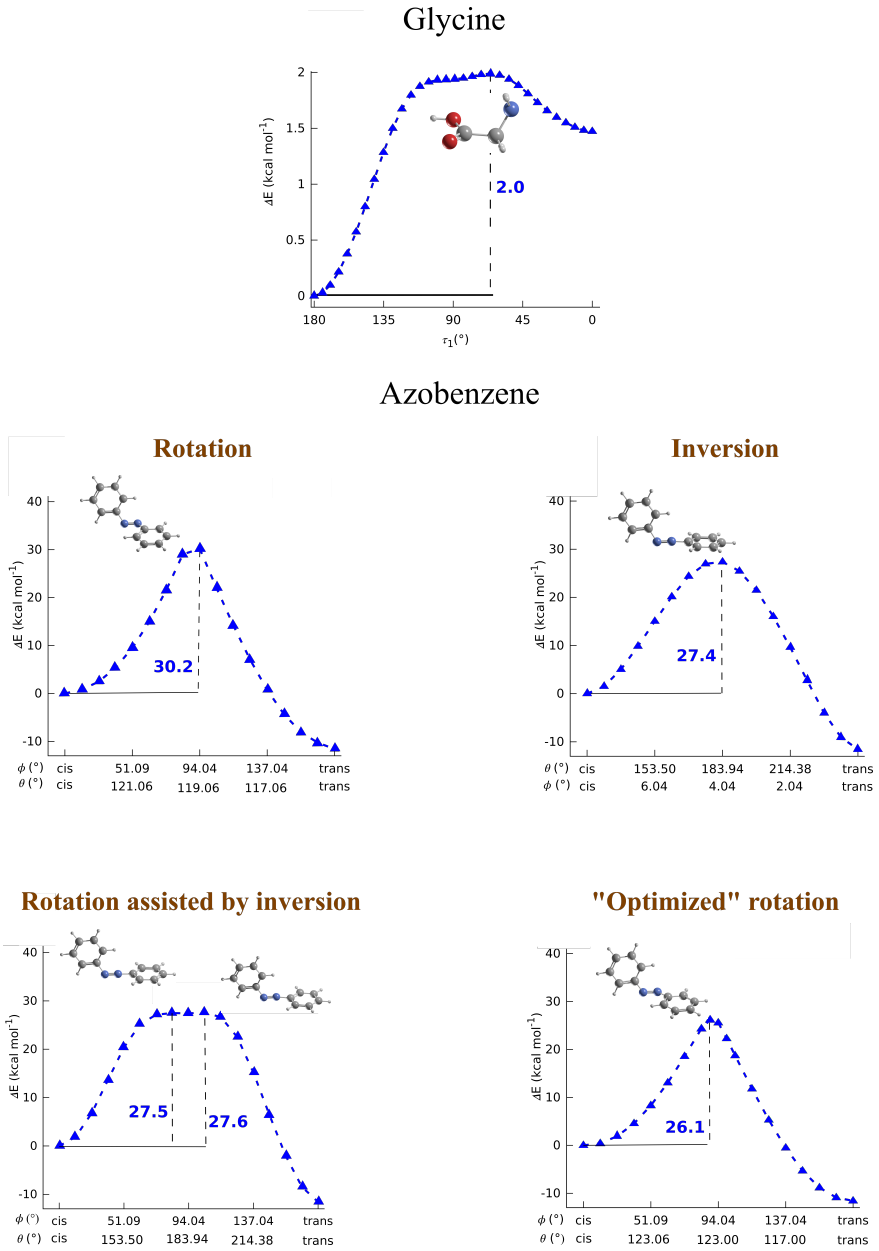

Figure 4.2: Energy profiles (in kcal mol^{-1}) for the paths constructed in this work: the glycine isomerization path, and rotation, inversion, rotation assisted by inversion and "optimized" rotation paths of azobenzene *cis* to *trans* thermal relaxation. The relative energy between the highest energy configuration on the path and the *cis* isomer is given.

Transition Paths

Although azobenzene is not much larger than glycine, all the transition pathways that can be found in the literature for this molecule are constructed manually. One can easily check that neither the string nor NEB methods converge to a reasonable path for *cis* to *trans* transition. Following the previous works,[220–224] we also constructed the transition pathways manually as follows:

- The rotation path, which is defined by a change in the torsional angle ϕ around the central double bond (see Fig. 4.1).
- The inversion path whose main feature is the bending of either θ or θ' (see Fig. 4.1).
- The rotation assisted by an inversion path, which is the combination of the first two.

Each path is comprised of 15 intermediate geometries linking the minima. In all cases, the molecule was forced to follow the desired mechanism by linearly interpolating the main degree(s) of freedom between both minima (Tables A.2-A.4 in the Appendices show important geometric data, and Fig. 4.2 shows the energy profiles). The obtained highest-energy geometries are in good agreement with those found elsewhere.[220–222]

Table 4.1 shows the relative energies (with respect to cisazobenzene) of the highest-energy structures found for each transition path. The rotation mechanism is the most favorable path at the initial and final steps of the isomerization, but it has the highest-energy barrier among the three transition paths considered here. The inversion mechanism is the one with the lowest-lying highest-energy structure within PBE+TS calculations. The rotation assisted by an inversion path is the least favored at the zones close to the minima and presents a plateau region at the top of the curve with two "peaks" with relative energies close to that of the highest-energy configuration in the inversion mechanism.

The pathways introduced in the previous paragraphs are just linear interpolations between the *cis* and *trans* geometries. This introduces constraints on how the different degrees of freedom can evolve through the transition. To obtain a path affected by the contributions of all the important degrees of freedom, the rotation path was "optimized" by choosing the values of θ, θ', and *a* and *b* (Fig. 4.1) that minimize the energy at each step. The optimized rotation path is the most favorable with the PBE+TS method with an energy barrier of 26.1 kcal mol^{-1}. The geometric details of this path can be found in Table A.5 in the Appendices.

Table 4.1: Relative energies (ΔE in kcal mol^{-1}) of the highest-energy structures on each mechanism computed with the PBE+TS method.

Mechanism	Rotation	Inversion	Rot+Inv
ΔE	30.2	27.4	27.5, 27.6

All the paths described here can be considered as good insights into the real isomerization process. It has been found that the activation barrier of the *cis* to *trans* thermal relaxation in *n*-heptane solution is between 22.7 kcal mol^{-1} and 25.1 kcal mol^{-1}.[242] Hence, in the gas phase, one would expect greater values like those presented here. In what follows, this study will focus on the optimized rotation (named simply rotation from now on) and the inversion mechanisms.

Datasets

Constructing a dataset for a molecule such as azobenzene requires a more elaborate procedure compared to the simpler glycine molecule. First of all, the transition process is a rare event at ambient conditions and cannot be easily accessed; second, there are more than one possible transition pathways.

Here, separate datasets are constructed for the rotation and inversion mechanisms. First, two types of MD simulations are combined: (a) long constant-temperature MD runs with a time step of 1 fs at 300 K at the PBE+TS/light level of theory starting from the equilibrium geometries, from which a configuration was selected every 25 steps and single-point calculations were carried out with the PBE+TS/tight method (around 3500 configurations were collected for each minimum); (b) constant-temperature MD runs of 300 steps with a 0.5 fs time step at 100 K starting from each of the intermediate steps of the rotation and inversion paths. From the energy distributions shown in Fig. 4.3a, one can conclude that the configurations visited during the MD simulations are bounded by the temperature to certain energy ranges, as indicated by the well-defined peaks representing *trans*- and *cis*-like configurations. The addition of configurations obtained from the rotation and inversion mechanisms had little impact on the energy distribution, which leads to the conclusion that the transition parts are still poorly sampled. Thus, additional configurations were generated by performing (a)

four constant-temperature MD simulations (of 2500 steps each) at 750 K with a time step of 1 fs starting from structures close to equilibrium and (b) constant-temperature MD simulations of 2500 steps at 50 K with a very small time step (0.025 fs) starting from steps 7 and 9 of the inversion path and the steps 8 and 10 of the rotation path (see Tables A.3 and A.5 for details). The former provide the data required to model the cooling down process from transition states to minima, which involves high kinetic energies. The latter allow us to include slow changes in the degrees of freedom during the transition process. Fig. 4.3b shows that the new reference geometries sample different energy distributions for close-to-equilibrium, rotation, and inversion datasets. The final datasets combine the results of all four types of simulations containing 26455 data points for the rotation and 25528 data points for the inversion mechanisms.

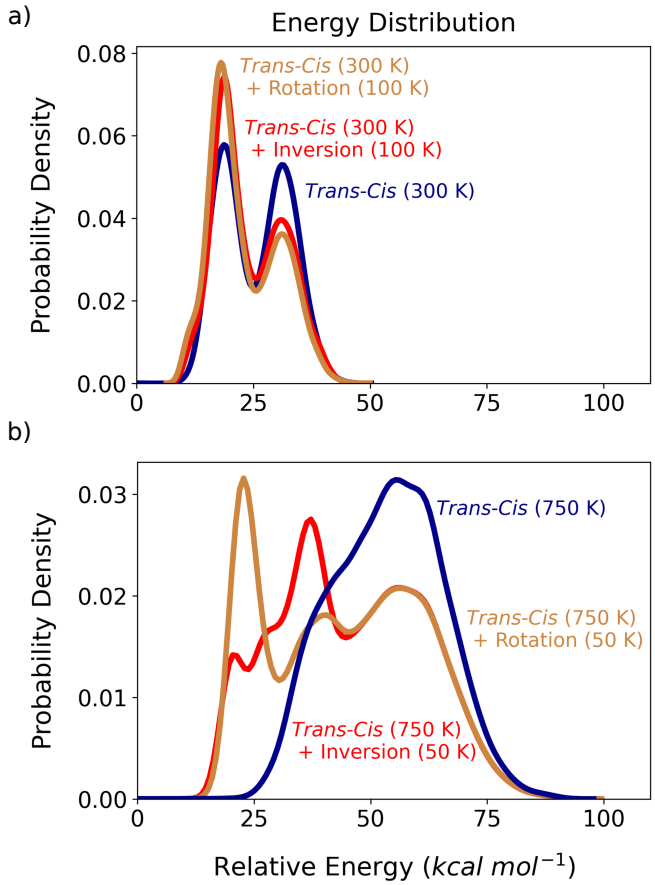

Figure 4.3: Distribution of relative energies (in kcal mol^{-1}, with respect to the optimized configuration of transazobenzene) of different sets of MD simulations: a) "close-to-equilibrium" configurations at 300 K and geometries close to the transition paths at 100 K and b) high-energy "close-to-equilibrium" configurations and properly sampled geometries close to the transition paths.

An alternative to the above-described procedure would be to use an enhanced sampling technique (e.g., metadynamics or umbrella sampling). The key component in such methods is the introduction of a bias potential that penalizes already visited system configurations. To do so, one defines a low dimensional projection of the

configurational space of the system, the so-called collective variables (CVs), where energy minima and transition paths are well distinguished. During the simulation, a repulsive (bias) potential is added every *n* steps in the CV space to prevent revisiting already sampled parts of the configuration space. As a result of this procedure, all configurations have an equal probability of being sampled, regardless of their potential energy. Thereby, broad regions of the PES containing multiple minima and transition pathways can be sampled. In this work, concern is focused in specific parts of the PES: the regions surrounding the *cis* and *trans* isomers of the azobenzene molecule and the transition pathways between them. While the two angles of azobenzene could be used as CVs, it is not clear that these two variables are sufficient to properly resolve different transition mechanisms. In addition, the transition state energies are high compared to the thermal energy of the molecule at ambient conditions. As a result, choosing the parameters of the bias potential for an efficient metadynamics simulation yields a dataset where the transition mechanisms are undersampled. For a more detailed analysis, see Sec. 4.5. On the contrary, setting the width and the amplitude of the bias potential that allows for a good resolution in the transition region makes the metadynamics simulation inefficient. Hence, the construction of datasets combining manual and MD-generated configurations is more efficient for sampling high-energy transitions, giving the opportunity to assess the performance of ML models on the transition paths of flexible molecules.

4.3 Advantages and Limitations of Different ML Methods

In Sec. 2.2 it was discussed that before applying any ML method, the molecular configurations must be encoded into a rotationally, translationally, and permutationally invariant representation or descriptor.[52–57,131,134,139–142,144–148] Now, it is important to remark that a descriptor must balance efficiency with accuracy; hence, different descriptors are applicable to different scenarios. One can divide them into local and global descriptors. For the former, NNs (e.g., SchNet[66,225,226]) and kernel-based potentials (e.g., when using descriptors such as FCHL[52,134] or SOAP[58]) assume locality through the introduction of a cutoff radius, and the interactions between atoms are modeled as a sum of individual atomic contributions. Conversely, global descriptors (such as inverse pair-wise distances[45,131,133]) can serve to build models where the prediction is obtained for the whole structure. Both approaches have their

own advantages: for instance, while local descriptors can identify similar neighborhoods in small molecules that can later be transferred to larger systems, global descriptors can capture all interactions of a given system whenever the reference calculations contain the relevant information. However, descriptors also have their pitfalls, some of which arise with large flexible molecules and might become problematic when dealing with complex processes, such as those happening along transition paths.

The first immediate issue that one can foresee is the limited reach of local descriptors, as imposed by the selection of the cutoff radius of the atomic environments. Fig. 4.4 shows the interatomic distance distribution in glycine and azobenzene rotation datasets. While for glycine, the largest distances remain below 6 Å and values lower than 4 Å are the most populated, in azobenzene, the distances present values of up to 12 Å, and distances between 5 Å and 8 Å are rather common. Thus, local descriptors might already face problems with molecules as large as azobenzene when relevant interactions fall outside their scope (see the results with GAP/SOAP in Sec. 4.4). An example of such interactions is the long-range ones, which play an important role in azobenzene isomerization as suggested from the paths constructed in Sec. 4.2. Specifically, vdW interactions decrease going from the *cis* to the *trans* configuration, and these interactions lead to an increment in the energy barriers of all paths of more than 1.0 kcal mol^{-1} (the details are shown in Tables A.2-A.5 in the Appendices). Increasing the cutoff radius appears to be a straightforward solution, but the potential gain in accuracy might lead to a significant loss in efficiency.

The second problem affects both local and global descriptors and is related to the scope of the descriptor itself. If some important features are not included or not properly represented in the selected descriptor, the method will fail to achieve appropriate performance (see the results with sGDML in Sec. 4.5). This could well apply to transition paths, where small changes in specific degrees of freedom of the molecule result in considerable energy variations. The addition of the relevant features to the descriptor might alleviate this issue but requires *a priori* knowledge of the studied system.

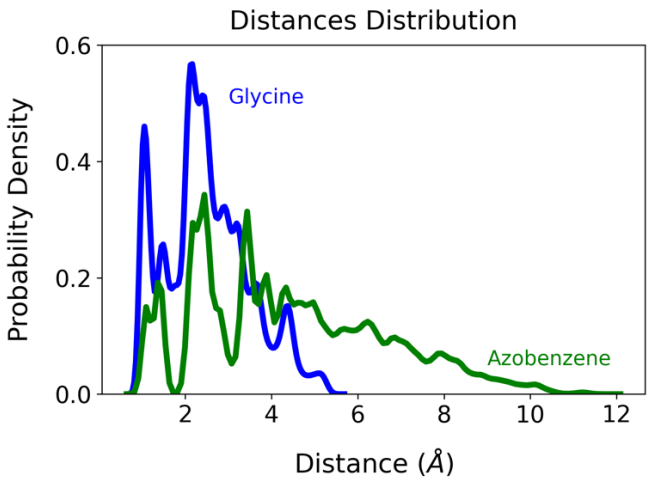

Figure 4.4: Distribution of interatomic distances (in Å) in glycine and azobenzene rotation datasets. The molecules present different length scales that will make the performance of ML methods variable from one system to another. The distributions were constructed using all the interatomic distances of all configurations in glycine and azobenzene rotation datasets, which, in turn, were built from MD simulations.

4.4 Accuracy of ML Models for Transition Paths

Although ML potentials have evolved successfully, there are many open challenges. Among others, the problem of building accurate and data-efficient ML models for flexible molecules describing equilibrium states and the transition pathways between them deserves special attention. Below, the performance of the state-of-art ML models (BPNN,[60,62] SchNet, [66,225,226] GAP/SOAP,[56–58] and sGDML[45–48]) is assessed on the PES of glycine and azobenzene molecules. Namely, (a) the glycine dataset and (b) the inversion and rotation datasets of azobenzene (see Sec. 4.2) were used.

The training and test sets were created as follows: first, the training set selection process of sGDML (which draws a sample from the dataset that preserves its original energy distribution) was used to create subsets from each of the datasets used in this work. The subsets have a size equal to five times the number of training points (e.g., for 1000 training points, a subset of 5000 configurations was constructed). Then, fivefold cross-validation was performed on each subset, using a single fold for training and the rest for testing. For instance, when using 1000 training points, the model was tested with the remaining 4000 configurations (which represents around 15% of the datasets of azobenzene and 40 % of the glycine dataset). The cross-validation tasks were randomly created while ensuring that in each fold the energy distribution of the

whole dataset was preserved. In this way, each of the folds is representative of the original dataset such that testing the models on these folds will give the same results as testing on the complete dataset. The proposed training/test set selection procedure is comparable to a default random scheme, as implemented in SchNet, while providing more reliable and accurate ML models.

Fig. 4.5 shows the energy and force prediction accuracy for the best models out of all cross-validation tasks. For SchNet, BPNN, and GAP/SOAP, the best model for a given training set size is the one showing the lowest energy RMSE in the test set because these methods are trained on energies and forces. In the case of sGDML, the best model usually is the one with the lowest force RMSE in the test set because sGDML is trained only on forces. However, if two or more sGDML models with similar force RMSE [$\sim$0.1 kcal (mol Å)$^{-1}$] present substantially different energy RMSEs ($\sim$1.0 kcal mol^{-1}), the one with the lowest energy RMSE was favored. The errors discussed in the following paragraphs correspond to those of the best models.

For the small glycine molecule (Fig. 4.5a), BPNN presents high errors, with a RMSE above 6.0 kcal (mol Å)$^{-1}$ for forces and around 2.0 kcal mol^{-1} for energies, even after using 1000 training points. sGDML and SchNet perform much better with errors below 1.0 kcal mol^{-1} and 1.0 kcal (mol Å)$^{-1}$ with 300 and 400 training points, respectively. GAP/SOAP also shows a good performance in energy prediction with errors below 1.0 kcal mol^{-1} after using 100 training points, although it is less accurate when predicting forces [errors remain around 1.5 kcal (mol Å)$^{-1}$ with 1000 training points]. Based on this analysis, BPNN is henceforth not considered as a valid candidate to reproduce a more complex PES of flexible molecules using a limited amount of training points, and the remaining analysis is only focused on GAP/SOAP, sGDML, and SchNet.

For the azobenzene datasets, the results for different models show high variability. GAP/SOAP obtains an error in energies below 1.0 kcal mol^{-1} for the inversion mechanism with only 200 training points (Fig. 4.5b). However, for the rotation mechanism, this performance is achieved with 400 training points (Fig. 4.5c). In addition, force prediction accuracy is worse for the rotation mechanism [remains above 2.4 kcal (mol Å)$^{-1}$ with 1000 training points] than for the inversion mechanism [remains above 1.8 kcal (mol Å)$^{-1}$ with 1000 training points] along the whole learning curves.

This means that the parts of the PES that are covered by each transition process cannot be evenly described with the same mapping. Indeed, SOAP learns the local information (within the selected cutoff radius), which quickly saturates with the increase in the training set, but cannot equally capture the relevant interactions of both mechanisms.

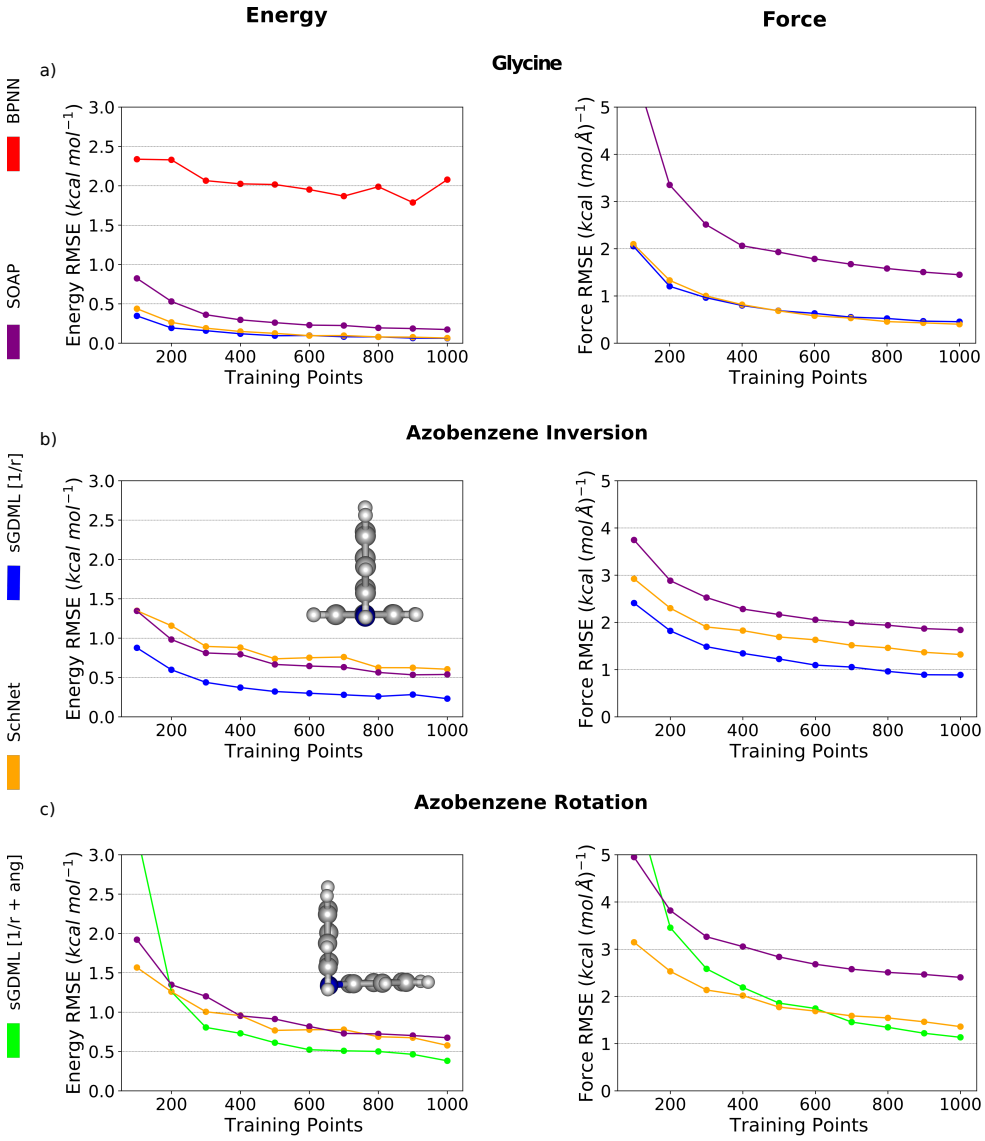

Figure 4.5: Energy (in kcal mol^{-1}) and force [in kcal (mol Å)$^{-1}$] prediction accuracy of the best models in terms of RMSE as a function of training set size. Prediction accuracy for a) glycine, b) inversion, and c) rotation datasets of azobenzene using the best models of BPNN, GAP/SOAP, SchNet, and sGDML with the default descriptor (sGDML [1/r]) or with the extended descriptor (sGDML [1/r + ang]) out of all cross-validation tasks. Only models with errors below 5.0 kcal (mol Å)$^{-1}$ and 3.0 kcal mol^{-1} are shown. b) and c) show a side view of the highest-energy structure on the inversion and rotation paths of azobenzene, respectively.

The sGDML model achieves an outstanding performance for both transition mechanisms. For the inversion dataset, errors in energy go below 1.0 kcal mol^{-1} with 100 training points and errors in forces go below 1.0 kcal (mol Å)$^{-1}$ with 800 training points (Fig. 4.5b). For the rotation dataset, 300 training points are needed to obtain an energy RMSE lower than 1.0 kcal mol^{-1} and the force RMSE gets close to 1.1 kcal (mol Å)$^{-1}$ after using 1000 training points. To achieve this performance, however, one requires different descriptors for different mechanisms: the default descriptor (inverse pair-wise distances) for the inversion mechanism and an extended descriptor [inverse pair-wise distances, and information on bonded angles and dihedrals in the form $D_\Theta = (1 - e^{-\Theta})^2 - 1$ and $D_\Phi = 1 + cos\Phi$, respectively]. The features in a given descriptor are not evenly important for all parts of the PES. Thus, a single descriptor might be unable to properly resolve all relations between the relevant degrees of freedom for the PES and each of its transition mechanisms.

SchNet also achieves a "chemical accuracy" of 1.0 kcal mol^{-1} for both inversion and rotation datasets, but, contrary to sGDML, it can deal with both mechanisms using the same settings (Figs. 4.5b, c). However, SchNet is less reliable than sGDML when predicting forces, with an overall RMSE of around 1.4 kcal (mol Å)$^{-1}$ for both the inversion and the rotation datasets with 1000 training points. Better performance can be expected with larger training sets. However, this is a trivial solution limited in practice by the increased computational costs of the reference data for larger molecular sizes. The reason that SchNet outperforms GAP/SOAP is also clear: even though SchNet primarily learns local features, it can learn other interactions by embedding such features into the local environments for different parts of the molecule.

4.5 Challenges for ML Models in Flexible Molecules

Even though GAP/SOAP, sGDML, and SchNet methods are able to learn the PES of the azobenzene molecule with chemical accuracy, there is a considerable difference between the predictions for different methods, as well as for different transition mechanisms within the same method. Below, it is demonstrated that these contrasting results are caused by imperfections of the implemented training set selection schemes (suboptimal for a complex PES with multiple local minima), as well as intrinsic

limitations of the employed descriptors (unable to equally capture all interactions). To do so, (i) the dependence of the performance of ML methods on the specific selection of a training set is explored by considering the results of all cross-validation tasks and (ii) the descriptors of GAP/SOAP and sGDML models are modified.

The prediction accuracy of GAP/SOAP models is not considerably affected by the particular choice of a training set of each cross-validation task. The average energy RMSE over all cross-validation tasks (see Fig. 4.6) is practically the same as the RMSE of the best model (see Fig. 4.5), with both errors going below 1.0 kcal mol^{-1} with 400 training points. The main shortcoming of GAP/SOAP models seems to be the learning of long-range interactions. This would explain the ~0.3 kcal mol^{-1} larger RMSE for the rotation mechanism (Fig. 4.6). Indeed, due to the different mutual orientations of the benzene rings (see geometries of the highest-energy structures on each transition path in Figs. 4.5b, c), the vdW energy contribution along the rotation transition pathway is, in average, larger by ~0.3 kcal mol^{-1} than that for the inversion one. To further explore this conclusion, the vdW correction to the energy and forces is removed from the rotation dataset (i.e., a dataset at the PBE level of theory was constructed) and GAP/SOAP models using the same cross-validation tasks were retrained. Unexpectedly, the prediction accuracy of all models was the same with or without the inclusion of long-range vdW interactions. This means that GAP/SOAP might face a different problem, which could be related to an issue in the mapping from configuration space to the PES for the rotation mechanism. Since this error is comparable to the vdW energy contributions, neglecting it is comparable to ignoring non-covalent vdW interactions. Increasing the cutoff radius while keeping the same number of basis functions does not resolve this issue. Fig. 4.7a shows the best energy and force RMSE as a function of the cutoff radius in GAP/SOAP models with 12 radial and 6 angular functions. In fact, both energy and force prediction become slightly worse. Increasing the number of basis functions to alleviate this issue would lead to computationally expensive ML models, impractical for realistic applications.

The sGDML model provides the most accurate and data-efficient FF but faces two important issues. The first one is the descriptor, as shown in Fig. 4.5, where one has to employ different descriptors for different transition mechanisms. Specifically, for the inversion mechanism, the default sGDML descriptor (inverse pair-wise distances) is sufficient. In contrast, a reliable description of the rotation mechanism requires the

inclusion of information about angles and dihedrals in the form $D_\Theta = (1 - e^{-\Theta})^2 - 1$ and $D_\Phi = 1 + cos\Phi$, where Θ and Φ are any bonded angle and dihedral of the molecule in radians, respectively. The performance of the default and extended descriptors is shown in Fig. 4.7b. The default descriptor for the rotation mechanism and the extended descriptor for the inversion mechanism present considerable oscillations in the energy error as a function of training set size (Fig. 4.7b), which is unacceptable behavior for a reliable ML model. The reason is simple: on the one hand, inverse pair-wise distances cannot correctly resolve the states along the transition path of the rotation mechanism, which are defined by changes in the dihedral angle ϕ. On the other hand, all angles and dihedrals are not equally representative of the inversion mechanism, and adding them misleads the model in this case.

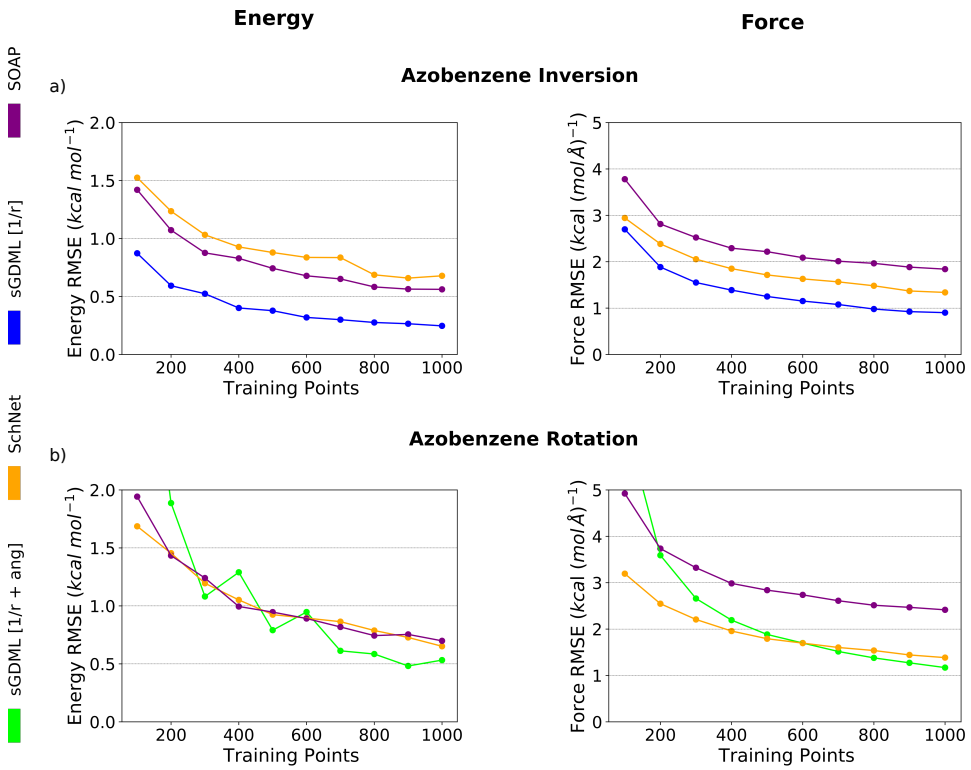

Figure 4.6: Mean of the energy (in kcal mol^{-1}) and the force [in kcal (mol Å)$^{-1}$] prediction accuracy over all cross-validation tasks in terms of RMSE as a function of training set size. RMSEs of models trained with SchNet, SOAP, and sGDML with the default descriptor (sGDML [1/r]) and with the extended descriptor (sGDML [1/r + ang]) for a) the inversion and b) the rotation datasets of azobenzene. Only models with errors below 5.0 kcal (mol Å)$^{-1}$ and 2.0 kcal mol^{-1} are shown. In the case of sGDML models, only the curve of the best performing descriptor for each mechanism is shown.

It is important to remark that even when using an appropriate descriptor, the training set selection is still crucial. Out of all the methods, sGDML is the most affected by the model selection in cross-validation tasks (compare Figs. 4.5c and 4.6b). While for 1000 training points, the difference between the average energy RMSE over all cross-validation tasks and the energy RMSE of the best model is of only 0.2 kcal mol^{-1}, for 400 training points, this difference is as large as 0.6 kcal mol^{-1}. Thereby, one needs to be very careful when selecting the best sGDML models.

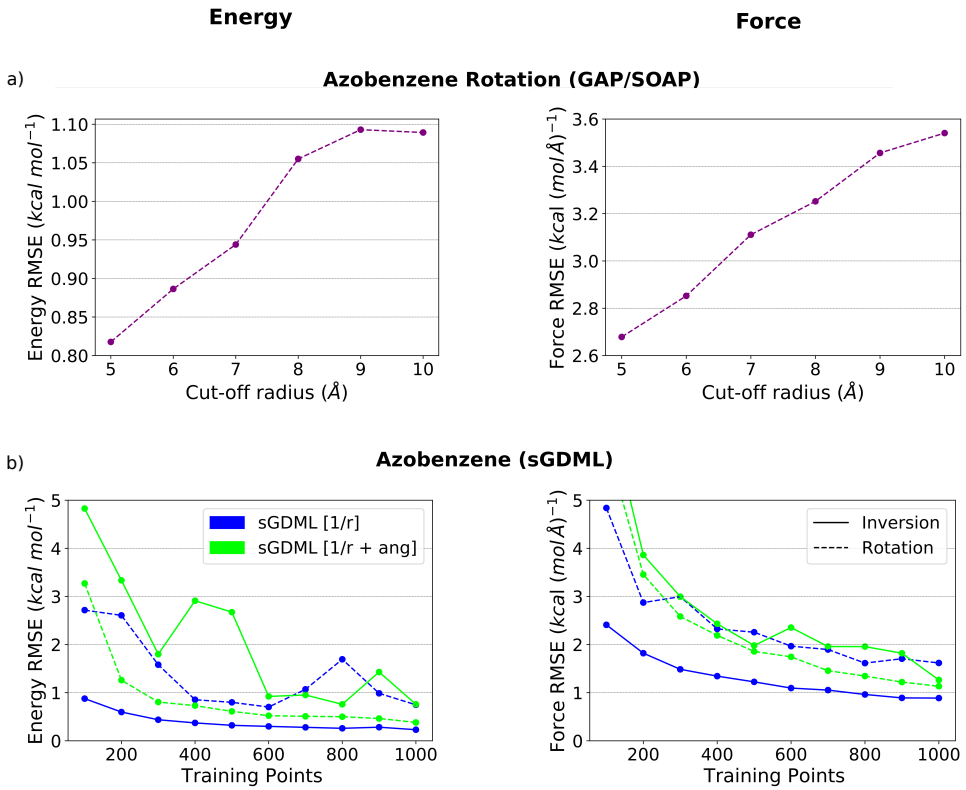

Figure 4.7: Energy (in kcal mol^{-1}) and force [in kcal (mol Å)$^{-1}$] prediction accuracy in terms of RMSE of GAP/SOAP models with different cutoff radius and sGDML model with different descriptors. RMSEs a) as a function of cutoff radius for GAP/SOAP models of the rotation dataset of azobenzene trained with 600 training points and b) as a function of training set size for sGDML with the default descriptor (sGDML [1/r]) and with the extended descriptor (sGDML [1/r + ang]) for the inversion and rotation datasets of azobenzene. Only models with errors below 5.0 kcal (mol Å)$^{-1}$ and 5.0 kcal mol^{-1} are shown.

Furthermore, the training set dependency often leads to models with a similar force RMSE but considerably different energy RMSEs. For instance, the average difference in energy RMSEs between the best and the worst models for the rotation mechanism using the extended descriptor (over all training set sizes considered here) is of 0.7 kcal mol^{-1}, while the average difference in force RMSEs is less than

0.2 kcal (mol Å)$^{-1}$. To understand this behavior, one needs to recall that sGDML models contain two hyperparameters apart from the regularization. One is the width of the kernel, which is defined by optimizing the force predictions. The other one is the constant shift for the energy, which is employed to minimize the difference between the prediction results and the energy values in the dataset. The energies of different conformers of flexible molecules are often degenerate. Consequently, even though all our training sets follow the energy distribution of the complete dataset, they represent different parts of the PES unequally. As a result, the energy shift hyperparameter obtained from a given training set can become suboptimal for the whole dataset. Hence, the force-based model selection scheme, as implemented in sGDML, may lead to large oscillation in energy prediction accuracy as a function of training set size (similar to those in the green solid line in Fig. 4.7b). To resolve this issue, one should consider both energy and forces to select the optimal model. For each cross-validation task, the training scheme does not change and still relies only on forces, but for selecting the best model out of many possibilities, the energy prediction accuracy is also taken into account. Summarizing, accurate and data-efficient models are achievable with sGDML, but both descriptors and training sets must be carefully selected.

SchNet is an optimal compromise between GAP/SOAP and sGDML models. Like GAP/SOAP, it does not heavily depend on the specific selection of a training set (see Fig. 4.6), while being capable to learn all kind of interactions, akin to sGDML (see Fig. 4.5). As a result, SchNet reproduces both transition mechanisms equally accurately using the same settings, with errors only slightly larger than those of the sGDML models. This is a consequence of the embedding of local features of different atoms through the interaction layers. To train the SchNet models, six interaction layers were employed with a 5 Å cutoff radius for local environment. This architecture guarantees covering all possible interatomic distances within an azobenzene molecule (Fig. 4.4), making the SchNet models effectively global. However, a good overall RMSE might not always mean a good ML model. Fig. 4.8a shows the energy and force prediction accuracy on different clusters of the rotation dataset of the best SchNet model out of the five cross-validation tasks with 1000 training points [RMSEs of 0.6 kcal mol^{-1} and 1.3 kcal (mol Å)$^{-1}$ over the complete rotation dataset]. Each cluster corresponds to different values of the dihedral angle ϕ (from the interval between 0º

and 10º for index 0 to the interval between 170º and 180º for index 17). One can see that the errors for close-to-equilibrium configurations are four times larger than those for the transition regions. Increasing the training set size and having information of the two mechanisms (i.e., adding the data of the inversion mechanism to the rotation dataset) does not change this ratio (Fig. 4.8b).

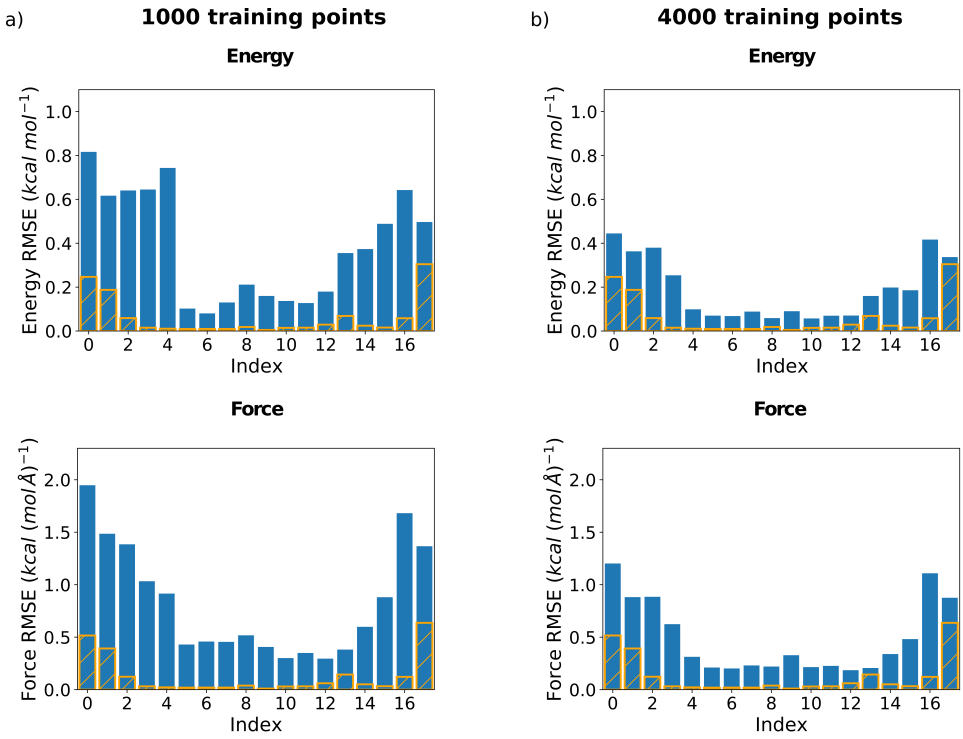

Figure 4.8: Energy (in kcal mol^{-1}) and force [in kcal (mol Å)$^{-1}$] prediction accuracy in terms of RMSE for different clusters of the rotation dataset using SchNet models. RMSEs of a) the best SchNet model obtained with 1000 training points of the rotation dataset [RMSEs of 0.6 kcal mol^{-1} and 1.3 kcal (mol Å)$^{-1}$ over the complete rotation dataset] and b) a SchNet model trained with 4000 configurations combining the rotation and inversion datasets [RMSEs of 0.4 kcal mol^{-1} and 0.8 kcal (mol Å)$^{-1}$ over the complete rotation dataset]. The configurations for each cluster were selected depending on the value of the dihedral angle ϕ, going from an interval between 0º and 10º for index 0 to an interval between 170º and 180º for index 17. Relative population of each cluster is also indicated (orange solid lines, arbitrary units).

It is important to mention that while building these models, one must expect to be able to use them for accurate simulations of the transition mechanism, computing reaction rates, and lifetime of *cis* configuration. In order to achieve this goal, one needs to ensure that the MLFF is equally accurate for all relevant parts of the PES. As one can see from Fig. 4.8, this requirement is not fulfilled by the obtained SchNet FFs. Importantly, by using a total of only 600 training points for both *trans*- and *cis*-like

configurations (300 for each isomer), one can train a single sGDML model that reproduces the performance of the 4k SchNet model on clusters 0, 1, 2, 16, and 17. Hence, while learning the entire PES of the azobenzene molecule is possible within a SchNet model, this approach is not particularly data efficient. There are two options to solve this problem: one is to use a training set optimization technique, flattening the prediction across the configuration space.[244] The second option would be to design schemes that combine a set of local models into a global one, finding optimal descriptors, training sets, and models for each part of the PES.

One might think of training ML models on a dataset constructed from metadynamics simulations, called ML_{meta} models hereafter. However, as mentioned in Sec. 4.2, metadynamics datasets of the same size would contain less information about the transition processes as compared to the rotation/inversion datasets. One expects the ML_{meta} models to be less accurate compared to ML models trained on the rotation/inversion datasets (for simplicity, henceforth, this models are referred as ML_{md} models). Fig. 4.9 shows the energy and force prediction accuracy of $SchNet_{meta}$ models when applied to reproduce the rotation dataset from Sec. 4.2. Each cluster corresponds to different values of the dihedral angle ϕ (from the interval between 0º and 10º for index 0 to the interval between 170º and 180º for index 17) to be compared to Fig. 4.8. The average RMSE of the resulting $SchNet_{meta}$ models is twice larger than that of $SchNet_{md}$ models presented in Fig. 4.8. In addition, the distribution of errors is drastically different. The largest errors are now in the transition region (clusters 4–13) and can be up to an order of magnitude larger compared to those of the $SchNet_{md}$ models. Moreover, the error distribution is not uniform, which means that the different parts of the PES responsible for the transition are not equally represented within the metadynamics dataset.

It is important to comment that several approaches exist to improve ML models to large and flexible molecules, such as multiscale GAP/SOAP,[143] the recently developed LODE,[245] or the addition of physical potentials.[246] These are all promising directions, and this study demonstrates how more advanced techniques are necessary even for relatively small molecules such as azobenzene. Nevertheless, multiscale approaches might lead to less data-efficient models. Indeed, a single GAP/SOAP model is the most computationally expensive method among those used in this work. Therefore, although a multiscale GAP/SOAP could solve the problem for azobenzene, it could

become infeasible when dealing with larger molecules. As for methods to learn long-range interactions, they need to be carefully integrated with local ML models since local properties used in long-range interaction models can vary substantially for different local chemical environments. This is supported by the results observed when comparing the performance of GAP/SOAP$_{md}$ models trained on PBE and PBE+TS reference data.

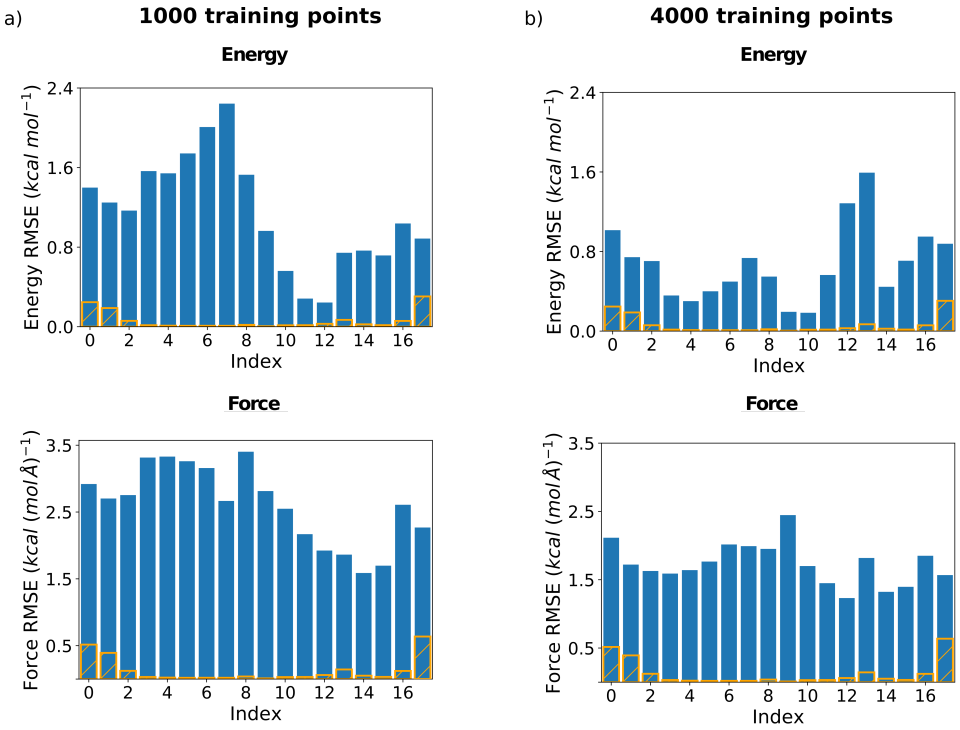

Figure 4.9: Energy (in kcal mol^{-1}) and force [in kcal (mol Å)$^{-1}$] prediction accuracy in terms of RMSE for different clusters of the rotation dataset using SchNet$_{meta}$ models. RMSEs ofa) a SchNet$_{meta}$ model trained with 1000 configurations and b) a SchNet$_{meta}$ model trained with 4000 configurations. The configurations for each cluster were selected depending on the value of the dihedral angle ϕ, going from an interval between 0º and 10º for index 0 to an interval between 170º and 180º for index 17. Relative population of each cluster is also indicated (orange solid lines, arbitrary units).

The final step to confirm the reliability of the ML$_{md}$ models is to demonstrate their applicability in MD simulations. Fig. 4.10 shows the values of the dihedral angle ϕ and the angle θ as a function of simulation time obtained with constant-temperature MD runs starting from structures close to the transition states. The best GAP/SOAP$_{md}$, SchNet$_{md}$, and sGDML$_{md}$ models trained on 1000 training points were selected to run with each method four MD simulations of 50k steps at 300 K with a time step of 0.5 fs. For the simulation that starts from step 7 of the inversion mechanism (Fig. 4.10a),

GAP/SOAP$_{md}$ and SchNet$_{md}$ models lead to the *trans* isomer, while the sGDML$_{md}$ model leads to the *cis* isomer. For the simulations that start from step 9 of the inversion path and steps 8 and 10 of the rotation path (Figs. 4.10b-d), all models lead to the *trans*, *cis*, and *trans* isomers, respectively. To keep local geometry fluctuations under control during the relaxation process, the friction coefficient for the Langevin thermostat was set to 2 fs. From Fig. 4.10, one can see that all ML$_{md}$ models are able to correctly reproduce the transition processes avoiding unphysical configurations, despite the large fluctuations in the angles caused by the thermostat. This demonstrates that the ML$_{md}$ models are reliable and stable, and the constructed datasets contain the information necessary for studying the *cis* to *trans* thermal relaxation of azobenzene.

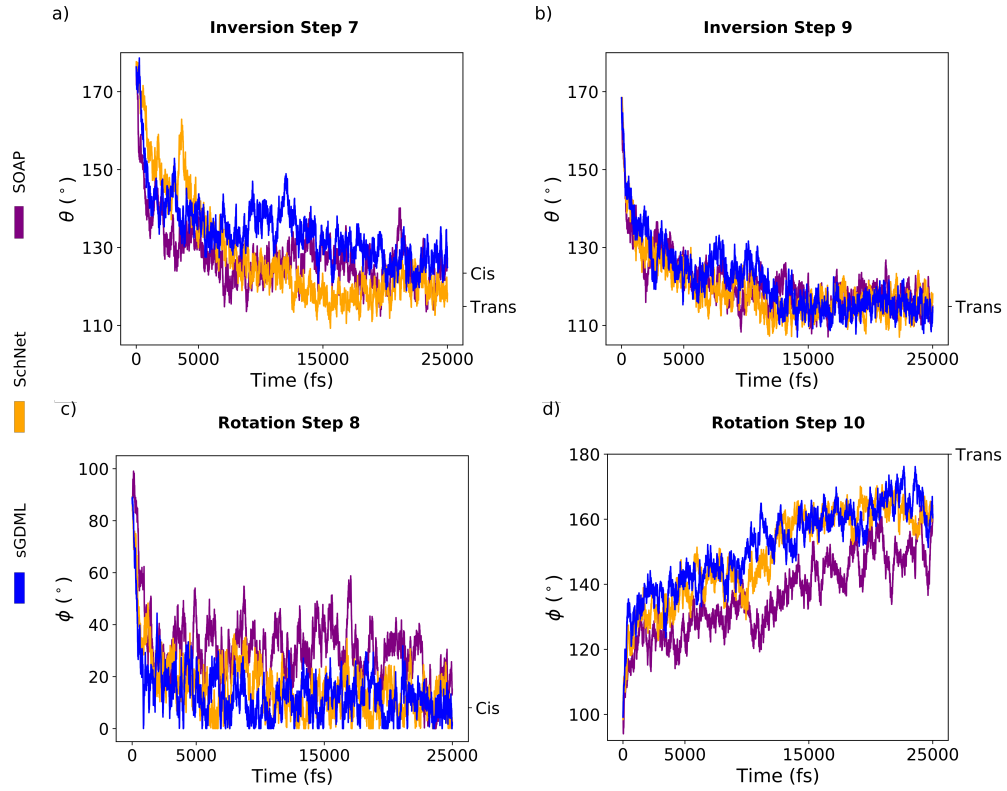

Figure 4.10: Change in the main degree of freedom (ϕ for the rotation mechanism and θ for the inversion one, in º) as a function of the simulation time (in fs) for MD simulations performed with ML$_{md}$ models. Change in a, b) θ and c, d) ϕ for the four different MD simulations done using the best GAP/SOAP$_{md}$, sGDML$_{md}$, and SchNet$_{md}$ models trained on 1000 training points. The initial geometry used for each simulation is indicated over the graph. Tick labels on the right mark the value of the degree of freedom in the *trans* isomer and the *cis* isomer.

Summarizing the results of this section, while the state-of-the-art ML models are capable of reproducing the complex PES of flexible molecules, this challenge is far from being solved in practice. Default approaches demonstrating excellent performance for small molecules or rigid systems struggle with increasing flexibility and dimensionality. Even the best performing models present difficulties to efficiently learn the PES in its entirety. Moreover, the way datasets are constructed is crucial for the performance of the resulting ML models. To overcome this challenge, one should ensure that the descriptors contain all relevant features to capture the complex geometrical transformations in the high-dimensional PES. In addition, training sets must represent all parts of configuration space, which exhibits high energetic degeneracy, which makes purely following an energy distribution ineffective. Although MD simulations executed with the constructed ML potentials are promising, one can suggest further developments of robust approaches for selecting training points, appropriate descriptors, or even using different models for different parts of the PES.

4.6 Conclusions

The challenge of modeling the PES of flexible molecules using state-of-the-art ML models when using limited sets of training data was discussed in the present chapter. The results show that methods based on local descriptors (e.g., BPNN and GAP/SOAP) saturate quickly with the increase in the number of training points, while not achieving the desired prediction accuracy. This is a consequence of the limitations imposed by the cutoff radius for the atomic environments. The ML methods based on global descriptors (e.g., sGDML) require careful model selection procedures when the reference datasets consist of several disconnected parts of the PES. The main challenges are the inability of simple training schemes to select appropriate training datasets in unbalanced reference data and the limitations of the standard molecular descriptors to pick up the features describing the complex geometric transformations in flexible molecules. Finally, end-to-end NNs (e.g., SchNet) do not reproduce all relevant parts of the PES with equal accuracy. Moreover, NNs require a larger amount of training data compared to kernel methods, which can result in a high computational cost when generating reference datasets. All the tested MLFFs in their current form can be further improved for quantitative studies of complex processes in flexible molecules. Important features of ML, such as descriptors and selection of training

points, could and should be revised. The varying performance of most of the ML models for the rotation and inversion mechanisms in azobenzene also suggests switching from learning the entire PES within a single task to the employment of multiple local models for different parts of the PES and then combining them into a global FF.

CHAPTER 5

Optimizing Descriptors for Accurate ML Models for Large and Flexible Molecules

In Chapter 3, the power of MLFFs to provide insights was shown by studying the impact of NQE on molecular interactions in small (semi-)rigid molecules. Nevertheless, from the results discussed in Chapter 4, there are still challenges for MLFFs that need to be solved in order to obtain similar insights for larger and more flexible molecules. Specifically, it has been stressed through this thesis that one of the main challenges is how to properly describe the interplay between local chemical bond fluctuations and non-local long-range interactions, prominent in such molecules.

In MLFFs, the description of interactions is handled to a greater or lesser extent by the descriptor used to encode the molecular geometries. Although different descriptors relying on different features have been successfully used to reconstruct PESs,[11,36,44–47,64–66,68–86] one cannot ensure that the features they contain optimally describe all important interactions. A common strategy to solve this issue is to construct bigger and/or more complex descriptors with additional features aiming to enable the description of the interactions missed previously (like it was shown in Chapter 4 for the rotation mechanism of azobenzene).[245] However, such a solution will quickly lead to computationally expensive descriptors with increasing size of the molecule. Already, the default SOAP descriptor contains over 10,000 features for a molecule such as the alanine tetrapeptide (Ac-Ala3-NHMe; 42 atoms) and default descriptors based on interatomic distances (e.g., the one of sGDML) exceed one thousand features with molecules of only 50 atoms. Hence, the inclusion of more features will lead to immense descriptors when applying ML methods to large molecules.

In practice, the use of descriptors containing a very high number of features compromises the performance of the ML models. Namely, an oversized descriptor i) spans a much larger space than what is effectively needed for an accurate prediction, making ML models harder to optimize, and ii) requires more computational time for its construction, leading to ML models that are inefficient for practical applications. In this context, even though many efforts have been made to optimize and reduce the size of descriptors,[136,247–251] to the best of our knowledge, an extensive assessment of the importance of different features for the modelling of PESs has not been performed to date. Such a profound analysis is a crucial step to tailor appropriate descriptors for building accurate and data-efficient ML models for large and flexible molecules.

In this chapter, this challenge is addressed by finding the most relevant features of a descriptor on the example of three large and flexible systems, for which a precise description of short- and long-range interactions is imperative: the alanine tetrapeptide (Ac-Ala3-NHMe; 42 atoms), the adenine-thiamine DNA base-pair dimer (AT-AT; 60 atoms), and the buckyball catcher (148 atoms). The chapter starts with a description of the computational details and the datasets used for obtaining the results. Then, a strategy to discard the least important features in a descriptor is introduced for optimizing the molecular representation. After that, the factors to be considered for finding the optimal descriptor of large and flexible molecules are discussed. Finally, the performance of models trained on a global, local (relying on a predefined cutoff radius) and optimized global descriptor is assessed in the description of interactions (from the context of the contribution of each atom to the prediction of forces in all other atoms) within the test molecules, as well as in the prediction of forces. For this, sGDML[45,46] and GAPs[56] using the SOAP representation[58] are employed.

The structure of the chapter is given as follows: in Sec. 5.1 the computational details and datasets are described. In Sec. 5.2, the procedure used to remove features for optimizing the descriptor is introduced. In Sec. 5.3, there is a discussion of the considerations that must be taken into account for determining the optimal descriptor of large and flexible molecules. Then, in Sec. 5.4, there is an analysis of the performance of typical global and local descriptors, and that of an optimized descriptor as proposed in this chapter, with special attention to the number of features in each descriptor and how they capture the interactions within the test systems. Sec. 5.5 contains the conclusions to this Chapter.

5.1 Computational Details and Datasets

5.1.1 Reference Datasets

MD was used to construct the reference datasets. All calculations were done using the DFT functional PBE[228] with many-body dispersion (MBD)[252,253] to account for vdW interactions. The calculations were carried out either with i-PI[185] wrapped with FHI-aims code[227] to compute forces and energies or with FHI-aims code alone. Table 5.1 includes all relevant information of the datasets.

Table 5.1: Settings of the MD simulations of the datasets used in the chapter. Temperature is given in K and the step size in fs. Coefficient refers to the friction coefficient (in fs) for the global Langevin thermostat, and to the effective mass (in cm^{-1}) for the Nosé-Hoover thermostat.

Molecule	Level of Theory	Temperature	Step size	Thermostat	Coefficient
Ac-Ala3-NHMe	PBE+MBD Tight	500	1	Global Langevin	2
AT-AT	PBE+MBD Tight	500	1	Global Langevin	2
Buckyball catcher	PBE+MBD Light	400	1	Nosé-Hoover	1700

5.1.2 ML Models

The ML models were built with sGDML[45,46] and GAPs[56] with the SOAP representation.[58]

sGDML models were trained using a numerical solver with an initial value of 70 inducing points. All models were validated using 1000 configurations and values for the hyperparameter σ from 10 (20 for the buckyball catcher) to 200 with a step of 20 were explored. No symmetries were considered in the models for a fair comparison between the default descriptor and those with a reduced size.

GAP/SOAP models were trained using 12 radial and 6 angular functions for the descriptor. The cutoff radius was set to 5 Å. δ was set to 0.25, the atom σ was set to

0.3, and the default σs for energy and forces were set to 0.001 and 0.1, respectively. All calculations were performed with the QUIP program package.[231]

5.1.3 Pairwise Contributions to Force Prediction

Given atoms i and j in a molecule, one can consider that the contribution of atom i to the prediction of the total force on atom j is related to the amount of interaction between the given atoms. To compute such pairwise contributions to the force prediction of a molecule (F_{ij}), the following was considered. From Eq. 10, one knows that the contribution of each training point M to the final prediction is given by each αK term in the predictor. Since the force predictor in sGDML has not only one but $3N$ values of α (one for each coordinate of each atom) for each training point, the contribution of a given atom i to the prediction of all atoms was obtained by calculating only the sum of the $Mx3$ αK terms corresponding to atom i. The result of this calculation is an $Nx3$ matrix with the contributions of atom i to the prediction of each force component on all atoms in the molecule. Then, the norm of each row in this matrix was computed to get the contribution of atom i to the total force on all N atoms. Finally, repeating this process for all other atoms in the molecule generated a F_{ij} matrix.

5.2 Theory

The reliability of an ML model depends, of course, on the set of samples used for training, but also on the descriptor used to encode the input samples. Although typical descriptors that are constructed before the training process (e.g., SOAP,[58] ACSFs,[60] inverse pairwise distances[45,131,133]) provide reliable MLFFs for different molecules and materials, there is no guarantee that a given descriptor is the optimal for all different instances of a given learning task. Furthermore, the computational cost of generating a descriptor steadily grows with the size of the system. For instance, the length of the input vector of descriptors based on interatomic distances scales as the square of the number of atoms, and other descriptors, such as SOAP, have a fingerprint vector that can contain tens of thousands of elements. So, the use of state-of-the-art descriptors can become impractical for constructing ML models for large molecules. A solution to this issue is the use of data-driven approaches to construct the descriptor. For

example, end-to-end NNs learn a representation directly from reference data using the atom types and positions of the system as input. The models resulting from these architectures are more accurate than those of NNs relying on predefined descriptors.[42,66,67,254] Unfortunately, such an end-to-end approach is not plausible with kernel-based methods because kernels cannot learn their own representation. In this regard, different approaches (useful for both NN and kernel-based methods) have been proposed to optimize descriptors with reference data, targeting specific properties of interest.[136,247–251] They have been successfully employed improving the accuracy of models trained using descriptors such as ACSFs and SOAP.[249–251] What is still missing, however, are optimized descriptors that not only provide more accurate and efficient MLFFs, but also an interpretable set of features that guarantees a precise description of all relevant interactions. An analysis of such descriptors is important to lay the foundations for constructing accurate and data-efficient models of large and flexible molecules and is the aim of this Chapter.

5.2.1 Reducing Descriptor Dimensionality

A typical procedure to assess the importance of different inputs in the results of a mathematical model is sensitivity analysis (SA).[255,256] In SA, one studies the relationship between inputs (i.e., features in the descriptor) and the uncertainty in the output(s) of the model. Namely, it quantifies how much the output of the model changes with respect to a change of each input variable. However, SA cannot explain the output of the ML model itself, but a variation of the outputs. This means that one could not use such an approach for evaluating the effects of varying/removing features on the prediction accuracy of the model. Because of this, here I describe a strategy to reduce the size of the descriptor that, like SA, assesses the importance of all features but, contrary to SA, it does it by taking the accuracy of the output as reference instead of its uncertainty.

The approach starts from a fully trained ML model (called $\mathrm{ML}_{\mathrm{original}}$ hereafter), considering a kernel-based method as the model architecture (Eq. 10) and a default (containing all n features) descriptor $\boldsymbol{x}$

$$\boldsymbol{x} = [x_1, x_2, \cdots, x_n]. \qquad (24)$$

Then, a subset of configurations in the dataset (avoiding samples used for training and validation) is randomly selected, and the energy and/or forces are computed using the $ML_{original}$ model. After that, the prediction of energies/forces of the subset of configurations is repeated n times with the $ML_{original}$ model, but now "masking" at each iteration one of the n features in the descriptor. The term "masking" in this procedure means that at iteration i the i-th feature in the descriptor is set to zero for all training and query configurations. This makes all configurations to have the same value for the i-th feature in the descriptor. Thus, the i-th feature does not affect the value of the kernel function for two given configurations.

Once one has the sets of energies/forces predictions with the $ML_{original}$ model masking each feature (for simplicity, henceforth, we refer to each of these sets as ML^{n}_{masked} predictions), the RMSE of each set of ML^{n}_{masked} predictions is computed with respect to the $ML_{original}$ one. Since the value of the regression coefficients (α_j in Eq. 10) is not changed when computing ML^{n}_{masked}, the lowest RMSE values should be implicitly related to less important features. However, it is not trivial to set a threshold under which one can consider a feature as irrelevant. Indeed, the values of the RMSEs depend on i) the predicted property, ii) the system(s) for which the model is trained, and iii) the reference data used for training. To solve this problem, a distribution with the force RMSEs of all ML^{n}_{masked} predictions is constructed, and the concept of percentile is used to select the lowest RMSEs values. A percentile is a score (e.g., an RMSE) at or below which a given percentage of scores in its distribution falls. For the sake of clarity Fig. 5.1 shows a hypothetical distribution of RMSEs and its 20th percentile. The 20th percentile is the RMSE at or below which 20 % of the RMSEs in the distribution may be found. So, if one selects all the RMSEs at or below this 20th percentile (shaded blue area in Fig. 5.1) and retrieves the features related to them, one would obtain, for a descriptor containing 100 features, the 20 least important features. In this chapter, different percentiles (in a range from 10 to 90) are used to properly study how the removal of different types of features in the descriptor affects an ML model.

It is important to remark that the RMSEs arising from "masking" a feature are by no means related to the accuracy of a model trained with a descriptor not containing such feature. The errors are a consequence of keeping the values of the regression coefficients of the $ML_{original}$ model unchanged, which might be suboptimal for the

masked descriptor. The real accuracy of a model with a descriptor of smaller size can only be obtained by retraining. Thus, as final step, a new model is trained after removing from the default descriptor (Eq. 24) all the features whose corresponding RMSEs are at or below the selected percentile.

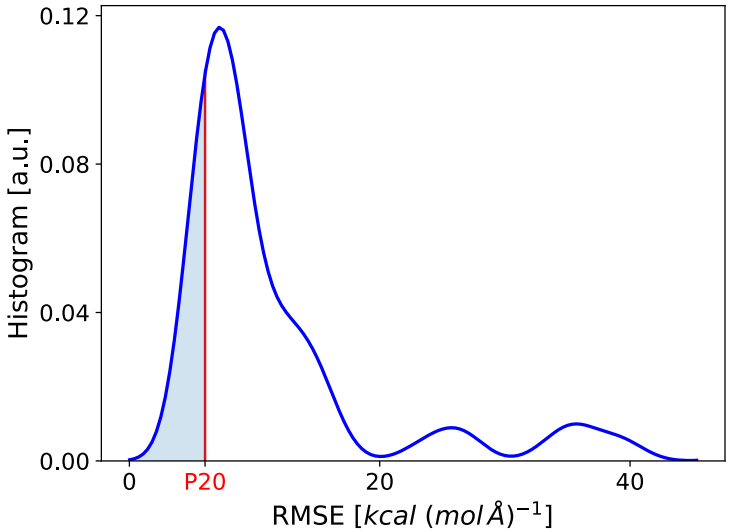

Figure 5.1: Hypothetical distribution of root mean squared errors (RMSEs) in kcal (mol Å)$^{-1}$. The red vertical line indicates the limit set by the 20th percentile of the distribution (P20). The shaded blue area between the distribution curve and the vertical line shows the area of the distribution where the RMSEs are lower than P20.

Important Considerations

This subsection discusses some considerations that are important when applying the aforementioned approach. The results presented here were obtained with the sGDML method using its default descriptor, but any other method or descriptor could have been used as well.

1. **The result of the dimensionality reduction approach is independent of the training set used for training the $ML_{original}$ model.** For example, 5 different models for aspirin (210 features in the descriptor) with 500 training points were trained, and 5 different subset of 3000 configurations (one for each model) were used to select the RMSEs under the 15th percentile. From the 32 removed features, 24 of them (75 %) were removed with all 5 aspirin models. This means that when applying the dimensionality reduction one only needs to ensure that the training set is representative of the dataset.

2. **The result of the dimensionality reduction approach is independent of the training set size used for training the $ML_{original}$ model.** For instance, 3 different models for alanine tetrapeptide (Ac-Ala3-NHMe; 861 features in the descriptor) with 100, 200, and 500 configurations were trained, and a subset of 3000 configurations (the same for all models) was used to select the RMSEs under the 65th percentile. From the 559 removed features, 443 of them (~79 %) were removed with all three models. This means that one does not need to start with a very accurate (probably computationally expensive) initial model.
3. **The result of the dimensionality reduction approach is independent of the size of the subset used for computing the $ML_{original}$ and the ML^{n}_{masked} predictions.** As an example, a model for an adenine-thiamine DNA base-pair dimer (AT-AT; 1770 features in the descriptor) with 100 training points was trained, and different subsets with 10, 100, and 1000 configurations were used for selecting the RMSEs under the 60th percentile. From the 1062 removed features, 1030 of them (~97 %) were removed with all subsets. This is advantageous because one can efficiently assess the importance of all features in the representation of a molecule, even if the descriptor has thousands of features due to the size of the molecule.

The reason why the overlap between removed features in the examples of points 1 and 2 (less than 80 %) is not as high as in the one of point 3 (~97 %) is simple. Most of the features that are not removed by all models (e.g., the remaining 8 features in each aspirin model discussed in point 1) involve an H atom. H atoms are the ones that fluctuate the most in MD simulations, which are the origin of the datasets. Thus, it is not surprising that the relevance of a given feature involving an H atom varies between different ML models (trained on different sets) without affecting the reliability of the resulting FF.

The efficiency and consistency of the descriptor dimensionalty reduction approach explained in this section give the opportunity to assess thoroughly the relevance of the features in the descriptor to construct accurate MLFFs, even for molecules with more than 100 atoms.

5.3 Size of the Optimal Descriptor for ML Models

Here, I discuss in detail the accuracy of MLFFs selecting different percentiles for removing features on the example of the default global descriptor of sGDML. Through an analysis of the performance of different ML models trained using different sizes of the descriptor, one can define the size of the optimal descriptor. For this section, Ac-Ala3-NHMe, the AT-AT dimer, the buckyball catcher, and aspirin were considered in order to compare results between large and flexible molecules, for which ML models are still suboptimal, and a rather small semi-rigid molecule, for which one can already build very accurate and data-efficient MLFFs.[45,46,52,66,67] For each of these systems, sGDML models with 300, 500, 800 and 1000 training points were trained using descriptors of different sizes (i.e. selecting different percentiles for removing features). For the larger molecules, percentiles from 10 to 90 in steps of 10 were explored, while for aspirin only percentiles from 10 to 50 in steps of 10 were selected (these models will be referred as ML_{PX}. X being the percentile selected to remove features). In Fig. 5.2, curves of energy and force RMSEs are shown for each training set size as a function of the size of the descriptor. The RMSEs were computed from predictions of all configurations in the dataset that were not used for training or validation. For selecting the optimal descriptor only the models trained with the highest amount of training configurations (i.e. models trained using 1000 training points) were taken into account.

For a small semi-rigid molecule as aspirin (210 features in its original descriptor), the optimal descriptor is the default global descriptor. The models present an increase in energy and force RMSEs with smaller number of features (see Fig. 5.2A). Conversely, for Ac-Ala3- NHMe (861 features), the AT-AT dimer (1770 features), and the buckyball catcher (10878 features), the optimal descriptor is different from the default global descriptor (see Figs. 5.2B-D). ML models for larger and more flexible molecules inherently have a reduced accuracy with respect to those of smaller ones (compare errors for aspirin in Fig. 5.2A and the rest of molecules in Figs. 5.2B-D). This is mainly caused by their complex PESs, more prone to incomplete training sets. For the test systems considered here, the ML_{P60} model is the optimal one. Only around 40 % of the original global descriptor (345 features for Ac-Ala3-NHMe, 708 features for the AT-AT dimer, and 4351 features for the buckyball catcher) is required to accurately describe the PES of these systems. For instance, when trained on 1000 training

points, ML_{P60} models show energy and force RMSEs that are up to 1.2 kcal mol^{-1} and 0.2 kcal $(mol\ Å)^{-1}$ lower than those of the models with the default global descriptor. A reliable selection of the optimal size of the descriptor is, however, not straightforward since one needs to use as reference an accurate model and to carefully assess the change in energy and force accuracy when reducing the size of the descriptor.

Models trained on a very limited amount of reference data might lead to wrong choices of the optimal descriptor size. This becomes evident when increasing the size and flexibility of the molecule of interest. While for the aspirin molecule (Fig. 5.2A) the behavior of the force RMSEs curves when reducing the number of features in the descriptor is qualitatively the same for all training set sizes, for the other molecules (Figs. 5.2B-D) the force RMSEs curves of the models trained on only 300 training points behave differently than those of the rest of the models. For Ac-Ala3-NHMe (Fig. 5.2B), the models trained on 300 training points show as optimal descriptor the one of the ML_{P70} model (259 features) and, for the AT-AT dimer and the buckyball catcher (Figs. 5.2C,D), models trained using 300 training points suggest that using less than 30 % (less than 531 features) and around 20 % (around 2170 features) of the original descriptor would produce more reliable models, respectively. Models that use a descriptor with a small number of features will tend to behave as if they were trained using typical local descriptors. Therefore, such models can even be transferred to configurations that are not well represented in the training set, but whose local environments are present in the training data. The advantages of these “more local” descriptors over global or optimized global descriptors vanish when more data is used for training (see curves for 500, 800, and 1000 training points). Hence, the use of training sets including a representative sample of the entire dataset is crucial for finding the optimal descriptor for large and flexible molecules.

The optimal descriptor must be the one that produces the models with the best accuracy. However, evaluating the accuracy of different sizes of a descriptor is not trivial. For instance, the constant energy shift used in sGDML for predicting energies is harder to optimize for large and flexible molecules. Molecules like the AT-AT dimer present many different minima (i.e., different ways each AT base pair is arranged respect the other), making difficult the selection of an appropriate training set for the constant.

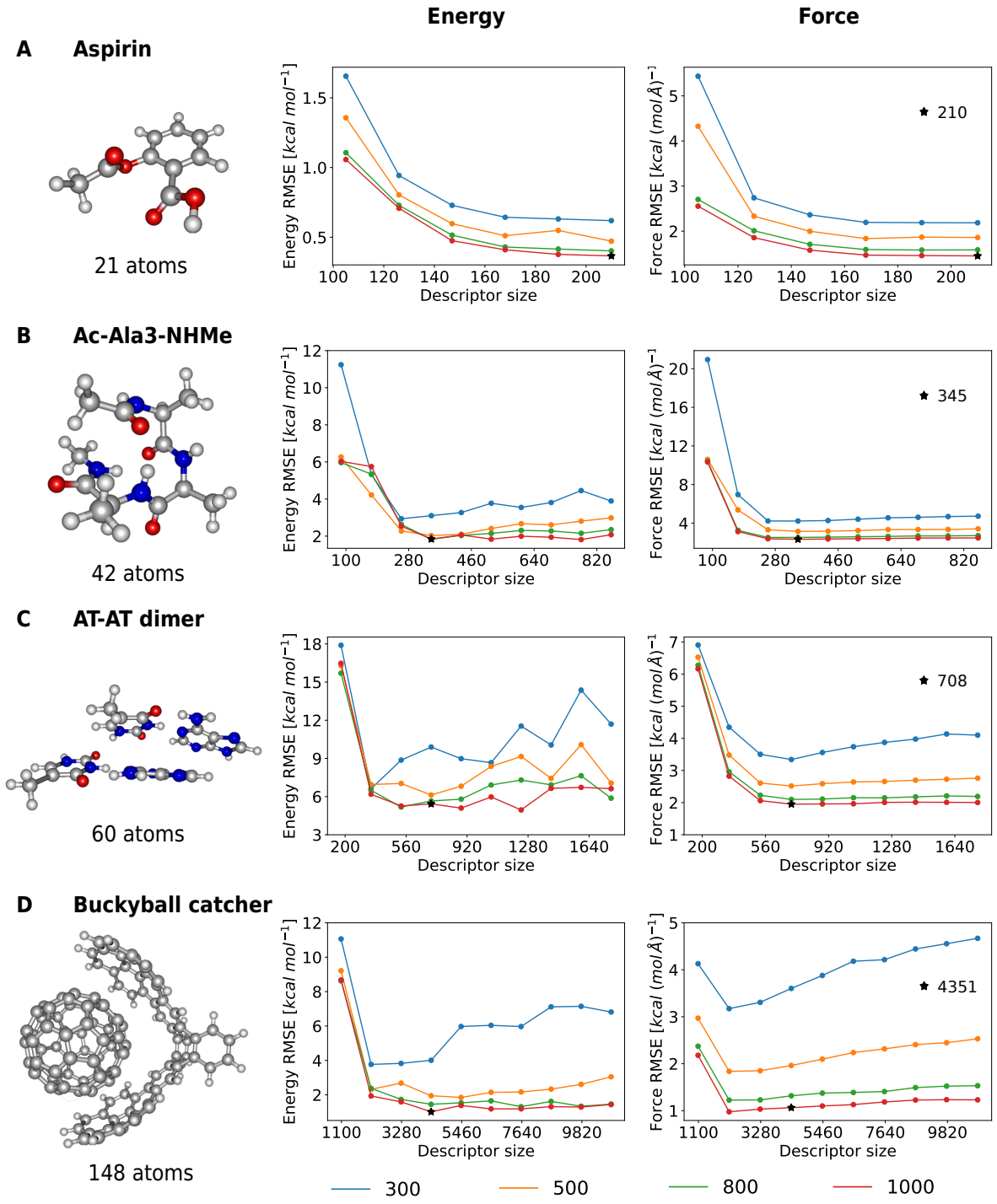

Figure 5.2: Energy (in kcal mol^{-1}) and force [in kcal (mol Å)$^{-1}$] RMSE as a function of the size of the descriptor. RMSEs of GDML models for **A**) aspirin, **B**) Ac-Ala3-NHMe, **C**) the AT-AT dimer, and **D**) the buckyball catcher trained with 300, 500, 800, and 1000 training. The black star on the curve of 1000 training points pinpoints the optimal descriptor size for each test molecule.

This results in the appearance of bumps on the energy prediction accuracy as a function of the size of the descriptor, like those observed in the energy RMSEs curves

of the AT-AT dimer in Fig. 5.2C. Hence, it seems that the RMSEs of the property used for training (force RMSEs in the case of sGDML) are always the best reference to optimize the descriptor, but one needs to be careful. Although for the peptide and the AT-AT dimer (Figs. 5.2B,C) this conclusion is true (the minimum force RMSEs for models trained on 1000 training points correspond to the ML_{P60} model, which also provides an accurate energy prediction), for optimizing the descriptor of the buckyball cátcher (Fig. 5.2D) force RMSEs are not reliable enough. Namely, while force RMSEs suggest that the descriptor of the ML_{P80} model (2176 features) is the optimal one, its energy RMSE is twice as large than that of the ML_{P60} model (which is the one with the lowest energy RMSE). Only considering force RMSEs to find the optimal descriptor leads to unphysical models whose force predictors are not related to the actual energy of the system. Thus, for selecting the optimal descriptor of large and flexible molecules one needs to consider the prediction accuracy of both forces and energy even if only one of these properties is used for training.

Finding an optimal (smaller) descriptor is not only advantageous to correctly describe complex PESs. It is also valuable to construct efficient MLFFs, which are crucial for practical studies of large and flexible molecules. In fact, the ML_{P60} models of Ac-Ala3-NHMe, the AT-AT dimer and the buckyball catcher trained on 1000 configurations are around 2 times faster than the models with the default global descriptor. All in all, one must consider optimizing the descriptor as an indispensable task for building ML models of large and flexible molecules.

5.4 Optimal Description of Interactions in Large and Flexible Molecules with ML Models

In Sec. 5.3, it was shown that the number of features in a typical global descriptor can be reduced down to 40 % of the original size, leading to an increase of the accuracy of the model. The improvement in accuracy when reducing the size of the descriptor should be a consequence of how well the ML model describes interactions within the system. Interactions in a molecule can be split into long- and short-range, and Sec. 4.3 introduced a classification of descriptors, into global and local ones, depending on how molecules are encoded, and which interactions can be captured. Although both classes of descriptors give a reasonably good description of different interactions and

have been applied to obtain reliable MLFFs for a myriad of different molecules and materials, each of them stem from certain approximations. On one hand, global descriptors assume that, for instance, all pairwise distances are relevant to model a given system, which might overdefine the learning problem and make harder the optimization of ML models for flexible and large molecules. On the other hand, local descriptors consider that all important interactions can be described within a cutoff radius that delimits the atomic environments. This might neglect relevant long-range interactions, crucial to describe the dynamics of complex systems. Therefore, it becomes essential to have a proper assessment of the reliability of a descriptor. For this reason, below the performance of different descriptors is assessed using as test systems the Ac-Ala3-NHMe molecule (42 atoms), the AT-AT dimer (60 atoms), and the buckyball catcher (148 atoms), for which a faithful description of both short- and long-range interactions is critical to obtain an accurate model. For each of these systems, three different sGDML models with 1000 configurations were trained using three different descriptors: i) the default global descriptor (all inverse pairwise distances, $\frac{1}{r}$), ii) a $\frac{1}{r}$ descriptor mimicking a local descriptor by removing all features involving distances greater than 5.0 Å (the typical value for cutoff radius in local descriptors) in at least one configuration in the dataset, and iii) the optimized $\frac{1}{r}$ descriptor obtained in Sec. 5.3 using the approach described in Sec. 5.2. Henceforward these models will be referred as ML_{global}, ML_{local}, and ML_{P60}, respectively.

This Section starts by analyzing how the different ML models predict the interactions between atoms. Namely, there is an analysis of the force prediction in terms of pairwise contributions in the molecule (F_{ij}, with i and j being any atom. The description of how the contributions were computed was given in the computational details in Sec. 5.1). Fig. 5.3 shows heatmaps of average F_{ij} values and of average interatomic distances for the three test systems. The averages were computed using 3000 configurations of the dataset for Ac-Ala3-NHMe and the AT-AT dimer, and 1000 configurations for the buckyball catcher.

For the ML_{global} models (containing 861 features for Ac-Ala3-NHMe, 1770 for the AT-AT dimer, and 10878 for the buckyball catcher) the contributions are more distributed among different pairs of atoms regardless of the distance between the atoms. This

allows the model to effectively capture long-range interactions but it might miss to optimally resolve all short-range ones. Conversely, the ML_{local} models (287 features for Ac-Ala3-NHMe, 296 for the AT-AT dimer, and 1626 for the buckyball catcher) only rely on the local environment of the molecule. This is confirmed by the contributions of the atoms to the force prediction of other atoms, which are directly related to the magnitude of the corresponding interatomic distances. Thus, the ML_{local} models offer a more adequate description of short-range interactions but completely neglect those interactions arising from distances greater than the selected cutoff. Finally, the ML_{P60} models (302 features for Ac-Ala3-NHMe, 708 for the AT-AT dimer, and 4351 for the buckyball catcher) offer a compromise between the ML_{global} and ML_{local} models.

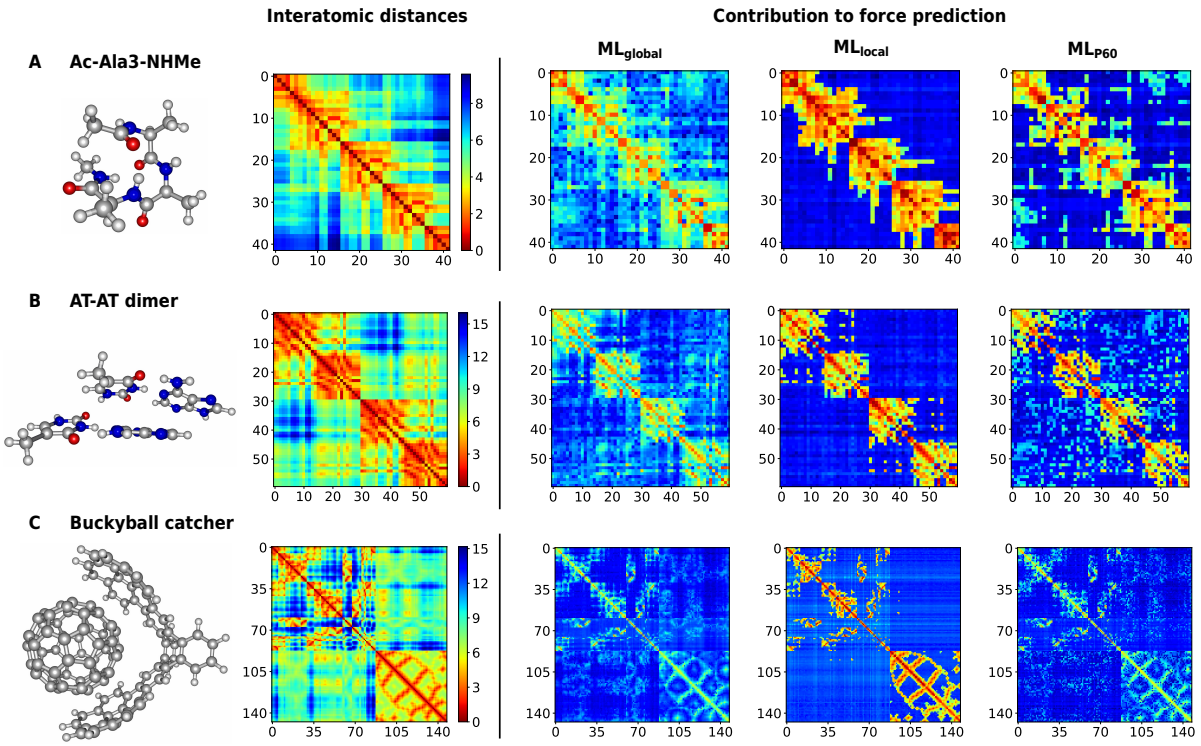

Figure 5.3: Heatmaps of average interatomic distances (in Å) and average contributions [in kcal (mol Å)$^{-1}$] of each atom to the force prediction of all atoms. Heatmaps of the averages from 3000 configurations of **A**) Ac-Ala3-NHMe, **B**) the AT-AT dimer, and 1000 configurations of **C**) the buckyball catcher. Each square in the heatmaps represents a given pair of atoms in the molecule (atom indices start from 0). The scale goes from red (short distances) to blue (long distances) for interatomic distances, while it goes from blue (small contributions) to red (big contributions) for the contributions to the force prediction.

Reduced ML_{P60} models achieve a satisfactory description of the local environment of the molecule but, at the same time, they keep the relevant information for describing

non-local interactions. The ML_{P60} models present the highest contributions to the force prediction from pairs of atoms that are separated by shorter distances. This is what one would expect from a physical point of view, since short-range interactions are those contributing the most to the total energy/forces of a given system. However, the models trained on the optimized descriptor also show important contributions from relevant long-range interactions. For instance, the interactions between one AT base pair (atoms 0 to 29 in Fig. 5.3B) and the other (atoms 30 to 59 in Fig. 5.3B) or the interactions between the catcher (atoms 0 to 79 in Fig. 5.3C) and the buckyball (atoms 79 to 147 in Fig. 5.3C).

The way interactions are described within an ML model must be related to the reliability of the predictions given by that model. In this regard, Fig. 5.4 extends the analysis of the ML_{global}, ML_{local}, and ML_{P60} models by comparing distributions of force errors for the three test molecules. Also, distribution of errors for GAP/SOAP models with a cutoff of 5 Å trained on 1000 configurations are shown for validating the results of the ML_{local} model against a state-of-the-art ML local model. The errors used for constructing the distributions were obtained from the force predictions of all configurations in the datasets that were not used for training or validation. On one hand, force error distributions show that the accuracy of the ML_{local} models, with respect the ML_{global} and ML_{P60} ones, is lower with increasing size and flexibility of the molecule. They start with an almost equal distribution as all other models for Ac-Ala3-NHMe (Fig. 5.4A) but show a wider and smaller peak than the ML_{global} and ML_{P60} models (i.e., the ML_{local} model has bigger errors) for the buckyball catcher (Fig. 5.4C). This conclusion is supported by the behavior of the GAP/SOAP model that presents considerably bigger errors than the other models for the buckyball catcher. On the other hand, the models trained with the optimized descriptor present almost the same population of small force errors [under an absolute value of 1.0 kcal (mol Å)$^{-1}$] as the ML_{global} model, while having a lower frequency of larger errors. For instance, errors above absolute values of 3.0 and 1.0 kcal (mol Å)$^{-1}$ for the AT-AT dimer (Fig. 5.4B) and the buckyball catcher (Fig. 5.4C), respectively, are more common with the ML_{global} model. Finally, although the force prediction accuracy of the ML_{local} models is comparable to that of ML_{P60} ones, the PESs reconstructed with the ML_{local} models might not be reliable enough. For example, as observed when optimizing the descriptor of the buckyball catcher in Fig. 5.2D, an insufficient number

of features in the descriptor produces a function that is accurate for predicting the forces but not the actual energy of the system. In other words, the integral of the force predictor does not yield an appropriate function for computing the PES of the molecule. Therefore, bigger errors in energy predictions can be expected from ML models trained on local descriptors. Training on both energies and forces would seem to be a possible solution to alleviate this problem. However, the resulting MLFF would produce better energy predictions at the cost of much less accurate force predictions. This assumption is confirmed by the distribution of force errors obtained with the predictions of the GAP/SOAP model (Fig. 5.4C), which are less reliable than those of the ML_{local} model.

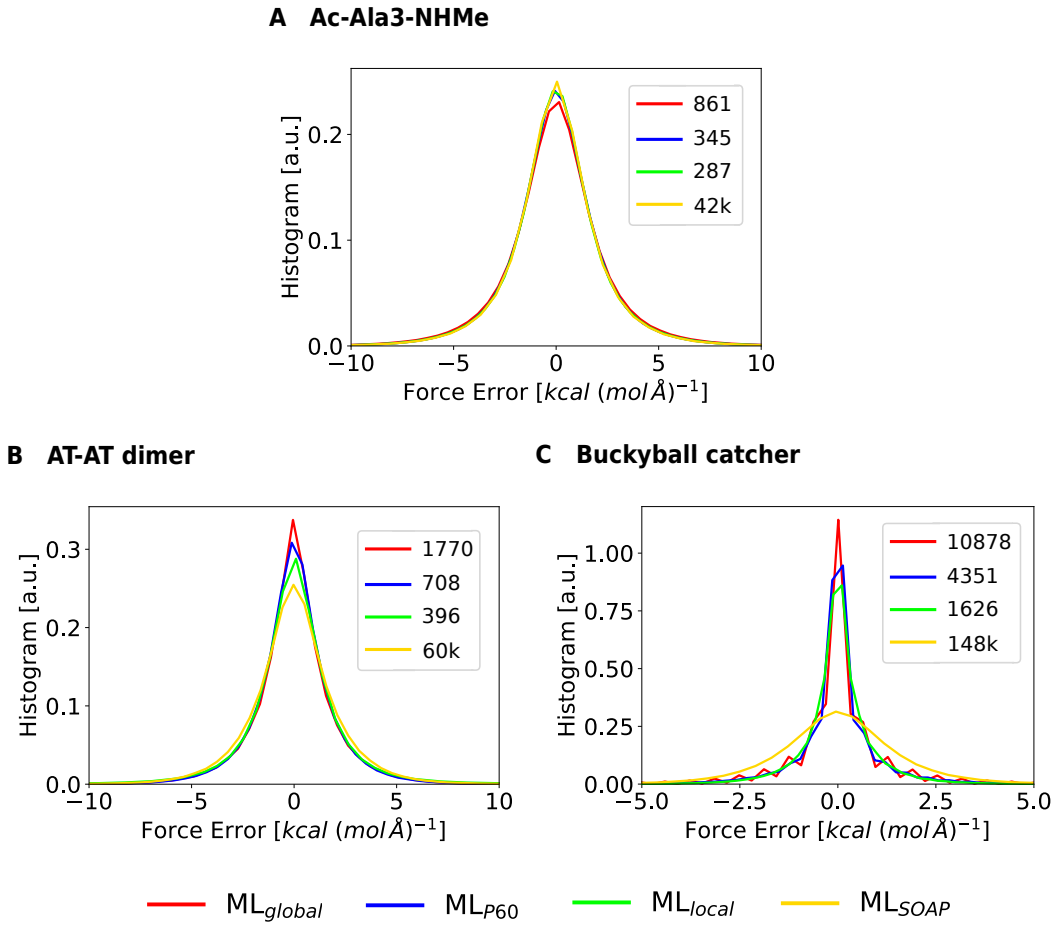

Figure 5.4: Distribution of force errors [in kcal (mol Å)$^{-1}$] of the ML_{global}, ML_{P60}, ML_{local} and GAP/SOAP (ML_{SOAP}) models. Distributions of **A**) Ac-Ala3-NHMe, **B**) the AT-AT dimer, and **C**) the buckyball catcher. The size of the descriptor of the models is given in the legend box of the figures on the right.

By now, the advantages of using an optimized descriptor have been consistently stressed in this Chapter: an increase in efficiency due to the reduction of the number of features and an improvement in accuracy because of an optimal description of all interactions in the molecule. However, little has been discussed about how the initial descriptor (i.e., the one to be optimized) should look like. One common assumption is to consider that everything within the local environment of an atom is relevant. Then, on top of this local information, one should only need to include the necessary features to account for non-local effects to optimally describe all interactions in the molecule. Nevertheless, in addition to local descriptors, such as SOAP, already containing a huge number of features that can compromise the efficiency of an MLFF (see descriptor sizes in Fig. 5.4), an inspection of the features removed for constructing the ML_{P60} models suggests that relying on a cutoff for constructing the local features is incorrect. Fig. 5.5 shows the example of three local features in Ac-Ala3-NHMe and the AT-AT dimer, whose information would be included in any local descriptor, that are irrelevant for constructing an accurate MLFF (see in Fig. 5.5 the blue squares corresponding to negligible contributions to the force predictions for the ML_{P60} models). These selected features correspond to interatomic distances between atoms that define bonded dihedral angles in the molecule. Also, all these distances have lengths shorter than 5 Å (the typical cutoff of local descriptors) for all configurations in the dataset as can be concluded from their contributions to the force predictions in the ML_{local} models. An interesting finding is that these interactions are part of different chemical environments. Namely, one can observe that these features belong to different pair of elements (H-H, C-N, C-O, C-H) and the atoms involved are embedded in different functional groups (e.g., methyl rotors, rings).

Summarizing the results of this section, while models trained on typical global or local descriptors are capable of achieving suitable predictions for big and flexible molecules, both descriptors are not optimal to accurately describe all interactions within such molecules. A global descriptor might overdefine the problem by including features that are not relevant, thus increasing the complexity of the learning task and compromising the final accuracy of the MLFF. This issue becomes quite prominent already when dealing with systems of 50 atoms, for which the size of the descriptor surpasses the thousand features. On the other hand, local descriptors neglect long-range effects, generating MLFFs with limited accuracy that might miss important interactions within

dimers or between different fragments of a system. To overcome these limitations, one should find a reduced set of features that allows an optimal description of all relevant interactions. Such an optimized set of features will depend on the molecule of interest and the information included in the dataset. The features that are removed from the original descriptor can be either local or non-local ones. Hence, although there is not only one strategy to obtain such an optimized descriptor, the results suggest starting from a global descriptor as the most advantageous one.

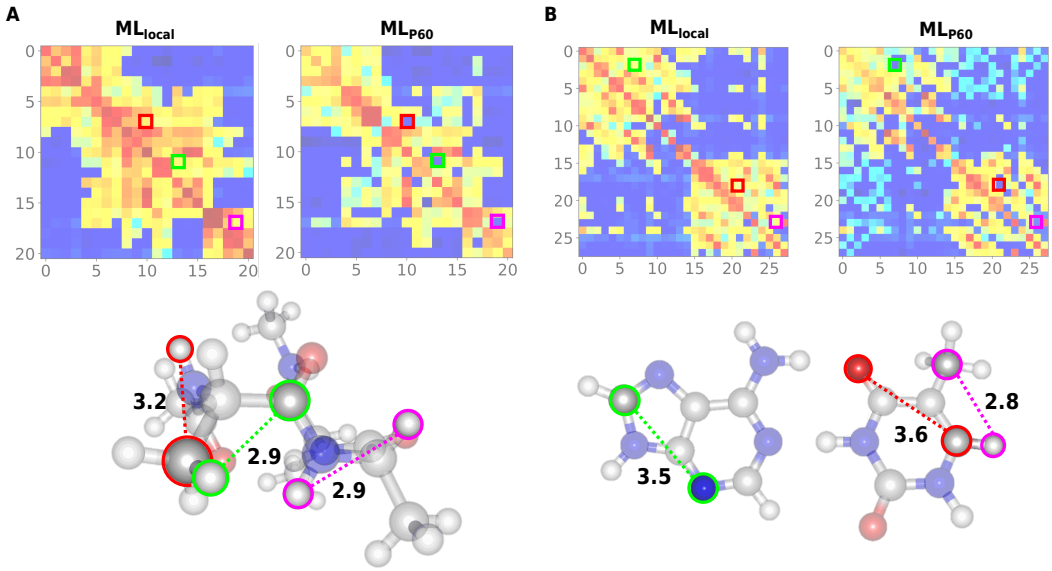

Figure 5.5: Examples of local features that are not important for constructing an accurate MLFF. A) Ac-Ala3-NHMe and **B)** the AT-AT dimer. Zoomed-in sections of the F_{ij} matrices (Figure 5.3) for the ML_{local} and ML_{P60} models show, inside colored frames (pink, red and green), the squares belonging to the selected features. Lengths of the distances (in Å) are given next to dotted lines connecting the pair of atoms involved in each feature. For the AT-AT dimer, only one AT base pair is shown for the sake of clarity.

5.5 Conclusions

In this chapter, an analysis of the challenges complex PESs pose in state-of-the-art descriptors was discussed. The results show that typical molecular representations used for constructing MLFFs are not optimal for such systems. Apart from having a huge number of features, descriptors do not properly resolve all relevant interactions in the molecules. Local descriptors (e.g., SOAP) ignore by construction all non-local interactions, leading to errors that are rather negligible for molecules of up to 50 atoms but that become unacceptable for larger molecules where interactions between different fragments play a significant role. On the other hand, despite global

descriptors (e.g., inverse pairwise distances) include information about all interactions in the molecule, they are prone to contain irrelevant or strongly correlated features that might cause a suboptimal description of both short- and long-range interactions. To overcome these limitations, MLFFs require descriptors that can be optimized to only include the features that are relevant for describing the interactions within each molecule. An optimized descriptor must optimally resolve all interactions in the molecule. This results in more robust ML models than those constructed with default global or local descriptors.

Overall, the performance of the optimized descriptor for Ac-Ala3-NHMe, the AT-AT dimer, and the buckyball catcher indicates that optimizing the descriptor must be an integral part of any MLFF. Our results clearly demonstrate that typically employed strategies to construct local descriptors, such as the selection of a cutoff radius, and global descriptors, which include all degrees of freedom, are insufficient when one needs to deal with complex PESs. An optimal descriptor should keep only some of the features included in local descriptors, together with the information about certain long distances (even beyond 10 Å) that are only present in global descriptors. Therefore, current approaches to construct descriptors could be replaced by more chemically inspired schemes, for example, based on concepts such as electrostatic potentials, electron affinities or molecular orbitals.

CHAPTER 6

Summary and Perspective

The use of machine learning (ML) in Computational Chemistry (CompChem) is commonplace nowadays because this synergy has strongly advanced the range of possible applications providing novel insights into chemical and physical processes in systems of practical interest.[91] In particular, ML force fields (FFs) have made possible a detailed reconstruction of potential-energy surfaces (PESs) with prediction errors under 1 kcal mol^{-1} with respect to the reference *ab initio* calculations at a fraction of computational cost. This achievement is, undoubtedly, a product of the joint effort of many groups and years of research. Despite the great success, MLFFs are still methods under development and, as such, certain challenges remain to be solved. The aim of the thesis was to investigate the capability of state-of-the-art MLFFs to provide simultaneously accurate and efficient models offering unprecedented insights into the dynamics of realistic molecules.

Firstly, a thorough exploration of PESs considering nuclear quantum effects (NQE) was done.[89] It was found that NQE induce the strengthening of molecular interactions that are ubiquitous in biological systems: the $E_{n\rightarrow\pi^*}$ interaction and the interactions related to methyl rotors. Furthermore, quantum fluctuations prevail over thermal contributions in the $E_{n\rightarrow\pi^*}$ interaction at the most important temperature range for the majority of practical applications (from 50 to 450 K). These results clearly show the necessity to go beyond single-point calculations for understanding the origins of interactions in molecular systems at realistic conditions and their effects on the measurable observables.

As a second important step, we analyzed the current limits of state-of-the-art MLFFs in reproducing complex PESs on the example of transition paths of flexible molecules.[90] We found that methods based on local descriptors (e.g., BPNN and GAP/SOAP) saturate quickly with the increase in the number of training points, while

not achieving the desired prediction accuracy. The MLFFs based on global descriptors (e.g., sGDML) require careful model selection procedures when the reference datasets consist of several disconnected parts of the PES. The main challenges here are the inability of simple training schemes to select appropriate training datasets in unbalanced reference data and the limitations of the standard molecular descriptors to pick up the features describing the complex geometric transformations in flexible molecules. Finally, end-to-end NNs (e.g., SchNet) do not reproduce all relevant parts of the PES with equal accuracy required, for instance, to compute transition rates. Moreover, NNs require a larger amount of training data compared to kernel methods, which can result in a high computational cost when generating reference datasets.

The revealed challenges inspired us to study in detail the role of a descriptor in the performance of an MLFF. Our results showed that the commonly employed strategies followed to construct both local and global descriptors need to be improved. On one hand, some features included in local descriptors are not required to accurately reconstruct a PES, while certain long distances (even beyond 10 Å) can be crucial. This suggests that the selection of a predefined cutoff radius to account for local environments should be revised. On the other hand, global descriptors can be reduced down to a 40 % of their original size for molecules containing 50 to 150 atoms without the loose of accuracy. Therefore, optimizing descriptors opens new avenues for building simultaneously more efficient and accurate ML models, which is crucial for large-scale applications.

In summary, the work presented in this thesis highlighted the potential of MLFFs to provide insights into molecular systems, offering knowledge on new phenomena in chemistry and physics. Also, it disclosed the current limitations of state-of-the-art ML methods that need to be overcome for extending the applicability of MLFFs to obtain novel insights of more realistic systems. Furthermore, it proposed the optimization of the description of short- and long-range interactions within a ML model as a valuable step towards more efficient and more reliable MLFFs of large and flexible molecules.

Overall, the results of this thesis suggest that the full workflow for building ML models, from training set selection to descriptor construction, still needs significant elaboration when dealing with complex systems. Appropriate optimization procedures should be implemented at the different stages in the construction of ML models. Training sets and descriptors tailored according to the problem at hand and the available data are

required to obtain more reliable MLFFs. In this regard, it is precisely the combination of supervised and unsupervised ML methods that offers a fully data-driven approach for this purpose. Such an approach would demand the integration of different techniques, like multiple-kernel methods,[143] active learning,[59,61,257] classification[258,259] and clustering[244,260,261] algorithms. Also, the difficulty of learning complex PESs, as revealed in this thesis, suggests that finding a single optimal model for the entire PES might not be possible without a drastic loss of efficiency of MLFFs. Therefore, we might need, for instance, to consider moving from learning the entire PES in its complexity within a single task to the employment of multiple local models for different parts of the PES and then combining them into a global FF.

In addition, descriptors constructed following more physically and chemically inspired approaches are needed to make more flexible MLFFs that can easily adapt to the task at hand. These descriptors must contain features based on concepts such as electrostatic potentials, electron affinities or molecular orbitals, and should be able to successfully describe local chemical bonding, as well as long-range electrostatics, polarization, and vdW dispersion interactions without incurring in excessive computational costs. Therefore, systematic approaches should be developed for finding the relevant chemical/physical patterns that define compositional (chemical arrangement of atoms in a molecule) and configurational (physical arrangement of atoms in space) degrees of freedom.

While further research is certainly needed, it is already clear that the combination of traditional CompChem and ML methods brings new horizons in atomistic modelling. By generating an unprecedented impact on our knowledge of physical and chemical phenomena, this synergy will ultimately help to solve multiple problems in different areas, including medicine, materials design, pharmacology, energy production, environmental sciences, among others

Appendices

Parts of Sec. A1 have been reproduced from the supporting material of Ref. 90

A1 Energetics and Geometric Data along the Transition Path of Glycine and Azobenzene

Table A.1: Angles (in º), and relative energies (ΔE and Δ(E+vdW)) (in kcal mol^{-1}) through the path of glycine at the PBE+TS level of theory.

Structure	τ_1	τ_2	Δ(E+vdW)	ΔE
Glycine Ip	180.00	0.00	0.00	0.00
Step1	174.89	-5.19	0.03	0.02
Step2	169.61	-10.56	0.10	0.09
Step3	164.16	-16.06	0.22	0.21
Step4	158.57	-21.67	0.38	0.37
Step5	152.84	-27.35	0.57	0.58
Step6	147.01	-33.07	0.80	0.81
Step7	141.10	-38.79	1.04	1.05
Step8	135.13	-44.48	1.28	1.30
Step9	129.16	-50.09	1.50	1.52
Step10	123.21	-55.59	1.67	1.70
Step11	117.35	-60.94	1.80	1.84
Step12	111.58	-66.18	1.87	1.94
Step13	105.88	-71.38	1.91	2.00
Step14	100.23	-76.63	1.93	2.03
Step15	94.62	-82.02	1.93	2.05
Step16	89.01	-87.60	1.94	2.06
Step17	83.38	-93.41	1.95	2.07
Step18	77.70	-99.46	1.96	2.08

Table A.1 (continued)

Step19	71.90	-105.77	1.98	2.08
Step20	66.00	-112.31	1.99	2.07
Step21	60.00	-118.99	1.97	2.05
Step22	53.93	-125.66	1.94	2.01
Step23	47.77	-132.31	1.88	1.95
Step24	41.58	-138.87	1.81	1.88
Step25	35.39	-145.31	1.73	1.80
Step26	29.22	-151.60	1.66	1.72
Step27	23.11	-157.70	1.60	1.66
Step28	17.10	-163.60	1.55	1.61
Step29	11.22	-169.29	1.51	1.56
Step30	5.50	-174.76	1.48	1.53
Glycine IIIp	0.00	-180.00	1.47	1.52

Table A.2: Angles (in º), and relative energies (ΔE and Δ(E+vdW)) (in kcal mol^{-1}) through the rotation path of azobenzene at the PBE+TS level of theory.

Structure	θ	θ'	ϕ	Δ(E+vdW)	ΔE
Cisazobenzene	123.43	123.43	8.04	0.0	0.0
Step1	122.56	122.56	18.79	0.8	0.8
Step2	122.06	122.06	29.54	2.5	2.3
Step3	121.56	121.56	40.29	5.4	5.0
Step4	121.06	121.06	51.04	9.5	9.0
Step5	120.56	120.56	61.79	14.9	14.3
Step6	120.06	120.06	72.54	21.5	20.7
Step7	119.56	119.56	83.29	29.0	28.1
Step8	119.06	119.06	94.04	30.2	29.5
Step9	118.56	118.56	104.79	22.0	21.2
Step10	118.06	118.06	115.54	14.1	13.2
Step11	117.56	117.56	126.29	7.0	5.8
Step12	117.06	117.06	137.04	0.7	-0.7
Step13	116.56	116.56	147.79	-4.3	-5.9
Step14	116.06	116.06	158.54	-8.1	-9.7
Step15	115.56	115.56	169.29	-10.3	-12.0
Transazobenzene	114.88	114.88	180.00	-11.6	-13.3

Table A.3: Angles (in º), and relative energies (ΔE and Δ(E+vdW)) (in kcal mol^{-1}) through the inversion path of azobenzene at the PBE+TS.

Structure	θ	θ'	ϕ	Δ(E+vdW)	ΔE
Cisazobenzene	123.43	123.43	10.25	0.0	0.0
Step1	130.67	122.56	7.54	1.5	1.3
Step2	138.28	122.06	7.04	5.1	4.7
Step3	145.89	121.56	6.54	9.9	9.2
Step4	153.50	121.06	6.04	15.0	14.2
Step5	161.11	120.56	5.54	20.2	19.2
Step6	168.72	120.06	5.04	24.4	23.2
Step7	176.33	119.56	4.54	27.0	25.7
Step8	183.94	119.06	4.04	27.4	26.0
Step9	191.55	118.56	3.54	25.5	24.0
Step10	199.16	118.06	3.04	21.5	20.0
Step11	206.77	117.56	2.54	16.0	14.4
Step12	214.38	117.06	2.04	9.6	8.0
Step13	221.99	116.56	1.54	2.7	1.1
Step14	229.60	116.06	1.04	-4.0	-5.7
Step15	237.21	115.56	0.54	-9.1	-10.8
Transazobenzene	244.83 (115.17)	114.88	0.00 (180.00)	-11.6	-13.3

Table A.4: Angles (in º), and relative energies (ΔE and Δ(E+vdW)) (in kcal mol^{-1}) through the rotation assisted by inversion path of azobenzene at the PBE+TS.

Structure	θ	θ'	ϕ	Δ(E+vdW)	ΔE
Cisazobenzene	123.43	123.43	10.25	0.0	0.0
Step1	130.67	122.56	18.79	1.8	1.5
Step2	138.28	122.06	29.54	6.7	6.2
Step3	145.89	121.56	40.29	13.6	12.8
Step4	153.50	121.06	51.04	20.4	19.5
Step5	161.11	120.56	61.79	25.2	24.1
Step6	168.72	120.06	72.54	27.2	25.9
Step7	176.33	119.56	83.29	27.5	26.1
Step8	183.94	119.06	94.04	27.4	26.1
Step9	191.55	118.56	104.79	27.6	26.3
Step10	199.16	118.06	115.54	26.6	25.3
Step11	206.77	117.56	126.29	22.5	21.1
Step12	214.38	117.06	137.04	15.2	13.7
Step13	221.99	116.56	147.79	6.4	4.8
Step14	229.60	116.06	158.54	-2.1	-3.7
Step15	237.21	115.56	169.29	-8.5	-10.2
Transazobenzene	244.83 (115.17)	114.88	180.00	-11.6	-13.3

Table A.5: Angles (in º), and relative energies (ΔE and Δ(E+vdW)) (in kcal mol^{-1}) through the optimized rotation path of azobenzene at the PBE+TS.

Structure	θ	θ'	ϕ	b	a	Δ(E+vdW)	ΔE
Cisazobenzene	123.43	123.43	10.25	1.25	1.43	0.0	0.0
Step1	123.06	123.06	18.79	1.25	1.42	0.4	0.3
Step2	123.06	123.06	29.54	1.26	1.42	1.9	1.6
Step3	124.00	124.00	40.29	1.26	1.41	4.6	4.1
Step4	124.00	124.00	51.04	1.26	1.41	8.3	7.7
Step5	124.00	124.00	61.79	1.27	1.39	13.0	12.3
Step6	125.00	125.00	72.54	1.28	1.38	18.5	17.6
Step7	125.00	127.00	83.29	1.28	1.37	24.3	23.4
Step8	125.00	125.00	88.67	1.28	1.37	26.1	25.2
Step9	123.00	123.00	94.04	1.27	1.36	25.6	24.8
Step10	121.50	122.00	99.42	1.27	1.36	22.3	21.5
Step11	120.00	121.00	104.79	1.27	1.36	18.7	17.9
Step12	119.00	119.00	115.54	1.27	1.37	11.8	10.8
Step13	118.00	118.00	126.29	1.27	1.38	5.3	4.1
Step14	117.00	117.00	137.04	1.27	1.39	-0.5	-1.9
Step15	116.00	116.00	147.79	1.27	1.40	-5.3	-6.8
Step16	115.17	115.17	158.54	1.27	1.41	-8.8	-10.4
Step17	115.17	115.17	169.29	1.26	1.41	-10.9	-12.6
Transazobenzene	114.88	114.88	180.00	1.26	1.42	-11.6	-13.3

Bibliography

[1] K. N. Houk and P. H. Y. Cheong, *Nature* **455**, 309–313 (2008).

[2] C. Wentrup, *Angew. Chemie - Int. Ed.* **57**, 11508–11521 (2018).

[3] M. Bachmann and W. Janke, *Phys. Rev. Lett.* **95**, 058102 (2005).

[4] D. Chakraborty and D. J. Wales, *J. Chem. Phys.* **150**, 125101 (2019).

[5] J. B. Cook, T. C. Lin, H.-S. Kim, A. Siordia, B. S. Dunn, and S. H. Tolbert, *ACS Nano* **13**, 1223–1231 (2019).

[6] Y. Zhang, J. B. Chou, J. Li, H. Li, Q. Du, A. Yadav, S. Zhou, M. Y. Shalaginov, Z. Fang, H. Zhong, C. Roberts, P. Robinson, B. Bohlin, C. Ríos, H. Lin, M. Kang, T. Gu, J. Warner, V. Liberman, K. Richardson, and J. Hu, *Nat. Commun.* **10**, 4279 (2019).

[7] C. M. Dobson, *Nature* **426**, 884–890 (2003).

[8] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17845–17850 (2012).

[9] J. B. Stiller, S. J. Kerns, M. Hoemberger, Y.-J. Cho, R. Otten, M. F. Hagan, and D. Kern, *Nat. Catal.* **2**, 726–734 (2019).

[10] D. J. Wales, *Annu. Rev. Phys. Chem.* **69**, 401–425 (2018).

[11] L. Bonati and M. Parrinello, *Phys. Rev. Lett.* **121**, 265701 (2018).

[12] F. Therrien, E. B. Jones, and V. Stevanović, *Appl. Phys. Rev.* **8**, 031310 (2021).

[13] J. L. Mancuso, A. M. Mroz, K. N. Le, and C. H. Hendon, *Chem. Rev.* **120**, 8641–8715 (2020).

[14] A. Von Hoegen, R. Mankowsky, M. Fechner, M. Först, and A. Cavalleri, *Nature* **555**, 79–82 (2018).

[15] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, *Chem. Rev.* **121**, 10142–10186 (2021).

[16] A. V. Akimov and O. V. Prezhdo, *Chem. Rev.* **115**, 5797–5890 (2015).

[17] J. A. Harrison, J. D. Schall, S. Maskey, P. T. Mikulski, M. T. Knippenberg, and B. H. Morrow, *Appl. Phys. Rev.* **5**, 031104 (2018).

[18] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, *Rev. Mod. Phys.* **73**, 33 (2001).

[19] J. A. Pople, M. Head-Gordon, and K. Raghavachari, *J. Chem. Phys.* **87**, 5968–5975 (1987).

[20] G. D. Purvis and R. J. Bartlett, *J. Chem. Phys.* **76**, 1910 (1982).

[21] R. O. Jones, *Rev. Mod. Phys.* **87**, 897 (2015).

[22] A. J. Cohen, P. Mori-Sánchez, and W. Yang, *Chem. Rev.* **112**, 289–320 (2012).

[23] D. J. Cole and N. D. M. Hine, *J. Phys. Condens. Matter* **28**, 393001 (2016).

[24] A. Jain, Y. Shin, and K. A. Persson, *Nat. Rev. Mater.* **1**, 15004 (2016).

[25] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, *J. Comput. Chem.* **30**, 1545–1614 (2009).

[26] P. K. Weiner and P. A. Kollman, *J. Comput. Chem.* **2**, 287–303 (1981).

[27] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 198–210 (2013).

[28] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. Van Gunsteren, *Eur. Biophys. J.* **40**, 843 (2011).

[29] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, *J. Comput. Chem.* **25**, 1656–1676 (2004).

[30] X. Daura, A. E. Mark, and W. F. Van Gunsteren, *J. Comput. Chem.* **19**, 535–547 (1998).

[31] W. L. Jorgensen, J. D. Madura, and C. J. Swenson, *J. Am. Chem. Soc.* **106**, 6638–6646 (1984).

[32] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).

[33] P. E. M. Lopes, O. Guvench, and A. D. Mackerell, in *Methods in Molecular Biology*, edited by A. Kukol, Vol. 1215 ,Humana Press, New York, NY, 2015, pp. 47–71.

[34] P. S. Nerenberg and T. Head-Gordon, *Curr. Opin. Struct. Biol.* **49**, 129–138 (2018).

[35] Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J. P. Piquemal, and P. Ren, *Annu. Rev. Biophys.* **48**, 371–394 (2019).

[36] H. E. Sauceda, M. Gastegger, S. Chmiela, K.-R. Müller, and A. Tkatchenko, *J.*

Chem. Phys. **153**, 124109 (2020).

[37] A. Mardt, L. Pasquali, H. Wu, and F. Noé, *Nat. Commun.* **9**, 5 (2018).

[38] J. Behler, S. Lorenz, and K. Reuter, *J. Chem. Phys.* **127**, 014705 (2007).

[39] Y. Zhang, C. Hu, and B. Jiang, *J. Phys. Chem. Lett.* **10**, 4962–4967 (2019).

[40] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, *Phys. Rev. Lett.* **120**, 143001 (2018).

[41] L. Zhang, J. Han, H. Wang, W. A. Saidi, R. Car, and E. Weinan, *Adv. Neural Inf. Process. Syst.* 4436– 4446 (2018).

[42] N. Lubbers, J. S. Smith, and K. Barros, *J. Chem. Phys.* **148**, 241715 (2018).

[43] J. S. Smith, O. Isayev, and A. E. Roitberg, *Chem. Sci.* **8**, 3192–3203 (2017).

[44] K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre, and J. Parkhill, *Chem. Sci.* **9**, 2261–2269 (2018).

[45] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, *Sci. Adv.* **3**, 1603015 (2017).

[46] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, *Nat. Commun.* **9**, 3887 (2018).

[47] H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, *J. Chem. Phys.* **150**, 114102 (2019).

[48] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, *Comput. Phys. Commun.* **240**, 38–45 (2019).

[49] J. Behler, *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).

[50] V. Botu and R. Ramprasad, *Phys. Rev. B* **92**, 094306 (2015).

[51] A. S. Christensen, F. A. Faber, and O. A. Von Lilienfeld, *J. Chem. Phys.* **150**, 064105 (2019).

[52] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. A. Von Lilienfeld, *J. Chem. Phys.* **152**, 044107 (2020).

[53] M. Rupp, R. Ramakrishnan, and O. A. Von Lilienfeld, *J. Phys. Chem. Lett.* **6**, 3309–3313 (2015).

[54] A. Glielmo, P. Sollich, and A. De Vita, *Phys. Rev. B* **95**, 214302 (2017).

[55] M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat, and L. Thiry, *J. Chem. Phys.* **148**, 241732 (2018).

[56] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).

[57] A. P. Bartók and G. Csányi, *Int. J. Quantum Chem.* **115**, 1051–1057 (2015).

[58] A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).

[59] Z. Li, J. R. Kermode, and A. De Vita, *Phys. Rev. Lett.* **114**, 096405 (2015).

[60] J. Behler, *J. Chem. Phys.* **134**, 074106 (2011).

[61] E. V. Podryabinkin and A. V. Shapeev, *Comput. Mater. Sci.* **140**, 171–180 (2017).

[62] J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).

[63] K. Jose, N. Artrith, and J. Behler, *J. Chem. Phys.* **136**, 194111 (2012).

[64] J. Behler, *J. Chem. Phys.* **145**, 170901 (2016).

[65] M. Gastegger, J. Behler, and P. Marquetand, *Chem. Sci.* **8**, 6924–6935 (2017).

[66] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K.-R. Müller, *J. Chem. Phys.* **148**, 241722 (2018).

[67] O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).

[68] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Nature* **559**, 547–555 (2018).

[69] V. L. Deringer, M. A. Caro, and G. Csányi, *Adv. Mater.* **31**, 1902765 (2019).

[70] U. Rivero, O. T. Unke, M. Meuwly, and S. Willitsch, *J. Chem. Phys.* **151**, 104301 (2019).

[71] T. Morawietz, A. Singraber, C. Dellago, and J. Behler, *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8368–8373 (2016).

[72] M. F. C. Andrade, H. Y. Ko, L. Zhang, R. Car, and A. Selloni, *Chem. Sci.* **11**, 2335–2341 (2020).

[73] V. L. Deringer and G. Csányi, *Phys. Rev. B* **95**, 094203 (2017).

[74] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, *Phys. Status Solidi Basic Res.* **245**, 2618–2629 (2008).

[75] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, *Phys. Rev. X* **8**, 041048 (2018).

[76] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, *J. Phys. Chem. Lett.* **9**, 2879–2885 (2018).

[77] S. Brickel, A. K. Das, O. T. Unke, H. T. Turan, and M. Meuwly, *Electron. Struct.* **1**, 024002 (2019).

[78] D. Lu, J. Li, and H. Guo, *CCS Chem.* **2**, 882–894 (2020).

[79] B. C. Sweeny, H. Pan, A. Kassem, J. C. Sawyer, S. G. Ard, N. S. Shuman, A. A. Viggiano, S. Brickel, O. T. Unke, M. Upadhyay, and M. Meuwly, *Phys. Chem.*

Chem. Phys. **22**, 8913–8923 (2020).

[80] S. Käser, O. T. Unke, and M. Meuwly, *J. Chem. Phys.* **152**, 214304 (2020).

[81] Q. Liu, X. Zhou, L. Zhou, Y. Zhang, X. Luo, H. Guo, and B. Jiang, *J. Phys. Chem. C* **122**, 1761–1769 (2018).

[82] M. Hellström, M. Ceriotti, and J. Behler, *J. Phys. Chem. B* **122**, 10158–10171 (2018).

[83] W. K. Chen, X. Y. Liu, W. H. Fang, P. O. Dral, and G. Cui, *J. Phys. Chem. Lett.* **9**, 6702–6708 (2018).

[84] J. Westermayr, M. Gastegger, M. F. S. J. Menger, S. Mai, L. González, and P. Marquetand, *Chem. Sci.* **10**, 8100–8107 (2019).

[85] J. Westermayr, M. Gastegger, and P. Marquetand, *J. Phys. Chem. Lett.* **11**, 3828–3834 (2020).

[86] N. Raimbault, A. Grisafi, M. Ceriotti, and M. Rossi, *New J. Phys.* **21**, 105001 (2019).

[87] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, *Nature* **589**, 59–64 (2021).

[88] F. Noé, S. Olsson, J. Köhler, and H. Wu, *Science* **365**, aaw1147 (2019).

[89] H. E. Sauceda, V. Vassilev-Galindo, S. Chmiela, K.-R. Müller, and A. Tkatchenko, *Nat. Commun.* **12**, 442 (2021).

[90] V. Vassilev-Galindo, G. Fonseca, I. Poltavsky, and A. Tkatchenko, *J. Chem. Phys.* **154**, 094119 (2021).

[91] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, and A. Tkatchenko, *Chem. Rev.* **121**, 9816–9872 (2021).

[92] P. Atkins and R. Friedman, *Molecular Quantum Mechanics*, 4th edition ,Oxford University Press, Oxford, 2005.

[93] I. N. Levine, *Quantum Chemistry*, 7th edition ,Pearson Advanced Chemistry Series, Pearson Education, Upper Saddle River, NJ, 2014.

[94] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry* ,McGraw-Hill, New York, 1989.

[95] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd edition ,John Wiley & Sons, Chichester, 2004.

[96] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* ,International Series of Monographs on Chemistry, Oxford University Press, New York, 1994.

[97] K. A. Peterson and T. H. Dunning, *J. Chem. Phys.* **117**, 10548 (2002).

[98] W. J. Hehre, R. F. Stewart, and J. A. Pople, *J. Chem. Phys.* **51**, 2657 (1969).

[99] A. Schäfer, H. Horn, and R. Ahlrichs, *J. Chem. Phys.* **97**, 2571 (1992).

[100] E. Van Lenthe and E. J. Baerends, *J. Comput. Chem.* **24**, 1142–1156 (2003).

[101] J. C. Slater, *Adv. Quantum Chem.* **1**, 35–38 (1964).

[102] A. H. MacDonald, W. E. Picket, and D. D. Koelling, *J. Phys. C Solid State Phys.* **13**, 2675 (1980).

[103] S. G. Louie, K. M. Ho, and M. L. Cohen, *Phys. Rev. B* **19**, 1774 (1979).

[104] D. R. Hartree and W. Hartree, *Proc. R. Soc. London. Ser. A - Math. Phys. Sci.* **150**, 9–33 (1935).

[105] J. C. Slater, *Phys. Rev.* **81**, 385 (1951).

[106] V. Fock, *Zeitschrift Für Phys.* **61**, 126–148 (1930).

[107] R. J. Bartlett and M. Musiał, *Rev. Mod. Phys.* **79**, 291 (2007).

[108] W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A113 (1965).

[109] C. Adamo and V. Barone, *J. Chem. Phys.* **110**, 6158–6170 (1999).

[110] U. C. Singh and P. A. Kollman, *J. Comput. Chem.* **5**, 129–145 (1984).

[111] J. W. Storer, D. J. Giesen, C. J. Cramer, and D. G. Truhlar, *J. Comput. Aided. Mol. Des.* **9**, 87–110 (1995).

[112] E. L. Mehler and T. Solmajer, *Protein Eng. Des. Sel.* **4**, 903–910 (1991).

[113] J. Chen and T. J. Martínez, *Chem. Phys. Lett.* **438**, 315–320 (2007).

[114] P. P. Poier and F. Jensen, *J. Chem. Theory Comput.* **15**, 3093–3107 (2019).

[115] A. K. Rappé and W. A. Goddard, *J. Phys. Chem.* **95**, 3358–3363 (1991).

[116] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell, *J. Comput. Chem.* **31**, 671–690 (2010).

[117] J. P. Piquemal and K. D. Jordan, *J. Chem. Phys.* **147**, 161401 (2017).

[118] I. Poltavsky, L. Zheng, M. Mortazavi, and A. Tkatchenko, *J. Chem. Phys.* **148**, 204707 (2018).

[119] M. Ceriotti, W. Fang, P. G. Kusalik, R. H. McKenzie, A. Michaelides, M. A. Morales, and T. E. Markland, *Chem. Rev.* **116**, 7529–7550 (2016).

[120] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Commun. ACM* **60**, 84–90 (2017).

[121] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, *Int. J. Comput. Vis.* **115**, 211–252 (2015).

[122] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* ,2015, pp. 1–9.

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Adv. Neural Inf. Process. Syst.* **31**, 5998–6008 (2017).

[124] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *ArXiv Prepr.* arXiv:1301.3781 (2013).

[125] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).

[126] D. M. Blei, A. Y. Ng, and M. I. Jordan, *J. Mach. Learn. Res.* **3**, 993–1022 (2003).

[127] C. E. Rasmussen, in *Advanced Lectures on Machine Learning*, edited by O. Bousquet, U. von Luxburg, and G. Rätsch ,Springer, Berlin, Heidelberg, 2004, pp. 63–71.

[128] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, 2nd edition ,Springer, New York, 2009.

[129] V. N. Vapnik, *The Nature of Statistical Learning Theory* ,Springer-Verlag, New York, 1995.

[130] S. Geman, E. Bienenstock, and R. Doursat, *Neural Comput.* **4**, 1–58 (1992).

[131] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301 (2012).

[132] F. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento, *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).

[133] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).

[134] F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld, *J. Chem. Phys.* **148**, 241717 (2018).

[135] H. Huo and M. Rupp, *ArXiv Prepr.* arXiv:1704.06439 (2017).

[136] J. Nigam, S. Pozdnyakov, and M. Ceriotti, *J. Chem. Phys.* **153**, 121101 (2020).

[137] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, *Phys. Rev. B* **89**, 205118 (2014).

[138] R. Drautz, *Phys. Rev. B* **99**, 014104 (2019).

[139] W. Pronobis, A. Tkatchenko, and K.-R. Müller, *J. Chem. Theory Comput.* **14**, 2991–3003 (2018).

[140] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von

Lilienfeld, A. Tkatchenko, and K.-R. Müller, *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).

[141] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).

[142] N. Artrith, A. Urban, and G. Ceder, *Phys. Rev. B* **96**, 014112 (2017).

[143] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, *Sci. Adv.* **3**, 1701816 (2017).

[144] K. Yao, J. E. Herr, and J. Parkhill, *J. Chem. Phys.* **146**, 014106 (2017).

[145] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. Von Lilienfeld, *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).

[146] A. Glielmo, C. Zeni, and A. De Vita, *Phys. Rev. B* **97**, 184307 (2018).

[147] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, *Phys. Rev. Lett.* **120**, 036002 (2018).

[148] Y. H. Tang, D. Zhang, and G. E. Karniadakis, *J. Chem. Phys.* **148**, 034101 (2018).

[149] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Nature* **323**, 533–536 (1986).

[150] Y. Lecun, *Proc. Cogn.* **85**, 599–604 (1985).

[151] C. M. Bishop, *Neural Networks for Pattern Recognition* ,Oxford University Press, New York, 1995.

[152] Y. Lecun, L. Bottou, G. B. Orr, and K.-R. Müller, in *Neural Networks: Tricks of the Trade*, edited by G. Montavon, G. B. Orr, and K.-R. Müller ,Springer-Verlag, Berlin, 2012, pp. 9–48.

[153] P. O. Dral, A. Owens, S. N. Yurchenko, and W. Thiel, *J. Chem. Phys.* **146**, 244108 (2017).

[154] K. Shao, J. Chen, Z. Zhao, and D. H. Zhang, *J. Chem. Phys.* **145**, (2016).

[155] A. J. Ballard, R. Das, S. Martiniani, D. Mehta, L. Sagun, J. D. Stevenson, and D. J. Wales, *Phys. Chem. Chem. Phys.* **19**, 12585–12603 (2017).

[156] B. Jiang, J. Li, and H. Guo, *Int. Rev. Phys. Chem.* **35**, 479–506 (2016).

[157] B. Kolb, B. Zhao, J. Li, B. Jiang, and H. Guo, *J. Chem. Phys.* **144**, (2016).

[158] J. Li, K. Song, and J. Behler, *Phys. Chem. Chem. Phys.* **21**, 9672–9682 (2019).

[159] B. Fu and D. H. Zhang, *J. Chem. Theory Comput.* **14**, 2289–2303 (2018).

[160] J. Chen, X. Xu, and D. H. Zhang, *J. Chem. Phys.* **138**, 154301 (2013).

[161] D. F. R. Brown, M. N. Gibbs, and D. C. Clary, *J. Chem. Phys.* **105**, 7597–7604

(1996).

[162] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, *J. Chem. Phys.* **148**, 241709 (2018).

[163] H. D. Merchant, K. K. Srvastava, and H. D. Pandey, *C R C Crit. Rev. Solid State Sci.* **3**, 451–504 (1973).

[164] V. Kirchner, H. Heinke, D. Hommel, J. Z. Domagala, and M. Leszczynski, *Appl. Phys. Lett.* **77**, 1434 (2000).

[165] P. Hermet, M. M. Koza, C. Ritter, C. Reibel, and R. Viennois, *RSC Adv.* **5**, 87118–87131 (2015).

[166] D. S. Kim, O. Hellman, J. Herriman, H. L. Smith, J. Y. Y. Lin, N. Shulumba, J. L. Niedziela, C. W. Li, D. L. Abernathy, and B. Fultz, *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1992–1997 (2018).

[167] T. E. Markland and M. Ceriotti, *Nat. Rev. Chem.* **2**, 0109 (2018).

[168] M. Rossi, W. Fang, and A. Michaelides, *J. Phys. Chem. Lett.* **6**, 4233–4238 (2015).

[169] X. Z. Li, B. Walker, and A. Michaelides, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6369–6373 (2011).

[170] B. R. Shrestha, S. Pillai, A. Santana, S. H. Donaldson, T. A. Pascal, and H. Mishra, *J. Phys. Chem. Lett.* **10**, 5530–5535 (2019).

[171] S. Habershon, T. E. Markland, and D. E. Manolopoulos, *J. Chem. Phys.* **131**, 024501 (2009).

[172] C. Schran, F. Brieuc, and D. Marx, *J. Chem. Theory Comput.* **14**, 5068–5078 (2018).

[173] I. S. Novikov, A. V. Shapeev, and Y. V. Suleimanov, *J. Chem. Phys.* **151**, 224105 (2019).

[174] A. Pérez, M. E. Tuckerman, H. P. Hjalmarson, and O. A. Von Lilienfeld, *J. Am. Chem. Soc.* **132**, 11510–11515 (2010).

[175] S. Hay and N. S. Scrutton, *Nat. Chem.* **4**, 161–168 (2012).

[176] W. Fang, J. Chen, M. Rossi, Y. Feng, X. Z. Li, and A. Michaelides, *J. Phys. Chem. Lett.* **7**, 2125–2131 (2016).

[177] A. Berger, G. Ciardi, D. Sidler, P. Hamm, and A. Shalit, *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2458–2463 (2019).

[178] T. Clark, J. Heske, and T. D. Kühne, *Chem. Phys. Chem.* **20**, 2461–2465 (2019).

[179] L. Wang, S. D. Fried, S. G. Boxer, and T. E. Markland, *Proc. Natl. Acad. Sci. U.*

S. A. **111**, 18454–18459 (2014).

[180] Y. Litman, J. O. Richardson, T. Kumagai, and M. Rossi, *J. Am. Chem. Soc.* **141**, 2526–2534 (2019).

[181] L. Wang, S. D. Fried, and T. E. Markland, *J. Phys. Chem. B* **121**, 9807–9815 (2017).

[182] H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, in *Lecture Notes in Physics*, edited by K. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, Vol. 968 ,Springer, Charm, 2020, pp. 277–307.

[183] S. Chmiela, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, in *Lecture Notes in Physics*, edited by K. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, Vol. 968 ,Springer, Charm, 2020, pp. 129–154.

[184] J. Wang, S. Chmiela, K. R. Müller, F. Noé, and C. Clementi, *J. Chem. Phys.* **152**, 194106 (2020).

[185] V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M. Wilkins, B. A. Helfrecht, P. Juda, S. P. Bienvenue, W. Fang, J. Kessler, I. Poltavsky, S. Vandenbrande, J. Wieme, C. Corminboeuf, T. D. Kühne, D. E. Manolopoulos, T. E. Markland, J. O. Richardson, A. Tkatchenko, G. A. Tribello, V. Van Speybroeck, and M. Ceriotti, *Comput. Phys. Commun.* **236**, 214–223 (2019).

[186] A. E. Reed, L. A. Curtiss, and F. Weinhold, *Chem. Rev.* **88**, 899–926 (1988).

[187] E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, P. Karafiloglou, C. R. Landis, and F. Weinhold, *NBO 7.0*.

[188] F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 73–78 (2012).

[189] F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1327 (2018).

[190] H. B. Bürgi, J. D. Dunitz, and E. Shefter, *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **30**, 1517–1527 (1974).

[191] S. K. Singh, K. K. Mishra, N. Sharma, and A. Das, *Angew. Chemie - Int. Ed.* **55**, 7801–7805 (2016).

[192] R. W. Newberry and R. T. Raines, *Acc. Chem. Res.* **50**, 1838–1846 (2017).

[193] A. Choudhary, K. J. Kamer, and R. T. Raines, *J. Org. Chem.* **76**, 7933–7937 (2011).

[194] B. C. Gorske, B. L. Bastian, G. D. Geske, and H. E. Blackwell, *J. Am. Chem. Soc.* **129**, 8928–8929 (2007).

[195] G. J. Bartlett, A. Choudhary, R. T. Raines, and D. N. Woolfson, *Nat. Chem. Biol.* **6**, 615–620 (2010).

[196] R. W. Newberry, B. Vanveller, I. A. Guzei, and R. T. Raines, *J. Am. Chem. Soc.* **135**, 7843–7846 (2013).

[197] C. E. Jakobsche, A. Choudhary, S. J. Miller, and R. T. Raines, *J. Am. Chem. Soc.* **132**, 6651–6653 (2010).

[198] N. A. Wenzell, H. K. Ganguly, A. K. Pandey, M. R. Bhatt, G. P. A. Yap, and N. J. Zondlo, *ChemBioChem* **20**, 963–967 (2019).

[199] C. Cabezas, J. L. Alonso, J. C. López, and S. Mata, *Angew. Chemie - Int. Ed.* **51**, 1375–1378 (2012).

[200] A. Rahim, P. Saha, K. K. Jha, N. Sukumar, and B. K. Sarma, *Nat. Commun.* **8**, 78 (2017).

[201] S. Blanco, A. Macario, and J. C. López, *Phys. Chem. Chem. Phys.* **21**, 20566–20570 (2019).

[202] Y. L. Zhao, N. Ullah, S. Chen, and R. Q. Zhang, *J. Phys. Chem. C* **123**, 13976–13982 (2019).

[203] J. Feher, in *Quantitative Human Phisiology: An Introduction*, edited by J. Feher, 2nd edition ,Academic Press, Boston, 2017, pp. 120–129.

[204] C. J. Harris, M. Scheibe, S. P. Wongpalee, W. Liu, E. M. Cornett, R. M. Vaughan, X. Li, W. Chen, Y. Xue, Z. Zhong, L. Yen, W. D. Barshop, S. Rayatpisheh, J. Gallego-Bartolome, M. Groth, Z. Wang, J. A. Wohlschlegel, J. Du, S. B. Rothbart, F. Butter, and S. E. Jacobsen, *Science* **362**, aar7854 (2018).

[205] K. M. Gough and B. R. Henry, *J. Phys. Chem.* **88**, 1298–1302 (1984).

[206] K. M. Gough, B. R. Henry, and T. A. Wildman, *J. Mol. Struct. THEOCHEM* **124**, 71–85 (1985).

[207] P. J. Breen, J. A. Warren, E. R. Bernstein, and J. I. Seeman, *J. Chem. Phys.* **87**, 1917 (1987).

[208] T. Liljefors and N. L. Allinger, *J. Comput. Chem.* **6**, 478–480 (1985).

[209] P. George, C. W. Bock, J. J. Stezowski, T. Hildenbrand, and J. P. Glusker, *J. Phys. Chem.* **92**, 5656–5666 (1988).

[210] T. Kundu, B. Pradhan, and B. P. Singh, *J. Chem. Sci.* **114**, 623–638 (2002).

[211] C. Peng and H. B. Schlegel, *Isr. J. Chem.* **33**, 449–454 (1993).

[212] C. Peng, P. Y. Ayala, H. B. Schlegel, and M. J. Frisch, *J. Comput. Chem.* **17**, 49–56 (1996).

[213] W. E, W. Ren, and E. Vanden-Eijnden, *J. Chem. Phys.* **126**, 164103 (2007).

[214] H. Jónsson, G. Mills, and K. W. Jacobsen, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by B. J. Berne, G. Ciccotti, and D. F. Coker ,World Scientific, 1998, pp. 385–404.

[215] G. Henkelman and H. Jónsson, *J. Chem. Phys.* **113**, 9978–9985 (2000).

[216] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *J. Chem. Phys.* **108**, 1964–1977 (1998).

[217] P. G. Bolhuis and G. Csányi, *Phys. Rev. Lett.* **120**, 250601 (2018).

[218] R. Meyer, K. S. Schmuck, and A. W. Hauser, *J. Chem. Theory Comput.* **15**, 6513–6523 (2019).

[219] L. Pattanaik, J. B. Ingraham, C. A. Grambow, and W. H. Green, *Phys. Chem. Chem. Phys.* **22**, 23618–23626 (2020).

[220] P. Cattaneo and M. Persico, *Phys. Chem. Chem. Phys.* **1**, 4739–4743 (1999).

[221] A. Cembran, F. Bernardi, M. Garavelli, L. Gagliardi, and G. Orlandi, *J. Am. Chem. Soc.* **126**, 3234–3243 (2004).

[222] L. Gagliardi, G. Orlandi, F. Bernardi, A. Cembran, and M. Garavelli, *Theor. Chem. Acc.* **111**, 363–372 (2004).

[223] L. Wang and X. Wang, *J. Mol. Struct. THEOCHEM* **806**, 179–186 (2007).

[224] P. Tavadze, G. Avendaño Franco, P. Ren, X. Wen, Y. Li, and J. P. Lewis, *J. Am. Chem. Soc.* **140**, 285–290 (2018).

[225] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, *Adv. Neural Inf. Process. Syst.* **30**, 991–1001 (2017).

[226] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, *J. Chem. Theory Comput.* **15**, 448–455 (2019).

[227] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, *Comput. Phys. Commun.* **180**, 2175–2196 (2009).

[228] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

[229] A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.* **102**, 073005 (2009).

[230] G. Henkelman, B. P. Uberuaga, and H. Jónsson, *J. Chem. Phys.* **113**, 9901–9904 (2000).

[231] A. P. Bartók, S. Cereda, G. Csányi, J. R. Kermode, I. Solt, W. Szlachta, C. Várnai, and S. Winfield, *http://www.libatoms.org*.

[232] PLUMED Consortium, *Nat. Methods* **16**, 670–673 (2019).

[233] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, *Comput. Phys. Commun.* **185**, 604–613 (2014).

[234] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello, *Comput. Phys. Commun.* **180**, 1961–1972 (2009).

[235] D. Freedman and P. Diaconis, *Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* **57**, 453–476 (1981).

[236] A. H. Larsen, J. Jø. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, *J. Phys. Condens. Matter* **29**, 273002 (2017).

[237] A. G. Csaszar, *J. Am. Chem. Soc.* **114**, 9568–9575 (1992).

[238] L. F. Pacios, O. Gálvez, and P. C. Gómez, *J. Phys. Chem. A* **105**, 5232–5241 (2001).

[239] G. S. Hartley, *Nature* **140**, 281 (1937).

[240] G. S. Hartley, *J. Chem. Soc.* **1938**, 633–642 (1938).

[241] E. W. G. Diau, *J. Phys. Chem. A* **108**, 950–956 (2004).

[242] H. M. D. Bandara and S. C. Burdette, *Chem. Soc. Rev.* **41**, 1809–1825 (2012).

[243] J. Dokić, M. Gothe, J. Wirth, M. V. Peters, J. Schwarz, S. Hecht, and P. Saalfrank, *J. Phys. Chem. A* **113**, 6763–6773 (2009).

[244] G. Fonseca, I. Poltavsky, V. Vassilev-Galindo, and A. Tkatchenko, *J. Chem. Phys.* **154**, 124102 (2021).

[245] A. Grisafi and M. Ceriotti, *J. Chem. Phys.* **151**, 204105 (2019).

[246] T. Bereau, R. A. DiStasio, A. Tkatchenko, and O. A. Von Lilienfeld, *J. Chem. Phys.* **148**, 241706 (2018).

[247] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, *Phys. Rev. Lett.* **114**, 105503 (2015).

[248] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, *Phys. Rev. Mater.* **2**, 083802 (2018).

[249] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, *J. Chem. Phys.* **148**, 241730 (2018).

[250] F. Musil, M. Veit, A. Goscinski, G. Fraux, M. J. Willatt, M. Stricker, T. Junge, and M. Ceriotti, *J. Chem. Phys.* **154**, 114109 (2021).

[251] R. K. Cersonsky, B. A. Helfrecht, E. A. Engel, S. Kliavinek, and M. Ceriotti, *Mach. Learn. Sci. Technol.* **2**, 035038 (2021).

[252] A. Tkatchenko, R. A. Distasio, R. Car, and M. Scheffler, *Phys. Rev. Lett.* **108**, 236402 (2012).

[253] A. Ambrosetti, A. M. Reilly, R. A. Distasio, and A. Tkatchenko, *J. Chem. Phys.* **140**, 18A508 (2014).

[254] K. T. Schütt, F. Arbabzadah, S. Chmiela, K.-R. Müller, and A. Tkatchenko, *Nat. Commun.* **8**, 456–468 (2017).

[255] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis. The Primer* ,John Wiley & Sons, 2008.

[256] A. Saltelli, *Risk Anal.* **22**, 579–590 (2002).

[257] B. Huang and O. A. von Lilienfeld, *Nat. Chem.* **12**, 945–951 (2020).

[258] P. Ambure, A. K. Halder, H. González Díaz, and M. N. D. S. Cordeiro, *J. Chem. Inf. Model.* **59**, 2538–2544 (2019).

[259] C. Merkwirth, H. Mauser, T. Schulz-Gasen, O. Roche, M. Stahl, and T. Lengauer, *J. Chem. Inf. Comput. Sci.* **44**, 1971–1978 (2004).

[260] B. Cheng, R. R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csanyi, *Acc. Chem. Res.* **53**, 1981–1991 (2020).

[261] A. Reinhardt, C. J. Pickard, and B. Cheng, *Phys. Chem. Chem. Phys.* **22**, 12697–12705 (2020).