Inferring Chemistry from Data with Atomistic Machine Learning: Applications to Potential Energy Surfaces and Chemical Space.

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel
von

Luis Itza Vazquez-Salazar

Basel, 2024

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Erstbetreuer: Prof. Dr. Markus Meuwly

Zweitbetreuer: Prof. Dr. Markus Lill

Externe Experte: Prof. Dr. Alexandre Tkatchenko

Basel, den 26. März 2024

Prof. Dr. Marcel Mayor
Dekan

Abstract

The influence of machine learning (ML) in chemistry is undeniable, and it is a powerful tool to obtain chemical insights from large amounts of data. In particular, ML is a perfect tool for exploring chemical space because it allows to obtain good results in a relatively short time. The quality of the results obtained with an ML model highly depends on the data used to train it. After introducing fundamental concepts in Chapters 1 and 2, Chapter 3 deals with the effect of training data on predicting a chemical property. Results show that adequate predictions require a large chemical diversity in the training set. This can be obtained by either using many chemical motives or employing an adequate number of conformers. Once the effect of the data is clear, the next aspect evaluated is the confidence in the predictions obtained with ML models. To this end, two uncertainty quantification strategies based on Bayesian statistics were implemented. The insights into the interplay between error, uncertainty and chemistry provide us with an essential understanding of how a chemical database can be constructed. The previous chapters deal with the use of data obtained from *ab-initio* calculations. Nevertheless, it is expected that a model can reproduce experimental results. Chapter 5 deals with improving a potential energy surface (PES) based on experimental results by employing a procedure called morphing. Continuing with the study of PES, Chapter 6 uses one of the models introduced in Chapter 3 to study a reactive process. In this case, the performance of detecting outliers through uncertainty quantification was evaluated and compared with the other two strategies. Finally, Chapter 7 plays with adding samples from the conformational space represented by a PES to chemical databases biased towards a chemical insight. The last chapter summarizes the different aspects of the relationships between data and chemistry for exploring chemical space or working with PES. Also, it provides insights into future extensions of the projects presented here.

Acknowledgments

Doing a PhD is, without a doubt, one of the craziest ideas I have ever had. However, it has also been an amazing journey, thanks to many, many people, so many that if I mentioned all, this section would be larger than the thesis.

First, I would like to thank my supervisor, Prof. Dr. Markus Meuwly, for allowing me to join his research group, giving me creative freedom that is heavily reflected in the chapters of this thesis, and always leading by example. I would also like to thank my second supervisor, Prof. Dr. Markus Lill and my external reviewer, Prof. Dr. Alexandre Tkatchenko, for kindly agreeing to be part of my committee.

Next, I would like to thank the past and current members of the Meuwly group (Meenu, Debasish, Mike, Taylan, Silvan) for making my years in Basel very enjoyable despite the COVID times; thanks for all the fish, guys! In particular, I want to thank my partners in crime, Eric Boittier and Dr Kai Töpfer, for their support, all the nice chats and for proofreading parts of this work. In that line, I also would like to thank my götti, Dr Marco Pezzella for his advice, help, and proofreading of this work. Special thanks also go to my collaborators from FU Berlin, Karl Horn, Prof. Christiane Koch, and Prof. David Wales from the University of Cambridge.

Quiero dedicar esta tesis con mucho amor y cariño a mi padre, Luis Felipe Vazquez Basulto (QEPD). Papá esta es la tercera tesis de las tres que te prometí. Desde hace 14 años que te fuiste y no te he dejado de extrañar un solo día.

Esta tesis está dedicada a mi madre, Caridad Salazar Guerrero, como agradecimiento a todos sus esfuerzos, sacrificios, apoyo y amor incodicional por tantos años. Yo no hubiera logrado nada de esto de no ser por ti, te quiero mamá. También, quisiera dedicarle esto a mi hermana, Ixchel Vazquez Salazar, esperando que le sirva de inspiración en los futuros retos de su vida.

Quiero dedicar y agradecer a mi hermosa e inteligente novia (esposa) MSc. Atzin Ruiz Lera por su amor y apoyo durante todo el proceso del doctorado. Particularmente durante estos ultimos meses que me dedique a escribir. Fuiste la unica constante en un mundo cambiante, sin tu amor, tu paciencia, tus regaños y tus abrazos no hubiera terminado este proyecto, ¡Te Amo!. Adicionalmente, quiero agradecer a Mtra. Araceli Lera por su apoyo y por confiarme a uno de sus tesoros.

Un agradecimiento muy especial a mis amigos en Basel, mis Travis' Pals, Travis, Hector, Juan Carlos y Cristela. Basel no hubiera sido tan divertido sin ustedes. También quiero agradecer a Rafa Ornelas por ser un gran amigo durante mis años en Europa y por sus chistes de humor negro.

This thesis comprehends the effort of the last four and a half years. However, behind it, there is also the unconditional support and help of family, friends, and different teachers during all my formative years. I apologize and want to thank those not explicitly mentioned here but who were also an important part of this effort. Thanks, Danke, Merci, Grazie, Gracias Totales!

Basel, March 2024

Luis Itza Vazquez-Salazar

Contents

Chapter 1

Introduction

> It is clear to me that AI will never replace physicians — but physicians who use AI will replace those who don't.
>
> Jesse Ehrenfeld, President of the American Medical Association

It is undeniable the effect that machine learning (ML) has in our daily lives, from digital assistants such as Siri or Alexa[1, 2] to algorithms that suggest a movie or a video to watch on Netflix[3] or Youtube[4] to large language models such as ChatGPT[5] or BARD[6]. Complementary to this, ML has had a profound impact on how science is done, with numerous and growing applications in different fields like healthcare[7, 8], medical imaging[9], biomedical engineering[10], cosmology and particle physics[11], quantum physics[12], astronomy[13], genetics and molecular biology[14, 15], and many others.

Consequently, it is unsurprising that ML had a profound impact on practically all branches of chemistry. This comes from the fact that ML methods are extremely powerful and promise to be an alternative to solve some of the major problems that chemists face daily. Some authors consider the use of ML to constitute a shift in the scientific paradigm[16, 17] or a revolution in how we understand and model matter[18]. A straightforward example of the power of ML in chemistry is the acceleration of molecular simulations, which allows the study of complex systems with sizes that were technically impossible before. An example of this is the team winner[19] of the 2020 ACM Gordon Bell prize for "*Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning*" or a recent example in biomolecular systems with millions of atoms[20].

The multiple capabilities of ML have also opened new opportunities for scientific discovery by allowing scientists to generate new hypotheses, design new experiments, and collect and analyse large amounts of data that previously would have been impossible[21]. Additionally, the use of large amounts of data allows the emergence of hidden patterns in data. Illustrations of how the use of ML methods has led to new findings such as the discovery of a new phase transition in liquid hydrogen[22] through the use of ML-aided simulations or the repositioning of Halicin originally designed to treat diabetes as an antibiotic[23]. In this work, we aim to use ML methods to better understand chemistry by applying ML methods to model chemical space and potential energy surfaces.

Using ML in chemistry involves the creation of models that can learn a functional relationship between a compound and a specific property. To this end, the model requires a set of examples, called the training dataset. The creation of the training dataset is not a trivial task and requires a considerable amount of computational resources. Nevertheless, ML methods require large amounts of data to obtain predictions with *chemical accuracy*[1]. However, the enormous size of chemical space with more than 10^{180} possible chemical compounds[26] makes it impossible to follow the typical approach in computer science that assumes that large amounts of data will beat the best algorithms[27].

Nonetheless, searching in chemical space is an essential step in view of the discovery of new chemical compounds or materials[28]. Therefore, there is a compromise to be made between the amount of data required to train a model and an adequate exploration of the chemical space of interest. In this sense, it is remarkable that there is no complete understanding of how the training data influences the prediction of a specific chemical property. After introducing basic theoretical concepts in Chapter 2, the interplay between initial training data and prediction is studied in detail to predict tautomerization energies in Chapter 3. This property is convenient to study because it involves small structural changes and the existence of public databases. A few databases that explore chemical space and chemical+conformational space were used to train an ML model, particularly a neural network (NN) model, for predicting tautomerization energies of pairs of molecules in a public database. The results indicate that common databases present redundancies that reduce the quality of prediction of an ML model. Additionally, it was found that, against the typical expectation, more data does not necessarily imply

[1] Chemical accuracy is defined by John Pople in his Nobel lecture as 1 kcal/mol for the prediction of heats of formation or ionization potentials[24]. Complementary spectroscopy accuracy can be defined as 1 cm^{-1} for vibrational spectroscopy[25]

better predictions. Finally, the interplay between chemical space and conformation space was found to be of critical importance, given that a rational augmentation of data from conformational space can compensate for a poor exploration of chemical space.

The understanding gained in Chapter 3 takes us to the next challenge: the rational construction of the training databases. The strategy followed to tackle the challenge is the uncertainty in prediction by the NN model. The aim was to use uncertainty as a guide for exploring chemical and conformational space. Therefore, it is desirable that the NN model can evaluate the uncertainty of its predictions to give us information on how to improve the training dataset. The usual approach implies the use of several models with a corresponding high computational price. Nevertheless, new developments [29–31] have provided simple methodologies for quantifying the uncertainty of prediction in neural networks.

Chapter 4 describes implementing a method called Deep Evidential Regression (DER)[30] on top of the PhysNet NN[32] architecture to allow the prediction of uncertainties on compounds across chemical space. The results shed light on noise and redundancy's effect on property prediction for molecules. Even in cases for which changes, such as double-bond migration in two otherwise identical molecules, are small. It was possible to extract insights on which information the model used to make a prediction and how it can be related to the predicted uncertainty. It was found that the model can be confused by adding several similar examples, and then it will assign a small uncertainty to a molecule that has been poorly predicted. An alternative to DER is the method called Regression Prior Networks[29]. This method was also implemented in PhysNet. However, its capability predictions were very poor, and they are briefly described in Chapter 4.

The results from chapters 3 and 4 are related to the prediction by an ML model of quantities obtained from *ab initio* calculation. Nevertheless, generating models that can accurately reproduce experimental quantities is important. Many experimental quantities can be obtained from the Potential Energy Surface (PES), a fundamental concept for characterizing the dynamics in the gas and condensed phase[33, 34]. The application of ML techniques to describe PES has been proven to be highly successful [35]. Nowadays, ML PES can be used to reproduce to certain accuracy the experimental results for quantum phenomena such as Feshbach resonances (FR)[36]. Nevertheless, integrating experimental information to refine PES obtained with ML is still unexplored. Chapter 5 discusses using information from FR for the "morphing" of a PES[37] obtained with the ML method, Reproducing Kernel Hilbert Space (RKHS)[38]. The results indicate that even the potential obtained at the highest level of electronic structure theory can be improved compared with experimental quantities. Additionally, the results improve the understanding of the origin of the experimental observables. In this case, it was found that FRs are sensitive to the long-range part of the PES.

Following the study of PES is also of interest the use of uncertainty quantification (UQ) methods such as the ones introduced in Chapter 4 to improve PES. This is a particularly challenging task because in the formulation of DER, it is not possible to obtain the uncertainty of forces ($\boldsymbol{F}$ given that those are the negative derivative of the potential energy (E), $\boldsymbol{F} = -\nabla_{\boldsymbol{R}} E$). Therefore, we aimed to understand the limitations of DER and other UQ methods, such as ensembles and Gaussian Mixtures Models (GMM), for detecting outliers. Finding such outliers or outlier regions helps to increase the trained model's robustness and further improves its accuracy and reliability. In Chapter 6, we apply different models for predicting outliers in a reactive potential. The construction of such potentials is particularly hard because it requires the sampling of rare events[2]. In consequence, an adequate detection of outliers is crucial to obtain reliable reactive PES.

Finally, Chapter 7 presents the union between chapters 3 and 4 with chapters 5 and 6 by exploring how the exploration of conformation space (PES) can be used to improve the exploration of chemical space. To that end, purposely biased databases were constructed by following chemical patterns. Then, we study different ways to perform data augmentation to improve the predictions of the constructed NN models. New data was added to the biased databases by adding structures from conformational space obtained with normal mode sampling. First, the best temperature for sampling was determined, followed by the number of structures required to observe an improvement. Additionally, the addition of structures generated within the Atoms-in-Molecules[40] fragment approach was also tested. The next method of addition was capitalising the uncertainty predictions to add samples from the conformational space of the molecules with the largest predicted uncertainty.

In this work, ML was employed in different flavours to create a better understanding of chemistry. From discerning the influence of data in predicting chemical properties to including experimental information to modify potential energy surfaces, the different ML methods explored in this work contribute to an enhanced understanding of how

[2] Rare events in mathematics are events that are expected to occur infrequently or, more technically, those that have low probabilities (say, order of 10^{-3} or less) of occurring according to a probability model.[39]

chemistry can be deducted from data. The interplay between chemical compound space and conformational space set the stage for better explorations of the first. On the other hand, including experimental information in the adjustment procedure of a PES is a step forward in the direction of the inverse problem of obtaining a PES from spectroscopical observables. Many new avenues of research can be explored, such as the interpretability of the ML models, the definition of chemical content by using information theory, the formulation of chemical exploration as an optimization problem, etc.

Chapter 2

Theoretical Background

Eigentlich weiß man nur wenn man wenig weiß; mit dem Wißen wächst der Zweifel.[a]

[a] We know accurately only when we know little; with knowledge doubt increases.
Johann Wolfgang von Goethe

This chapter introduces the basic concepts behind the techniques used in the thesis, which are necessary to understand the following chapters.

Parts of this chapter have been previously published in: Dig. Disc., 2023, 2, 28-58.

2.1 Chemical Space

In 1905, Swiss Nobel Prize winner and one of the fathers of modern chemistry Alfred Werner described the mission of chemistry as[41]: *"Die Chemie muss zur Astronomie der Molekulare Welt werden"*[1]. This quote is the first reference to what will be later called chemical space(CS)[42]. The definition of chemical space is the set of all possible molecules or materials[43]. This implies that the size of CS is extraordinarily large. The total number of particles in the universe is estimated to be 10^{80}, from which 7×10^{76} are atoms [44, 45]. Therefore, the possible number of substances that can be theoretically obtained is[26]:

$$C = \sum_{k=1}^{10^{76}} \binom{k + 10^{76} - 1}{10^{76}}$$

[1] Chemistry must become the astronomy of the molecular world

here, we consider that there are k numbers of choosing atoms from the total numbers without considering the order and allowing for repetitions. Of course, not all combinations of atoms are allowed. By applying physical constraints and restricting the possible compounds to the elements C, N, O, P, S, F, Cl, Br, and I with a molecular weight of less than 1000 daltons, David Weininger[2] hypothesised that the number of possible substances is about 10^{200}[26, 46]. This number is known as Weininger number or one Weinamol. Later, it was estimated that only 1 in 10^{20} compounds could be physically and chemically stable, reducing the number to 10^{180}[46]. This number is larger than the total amount of information in the visible universe (10^{123})[45]. Nevertheless, if we ignore this limitation and consider the trend of exploration of chemical space continues as of now, it is estimated that it would take 10300 years to discover one Weinamol[26].

Despite the huge size of CS, its exploration is a task that chemists have been doing actively for the last two centuries. This is proven by the fact that the number of chemical substances reported has been constantly doubling every 16 years since 1800[26]. A large part of the exploration of CS has been led by theoretical and computational chemists who have applied mathematical and computational tools for the comprehensive enumeration of chemical structures. For example, the Reymond group from the University of Bern, Switzerland, enumerated 166.4 billion chemical structures with up to 17 heavy atoms by using purely graph methods[47–50]. Although this effort has considerably helped to enrich our chemical knowledge, some of the generated molecules are unrealistic or tend to be biased by construction[51]. Other examples are the PubChem database[52, 53] that in 2023 contained information on 116 million pure and characterized compounds.

Despite these great efforts, CS is too large to be exhaustively explored. For this reason, a more effective approach is reducing the number of compounds based on their structure (i.e. peptides, proteins, etc.) or the properties (and possible functionalities) of those molecules[43]. Therefore, subsets of CS can be defined, such as ’drug-like chemical space’. Unfortunately, this is not an easy task because CS is very large and sparse, with regions densely populated and others empty[54]. Thus, it is desirable to make a targeted exploration of chemical space by considering together with the structures the properties that each point in chemical space has (Figure 2.1). Complementary to this, we could extend the definition of chemical space as suggested by von Lilienfeld *et al.*[55, 56] to consider all feasible metastable atomic configurations, including conformational isomers, reactive intermediates or minima in electronically excited states[56]. By

[2] David Weininger was a cheminformatician from the USA known for the invention of common linear chemical representations such as SMILES, SMARTS, and SMIRKS.

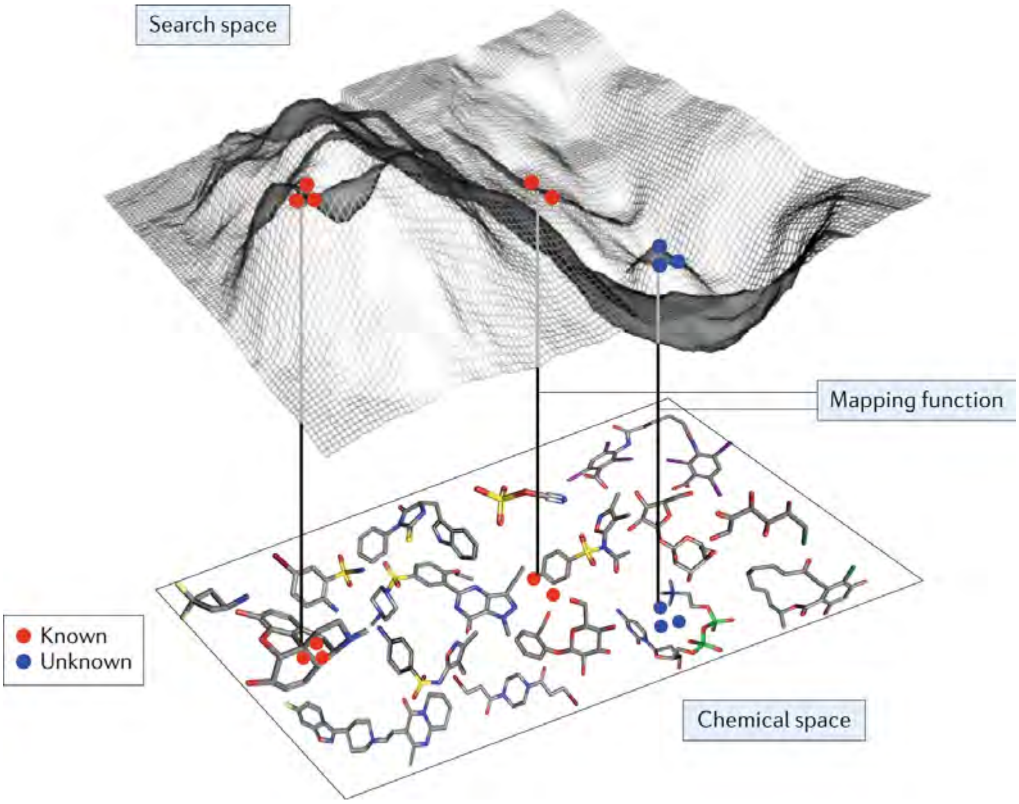

Figure 2.1: **Chemical space**. 3D representation of the relationship between property space and the chemical compound space. The z-axis represents the value of a physical property while the plane xy are features to describe the chemical space in 2D. The mapping function illustrated can be a machine learning model, while the property space is the search space. The red points in the figure correspond to known compounds, while the blue ones are unknown species. Adapted with permission from Ref. [54]. Rights reserved by the authors and Springer Nature.

extending the definition, we could use the non-equilibrium structure as a smooth link between samples in chemical space[55]. The described link is the key behind the connection between chemical space and potential energy surfaces that will be described in the following chapters. Then, the exploration of chemical space by mapping functions can be done by using ML models as those mapping functions (Figure 2.1).

2.2 Potential Energy Surfaces

The energetics of a molecular system can be described by solving the electronic Schrödinger Equation (SE). Unfortunately, the SE can only be solved exactly for simple, single-electron atomic systems. In order to obtain solutions for many-electron systems, it is necessary to introduce approximations. The Born-Oppenheimer approximation (BOA)[57], also called *the most important approximation in quantum chemistry*,[58] assumes that the coupling between the nuclear and electronic motion can be neglected because the mass of the nuclei is several orders of magnitude larger than the mass of the electrons. Under this assumption, it is possible to rewrite the total wavefunction Ψ, which is a solution of the SE, as the product of a nuclear wavefunction $\chi(\boldsymbol{R})$ with nuclear positions $\boldsymbol{R}$ and the electronic wavefunction $\psi(\boldsymbol{r};\boldsymbol{R})$ with electron coordinates $\boldsymbol{r}$ for a fixed configuration of nuclear positions

$$\Psi(\boldsymbol{r},\boldsymbol{R}) = \psi(\boldsymbol{r};\boldsymbol{R}) \cdot \chi(\boldsymbol{R}). \tag{2.1}$$

As a consequence, the electronic wavefunction can be obtained by solving the electronic time-independent SE:

$$\hat{H}_{\mathrm{e}}\psi_\lambda(\boldsymbol{r};\boldsymbol{R}) = \left[\hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{ee}}\right]\psi_\lambda(\boldsymbol{r};\boldsymbol{R}) = \epsilon_\lambda(\boldsymbol{R})\psi_\lambda(\boldsymbol{r};\boldsymbol{R}) \tag{2.2}$$

Here, $\hat{H}_{\mathrm{e}}$ is the electronic (spin-free) Hamiltonian describing the kinetic energy of the electrons $\hat{T}_{\mathrm{e}}$, the Coulomb interaction between the nuclear and electron charges $\hat{V}_{\mathrm{ne}}$ and the electron-electron interaction $\hat{V}_{\mathrm{ee}}$. The solution to the eigenvalue problem is the electronic wavefunction ψ_λ and electronic energy ϵ_λ for the electronic state λ. The so-called adiabatic Potential Energy Surface (PES) of an atomic system $E_\lambda(\boldsymbol{R})$ in electronic state λ constitutes an effective potential for the nuclear dynamics. It is obtained by the sum of the Coulomb repulsion V_{nn} between the nuclei with nuclear charge Z_i for the total number of atoms N_{atom}, and the respective electronic energy at the associated nuclear positions[59].

$$E_\lambda^{\mathrm{BO}}(\boldsymbol{R}) = V_{\mathrm{nn}}(\boldsymbol{R}) + \epsilon_\lambda(\boldsymbol{R}) \tag{2.3}$$

Equation 2.3 defines a PES as a $(3N-6)-$dimensional function that can be approximated as an analytical function, which is, however, a challenging task. Often, one can only report low-dimensional cuts of such high-dimensional hypersurfaces and one example is shown in Figure 2.2. Alternatively, equation 2.3 suggests that there should be a mapping between the total electronic energy of a molecular system and the combination of position of the nuclei and the set of nuclear charges $(\{Z_i\}_{i=1}^N)$, $f : \{Z_i, \boldsymbol{R}_i\}_{i=1}^N \rightarrow E_\lambda$. This is the starting point for describing a PES using an ML method.

PESs lie at the heart of computational chemistry[60] because it contains all the information about the many-body interactions of a molecular system, including stable and metastable structures[61]. The relationship between structure and potential energy E_λ allows to derive many molecular properties by taking derivatives with respect to a perturbation such as atomic positions $\boldsymbol{R}$, an external electric $\vec{\mathcal{E}}$ or magnetic field $\vec{\mathcal{B}}$, which

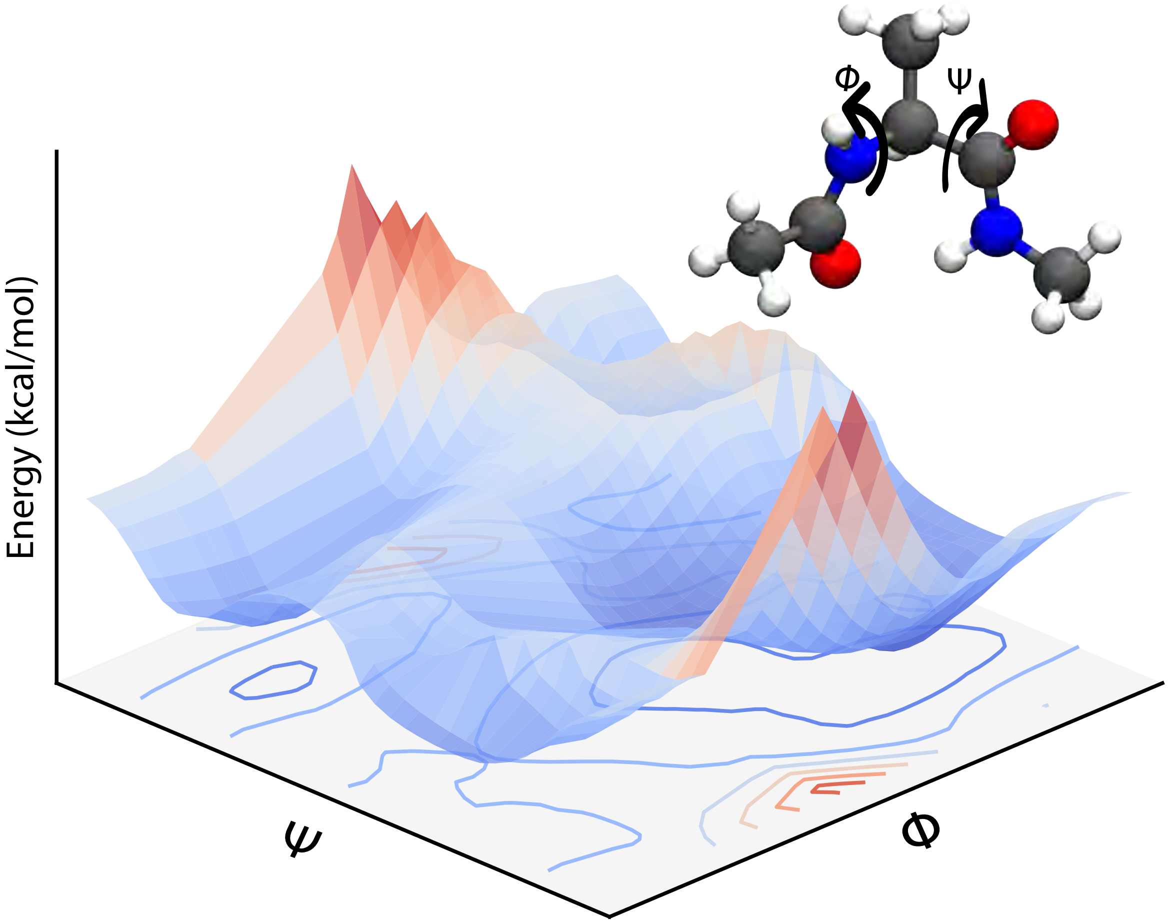

Figure 2.2: **Example of a potential energy surface**. A two-dimensional PES for the dialanine molecule calculated at the MP2 level with the 6-31G** basis set along dihedral angles Φ and Ψ. A representation of the molecule (ball and stick) indicating the dihedral angles (Φ, Ψ) calculated is given as well. The bottom gives the projection of the 2D PES.

require additional coupling terms in the Hamiltonian and an analytical representation of the PES.[59] Following this, a general response property takes the form

$$\text{Property} \propto \frac{\partial^{(n+m+l)} E_\lambda}{\partial \boldsymbol{R}^n \partial \vec{\mathcal{E}}^m \partial \vec{\mathcal{B}}^l} \tag{2.4}$$

where n, m, l indicate the order of the derivative with respect to the perturbation. Derivatives of Equation 2.4 provide, e.g., the forces $\boldsymbol{F} = -\partial E_\lambda / \partial \boldsymbol{R}$ that constitute the foundation of MD simulations and structure optimization schemes. The second derivatives $\partial^2 E_\lambda / \partial \boldsymbol{R}^2$ gives access to the Hessian matrix from which the harmonic frequencies of molecular vibrations can be obtained. Other properties such as the dipole moment ($\vec{\mu} = -\partial E_\lambda / \partial \vec{\mathcal{E}}$) or the molecular polarizability ($\vec{\alpha} = -\partial^2 E_\lambda / \partial \vec{\mathcal{E}}^2$) are directly related to experimental observables such as the infrared (IR) or Raman spectra.[62] Mixed derivatives also provide IR absorption intensities ($\partial^2 E_\lambda / \partial\vec{\mathcal{E}}\partial\boldsymbol{R}$) or the optical rotation in circular dichroism ($\partial^2 E_\lambda / \partial\vec{\mathcal{E}}\partial\vec{\mathcal{B}}$)[59].

2.3 Machine Learning

In general, Machine Learning (ML) is a subfield of artificial intelligence focused on the design and analysis of algorithms that allow computers to learn[63]. Although the use of these methods has increased considerably in the last decade[35], the mathematical and theoretical foundations of ML techniques can be traced back to the decades of '40s and '60s of the last century with the introduction of fundamental concepts such as the Turing machine[64] and the perceptron by Rosenblatt[65]. In chemistry, the first applications appeared at the end of the '60s in synthesis prediction[66, 67]. Thirty years later, many technical and theoretical limitations were overcome for new applications in chemistry to appear in the '90s in analytical and medicinal chemistry[68]. Around the same years, the first applications in PES appeared[69, 70].

The main mission of ML methods is obtaining algorithms that can infer a function that maps a collection of inputs to an observed outcome[71]. This implies that ML methods improve the quality of their results as a function of the amount of information the algorithm receives, a consequence of the fact that these are based on statistical methods. ML methods can be broadly classified into three types[72]:

- **Supervised methods**: In these methods, a dataset composed of input-output pairs is available and is used to train a model to obtain property predictions. The goal of these methods is to generate predictions for unseen input values.

- **Unsupervised methods**: In this case, there is no specific output that the method needs to predict. Those methods are used to extract information from the input values. Then, their main applications are in pattern search, trend identification or information reduction. A review of applications of these methods in molecular simulations can be found in Ref. [72].

- **Reinforcement learning methods**: These methods assume a setup in which the model obtains data, learns from its environment and executes actions with the goal of maximising a reward[73]. Learning occurs by "trial and error" through continuous interaction with the environment. Recently, applications of these methods have been reviewed in Ref. [73].

In addition to the methods mentioned above, there are the so-called generative methods. In those methods, the model learns a data distribution and generates new samples that are similar to the ones on the training distribution[74]. More formally, given a collection of data points, $\boldsymbol{X} = \{X_i\}$, in a space $\mathcal{X}$, a model is trained to match the data distribution P_X by means of a generative process P_G, such as $Y \approx P_G$ resembles the real data, $X \approx P_X$[75]. These models are particularly useful in chemistry because they can be used to generate new compounds while following chemical rules. Several reviews of the use of these models can be found; we refer the interested reader to Refs. [76] and [77] for further details about generative models.

In chemistry, the most commonly used methods are of the supervised type. To use a supervised model, you usually need to follow seven steps; see Figure 2.3. Those are:

1. **Define the objective or problem to be solved**. First of all, it is important to delimit the task to be performed, as the next steps depend on a good definition of it. It is important to keep in mind that not for every problem of chemical interest, it is possible to apply ML methods. In some cases, using an ML method could be an excess, especially if simpler and/or more robust methods are available. Therefore, a proper review of the existing literature is necessary before starting a new project. At this point, the chemist must answer the question, "What do I want to achieve by using an ML method?" and then answer: "What technique can I use to achieve my goal?".

2. **Data collection, curation, and cleaning**. This point is critical for obtaining a functional model. In this step, we include the generation process of a database, its posterior cleaning and validation. For an ML method, being able to be trained with the necessary information is as important as having the appropriate reagents in a chemical synthesis[78]. It is important to emphasise that if raw data are used for training and ML without having been previously curated and cleaned, it is impossible to obtain correct predictions[79]. In ML, the general rule of thumb for training ML models is that the more information we use, the better our predictions will be. However, this is not always correct, as will be discussed in Chapter 3.

3. **Design and selection of descriptors**. This step corresponds to a very active area of research[80–82]. Here, we consider the process in which molecules are transformed from the common notations used by chemists into values that can be interpreted by the machine, a process known as encoding[83]. There is no unique solution to the problem of how to encode a molecule. In turn, the possible descriptors vary in complexity, so the level of description desired and the type and

size of chemical structures (organic, inorganic, biomolecules, macromolecules, etc.), among other factors, must be considered.

4. **Algorithm selection**. In this step, a method must be selected to process the data obtained from the previous step to obtain a mapping between inputs and outputs. It is important to note that algorithm selection must be done by considering the amount and reliability of available data, computational resources, and the level of accuracy desired, among other factors. Using the most sophisticated algorithm (e.g. a neural network) is not advisable if you do not have an adequate amount of data or if it is possible to obtain similar results with simpler models (Occam's Razor). The selection of an ML algorithm should be justified by the nature of the problem to be solved and with concrete scientific reasons[79]. The use of fashionable methods without having a deep understanding of the scope and limitations of the algorithm should be avoided at all costs, as this could lead to incorrect results.

5. **Training the model**. Once an algorithm has been selected and the data to be used has been collected, it is necessary to 'train' the model. The training step corresponds to the real 'learning' of the model. The algorithm will adjust the model's parameters based on a training dataset to do this. The goal is to minimize a loss function $\mathcal{L}$ that measures the accuracy of the fit. It is important to mention that the quality of the fit must be independently evaluated in a subset of data that was not used for model fitting. This process is called validation. The training process should be stopped when it is considered that a minimum value of the loss function evaluated in the validation subset is reached.

6. **Testing the model**. Once a model has been trained and validated. A final step of testing is necessary to assess the quality of the model. This is done on a third subset of data called the test set. As in the validation case, the test set should contain information completely external to the information used for training and validation. This step allows us to evaluate the predictive capability of the model. Also, in this step, other tests of the model could be considered, such as data outside distribution or its use in molecular dynamics simulations. If the results of this step are not satisfactory, the model needs to be improved by one of the methods described in the next step.

7. **Refinement and update of the model**. ML models are not definitive, so as they are used, limitations or shortcomings of the model are discovered. Then, after a certain time, the ML model would need to be updated in light of new information

or to extend its predictive capacity. There are different ways to update or refine a model. One of the most common is the addition of new data points to the training set based on a performance measure. This procedure is called active learning[84]. On the other hand, if the model wants to be reused in a specific target for which information is available, it is possible to retrain the model with only the new data from the previously trained model. This procedure is called transfer learning[85].

In the next sections, we will review the theory behind the two ML methods that were employed in this thesis: kernel methods and neural networks.

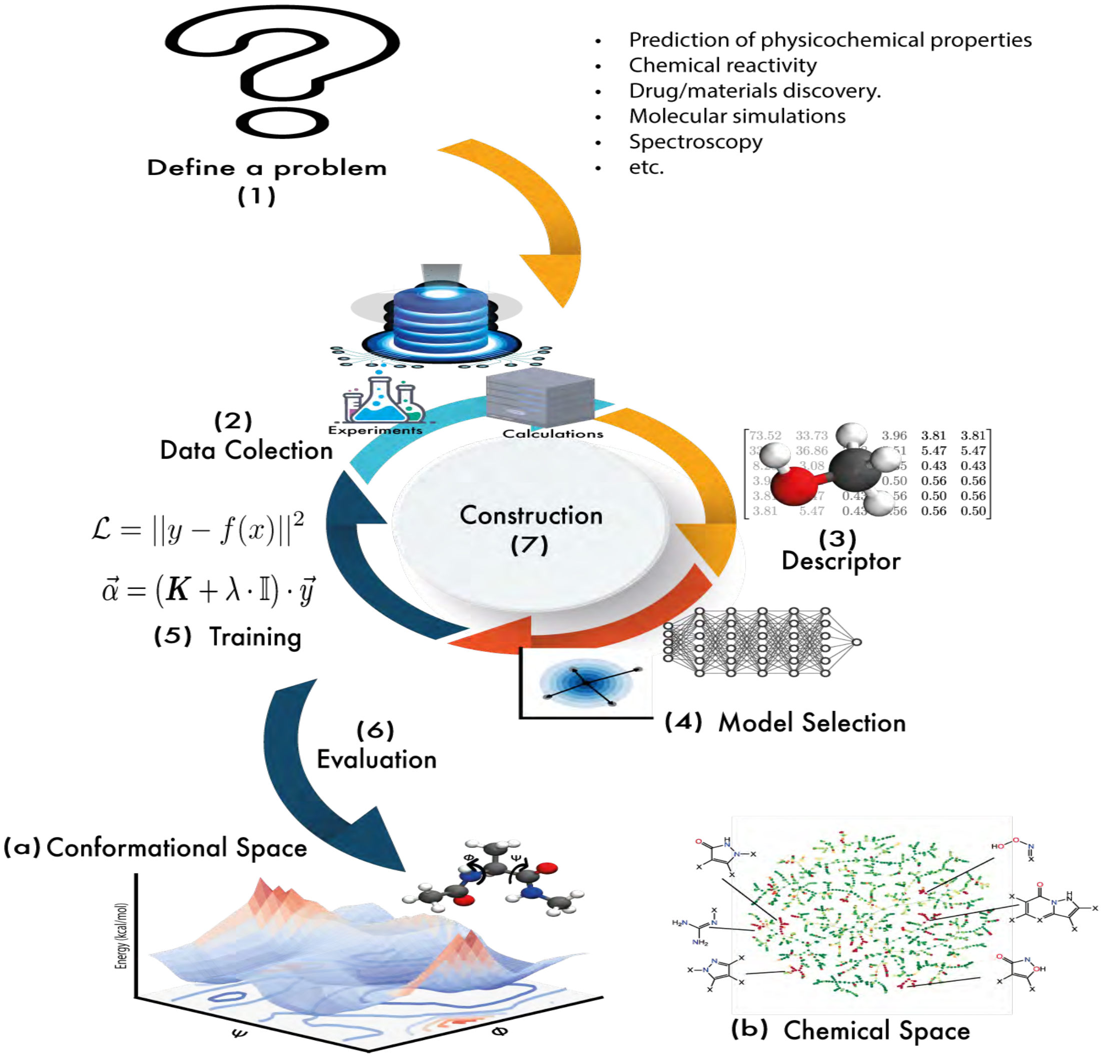

Figure 2.3: **Machine learning pipeline** The cycle consists of seven steps necessary for constructing a machine learning (ML) model to be applied in chemistry. The (1) first step is defining the problem to solve; examples of problems that can be solved with ML are enumerated. In the next step, (2) data needs to be collected. This data can come from experimental results or computer simulations. The third step is the transformation of the recollected data into values that can be understood by the machine; in panel (3), the molecule of methanol is shown with the corresponding Coulomb matrix as background. The fourth step is the algorithm selection (4). In this step, the amount of data collected from step 2 and the amount of available computational resources should be considered for selecting an adequate algorithm (4). The most common ML methods in chemistry are kernel-based algorithms (left) and neural networks (right). The next step is the training procedure (5), which minimises a loss function by adjusting the model's parameters to reproduce the reference values. Steps 2-5 are represented as a cycle, given that the model needs to be updated to increase its predictive capabilities (7). Finally, the model can be used to evaluate (6) different molecules in chemical space or different geometries in conformational space. The chemical space is represented with the T-Map[86] of the TautoBase[87] (a) where the colours represent the error in prediction of energy with a ML model[88]. For the conformational space (b), the potential energy surface of dialanine is represented.

2.4 Kernel Methods

Kernel methods are an ML algorithm used to find convex solutions to non-linear optimisation problems[89]. In these methods, the collected information is transformed into a new space (Figure 2.4A), called "feature space", through a function (kernel) that is defined by the user. This new space, created by the kernel, encodes the similarity between different points (Figure 2.4C), facilitates the learning, and guarantees an optimal generalisation[89]. Kernel methods can be used as unsupervised or supervised ML methods. Here, we will focus on the second type. Details of using kernels as unsupervised ML methods can be found in Ref. [90].

The construction of a kernel method consists of two steps. The first is to find a representation of the data that encodes the distribution of information in a complete, unique, and efficient way[92]. Multiple representations have been used for kernels; a complete review can be found in Ref. [93]. The second step is selecting a kernel that creates a map between the selected representation and the feature space; examples of common functions are shown in Figure 2.4B. The kernel function (ϕ) , in general $\phi : \mathbb{R}^{n_\mathrm{i}} \to \mathbb{R}^{n_\mathrm{o}}$, transforms representations with dimension n_i to the corresponding feature space with dimension n_o. The mapping is guaranteed by Mercer's theorem[94] for functions on a Hilbert space ($\mathfrak{L}^2$) defined on a compact set[89]. Another condition that the kernel function must follow is the reproducing property, which stays that the map ϕ exists if and only if[95]:

$$\forall \quad \boldsymbol{x_i}, \boldsymbol{x}_j \in \mathfrak{L}^2 : \boldsymbol{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \int \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j) d\boldsymbol{x}_j = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle \tag{2.5}$$

In Equation 2.5, it is not necessary to explicitly know ϕ or $\mathfrak{L}^2$. Their existence is a sufficient condition.

Kernel methods are usually applied in chemistry to solve regression problems. In that case, the problem can be set up as follows. Given a data set $\{\vec{y}, \boldsymbol{x}\} = \{y_i; x_i\}_{i=1}^{M}$ in which $y_i \in \mathbb{R}$ are the M reference values and $x_i \in \mathbb{R}^{n_\mathrm{o}}$ are the M values of the inputs encoded in the chosen representation, the function that is wished to approximate is defined according to the approximation theorem[95, 96] as:

$$\vec{y} = f(\boldsymbol{x}) + \epsilon \tag{2.6}$$

Where ϵ is measurement noise, the function $f(\boldsymbol{x})$ is approximated as a linear combination of the kernel evaluated at each of the M input points.

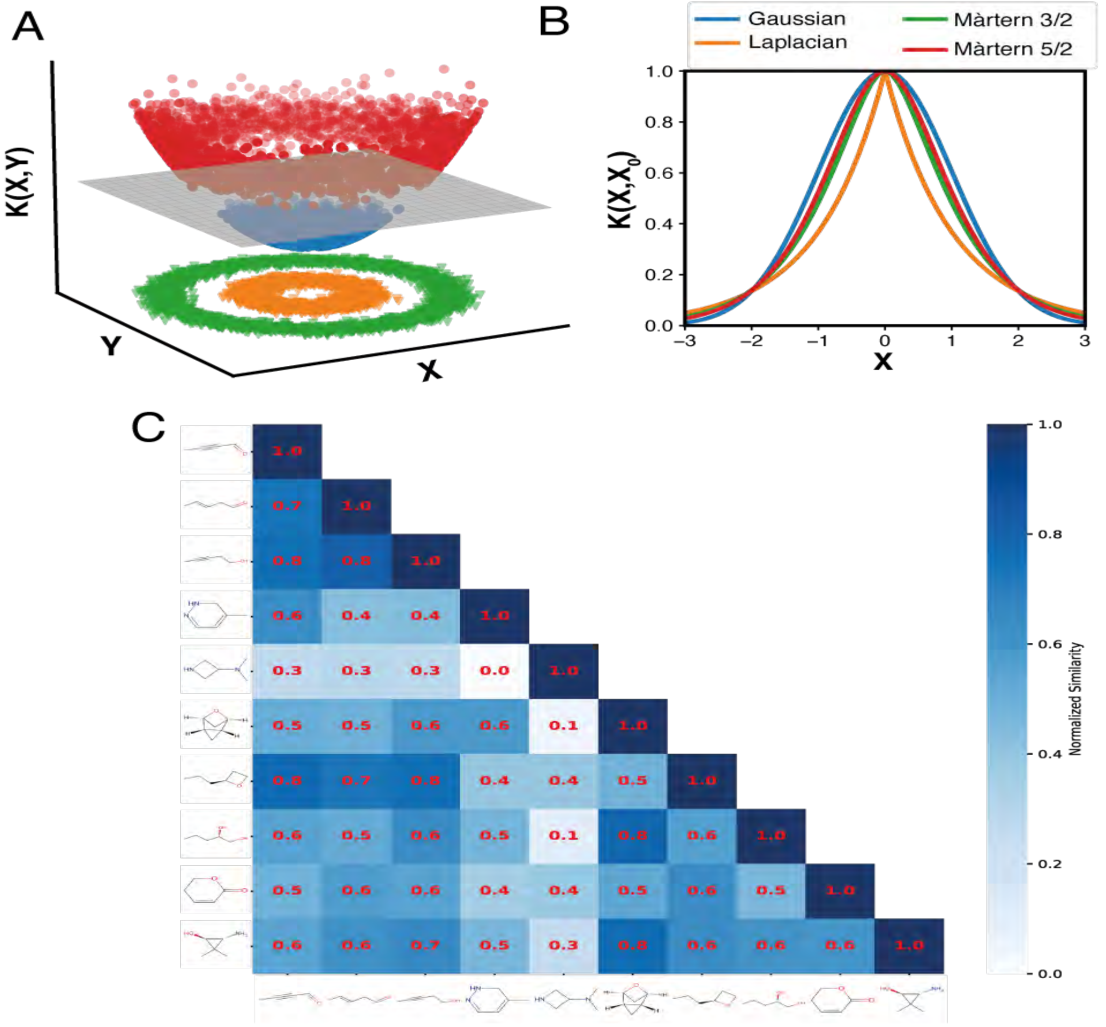

Figure 2.4: **Kernel methods**. Panel A illustrates the transformation of initial data to the feature space. The initial data represented by green and orange points is not separable on two dimensions. After the application of a polynomial kernel ($K(x, y) = x^2 + y^2$), the data is separated in the feature space. The decision plane that separates the data in feature space is illustrated in grey, while the transformed data is represented with blue and red points. Panel B displays examples of typical kernel functions. Panel C shows the normalized similarity between 10 random molecules from the QM7 database[91]. The molecules are encoded using the Coulomb matrix representation. Then, the encoded molecules are passed to a Gaussian kernel with $\sigma = 4000$. The kernel function encodes the distance in chemical space between the selected molecules.

$$f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i K(x, x_i) \tag{2.7}$$

In equation 2.7, α_i are the coefficients of the expansions, and $K(x, x_i)$ are the kernel functions (typically a nonlinear, symmetric and positive semidefinite function[96]). Figure 2.4B shows examples of kernel functions. The coefficients $\vec{\alpha} = \{\alpha_i\}$ can be obtained through the minimum-squares method by minimizing a loss function defined as:

$$\mathcal{L} = \min_{\alpha} \sum_{i=1}^{N} \left[\left(\sum_{j=1}^{M} \alpha_j K(x_i, x_j) - y_i \right)^2 + \lambda \cdot \left\| \sum_{j=1}^{M} \alpha_j K(x_i, x_j) \right\|^2 \right] \tag{2.8}$$

where λ is a hyper-parameter that helps to obtain numerically stable solutions. This parameter is known as a *regularizator* and usually corresponds to a tiny number[90].

Equation 2.8 can be solved to obtain the values of α_j by using linear algebra techniques such as Cholesky decomposition. Consequently, we should write Equation 2.8 in matrix form. First, the kernel matrix is defined as:

$$\boldsymbol{K}(x_i, x_j) = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_M, x_1) & \cdots & k(x_M, x_N) \end{pmatrix}$$

by replacing the kernel matrix in equation 2.8, we obtain the matrix version of it.

$$\mathcal{L} = (\boldsymbol{K}\vec{\alpha} - \vec{y})^\top \cdot (\boldsymbol{K}\vec{\alpha} - \boldsymbol{y}) + \lambda(\vec{\alpha}^\top \boldsymbol{K}\vec{\alpha}) \tag{2.9}$$

To obtain the values of $\vec{\alpha}$, we take the gradient of equation 2.9 with respect to $\vec{\alpha}$ and set it equal to 0 (i.e. $\nabla_{\vec{\alpha}} \mathcal{L} = 0$). Then, the values of α are:

$$\vec{\alpha} = (\boldsymbol{K} + \lambda \cdot \mathbb{I})^{-1} \cdot \vec{y} \tag{2.10}$$

here $\mathbb{I}$ is the identity matrix.

In this work, the kernel method studied is the reproducing kernel Hilbert space (RKHS)[38, 97]. In this method, the kernel matrix in D dimensions is constructed as a product of $1d$ polynomial kernels as follows:

$$\boldsymbol{K}(x_i, x_j) = \prod_{d=1}^{D} k^d(x_i^{(d)}, x_j^{(d)}) \tag{2.11}$$

When representing a PES, the uni-dimensional kernels should decay asymptotically to zero at large distances. Then, a general formula for bonds can be obtained. Here, we only provide the expression for the complete derivation the reader is referred to Ref. [97].

$$k^{[n,m]}(x_i, x_j) = n^2 x_>^{-(m+1)} B(m+1, n) {}_2F_1(-n+1, m+1; n+m; \frac{x_<}{x_>}) \tag{2.12}$$

in this expression $x_>$ and $x_<$ are the larger or smaller value between x_i and x_j, $B(a, b)$ is the beta function defined as:

$$B(a, b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

and ${}_2F_1(a, b; c; x)$ is the Gauss hyper-geometric function that has the series expansion[98]:

$${}_2F_1(a, b; c; x) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{x^n}{n!}$$

where $(a)_n$ is the Pochhammer symbol defined by:

$$(a)_n = a(a+1)(a+2) \cdots (a+n-1) \quad ; (a)_0 = 1$$

Values of m and n are integers related to the asymptotic behaviour and the smoothness of the kernel function, respectively.

Complementary to the bond expression, a general form for angles will be given. Again, we refer the interested reader to the literature for the detailed derivation[97].

$$k_2^{[n]}(x_i, x_j) = \sum_{i=0}^{n-1} x_>^i x_<^i + n x_<^n x_>^{n-1} {}_2F_1(-1, -n+1; n+1; \frac{x_<}{x_>}) \tag{2.13}$$

It should be noticed that the last expression works in a closed interval $[0, 1]$. Then, the angle θ that works in an interval $[0, \pi]$ is converted to a new coordinate, z, through the transformation:

$$z(\theta) = \frac{1 - \cos(\theta)}{2} \tag{2.14}$$

A general problem that limits the application of kernel methods is the amount of information that can be handled. This is because the amount of computational resources to use them scales with the size of the kernel matrix. By considering a kernel matrix of size $N_{\text{train}} \times N_{\text{train}}$, where N_{train} is the size of the training data, the amount of equations to solve is on the order of $\mathcal{O}(N_{\text{train}}^3)$[90]. Additionally, larger kernel matrices experience problems while loading to the RAM memory. In light of the previous problems, kernel methods are constrained to the study of databases of small or medium size. Nevertheless, solutions to circumvent these problems have been proposed, among them the use of reduced databases or the incorporation of the gradient during the training procedure[99, 100].

2.5 Neural Networks

The second ML method reviewed in this work is neural network(NN). These algorithms are inspired by the networks formed in the brain and how information is processed by them[101]. This is why these models are often called artificial neural networks (ANN). ANNs have gained popularity in the last decade[35] and consequently have found diverse applications in different fields. The popularity of ANNs is due to their ability to obtain multidimensional nonlinear relationships from large amounts of data in a computationally efficient manner[102]. In turn, this capability is a consequence of their construction in which many small computational units called neurons (Figure 2.5A) are interconnected to form complex predictions[103, 104]. There are different types of NNs; here, we will describe four types: fully connected (Figure 2.5C), convolutional (Figure 2.6), graph (Figure 2.7) and the so-called transformers (Figure 2.8). Transformers are the key behind the success of Large Language Models (LLM) such as ChatGPT[5] or BARD[6].

Regardless of the type of NN to consider, in an abstract way, a NN can be thought of as a function, $f : \mathbb{R}^d \times \Theta \to \mathbb{R}$, which takes as input a point x and a vector of parameters $\theta \in \Theta$. The parameters Θ will be *learned* from the information supplied to the model[104]. The output of the NN, $f(x, \theta)$, is a real value prediction of x, which, after being processed, is transformed into an image, text or chemical property. As mentioned above, the function f is structured as an interconnection of multiple non-linear functions (neurons) organised in layers where the input of any layer can be the output of the previous one. The layers can be divided into input, hidden, and output layers. Depending on the number of hidden layers, the model can be classified as shallow (one or two hidden layers) or deep (more than three hidden layers). The term deep learning comes from the use of models that have more than three hidden layers[105].

The basic unit of any NN are the deep layers or hidden layers. These structures linearly transform, through an affine transformation, an input vector x into an output vector y

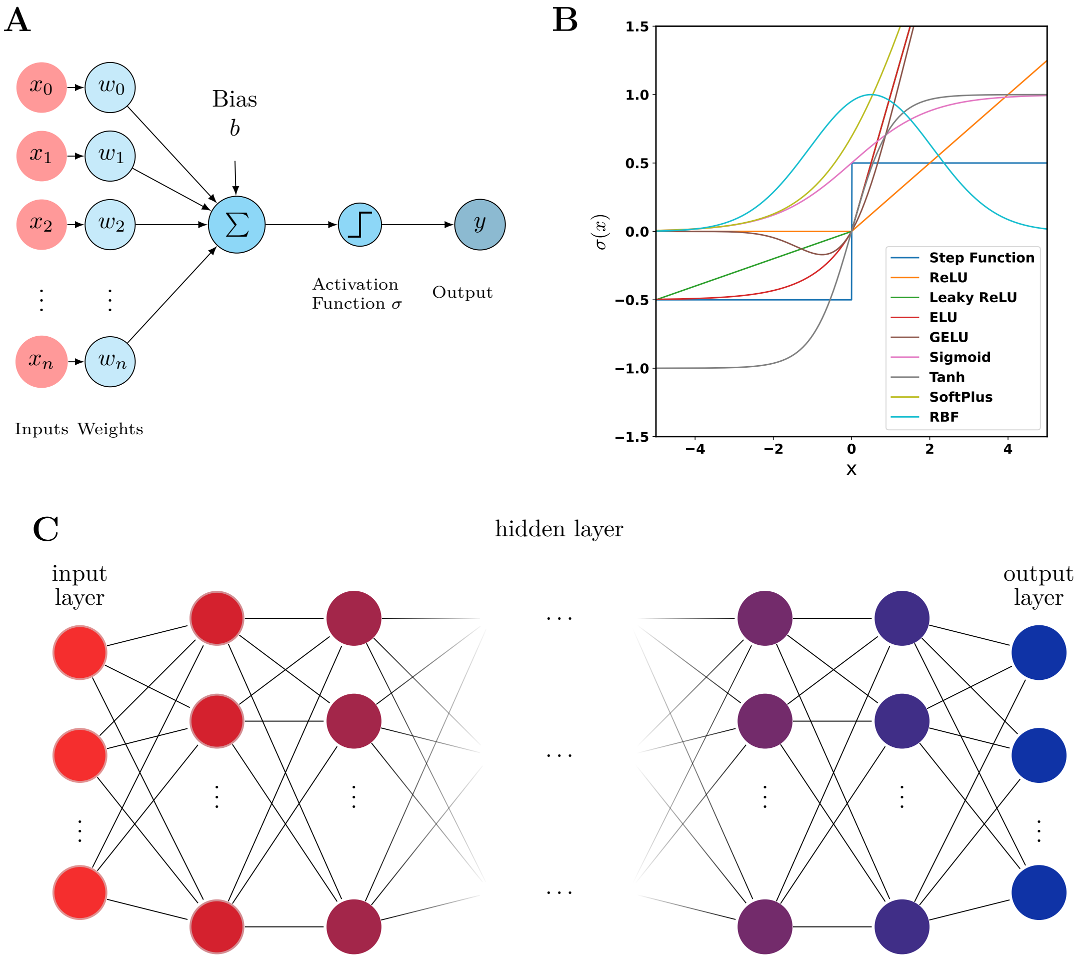

Figure 2.5: **Basic components of a neural network.** Panel A shows the diagram of the information process by a single neuron. An input is passed through a layer of weights that later are summed, and a bias value is added. Next, the sum is passed through a nonlinear function called activation function to obtain an output. Panel B shows some typical activation functions used in NN. Mathematical expressions for these functions are detailed in Table 2.1. Panel C displays the architecture of a fully connected neural network. The input is taken by an initial layer, called input layer, that passes to n-hidden layers to obtain a prediction by the last layer, called the output layer.

using the following expression:

$$y = \mathbf{W} \cdot x + \mathbf{b} \tag{2.15}$$

Here, $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{N,M}$ and $\mathbf{b} = \{b_i\}_{i=1}^{N}$ are the weights (a matrix) and biases (a vector),[96] M is the dimension of the input, and N is the number of nodes. This step has established a linear relationship between input and output; still, the ability to obtain non-linear relationships in NNs is thanks to the use of an activation function (Figure 2.5B and Table 2.1). After applying the activation function to Equation 2.15, it can be rewritten as:

$$\mathbf{h}_i = \sigma\left(\mathbf{W}_i \mathbf{x} + \mathbf{b}_i\right) \tag{2.16}$$

The function $\mathbf{h}_i$ in Equation 2.16 is called a *perceptron* and was originally proposed by Rosenblatt to model the synaptic process in the human brain[65].

Table 2.1: Examples of typical activation functions[3] used by neural network models. A graphical representation can be seen in Figure 2.5.

.

Name	Equation
Step function	$\sigma(x) = \begin{cases} -0.5 & x < 0 \\ 0.5 & x \geq 0 \end{cases}$
ReLu (Rectifier Linear Unit)	$\sigma(x) = \max(0, x)$
Leaky ReLu	$\sigma(x) = \begin{cases} a \cdot x & x < 0; 0 < a < 1 \\ x & x \geq 0 \end{cases}$
ELU (Exponential Linear Unit)	$\sigma(x) = \begin{cases} a \cdot (e^x - 1) & x < 0; 0 < a < 1 \\ x & x \geq 0 \end{cases}$
GELU (Gaussian Error Linear Unit)[4]	$\sigma(x) = 0.5x \cdot \left(\text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$
Sigmoid	$\sigma(x) = \frac{1}{1+e^{-x}}$
Hyperbolic Tangent	$\sigma(x) = \tanh(x)$
SoftPlus	$\sigma(x) = \ln\left(1 + e^x\right)$
RBF (Radial Basis Function)	$\sigma(x) = \exp\left(-a(x-c)^2\right.$

Equation 2.16 is the building block of any NN. By putting together two hidden layers with their respective activation functions, we obtain a shallow NN:

$$\mathbf{y} = \mathbf{W}_{i+1}\sigma\left(\mathbf{W}_i\mathbf{x} + \mathbf{b}_i\right) + \mathbf{b}_{i+1} = \mathbf{W}_{i+1}\mathbf{h}_i + \mathbf{b}_{i+1} \tag{2.17}$$

In principle, if enough neurons are used, this model can represent any continuous unidimensional function on a compact subset of the real line to arbitrary precision[105] as guaranteed by the universal approximation theorem[106–108]. However, deeper variants are usually preferred due to improved performance and parameter-efficiency.[109–112]

The functional form of a deep NN is characterized by the number of layers L and number of nodes N in a given layer. With increasing L and N, the functional form becomes more flexible. However, there is a larger risk of overfitting[5]. Careful attention should be given to this phenomenon since the obtained form of an NN has no underlying physical meaning.[114]. Mathematically, a fully connected deep NN (Figure 2.5C) is given by the following relation

$$\mathbf{y} = \mathbf{W}^L_{i_L}\sigma(\mathbf{W}^{L-1}_{i_L i_{L-1}}\sigma(\cdots\sigma(\mathbf{W}^1_{i_2 i_1}(\sigma(\mathbf{W}^0_{i_1 i_0}\mathbf{x} + \mathbf{b}^0_{i_0}) + \mathbf{b}^1_{i_1})\cdots) + \mathbf{b}^{L-1}_{i_{L-1}}) + \mathbf{b}^L_{i_L} \tag{2.18}$$

The output of Equation 2.18 is usually followed by a linear transformation (Equation 2.15 in the final output layer to yield the prediction $\mathbf{y}_{L+1}$. This type of NN is known by different names; examples are Multilayer Perceptron (MLP), Fully Connected Neural Networks (FCNN) or Feed-Forward Neural Networks (FFNN). An important aspect of FCNN is that the processed information moves in a unique direction through the different layers of the model. Alternatively, the output of a given layer can be feedback to itself. In that case, the model is called Recurrent Neural Network (RNN)[113].

Convolutional Neural Networks

FCNN present practical complications regarding the amount of information required for training and memory management because of the large number of parameters required

[3] Usually only the Sigmoid activation function is denoted as $\sigma(x)$, while others are denoted as $f(x)$. However, to keep the notation general in the following, $\sigma(x)$, denotes any activation function.

[4] The error function is define as[98]: $\mathrm{erf(x)} = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$. Note that $\mathrm{erf}(\infty) = 1$.

[5] Overfitting is a phenomenon that appears when the model describes the statistical peculiarities of the training data[105]. Another way to measure overfitting is when the difference between the training error and the test error is too large[113]

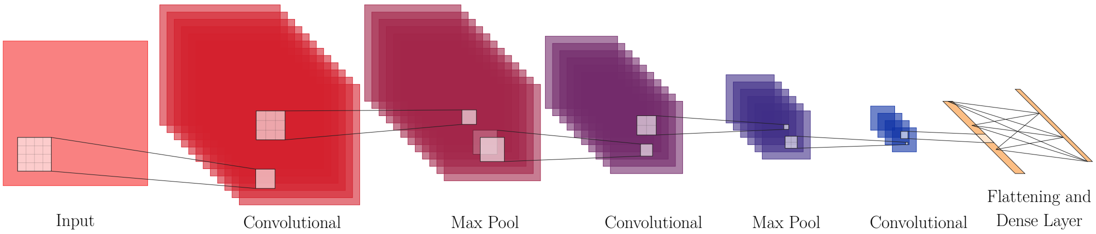

Figure 2.6: **Convolutional neural network.** The more common operations (Convolution, Max Pool, Flattening and Dense layer) are illustrated.

for training[105]. A way around this complication comes from the field of image processing with the invention of the Convolutional Neural Network (CNN). In this type of NN, the number of parameters is reduced by extracting predictive information from the input vector by applying a convolutional filter[74]. Nevertheless, the main difference between FCNN and CNN is that in the latter, a convolution is used in at least one layer instead of general matrix operations[113]. CNNs were designed to work with grid-structured inputs, which have strong spatial dependencies in local regions of the grid[115]; this is why these models have been mainly used in image processing.

A CNN usually contains multiple layers that perform three essential operations: convolution, pooling and activation (Figure 2.6). This last one uses the ReLU activation function. First, we will describe the convolutional operation[6]. In this operation, an input vector x is transformed to an output vector z so that each element z_i of the output vector is a weighted sum of the nearby points[105]. Mathematically, this is written as:

$$z_i = (x \star w)_i = \sum_{j=1}^{N} w_{ij} x_j \tag{2.19}$$

In Equation 2.19, w is the weight vector, sometimes refer as kernel, that transforms the input vector x. The output of Equation 2.19 is called feature map[113]. It must be notice that here the convolution is denoted with $\star$ although this symbol is reserved for cross-correlations. This notation is used to be congruent with software implementations[102].

As mentioned before, the convolutional operation reduces the number of parameters used by the NN model, known as sparse interactions (or sparse connectivity). Parameters are reduced by using a kernel matrix smaller than the input vector[113]. The application of a smaller kernel matrix results in a loss of information. Multiple convolutions are used in parallel to avoid this, creating the so-called channels[74]. A second consequence of the loss of information is the reduction of the grid size. The solution

[6] It must be mentioned that in signal processing, a continuous convolution is defined as $s(t) = (x \circledast w)(t) = \int x(a)w(t-a)da$. However, in ML, the discrete convolution is used.

to this complication is the operation of padding, which adds information around the borders of the feature map to keep a constant grid size[115]. During the convolution operation, each kernel element can be used in different positions. This implies that the parameters of the NN are shared, which implies a considerable gain in terms of memory management and statistical efficiency[113]. However, to guarantee that the parameters can be shared, the convolutional layer needs to be equivariant[7] (covariant) to translation.

The second operation in a CNN is pooling. The pooling operation works on small grid regions of size $P_q \times P_q$ in each layer. The output of this operation is the maximum value inside the sub-region of the grid in which the operation is applied. This is called max-pooling[115]. The pooling operation can be understood as a statistical summary of responses in a neighbourhood. Therefore, translations in the input vector do not modify the final output, implying that the pooling operation makes the model invariant to translations[113].

Graph Neural Networks

As mentioned before, CNNs were designed to work in regular grids. This is a problem when applying CNN in chemistry because chemists usually work with datasets ordered on irregular grids. This implies that the data is sparsely distributed over the grid. A solution to this problem is considering the data is organized in a graph. A graph, formally defined, is a tuple $G = (V, E)$ of a set of nodes V and a set of edges E, where each edge $e \in E$ connects pairs of nodes in $v \in V$[83]. Nodes are usually related to atoms, while edges are related to the bonds in the molecule. Graphs are very natural in chemistry because this type of representation has been used since the 19th century. By describing the data in a graph, CNN can be generalised to irregular domains, giving place to Graph Neural Networks (GNN).

Depending on how the information is shared through the nodes and their special distributions, there are several types of GNN. These are Graph Convolutional Networks, Graph Attention Networks, and Message-Passing Neural Networks (MPNN)[116]. The expressivity of the NN increases with the complexity of the information transmission. However, this is at the cost of interpretability, scalability, or learning stability[116]. In

[7] Equivarance means that if the input changes, the output changes in the same way. Specifically, an function $(f(x))$ is equivariant to the operation g if $f(x)$ satisfy the condition that:

$$f(g \cdot x) = g \cdot f(x)$$

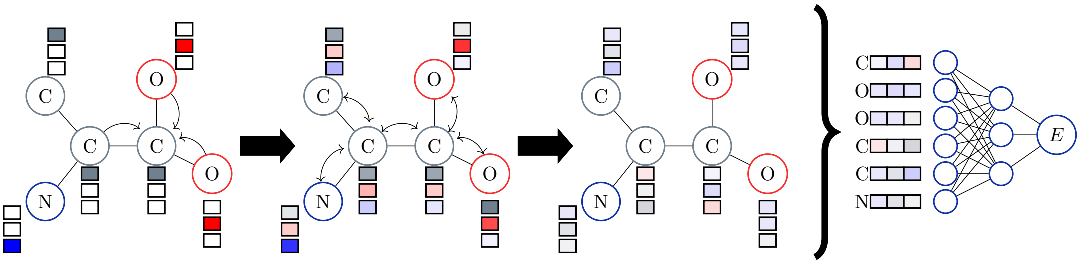

Figure 2.7: **Graph neural network.** Schematic representation of the graph neural message passing process. In the initial step, each atom in the alanine molecule is initialized with an embedding vector. Then, the message is created by the sum of the message from all atoms (cf. Equation 2.20). In the next step, the message is updated between all atoms for the embedding vectors. Finally, in the readout phase, the message obtained is passed through an MLP that is used to predict a property.

chemistry, MPNNs are the most commonly used so that we will explain more about them in the following.

In MPNNs[117], each atom belongs to one of the nodes in the graph (Figure 2.7). Each node is associated with an embedding vector (node characteristics). The embedding vector, x_v, may contain atom properties such as its type or charge[118] or it can be randomly initialized[32, 119]. The initial embedding vector is also called the initial hidden representation, usually represented as $h_v^0 = x_v$. Complementary molecular bonds are represented by the edges of the graph, e_{vw} (edges characteristics). Vector e_{vw} might be the bond order or a quantity that depends on the interatomic distances. Quantities, x_v and e_{vw}, build the graph $G = (x_v, e_{vw})$ over which the MPNN algorithm operates.

The algorithm of MPNN consists of two phases: a message-passing phase and a readout phase (Figure 2.7) [120]. The first phase calculates the message, m_v^{t+1}, where t is the number of steps to obtain the message. Using the hidden representation and the edges vectors, the message is defined as the of all neighbours of atom v in the graph G, $N(v)$

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \tag{2.20}$$

Here, M_t is the message function, which can be an MLP, a linear function, a concatenation, or a max pooling[74]. The next step of the message-passing phase is the update of the hidden representations. This step uses an update vertex function U_t. Then, the hidden representation is updated according to:

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \tag{2.21}$$

where m_v^{t+1} is the message obtained from Equation 2.20. The function U_t is also an MLP, which can be refined during training.

The last phase of the MPNN algorithm is the readout. In this phase, an embedding vector for the complete graph is obtained by using a function R defined as[120]:

$$y = R(h_v^T | v \in G) \tag{2.22}$$

The function R in equation 2.22 operates over the set of nodes. Then, it must be invariant to node permutations in order to make the MPNN model invariant to graph isomorphisms[8]. In summary, the MPNN algorithm is equivalent to applying a convolutional operation over each of the nodes of the graph[122].

So far, we have treated graphs as an abstract object. Nevertheless, it should be noticed that a graph is a purely topological object that specifies how nodes are connected but does not have information about their spacial arrangement (geometry)[123]. This looks like a downside for GNN. However, it becomes a large advantage because it is possible to encode the physics of the system of interest by adding additional information about the geometry of the system to the embeddings. Adding physical constraints helps construct more data-efficient models; additionally, the prediction and generalisation capacities increase[34, 123]. The conjunction at the nodes of a graph of an embedding vector and coordinates is known as a geometric graph[116]. The use of this new object by NN models leads to the surge of a new branch of deep learning called geometric deep learning [124, 125]. This new branch has found a lot of applications in chemistry and drug discovery[126] because the use of graphs is very natural for chemistry. However, the use of only a graph is not sufficient for property prediction[122], so the use of geometric graphs in conjunction with an MPNN is necessary.

The addition of geometric constraints helps to enforce symmetry conditions. In quantum chemistry, the models must follow some symmetric conditions. In general, the model is required to be equivariant (covariant) to the Euclidean transformations (translation, rotation, and reflection) of the Euclidian group $E(3)$[116]. This is achieved by providing geometric information of the system to the model. The first MPNN models were

[8] An isomorphism is defined in group theory if a map between two groups G and H, $f : G \rightarrow H$ conserves the multiplicative order of G, this is $f(g1)f(g2) = f(g1g2)$, with one-to-one mapping[121]. In particular, two graphs can have the same connectivity but different orders.

invariant to the transformations of the $E(3)$ group because initially, only scalar properties (e.g. Energy) that are independent of the reference systems were predicted. Then, the hidden representation is built from scalar quantities like intramolecular distances, angles or dihedrals that will be later passed through the operations of the MPNN[123]. Examples of this type of NN model are SchNet[119] or PhysNet[32] which encode intramolecular distances. This was later extended by DimeNet[127], which includes angular information, and then GemeNet[128], which rounds the description by adding dihedral angles.

New developments have focused on predicting vectorial properties (atomic forces) and tensorial values (atomic polarizabilities, dipole and quadrupole moments). In that case, the condition of equivariance must be enforced. The models that fulfil this condition are called equivariant neural networks (ENN). There are two ways to construct ENN. The first is by modifying the message to include directional information[129, 130]. The second option is to use spherics harmonics for the construction of the hidden representation[131]. Finally, the latest developments that combine the two techniques mentioned and add extra information to represent many-body interactions is the MACE model[132].

Transformers

The last NN model that will be discussed in this work is the so-called transformers. Abstractly, a transformer can be considered a special case of a GNN in which the graph is fully connected[116, 133]. This type of NN architecture was designed to work with sequential information and in automatic language processing[105]. A transformer consists of two parts (Figure 2.8). The first is an encoder that embeds an input vector. The second part is a decoder that transforms the embedding vector into an output.[134] The key concept behind the huge success of the transformer model is the attention mechanisms[135].

To understand the attention mechanism, we must go back to the first applications of transformers in natural language processing. A characteristic of language is that words are distinctively connected, and this connection depends on the words themselves[105]. In a sentence, certain words have more importance than others, and words only gain value if they are in a certain position. Considering this, the NN model must give more *attention* to certain words than to others while considering their order. This is impossible to achieve with an FCNN because the number of required parameters scales very fast[105]. In addition, in FCNN, the correlation between words can not be conserved after training[136]. The proposed solution to these complications is the attention mechanism, which preserves the correlation between different words in a text by creating an embedding space in which a vector can be translated to a location that depends on the other vectors in a sequence[136]. In a chemical example, we can think of a protein as a one-dimensional sequence of amino acids. Once the protein is folded, amino acids, which in a one-dimensional sequence are separated, may get close enough to interact and induce characteristic stable structures. It is also known that a protein will have a function that depends on the type and position of the amino acids. This is the reason why transformers played a key role in AlphaFold[137].

The attention mechanism can be formulated as follows. Let us consider a set of input vectors $X = \{x_1, x_2, \ldots, x_n\}$ in an embedding space that are wished to be transformed into a set of output vectors $Y = \{y_1, y_2, \ldots, y_n\}$. The transformation must follow the condition that the element y_n depends on all the elements of the set X. The simplest way to achieve this is by making the output y_n a linear combination of the vectors in the set X.

$$y_n = \sum_{m=0}^{N} \alpha_{nm} x_m \tag{2.23}$$

In equation 2.23, the α_{mn} values are known as attention weights. These can be interpreted as the attention that the output vector y_n puts on the input vector x_m[105]. The attention weights should follow the conditions that: i. $\alpha_{nm} > 0$ and ii. $\sum_{m=1}^{N} \alpha_{mn} = 1$.

Equation 2.23 describes a linear relationship. However, in practice, the transformation must be nonlinear. The nonlinearity is induced by applying extra operations to the input vector. Then, two transformations, such as the ones in Equation 2.15, are applied separately to the input vector. After this, two new quantities called queries vector (q_n) and keys vector[9] (k_n) are obtained. In the next step, the dot product between queries and keys is computed, followed by applying a special activation function called *softmax*[113] that ensures the constraints of the attention weights. Then, the coefficients α_{nm} in Equation 2.23 are obtained as:

$$a_{mn} = \mathrm{SoftMax}(k_m^\top \cdot q_n) = \frac{e^{k_m^\top \cdot q_n}}{\sum_{j=1}^{N} e^{k_j^\top \cdot q_n}} \tag{2.24}$$

The interpretation of the dot product between the keys and queries vector is a similarity measure of each element in the queries vector with each element on the keys vector. In place, the softmax function makes the values of the query vector compete to have a

[9] The names of keys and queries are inherited from the information retrieval field.

larger contribution to the final value[105].

The expression in Equation 2.24 can be generalized to all the values in the set X. Therefore, writing 2.24 as a matrix is more convenient. Then, we redefine Equation 2.24 as:

$$Y = \mathrm{SoftMax}\left(\frac{\boldsymbol{K}^{\top} \cdot \boldsymbol{Q}}{\sqrt{D_{q/k}}}\right) \tag{2.25}$$

Where K and Q are the keys and query matrices, respectively. The denominator $\sqrt{D_{q/k}}$ is the number of rows in K and Q. Equation 2.25 is known as self-attention because the same values are used to determine the matrices of keys and queries[136]. The denominator in Equation 2.25 avoids large values, which may dominate the output and make the training step easier[105].

Equation 2.25 is also known as an attention head in which all outputs depend on the input vectors. A problem that arises by using a single attention head is that the attention coefficients are more focused on the entries than in the context, resulting in an averaging of the correlation between inputs and a loss of individual effects[136]. The solution to this problem is the use of multiple attention heads. In that case, the input vector is split into n parts for which the attention coefficients are obtained. The final output is the concatenation of the results for all attention heads.

The described multi-head attention layer is the heart of the transformer architecture; see Figure 2.8. However, other operations are required to complete a transformer layer. The first operation required is the input embedding. As in the case of GNN, each input vector element must be transformed into an embedding vector. This is done in a process called tokenization. In this process, the input vector is decomposed into small units (tokens) of a large vocabulary of possible tokens. In chemistry, the vocabulary can be defined as the organic functional groups in a molecule or the amino acids in a protein[134]. The obtained tokens are passed through a positional encoding layer. This step is necessary to conserve the positional correlations between tokens because the multi-head attention layer is covariant to permutations. Then, the results from this layer would be the same independently of the token's positions [105]. In the positional encoding layer, a vector of positions handcrafted or learned by the NN is added to the vector of tokens.

The next operation on a transformer is normalisation and addition. In the addition layer, a special type of connection called residual connection is used. In a residual connection,

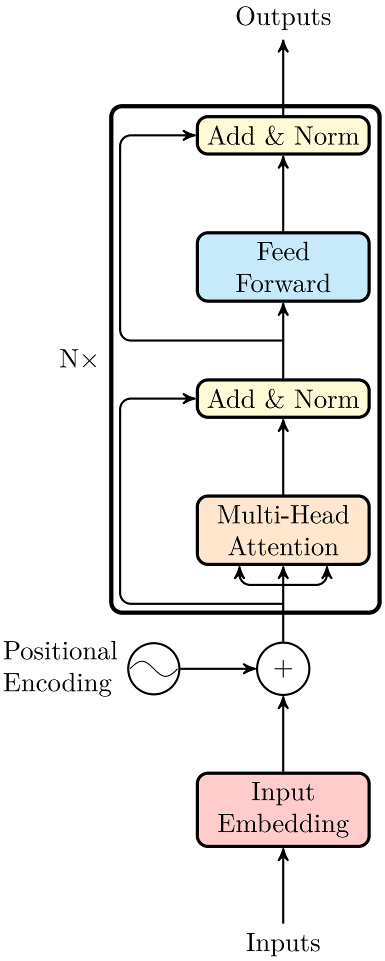

Figure 2.8: **Transformer architecture**. This diagram represents the basic operations of a transformer model from the initial encode, which consists of input embedding and positional embedding. The embedded input is passed to the multi-head attention together with a residual layer. The output of the multi-head attention is then passed through a residual layer with an FCNN and two addition and normalization layers. This basic unit can be repeated N times. For example, GPT3 uses 96 of these layers stacked on top of each other.

the output value of the previous operations is added to the initial value. Next, the normalisation layer adjusts the vectors by taking the average and standard deviation of the weights in the layer so that they have variance equal to unity and average equal to zero. Both operations facilitate the training of the model and function as regularisers. The output of the normalisation layer is passed to an MLP, which increases the model's flexibility since the output vectors are kept within the subspace created by the input vectors. Finally, the residual and normalisation operation is repeated to obtain a transformer layer. Generally, a transformer contains multiple transformation layers stacked on top of each other. Let us examine the case of GPT3, which contains 96 transformation layers, each with 96 multi-headed attention layers with 175 billion parameters in total with 300 billion tokens[5, 105]. Due to their large size and diversity of information on which they are trained, models like GPT3 are known as Foundational Models[138]. In chemistry, the first example of a foundational model has recently been released with more than 30 applications in materials science using the MACE model[132] and trained on 150,000 crystal structures[139].

Usually, transformers are trained with a large amount of information to create models that are as general as possible. However, if these models will be applied to specific tasks, they need to be fine-tuned. For this purpose, some model parameters are optimised specifically for the task to be solved[134]. Another type of tuning strategy is context learning. In this case, the model learns to solve a task after a small set of examples is presented to it as a demonstration[140]. In chemistry, this has been exploited by using GPT3 to solve various tasks where information is scarce[141] or to predict functional and electronic properties of organic molecules[142].

Training

The last aspect of neural networks that we will review is their training, which is the process in which "learning" occurs. This step is independent of the type of architecture used and consists of adjusting the different parameters of the NN to reproduce the reference values provided. This is achieved by iteratively minimising a so-called Loss Function(LF). The LF to be used depends on the task to be solved. In regression settings, it takes the general form[104]:

$$\mathcal{L} = \frac{1}{N_{\text{data}}} \sum_{n=1}^{N_{\text{data}}} [\mathbf{y}(\mathbf{x_n};\theta) - \mathbf{t_n}]^m + \omega \tag{2.26}$$

Here N_{data} are the number of samples used during the training procedure, $\mathbf{y}(\mathbf{x_n};\theta)$ is the NN model that takes an input $\mathbf{x_n}$, and parametrically depends on θ, the NN parameters.

Values t_n are the reference values provided that the model should reproduce. The values of the exponent m can be 1, which corresponds to the absolute value and is commonly known as L^1 loss function, or 2, which is the squared error and corresponds to the L^2 loss function. The value ω is a regularization term that helps improve the model's generalizability and prevent overfitting. Besides the form in Equation 2.26, many others can be used; see, for example, Ref. [143]. The interested reader is referred to Ref.[144] for a complete overview of different loss function types.

The selected loss function to train the model must be optimised to obtain the optimal values of the model parameters. The optimization is a complicated process since the LF is a multidimensional object because each parameter creates a new optimization dimension. Additionally, the loss function is non-linear. As a consequence of these facts, the landscape of the loss function is very complex, with many local minima. As such, the optimization process must be done iteratively using numerical methods. Obtaining a global minimum requires exponential time and, in general, is very complicated[62]. In the same way, it is not desirable to obtain a global minimum as it would, most likely, correspond to an overfitted solution[145].

The preferred algorithm to optimise the chosen loss function is the so-called Gradient descent. It consists of two steps. In the first step, the derivatives of the loss function with respect to each of the parameters are calculated.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Theta}} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial \theta_1} \\ \frac{\partial \mathcal{L}}{\partial \theta_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \theta_n} \end{pmatrix}$$

In the second step, the parameters of the model $\boldsymbol{\Theta}$ are updated according to the following rule:

$$\theta_i^{t+1} = \theta_i^t - \alpha \left(\frac{\partial \mathcal{L}}{\partial \theta_i} \right) \tag{2.27}$$

Where α is a positive constant called learning rate. Equation 2.27 can not be used directly because it is inefficient with large amounts of data. Therefore, a sequential version called stochastic gradient descent is used in practice. In this version, a batch, which is a subset of the dataset used for training, is used. Then, the gradient is calculated only for the examples in the batch[105]. So, Equation 2.27 can be rewritten as:

$$\theta_i^{t+1} = \theta_i^t - \alpha \sum_{j \in \mathcal{B}_t} \left(\frac{\partial \mathcal{L}_j}{\partial \theta_i} \right) \tag{2.28}$$

Where $\mathcal{B}_t$ is the batch that contains the examples after all, batches in the training set are used, the process is repeated, and it is called an epoch. In each step, white noise is added to the gradient to help the algorithm escape local minima[105].

Further modifications to the gradient descent algorithm include the addition of momentum. This is convenient because information on the gradient of the previous step is added to the calculation. In that case, Equation 2.28 is modified as:

$$m_{t+1} = \beta m_t + (1-\beta) \sum_{j \in \mathcal{B}_t} \left(\frac{\partial \mathcal{L}_j}{\partial \theta} \right)$$

$$\theta^{t+1} = \theta^t - \alpha m_{t+1}$$

Here $\beta \in [0, 1]$ controls the degree to which the gradient is softened. Another popular modification to the stochastic gradient is the popular ADAM (Adaptative Moment Estimation)[146], which has become the standard way to train NN models.

A key aspect of the training of the NN is the acquisition of the derivatives of the loss function with respect to the parameters of the NN. This is done using the backpropagation algorithm, which is an iterative application of the chain rule of differential calculus and is inspired by dynamic programming. This algorithm consists of two steps. The first, called forward, is the evaluation of the NN on a given input. This is a problem because the weights and bias of the model need to be initialized to a certain value. There are several ways to do this. However, the rule of thumb indicates that the values should be randomly initialized with values close to zero. It must avoid zero values (derivatives become zero, and the parameters do not change) or very large values that end up in poor solutions[145]. An incorrect initialisation leads to instabilities in the backpropagation step, which can result in very large gradients, called the exploding gradients problem, or very small gradients, called the vanishing gradients problem[105]. The values obtained from the forward step are used to evaluate the loss function.

The next step of the algorithm is called backward because the error is fed back to the multiple layers of the NN[104]. In this step, the chain rule is used. The objective is to obtain the derivative of the loss function with respect to the neural network parameters.

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial y\left(x; \theta\right)} \cdot \sum_i \left(\frac{\partial y\left(x; \theta\right)}{\partial h_i} \right) \cdot \left(\frac{\partial h_i}{\partial \theta} \right) \tag{2.29}$$

here h_i is the i- hidden layer. For simplicity, indices for the parameters of the NN were omitted. By replacing equation 2.29 in 2.27, the parameters of the NN are updated. In practice, this step is put in matrix terms to facilitate the calculation of the derivatives. Packages such as PyTorch[147], Tensorflow[148] or Jax[149] contain an automatic differentiation mechanism, where the backward calculations are automatically generated. Consequently, the backward pass does not have to be implemented manually.

Chapter 3

How Data Affects Predictions of Chemical Properties by Atomistic Neural Networks

> It's more interesting to work on challenges where you don't know the answer. In chemistry, you should enter into an adventure with molecules.
>
> Ben Feringa

This chapter presents a comprehensive assessment of the effects of the database composition in chemical and configurational space for predicting a specific chemical property (i.e. Tautomerization energy). Different characteristics of the databases used for training are followed, and their impact on the prediction is evaluated. The results obtained in this study show that contrary to usual expectations, increasing the amount of data does not necessarily lead to better predictions. A second key conclusion from this work is that the lack of exploration in chemical space can be compensated with an adequate sampling of conformational space. The rest of the chapter is organized as follows: First, an introduction to the topic will be presented, followed by a description of the methods used. Next, results and discussion of those are presented. Finally, some of the more salient conclusions of the work are discussed.

The results presented in this chapter have been previously published in J. Chem. Theor. Comp., 2021, 17(8), 4769–4785

3.1 Introduction

In the last decade, the application of machine learning (ML) techniques in chemistry has significantly increased[35, 61, 96, 150]. This has occasionally been related to a paradigm shift, revolutionizing the available techniques to understand and simulate chemistry[16, 18]. The excitement is seemingly justified, given the outcomes of ML techniques' central promise that, by using a sufficiently large number of examples and a rule-discovery algorithm, it is possible to obtain a scientific understanding of the underlying relationships covered by the data[35, 150]. Furthermore, ML techniques are fast compared with quantum chemical methods, while also reaching comparable accuracy[91, 119, 151–159].

On the other hand, application of quantum ML methods to concrete problems requires large amounts of data which first need to be generated from electronic structure calculations[160–162]. Consequently, data generation is computationally demanding. An essential challenge for the extension of ML methods' applicability in chemistry is understanding how suitable databases can be constructed to maximize accuracy and transferability of the models. An important ingredient for this step is the degree and confidence with which a human can understand the relationship between cause (starting database and model) and result or observation (applying the model to a new task)[163, 164]. This process has also been called "interpretability" and it can be used to assess the relationships learned by the model or contained in the data used for training it[165, 166]. The present work quantifies the relationship between the composition of general-purpose quantum chemical databases trained with a ML model (Neural Network) with the performance for predicting a property of interest (here: tautomerization energy) on a set of unseen samples.

To test the effect of different databases on the reliability of the ML model, the problem of predicting tautomerization energies is considered. Tautomerism is a form of reversible isomerization involving the rearrangement of a charged leaving group within a molecule[167] (e.g. Figure 3.1A). One isomer transforms into the other by a heterolytic splitting followed by a recombination of the fragments formed[168]. This process involves the migration of one or more double bonds and atoms or groups. The isomers (i.e. tautomers) generated in this reaction are chemically independent species with defined properties[169]. It is known that this type of reaction is of importance for biological molecules such as amino acids[168], DNA[170, 171], RNA[172], and atmospheric processes.[173] Additionally, it is estimated that tautomerism can occur in up to two thirds of small molecules[174], and a majority of commercial drugs[175, 176].

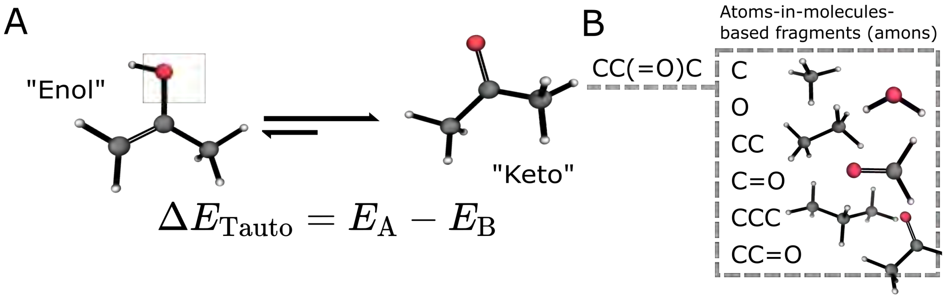
A
"Enol"
"Keto"
$\Delta E_{\text{Tauto}} = E_{\text{A}} - E_{\text{B}}$
B
Atoms-in-molecules-based fragments (amons)
CC(=O)C
C
O
CC
C=O
CCC
CC=O

Figure 3.1: (A) Tautomerism is a form of reversible isomerization involving the rearrangement of a charged leaving group within a molecule. The keto-enol tautomerism of acetone, an equilibrium which heavily favours the keto side, is shown as an example. (B) Chemical space can be decomposed systematically through the use of atoms-in-molecules-based fragments (amons).[40] The amons present in acetone (SMILES: CC(=O)C), as well as their corresponding SMILES are given as an example.

Despite its widespread occurrence and importance, quantitative studies of tautomerism are still challenging because small changes in molecular structure or solvent environment can dramatically change the tautomeric equilibrium[169, 177]. Moreover, small free energy differences between two tautomers in solution make the use of high level theoretical methods and an adequate basis set mandatory which limits its use for calculations of tautomerization energies and ratios. [177, 178] As an example, tautomerization in malonaldehyde (MA) is considered. MA has served as a prime example to develop and test computational methods for a realistic description of hydrogen transfer in small molecules.[162] Experimentally, the ground state tunneling splitting is 21.58314 cm^{-1} which has been determined by different experiments with very high accuracy. [179, 180] Furthermore, proton transfer rates in a di-imine derivative have been determined with nuclear magnetic resonance (NMR) spectroscopy. [181] Such experiments provide direct information on the barrier height separating the two tautomeric states “A” and “B”. Using a state-of-the art full-dimensional potential energy surface at the near basis-set-limit frozen-core CCSD(T) level of theory,[182] the tunneling splitting from quantum simulations was determined as 23.4 cm^{-1}.[183] Alternatively, using a reduced dimensionality Hamiltonian, the barrier height for proton transfer in a parametrized molecular mechanics with proton transfer (MMPT) potential was found to be 4.34 kcal/mol which yields a tunneling splitting of 21.2 cm^{-1}, consistent with experiment.[184, 185] This barrier height is close to the value from CCSD(T) calculations which yield 4.1 kcal/mol.[182] These examples illustrate that calculations at the highest levels of theory are required for quantitative studies of the energetics underlying tautomerization.

In the last decade, development of ML models has allowed the design of robust models that can routinely reach prediction errors lower than 1 kcal/mol with respect to the reference data at low computational cost[32, 119]. However, there have been few discussions on how databases can be improved/designed to obtain better predictions from the ML model. Ideally, the combination of a robust ML model and an adequate database will result in quantitative results for the prediction of a property of interest. The availability of public databases of tautomers[87, 186] makes the prediction of tautomerization energies using these ML models an ideal test case to study how different training databases influence the accuracy of ML methods.

The present work is structured as follows. First, the methods, databases, and the analysis performed are introduced. Next, the results for the tautomerization energy predictions using models trained on the different tested databases are presented. Additionally, prediction errors for tautomerization energies are analyzed. The effect of different characteristics of the training data on predicting the tautomerization energy and the individual molecules' energy are evaluated. Finally, the results are discussed and conclusions regarding the findings and interpretability of broadly conceived and learned ML models applied to a specific chemical question are drawn.

3.2 Methods

3.2.1 Machine Learning

PhysNet was used for the representation and evaluation of the data sets, using the hyperparameters from the original publication[32]. For training, only the nuclear charges (Z), the energies of the molecules (E) and their coordinates ($\mathbf{R}$) were considered. The energies used for training were those reported by the different databases minus the atomization energy at the given level of theory to ensure that energies of molecules are referred to the same zero of energy. In all cases, a training, testing and validation split of 8 : 1 : 1 was used. The loss function was

$$\mathcal{L} = w_{\mathrm{E}}|E - E^{\mathrm{ref}}| + \lambda_{\mathrm{nh}}\mathcal{L}_{\mathrm{nh}} \tag{3.1}$$

where E^{ref} is the reference energy, $w_{\mathrm{E}} = 1$ is the weighting hyperparameter for the energy, $\lambda_{\mathrm{nh}} = 10^{-2}$ is a regularization hyperparameter and the term $\mathcal{L}_{\mathrm{nh}}$ is a non-hierarchicality regularization penalty[187]

$$\mathcal{L}_{\mathrm{nh}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=2}^{N_{\mathrm{module}}}\frac{(E_i^{(m)})^2}{(E_i^{(m)})^2 + (E_i^{(m-1)})^2} \tag{3.2}$$

that penalizes higher order (m) many-body interactions over lower order ($m-1$) ones, i.e. ensures that e.g. the magnitude of 2-body interactions is overall larger than 3-body interactions. The loss function (Eq. 3.1) was minimized using AMSgrad with a learning rate of 10^{-3}. Overfitting was prevented by the use of early stopping, the convergence criteria considered is the saturation of the validation-loss function[32].

The NN architecture used here (PhysNet) belongs to a more general class of deep neural networks (DNNs) that have been successfully applied to ML of quantum chemical properties, including (but not limited to) SchNet [119], ANI-1[188], or HDNN.[151, 159] All of them perform well on typically used databases such as QM9. As the present work focuses on characterizing the performance of NN-based approaches for different databases, PhysNet[32] was used as a representative DNN.

3.2.2 Database Selection

For training the NNs, four widely used databases for benchmarking predictive models of DFT-based energies are employed, namely the QM9[189], PC9[51], ANI-1[190] and ANI-1x databases (not to be confused with the neural network potential carrying the same name).[191] An additional database, ANI-1E (where "E" stands for equilibrium) containing only the equilibrium structures of the ANI-1 database, was generated at the ωB97x[192]/6-31G(d) level of theory. These databases can be divided into categories based on the type of geometries they contain. The datasets consisting solely of equilibrium structures are QM9, PC9 and ANI-1E, and sample only chemical space. Contrary to that, the ANI-1 and ANI-1x databases contain equilibrium and non-equilibrium structures which sample chemical and conformational space.

Training sets: The QM9 data set[189] was generated as a subset of the GDB-17 chemical universe[49], consisting of 133885 molecules, containing less than or equal to 9 heavy atoms (either C, N, O or F). Reference energies were computed at the B3LYP[193, 194]/6-31G(2df,p) level of theory. For the present work, QM9 was filtered to include only molecules which passed a geometry consistence check[189] and considering only those containing carbon, nitrogen or oxygen atoms. The final size of the QM9 training dataset used here consisted of 110426 molecules.

The PC9 dataset was created as an alternative to QM9 to improve coverage of chemical space.[51] It is a subset of the PubChemQC[195] and is limited to molecules with 9 heavy atoms or less ($n_{\text{atoms}} \leq 9$). This database consists of 99234 molecules, calculated at the B3LYP[194]/6-31G(d) level of theory, and excludes enantiomers, tautomers, isotopes as well as other specific artifacts in PubChemQC[51]. PC9 also contains 5325 molecules with an electronic state different from a singlet which were removed for the present work. As in the case of QM9, molecules which contain fluorine were removed. The final size of this dataset was 85875 molecules.

ANI-1[190] consists of 24 million geometries generated using normal mode sampling from 57462 unique molecules. ANI-1 is a subset of the GDB-11 chemical universe[47, 48]. A related dataset, ANI-1x [191], was created using an active learning[196] procedure which reduced the original ANI-1 database to 5 million structures. Starting from the ANI-1 database[188, 190] the ANI-1E dataset was generated and consists only of the corresponding equilibrium structures. The new ANI-1E database contains 57462 molecules limited to eight heavy atoms (either C, O or N). The generation of this database is further described below in the subsection of electronic structure calculations.

Tautomerization energy evaluation set: The performance of the NN models described above was evaluated on a subset of molecules from Tautobase[87], a public database of 1680 tautomer pairs. The Tautobase was filtered to molecules only containing hydrogen, carbon, nitrogen, or oxygen atoms. The size of the final test set was 1257 tautomer pairs (2514 molecules). Of those 2514 molecules, 118 appear in two or more pairs. However there is no influence on the results for the prediction of $\Delta E_{\mathrm{Tauto}}$ because all the pairs are unique. The geometry generation and structural optimization for these molecules is described below.

3.2.3 Initial Geometry

To investigate the effect of the geometry of the molecules passed to the NN model on its performance, a second set of geometries for the Tautobase was also evaluated. These geometries were generated from the SMILES representation using OpenBabel[197] and were optimised with the MMFF94 force field[198].

Additionally, a subset of the test set composed of 34 tautomeric pairs which were part of the SAMPL2 challenge[199] were considered. Those 34 pairs (68 molecules) were optimized by six popular general atomistic force fields: CHARMM27[200], GAFF[201], OPLS[202], UFF[203], Gromos[204], and Ghemical[205]. Details on the generation of the geometries are reported in the SI.

3.2.4 Electronic Structure Calculations

Generation of ANI-1E: Starting from the ANI-1 database[188], a new data set, ANI-1E, was generated. From the SMILES strings provided by [188] initial geometries using OpenBabel[197] were generated. Subsequently, geometries were optimised using PM7[206] implemented in MOPAC2016[207], before a final geometry optimization and frequency calculation at the ωB97x[192]/6-31G(d) level of theory was performed using Gaussian09[208]. Finally, it was verified that the optimized structures did not exhibit imaginary frequencies. The calculations were performed in an iterative way. First, default thresholds of Gaussian09 were used. Molecules that did not converge in this first round were run with a calculation of force constants (*Opt=CalcFC*), increasing the number of maximum cycles (*maxcycles=1000*) and using quadratic convergence for the self-consistent field procedure (*SCF=QC*). In a third step, molecules that did not converge up to this point were computed with a tight convergence criterion (*Opt(tight)*) and an increase of the number of maximum cycles (*maxcycles=1000*). Structures that still could not converge were computed again with the option *Opt=CalcAll* that computes the force constants at each step. Additionally, an ultrafine integration grid (*int=ultrafine*) together with quadratic convergence (*SCF=XQC*). Finally, for problematic structures that still did not converge *verytight* convergence criteria and a maximum number of cycles of 5000 were used. With this procedure, minima for all 5000 structures without imaginary frequencies were obtained.

Tautomerization evaluation set: The molecules used for the evaluation of the NN models were generated from the SMILES provided in Ref. 87 using the OpenBabel software[197]. These structures were then optimized at the all levels of theory used to conceive the databases (QM9: B3LYP/6-31G(2df,p), PC9: B3LYP/6-31G(d), ANI: ωB97x/6-31G(d)) using Gaussian09[208]. Thus, when working with PC9, geometries optimized at the B3LYP/6-31G(d) level of theory were passed to the NN trained with PC9 to obtain single isomer energies. Those values were used to compute the tautomerization energy ΔE_{Tauto} defined as the energy difference between tautomers A and B in their optimized structures, see Figure 3.1. This was repeated in the same fashion for all other databases.

Table 3.1: Overview of the training datasets used in this work. QM9, PC9 and ANI-1E contain equilibrium structures and can be considered to only sample chemical space, whereas ANI-1 and ANI-1x also contain non-equilibrium geometries which sample conformational space. The number of molecules refers to the total number of data in each dataset.[a]Structures generated through normal mode sampling. [b]Training set selected using active learning.

Database	Number of Molecules	Level of Theory	Parent Universe
QM9	128908	B3LYP/6-31G(2df,p)	GDB-17
PC9	85870	B3LYP/6-31G(d)	PubChemQC
ANI-1E	57462	ωB97x/6-31G(d)	GDB-11
ANI-1[a]	24 million	ωB97x/6-31G(d)	GDB-11
ANI-1x[b]	5 million	ωB97x/6-31G(d)	ANI-1

3.2.5 Comparison of Structural Properties of Different Databases

As a way to compare the composition of the different datasets evaluated in terms of structural properties (e.g. bond lengths), a Gaussian kernel density estimation[209] of their distributions was generated, see Figures S7 to S10. The similarities between the distributions used to train the NN models and those from the test set of tautomers was quantified by computing the relative entropy (or Kullback-Leibler (KL) divergence)[210]

$$D(p \parallel q) = \int_{-\infty}^{\infty} p(x) log \left(\frac{p(x)}{q(x)} \right) dx \tag{3.3}$$

This metric quantifies the overlap between a reference distribution $p(x)$ and a target distribution $q(x)$. Because the KL divergence is not symmetric ($D(p \parallel q) \neq D(q \parallel p)$), it is important to specify which distribution is used as the reference. In the present work, the Tautobase is the target distribution and QM9, PC9, ANI-1E are the reference distributions. The KL divergence allows quantification of how much information of the reference databases (i.e. QM9, PC9, ANI-1E) is 'missing' to best cover the information contained in Tautobase. If the two distributions are identical, $D(p||q) = 0$. On the other hand, if the reference database $p(x)$ (here QM9, PC9, ANI-1E) contains more information than the target set $q(x)$ (Tautobase), $D(p||q) > 0$, and if specific information is missing, $D(p||q) < 0$. Hence, cases for which $D(p||q) < 0$ are of particular relevance if improvements of the reference databases are sought for better capturing ΔE_{Tauto}.

3.2.6 Chemical Space “Coverage” from Fragment Analysis

The ‘coverage’ of chemical space contained in the Tautobase, QM9, PC9 and ANI family of databases was analysed by considering their atom-in-molecule-based fragments, referred to as “amons”.[40] The amons are generated from the SMILES representation of the molecule, see Figure 3.1B. This representation is used to construct a molecular graph from which sub-graphs to a maximum number of atoms (excluding hydrogen) are generated. All sub-graphs are checked to be valid and unique. Here, amons up to and including a maximum of five heavy atoms were generated by an in-house script following the published algorithm[40] which is available on GitHub. The official implementation is given in Ref. 211.

3.2.7 Visualization of Chemical Space

Visualization of chemical space is a complex task given the high dimensionality of it. On this work it would be used a recent development is the TreeMAP (TMAP) algorithm, which allows for an interpretable, low dimensional representation of the test set’s chemical space[86]. The TMAP algorithm constructs a weighted graph that is efficient for compact representations of high dimensional data. The necessary weights are based on the Jaccard distance, which measures the dissimilarity between the fingerprint of two structures. This graph is then pruned to the minimum spanning tree, a fully connected, acyclic sub-graph containing all nodes of the parent graph, and retaining only essential edges which minimize the weights. This organizes the compounds contained in a database into a tree, putting related structures on nearby branches. From this, groups of related moieties can be identified which are potentially detrimental to predicting the quantity in question, here ΔE_{Tauto}.

3.3 Performance of the NN on the Tautobase

3.3.1 Overall Performance

The mean absolute errors for the tautomerization energies ΔE_{Tauto} range from 1.68 kcal/mol (ANI-1x) to 4.59 kcal/mol (ANI-1). The results are summarized in Table 3.2 for all molecules in the test set and graphically reported in Figures 3.2 (datasets with equilibrium structures) and 3.3 (datasets with both equilibrium and non-equilibrium structures). The prediction errors for the energy of single isomers, E_{SI}, with respect to DFT energies are also reported (Table 3.2). Note that because the tautomerization energy is defined as the difference between the isomer energies, predictions of tautomerization energies are often more accurate due to cancellation of systematic errors. On the other hand, the energies for single isomers are considerably larger and span a much wider range because they scale with the number of atoms that make up a molecule. Therefore, the NN-based energies for larger molecules are expected to be associated with considerably larger errors.

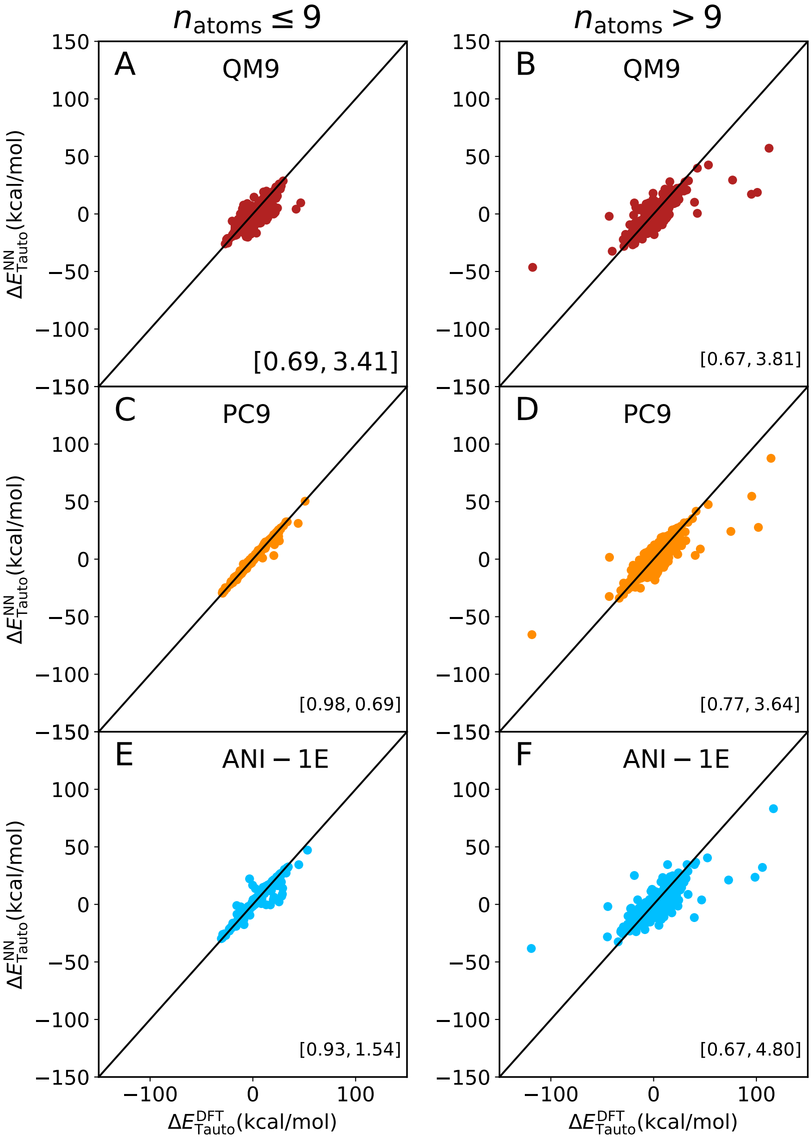

Figure 3.2: Correlation between the calculated (DFT) and predicted (NN) tautomerization energies ($\Delta E_{\text{Tauto}} = E_{\text{A}} - E_{\text{B}}$) for molecules with $n_{\text{atoms}} \leq 9$ (left) and $n_{\text{atoms}} > 9$ (right) for the models trained on the datasets which only cover chemical space. Pearson correlation coefficients (r^2) and the Mean Absolute Error (MAE) values are reported in brackets as [r^2, MAE].

To assess whether the accuracy for predicting tautomerization energies correlates with the performance of the trained NNs on the chemical databases, the QM9, PC9, and ANI-1E models are considered. The MAEs on held-out test sets for each of the respective training runs are 0.10 kcal/mol (QM9), 0.69 kcal/mol (PC9), and 0.27 kcal/mol (ANI-1E) which is comparable to values ranging from 0.19 kcal/mol to 0.30 kcal/mol for the QM9 data set (depending on the sizes of the training and validation sets used).[32] However, when applying these trained NNs to evaluate Tautobase, the MAEs are 3.67 kcal/mol (QM9), 2.60 kcal/mol (PC9), and 3.66 kcal/mol (ANI-1E), respectively. Hence, there appears to be no correlation between the quality of the trained NNs, defined as the MAE on the test set split from the training database, and their performance on Tautobase.

Table 3.2: Mean Absolute (MAE) and Root-Mean-Squared Error (RMSE) for the prediction of tautomerization energy ΔE_{Tauto}, and the single isomer energies, E_{SI}, for the entire Tautobase (1257 tautomeric pairs) for each of the datasets. The 95 % confidence interval (given in brackets) was computed from bootstrapping.

	ΔE_{Tauto}		E_{SI}	
Database	MAE	RMSE	MAE	RMSE
QM9	3.67 (3.32,3.99)	7.12 (5.50,8.18)	5.00 (4.77,5.29)	8.40 (7.87,8.92)
PC9	2.60 (2.33,2.86)	5.41 (4.04,6.28)	6.90 (6.50,7.38)	13.20 (12.39,13.98)
ANI-1E	3.66 (3.51,3.98)	7.09 (5.50,8.14)	15.20 (14.95,15.59)	17.50 (16.83,17.99)
ANI-1	4.59 (4.26,4.92)	7.56 (6.56,8.29)	13.40 (12.99,13.80)	17.00 (15.16,18.26)
ANI-1x	1.68 (1.55,1.81)	2.85 (2.30,3.15)	1.80 (1.67,1.91)	3.60 (3.13,3,89)

Next, the performance of the trained models for predicting ΔE_{Tauto} and E_{SI} depending on the number of heavy atoms is assessed. For this, results for the subset of molecules with $n_{\text{atoms}} \leq 9$, referred to as "SetLE9" in the following, is considered separately from those with $n_{\text{atoms}} > 9$, which is "SetG9". This distinction is motivated by the fact that the PC9 and QM9 databases contain structures with only up to 9 heavy atoms, i.e. models need to extrapolate for larger structures. For SetLE9, the PC9 (Figure 3.2C) and ANI-1x (Figure 3.3C) data sets perform best. Both achieve a MAE < 1 kcal/mol with respect to the DFT values for ΔE_{Tauto}.

The extrapolation to SetG9 increases the prediction errors for most of the databases studied. Again, the ANI-1x database performed best for ΔE_{Tauto} with a MAE of 2.20 kcal/mol, followed by PC9, QM9, ANI-1E and ANI-1 with a MAE of 6.29 kcal/mol. The number of atoms in the database also influences the Pearson correlation coefficient r^2 for the same number of points for both sets studied here. A better correlation is observed when the size of evaluated molecules is in the range covered by the training database, i.e. for SetLE9. Here, the correlation coefficients range from 0.69 (QM9)

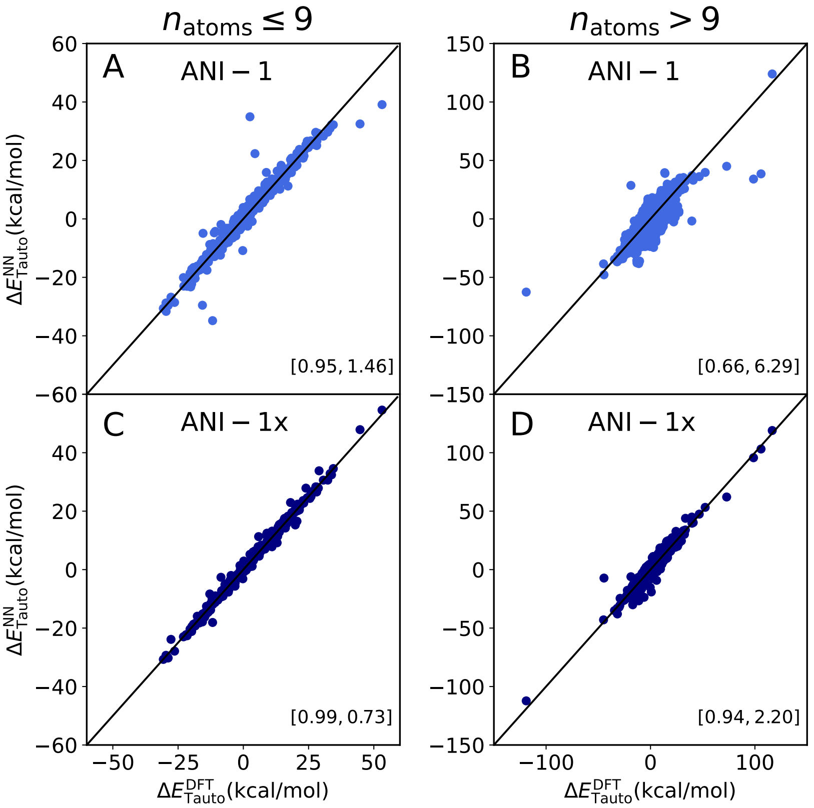

Figure 3.3: Correlation between the calculated (DFT) and predicted (NN) tautomerization energies ($\Delta E_{\text{Tauto}} = E_{\text{A}} - E_{\text{B}}$) for molecules with $n_{\text{atoms}} \leq 9$ (left) and $n_{\text{atoms}} > 9$ (right) for the models trained with the datasets which cover chemical and conformational space. Pearson correlation coefficients (r^2) and Mean Absolute Error (MAE) values are reported in brackets [r^2, MAE].

to 0.99 (ANI-1x). For SetG9, the r^2 values are significantly lower. A particularly noteworthy case is the QM9 database, which shows almost the same MAE and r^2 for both subsets of the Tautobase (Figures 3.2A and B). The performance for SetLE9 and SetG9 also differs in the number and magnitude of outliers, see Figures 3.2 and 3.3.

The RMSE in Table 3.2 show that the spread for E_{SI} could be a reason for large outliers (see Figure S1). It is likely that there is some error cancellation when predicting tautomerization energies (i.e. energy differences). For example, if the trained NN predicts too large energies for both isomers, this systematic error cancels when their energy difference is computed.

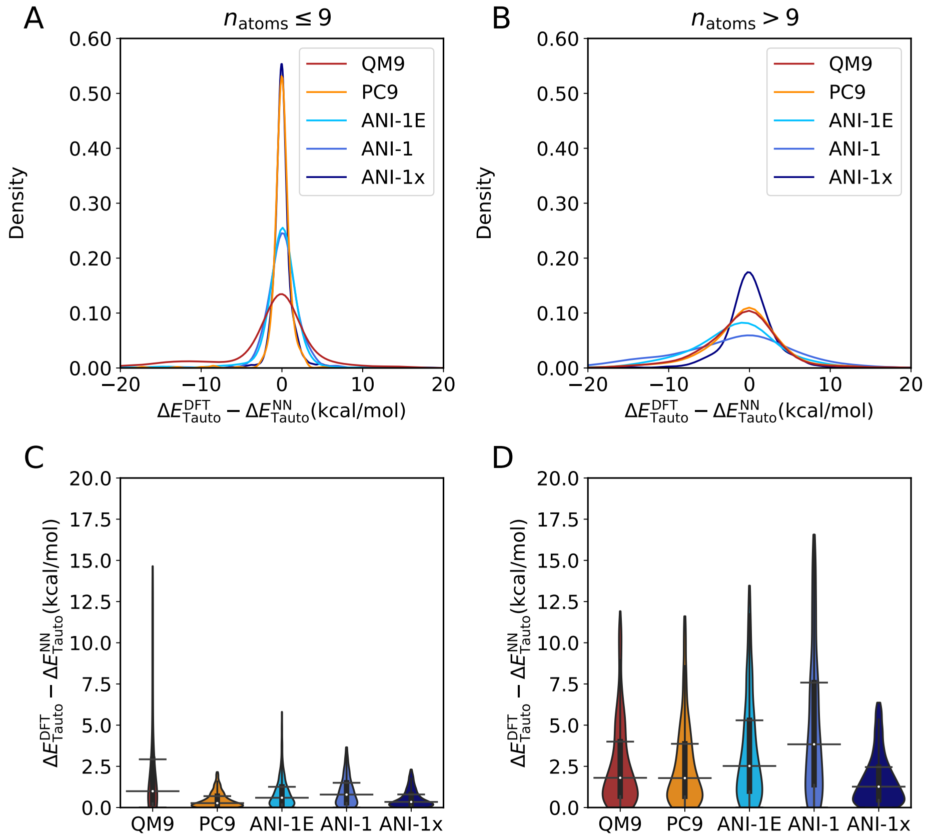

Figure 3.4: Error analysis on the prediction of the tautomerization energies. Panels A and B: Kernel density estimate of the error on the prediction of the tautomerization energies for the different databases evaluated in this work. Panels C and D: Normalized error distribution up to the 95% quantile of the different datasets for the errors in the tautomerization energy. The blackbox inside spans between the 25% and 75% quantiles with a white dot indicating the mean of the distribution. The whiskers indicate the 5% and 95 % quantiles. The panels on the left and right are for SetLE9 and SetG9, respectively.

3.3.2 Error Analysis

Next, prediction errors for ΔE_{Tauto} and E_{SI} are analyzed and discussed for all trained models. In Figures 3.4A and B, the kernel density estimate of the error distribution is reported. The violin plots in Figures 3.4C and D show the spread of errors, which helps to identify large outliers.

In all cases studied, the distribution of errors for $\Delta\Delta E_{\text{Tauto}} = \Delta E^{\text{DFT}}_{\text{Tauto}} - \Delta E^{\text{NN}}_{\text{Tauto}}$ is centered around zero. This indicates that the highest probability is to obtain a correct prediction. The width of the distribution depends on the reference dataset and on the number of atoms in the molecule. For SetLE9 (Figure 3.4A), the error distributions for PC9 and ANI-1x suggest high probabilities ($p(\Delta\Delta E_{\text{Tauto}}) = p(\Delta E^{\text{DFT}}_{\text{Tauto}} - \Delta E^{\text{NN}}_{\text{Tauto}})$), around 60 %, to obtain a small error. Conversely, QM9 performs worst with a maximum height of only $p(\Delta\Delta E_{\mathrm{Tauto}}) \sim 15$ % and a faint maximum below $\Delta\Delta E_{\mathrm{Tauto}} = -10$ kcal/mol to predict such energy differences with a larger probability than a positive value. The error distributions for ANI-1E and ANI-1 are similar in shape which indicates that their performance is comparable although the number of structures in ANI-1E is one order of magnitude smaller than that in ANI-1. Hence, adding additional structures (ANI-1 vs. ANI-1E) does not necessarily improve performance.

On the other hand, for SetG9 (Figure 3.4B), the performance of QM9 and PC9 is comparable given the similar shape of their distribution of $p(\Delta\Delta E_{\mathrm{Tauto}})$, ANI-1x gives the best predictions with a maximum height of around 15 % to obtain an error for $\Delta E_{\mathrm{Tauto}}$ close to zero. All other reference data sets perform inferior with ANI-1 reaching only a 5 % $p(\Delta\Delta E_{\mathrm{Tauto}})$ for a prediction close to zero. In addition, for most of the data sets the error distribution is asymmetric with an increased probability to predict a negative value for $\Delta\Delta E_{\mathrm{Tauto}}$ compared to a positive value.

Results for the normalized error distributions $p(\Delta\Delta E_{\mathrm{Tauto}})$ are shown in Figures 3.4C and D. For SetLE9, PC9 and ANI-1x show the smallest outliers by magnitude with an error below 2.5 kcal/mol. On the other hand, QM9 has the largest outliers with some errors larger than 15 kcal/mol. The average error for all reference distributions is around or below 1 kcal/mol for the 75 % quantile. For molecules in SetG9, ANI-1 has the largest outliers, followed by ANI-1E, QM9, PC9, and ANI-1x performing best with a maximum error of around 5 kcal/mol, see Figure 3.4D.

For completeness, error distributions $p(\Delta E_{\mathrm{SI}}) = p(E_{\mathrm{SI}}^{\mathrm{DFT}} - E_{\mathrm{SI}}^{\mathrm{NN}})$ for individual molecules and their normalized variants are also reported in Figures S1A to D. For SetLE9, the distributions for PC9 and ANI-1x are centered around zero with peak heights at 80 % which decreases to 25 % for ANI-1E. For ANI-1 it is shifted to negative and for QM9 to positive values. For SetG9 (Figure S1B), all error distributions are asymmetric and extend to large negative values of ΔE_{SI}. The best and worst performing reference distributions are ANI-1x and ANI-1, respectively. The normalized error distributions (Figures S1C and D) for both sets are strongly peaked. For SetLE9 (Figure S1C) the maxima for PC9 (2.5 kcal/mol) is the lowest whereas ANI-1 has the largest errors. For SetG9 (Figure S1D), the outliers are even more pronounced with $|\Delta E_{\mathrm{SI}}| > 100$ kcal/mol for ANI-1E and ANI-1. In general, the performance of ANI-1E is better than that of ANI-1 with a smaller MAE, outliers of smaller magnitude and a more compact distribution. These results are surprising, given the large difference between the size of both datasets (ANI-1E (≈ 57k) and ANI-1 (≈ 20M)) and confirm the earlier observation that addition of new structures to a database does not necessarily improve performance.

In summary, for SetLE9 the database with broader chemical diversity (PC9) and the database with the widest sampling of chemical and conformational space (ANI-1x) perform best. Hence, chemical diversity is essential for faithful prediction of ΔE_{Tauto} but it can be substituted to some extent with adequate sampling of conformational space. For larger molecules (SetG9), the best results are obtained by ANI-1x which suggests that sampling of conformational space improves extrapolation to larger molecules. Datasets containing only equilibrium structures perform similarly for predicting ΔE_{Tauto}.

3.4 Effect of Different Database Characteristics on Predictions

This section analyzes the predictive power of the NNs trained on the five different training databases for tautomerization energies by considering various chemical properties such as the number of heavy atoms, the number of atoms of a given element, or the type of chemical bonds. Given the non-linear nature of the NN, the relationships between these characteristics and whether/how they are related to the performance of the model is a challenging task. The features studied here were selected because they might be considered for the selection of a training database for the prediction of a chemical property (in this case the tautomerization energy) or because they can be optimized for the generation/enhancement of datasets used to train models for specific purposes.

3.4.1 Number of atoms

The first characteristic considered was the number of heavy atoms (C, N and O) contained in the reference data sets and how this affects the prediction quality on Tautobase. For SetLE9 the MAE for ΔE_{Tauto} typically decreases with increasing molecular size, see Figure 3.5A, for all five reference data sets. This can be broadly related to the increase in the number of samples with the number of heavy atoms contained in the reference databases used for training (see Figure S2). For all data sets except for QM9, the MAE decreases to levels below 0.5 kcal/mol as the number of heavy atoms increases.

For larger molecules (SetG9) in the Tautobase, the MAEs increase significantly, see Figure 3.5B. Broadly speaking, for up to 25 heavy atoms the MAE is still within 5 kcal/mol but increases considerably for larger molecules. ANI-1x performs best with $\mathrm{MAE} < 1$ kcal/mol up to $n_{\text{atoms}} = 25$ but errors increase above 10 kcal/mol beyond that. This is followed by QM9 and PC9 which, on average, have MAEs of ~ 2 kcal/mol followed by ANI-1E and ANI-1. Typically, the MAE for ΔE_{Tauto} is around or below 5 kcal/mol irrespective of the number of samples, see Figure S3. However, for the largest molecules for which there are only few samples, and for a small number of mid-sized molecules ($n_{\text{atoms}} \sim 20$) the MAE exceeds 10 kcal/mol.

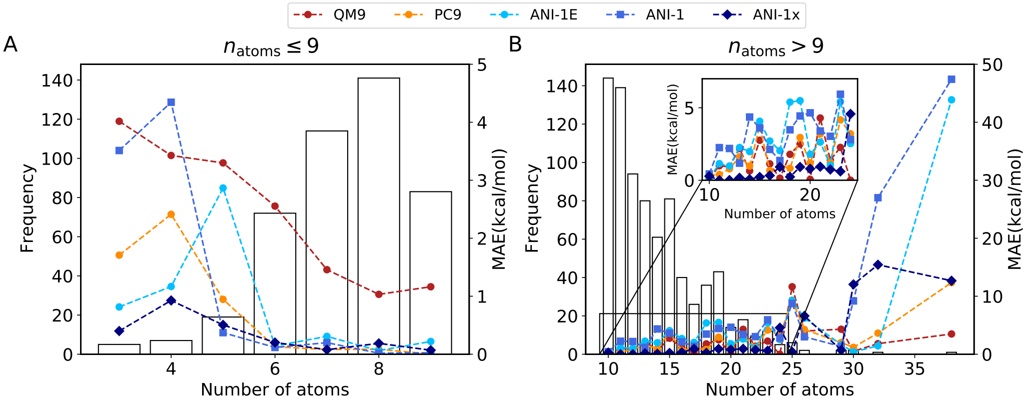

Figure 3.5: Mean Absolute Error (MAE) by number of heavy atoms (C,O,N) on the molecule for the tautomerization energy. The number of molecules for increasing number of heavy atoms is shown as a histogram. Panel A: Results for SetLE9, i.e. molecules with $n_{\text{atoms}} \leq 9$. Panel B: for SetG9. The inset in panel B shows the MAE for ΔE_{Tauto} for $10 \leq n_{\text{atoms}} \leq 25$.

The MAE for predicting E_{SI} for SetLE9 (Figure S4A) is large for molecules with 3 and 4 heavy atoms which differs from the findings for ΔE_{Tauto}. With increasing size the error decreases. This is most pronounced for PC9 which eventually achieves the same quality as ANI-1x. For E_{SI} the overall shape of MAE vs. n_{atoms} for databases which only contain equilibrium structures is related but the magnitude of the MAE differs. This is a consequence of the chemical diversity of the databases, see subsection 3.4.4. For ANI-1 and SetLE9, the MAE is smallest for $n_{\text{atoms}} = 5$ and then starts to grow again. For SetG9 (Figure S4B), the MAE displays a steady increase with the number of heavy atoms.

In summary, ANI-1x dataset performs best across all values of n_{atoms} for the Tautobase, followed by PC9 across most values for n_{atoms}. For SetG9, QM9 is quite reliable whereas ANI-1 and ANI-1E perform worst which reiterates the earlier finding that adding perturbed structures to a data set does not necessarily improve the quality on the task at hand (which is the estimation of ΔE_{Tauto}). Consequently, the results can be worse than those obtained when training only on equilibrium structures.

The analysis can be refined by considering predictions for ΔE_{Tauto} depending on the number of C-, N-, and O-atoms contained in the molecules of the reference database (Tautobase), see Figure 3.6. For SetLE9 the MAE tends to decrease (except for QM9) with increasing number of carbon atoms as shown in Figure 3.6A whereas for SetG9 it increases to different extents depending on the reference database considered, see Figure 3.6B. For nitrogen and oxygen atoms and ANI-1x all MAEs for SetLE9 and SetG9 are small (~ 1 kcal/mol), except for the largest numbers of N-atoms, see Figure 3.6F. For the PC9, ANI-1E, and ANI-1 databases and SetLE9 all MAEs are below or around 1 kcal/mol whereas for QM9 they can be larger. For SetG9, the MAEs are up to 5 kcal/mol for molecules for which at least tens of representatives are contained in Tautobase, but start to increase significantly below that, see Figure 3.6D and F.

Considering the MAE for E_{SI} confirms these general findings, see Figure S5. Molecules with a small number of atoms of a given element have fewer different chemical environments (see Figure S6). This makes it more difficult to predict ΔE_{Tauto} if that chemical environment is present in the target data set (Tautobase) but not sampled in the reference sets. Consequently, larger errors are observed for molecules with few atoms of a given element.

3.4.2 Structural Composition of the Chemical Databases

The structural diversity of the databases can also be quantified in terms of the bond types that are covered. It can be assumed that the NN model learns that specific composition, and consequently, if the database used for training a NN model covers a large range of bond lengths, better results are expected. In the following, bond length distributions in the reference databases of equilibrium structures (PC9, QM9, and ANI-1E) are compared with the distributions contained in Tautobase. Figures S7 to S10 show that the reference and target distributions have a different coverage of bond lengths. The general finding is that for SetLE9 the overlap between reference and target distributions is better than for SetG9.

Figure S7 shows that C-C single bonds between C(sp^3) atoms are well covered for the three reference databases compared with Tautobase. The C(sp^2)-C(sp^2) double bonds

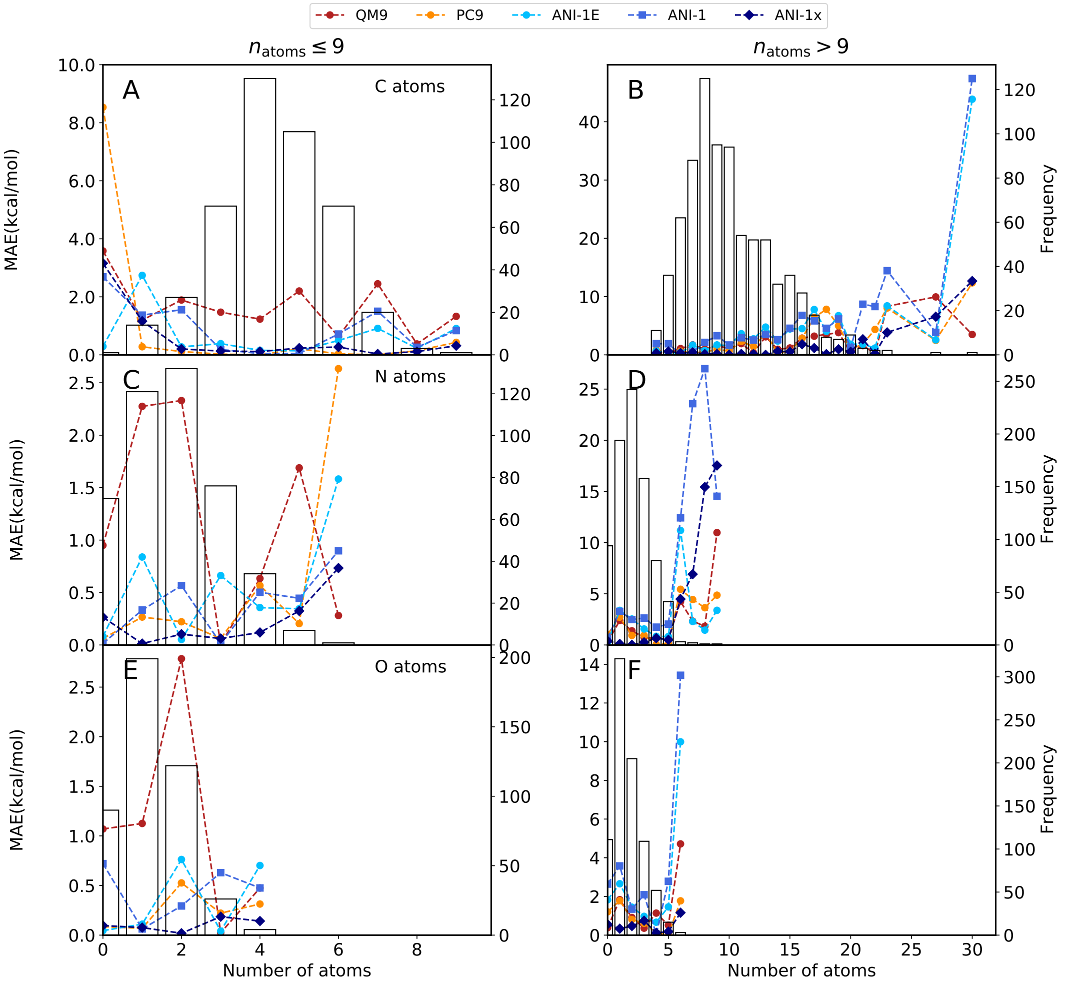

Figure 3.6: Mean Absolute Error (MAE) by number of atoms of a given element for the tautomerization energy. A histogram of the number of molecules for different numbers of heavy atoms is shown in the background. Panels A and B show the results by number of carbon atoms. Panels C and D shows the results by number of nitrogen atoms. Finally, panels E and F show the results by number of oxygen atoms. Left panel shows results for molecules with $n_{\text{atoms}} \leq 9$ the right for molecules with $n_{\text{atoms}} > 9$.

are covered differently for the reference datasets: QM9 has the fewest examples of this type of bond, whereas ANI-1E shows the best coverage. Such bonds are important for large molecules ($n_{\text{atoms}} > 9$) because of the presence of aromatic rings (Figure S8). Double C(sp^2)-C(sp^2) bonds close to hetero atoms are poorly covered by all reference datasets. Those bonds are crucial because they are the main origin of tautomerization rearrangement.

C(sp^2)-N double bonds (Figure S7) are abundantly present in the Tautobase. However, the coverage of the reference datasets of that type of bond is heterogeneous; ANI-1E shows the best coverage followed by PC9 and QM9. On the other hand, C(sp^2)-N bonds close to a heteroatom, more prevalent in larger molecules, are better covered by QM9 than PC9 whereas C(sp)-N bonds are well covered by all three databases. Carbon-Oxygen bonds for carbonyl groups are more predominant in SetLE9 and are well covered for the reference databases. Bonds for enols, esters and others are important for the Tautobase; PC9 covers such C-O bonds sufficiently but it is poorly sampled for QM9. Lastly, while C-O bonds of the type of alcohols and dialkyl ethers are most sampled for the reference databases they are least important for the Tautobase.

A quantitative measure for the overlap of two distributions is the KL divergence $D(p||q)$, see Equation 3.3. The KL divergence analysis indicates that the coverage of the reference sets is heterogeneous, see Tables S1, S2 and Figure 3.7. There are several types of C-C bonds that are insufficiently covered, such as C(sp^2)-C(sp^2) single and double bonds, or C(ar)-C(ar) bonds. Also, certain types of C-N bonds would require more data as the bonds involving C(sp^3) and C(sp^2) with different types of nitrogens. Coverage of C-O bonds by the reference databases displays a bias toward alcohols, ethers and esters. Finally, N-N are the types of bond that show a more diverse coverage between databases, with some cases for which QM9 has a good coverage (N(3)-N(2) and N(2)-N(2)(aromatic)) but a poor coverage for N(3)-N(3). Interestingly, there are cases for which QM9 has a good coverage, whereas ANI-1E and PC9 are deficient. Figure 3.7 shows that none of the reference databases covers all of the predominant types of bonds present in the Tautobase.

Next, the MAE for a specific number of a particular type of bond (e.g. C-C, C-O, or C-N) was determined for single isomer energies, see Figure 3.8. The results in Figure 3.8A show that for C-C bonds and SetLE9 the error for PC9 (orange) and ANI-1x (black) is constant and well below 1 kcal/mol. On the other hand, for QM9 (red) the error oscillates without following a clear trend. ANI-1E (light blue) and ANI-1 (dark blue) behave similarly to one another with a smaller MAE for ANI-1E than the one for ANI-1.

For C-O bonds, the MAE of the prediction of E_{SI} slowly increases for PC9 but remains well below 1 kcal/mol, whereas for QM9 it starts at above 5 kcal/mol and decreases to below 1 kcal/mol but always remaining above that for PC9, see Figure 3.8C. The error for the database of the ANI family is largely constant over the number of bonds. For ANI-1 the MAE oscillates between 1 kcal/mol and 2 kcal/mol, whereas for ANI-1E and ANI-1x the MAE is well below 1 kcal/mol, except for zero C-O bonds and ANI-1E. Considering C-N bonds (Figure 3.8E) it is found that their maximum number is larger than that for C-O bonds. The magnitude for the MAE for this bond type is at least a

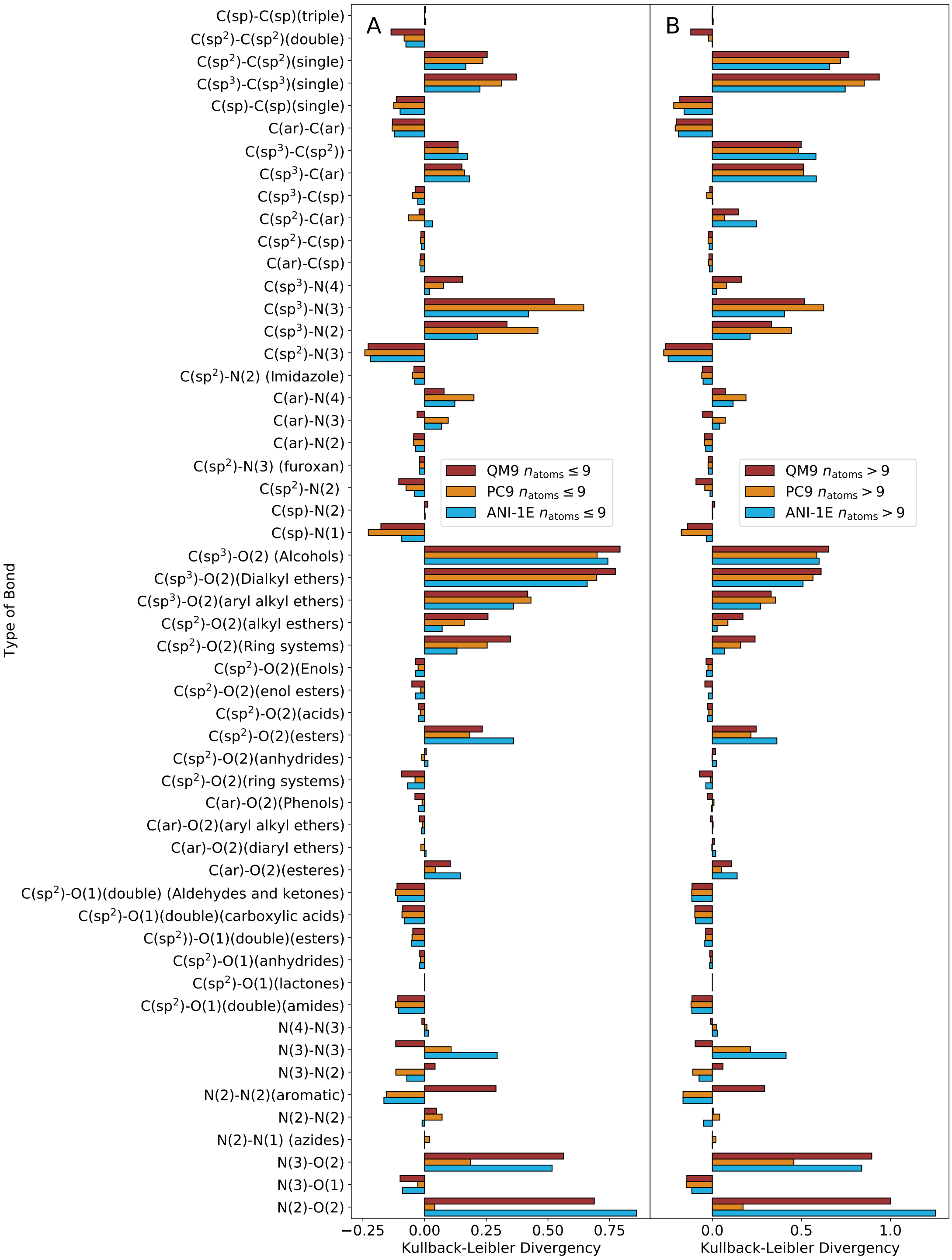

Figure 3.7: Values of the Kullback-Leibler divergence for different types of bonds present in the reference databases (QM9,PC9 and ANI-1E) compared with Tautobase.Panel A corresponds to SetLE9 and Panel B corresponds to SetG9.

factor of three larger than that for the C-C and C-O bonds, respectively. Again, PC9 and ANI-1x perform best, followed by ANI-1E (except for molecules with only one C-N bond). The MAE for QM9 slowly decreases whereas that for ANI-1 is constant at below 2 kcal/mol up to five C-N bonds after which it sharply increases.

Regarding C-O and C-N bonds, it is clear that the good coverage of ANI-1E helps to reach small MAE when the number of bonds increases (Figure S7). These results show that PC9 has a good overall performance because there is an adequate coverage of different chemical bond types whereas QM9 and ANI-1E have biases toward some types of bonds (Table S1 and S2). It should be stressed though that such an analysis excludes the fact that the same type of bond can behave differently given different chemical environments.

For SetG9 the increase of the error with decreasing number of samples is more apparent. As discussed before, the MAE observed for larger molecules ($n_{\text{atoms}} > 9$) grows proportional with the number of bonds (See Figure 3.8 B, D and F). In this regard, ANI-1 and ANI-1E are the databases with largest growth rate, followed by PC9 and finally QM9 and ANI-1x. The low MAE for E_{SI} by number of bonds obtained with ANI-1x is a consequence of the addition of an adequate number of non-equilibrium structures. This suggests that a lack of chemical diversity can be partially compensated by including non-equilibrium structures in a database.

3.4.3 Initial Geometry

In the previous sections, the energy of the molecules was computed using the equilibrium geometry of the tautomeric pairs computed at the level of theory of the various reference databases used for training the NNs. However, in practice, it would be of interest to sidestep the computationally rather expensive optimization of the structures in the reference dataset (here Tautobase) at the density functional or even higher level of quantum chemical theory. For this, using empirical force fields is a possibility. A recent study for 3271 small organic molecules ($n_{\text{atoms}} < 50$), similar to those contained in Tautobase, found typical RMSDs of 0.25 Å to 1 Å between optimized structures at the B3LYP/6-31G* level of theory compared with those optimized by nine different force fields, including GAFF, MMFF94, OPLS and others.[212] Considering this, it is interesting to assess the performance of the NN-based models on FF-optimized geometries.

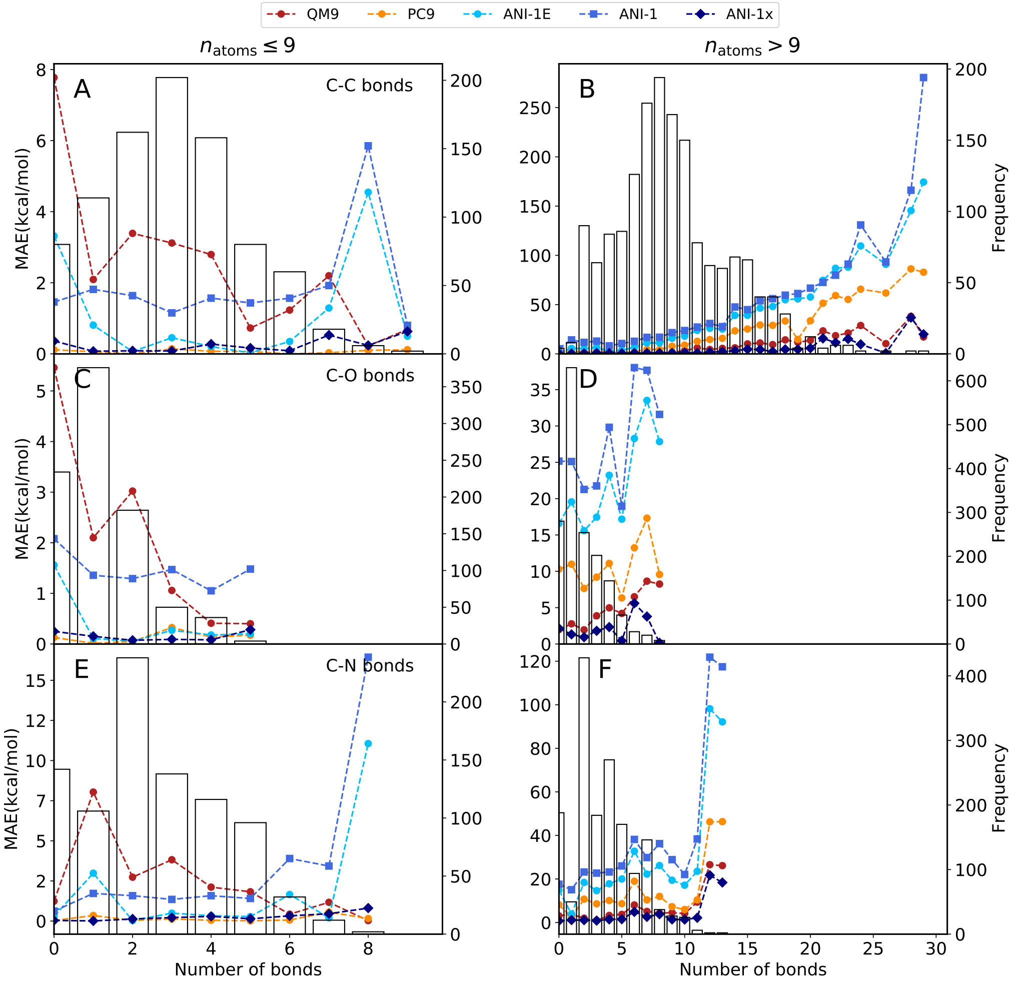

Figure 3.8: Mean Absolute Error (MAE) of the prediction of the energy of single isomers by number of bonds involving carbon atoms. The number of molecules for different numbers of bonds is shown as a histogram. Panels A and B correspond to the results of C-C bonds, panels C and D show the results for C-O bonds and panels E and F show the results for C-N bonds. Left panel for SetLE9 and right panel for SetG9.

For this analysis, the geometries for molecules from Tautobase were optimized as described in section 3.2.3 with the MMFF94 force field and then used to evaluate the tautomerization energies using the five trained NNs. Table S3 shows that the MAE for the tautomerization and single isomer energy increases for all evaluated models when the geometry used to evaluate the energies differs from geometries optimized with the respective *ab initio* method. In all cases the MAE for ΔE_{Tauto} increases by a factor between 1.5 and 3 compared with the error obtained using the optimized geometries at the quantum chemical level of theory used to train the NN (see Table 3.2). A similar effect is observed for E_{SI}. It is noticeable that this geometry effect is less pronounced for databases which contain non-equilibrium structures: ANI-1 shows the smallest increase of the MAE for ΔE_{Tauto} compared with results from using optimized geometries at the appropriate level of theory.

Normalized distributions for $\Delta E_{\mathrm{Tauto}}$ using MMFF94 geometries for SetLE9 (Figure S11C) indicate that the datasets which only cover chemical space (QM9, PC9 and ANI-1E) perform similarly with the highest values of the outliers close to 15 kcal/mol. Conversely, the ANI-1 and ANI-1x databases have a more compact distribution with the maximum values for outliers around 10 kcal/mol. The most challenging case for predicting tautomerization energies is for SetG9 with geometries generated with MMFF94 (Figure S11D). This demonstrates that the geometry that is used to evaluate a trained ML model strongly influences the performance. This is also found for all other datasets with outliers larger than 20 kcal/mol, except for ANI-1x with a maximum of 10 kcal/mol. In conclusion: scoring a model trained on minimum energy structures computed at a given level of quantum chemical theory can not be done using optimized structures from an empirical force field (or from structures at a sufficiently different level of quantum chemical theory).

To confirm this finding, molecules from the SAMPL2 challenge[199] were evaluated using geometries generated by six popular force fields (see Section 3.2). The results (see Figure S12) do not show a correlation between the RMSD between the reference structure from DFT calculations and the FF-optimized structure and the corresponding energy differences. There are molecules with $\mathrm{RMSD} \leq 0.1$ Å which display a significant error (> 5 kcal/mol) in the energy predicted by the NN and vice versa. This, again, confirms the finding that using FF-based (minimum energy) structures for evaluating a NN-based model trained on (minimum energy) quantum chemical reference data can lead to considerable errors.

3.4.4 Visualization of Chemical Space

To understand the influence of the different databases studied on the performance of the models, it is of interest to analyze the coverage of ‘chemical space’. Firstly, molecules in the databases were deconstructed and their constituent amons (unique chemical fragments) were enumerated (Figure 3.1B). PC9 contained the largest number of unique amons (8424), followed by QM9 (3929) and, finally, the ANI family of databases (1663). There is significant overlap of common amons between the datasets (Figure 3.9A, which suggests that they cover similar regions of chemical space. Regarding the overlap of the test set (Tautobase) with the databases tested, PC9 is the one which covers the most amons by number in the reference set, followed by ANI-1E and, QM9

(Figure 3.9B).

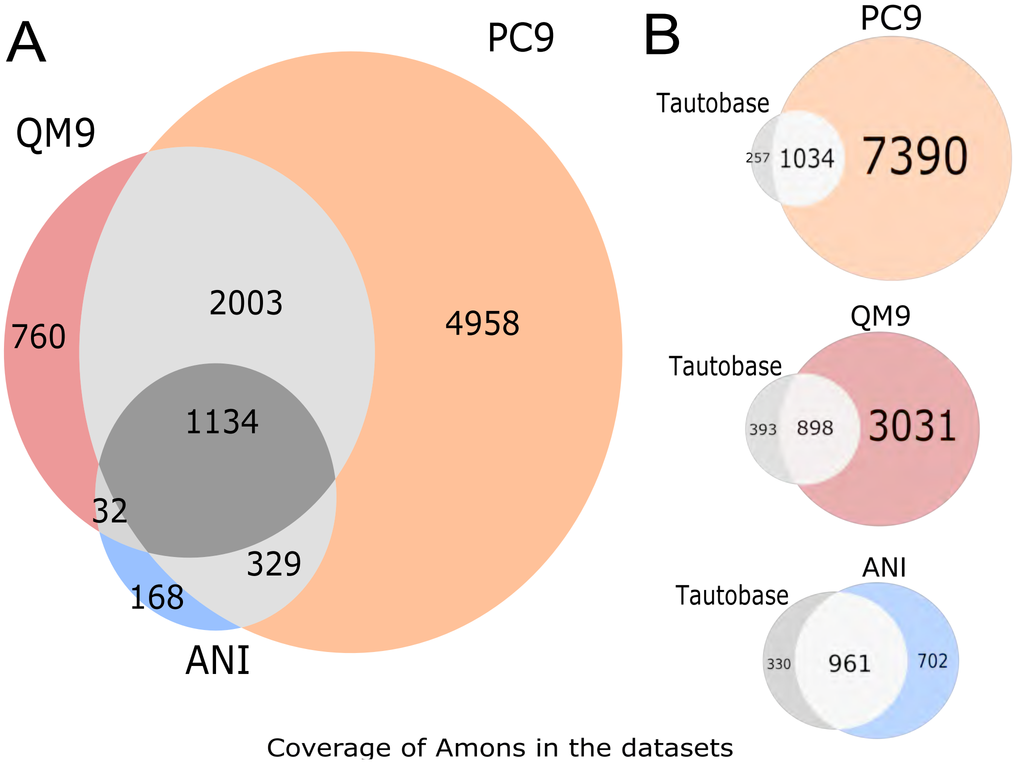

Figure 3.9: A) Venn diagram showing the overlap of amons between the QM9[189], PC9[51] and ANI[190] family data sets. B) Overlap of amons up to length five between three popular quantum chemistry datasets (QM9 (69.5 %), PC9 (80.1 %) and ANI-1E (74.4 %)) and the 'Tautobase'[87], a collection of experimentally observed tautomers. Molecules containing other atoms than hydrogen, carbon, nitrogen, and oxygen were filtered from all datasets.

It is also of interest to consider coverage of chemical space (as quantified by the overlap of amons between reference and target sets) and to compare this with the prediction errors for ΔE_{Tauto}. The database that contains the largest number of amons of the Tautobase is PC9 (1034) which also has the smallest MAE for ΔE_{Tauto} (2.60 kcal/mol). This compares with ANI-1E (961 amons, $\Delta E_{\text{Tauto}} = 3.66$ kcal/mol) and QM9 (898 amons, $\Delta E_{\text{Tauto}} = 3.67$ kcal/mol). Considering SetLE9 and SetG9 separately, PC9 performs best for both, with MAEs of 0.69 kcal/mol and 3.64 kcal/mol, respectively. For SetG9, containing larger molecules, it is noted that the MAE for QM9 (898 amons) is 3.81 kcal/mol, compared with 4.80 kcal/mol for ANI-1E (961 amons) which can probably be explained by the fact that QM9 contains more large molecules (Figure S2). Results in Figure 3.9 show that the databases of the ANI family cover the relevant chemical space well but the addition of non-equilibrium structures can further improve performance.

For ANI-1x the MAE on the full dataset is 1.68 kcal/mol, compared with 3.66 kcal/mol for ANI, whereas for ANI-1 it is 4.59 kcal/mol which is larger than for any of the other databases. The difference between ANI-1x and ANI-1 is the way by which such non-equilibrium structures were added to the dataset which apparently can have an important influence on the final performance of the data set and ML models derived from it.

A more detailed analysis is possible by considering if the amons of the isomers in the Tautobase are present (or not) in the training databases. For this, a set 'seen amons' (all constituent amons included in the reference database) and 'unseen amons' (molecules for which one or several amons were missing from the reference set) was defined. The error distributions for both sets were determined and are reported in Figure 3.10. Perhaps unsurprisingly, the 'seen amons' had a larger probability of obtaining a small error compared to the 'unseen amons'. Interestingly, PC9, which provides the broadest sampling of chemical space as quantified by the number of amons in the database, showed a similar probability error distribution for SetLE9 and SetG9. The errors for 'unseen amons' using the NN trained on QM9, a significantly smaller dataset, shows a larger and more right-skewed distribution of errors. One possible explanation may be that a better exploration of chemical space helps when predicting energies for molecules containing chemistry outside that covered by the database.

Regarding the ANI family of databases, the ANI-1 and ANI-1E results have similar error distributions. However, ANI-1x shows a smaller mean error for both the seen and unseen sets. These results are another indication that a random sampling of conformational space does not help improve the NN model predictions. On the contrary, it makes it worse than when only equilibrium structures are considered. Another notable finding is that ANI-1x shows similar performance for molecules with seen and unseen amons. This can be explained given the good sampling of chemical space, which is the same as for ANI-1E, but combined with a broad exploration of conformational space by a refinement from ANI-1 using active learning [196].

Rational detection of systematic deficiencies in quantum chemical databases is challenging because of the high dimensionality of chemical space. For this reason, methods to visualise chemical space in a digestible way, such as the TMAP algorithm used here, are highly desirable.

For instance, coloring the nodes of Tautobase TMAP by error of tautomerization energy (Figure 3.10 F) reveals that structures with azoles containing N-N and N-O bonds corre-

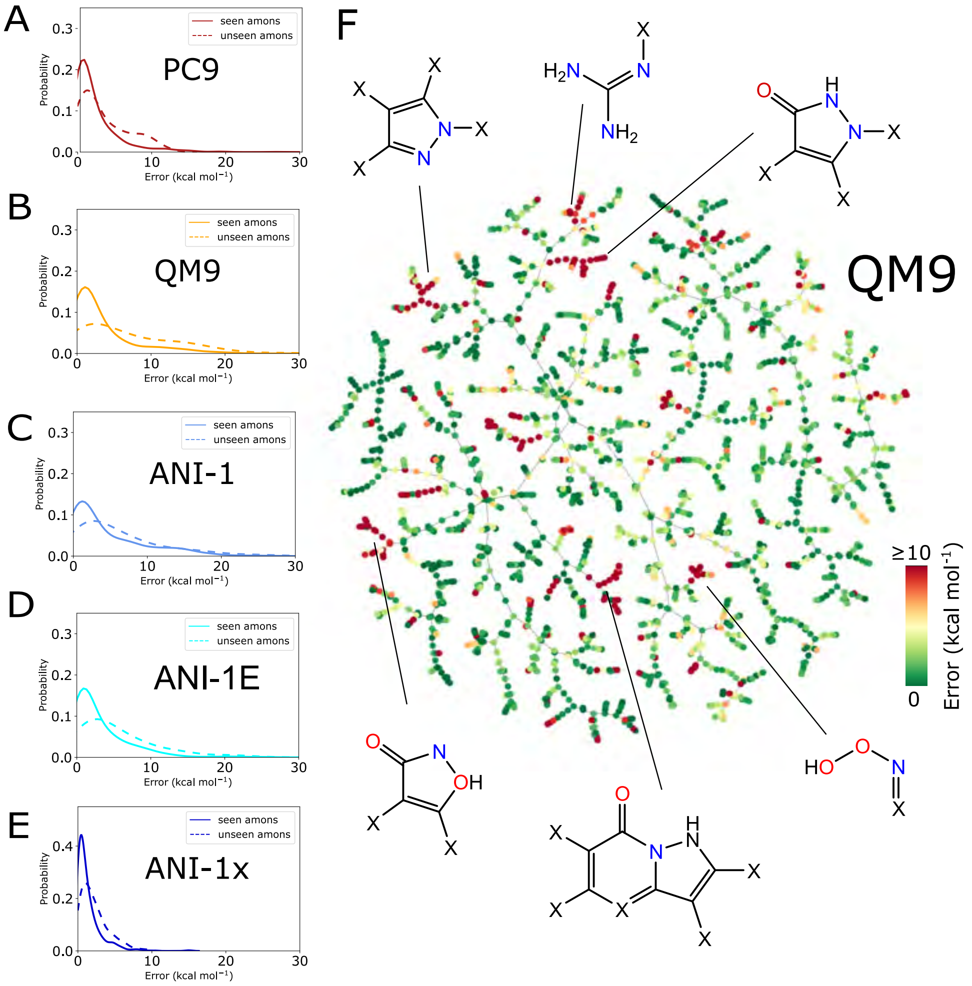

Figure 3.10: The relationship between chemical space and model reliability. Panels A to E: The mean average absolute error for molecules from Tautobase with one or more amons outside the training set is larger than if all amons are present, a trend observed for all databases. Panel F: Projection of the Tautobase using the TMAP algorithm identifies 'branches' of chemical space that are poorly predicted by the neural network models. The error displayed here is for the QM9 database. TMAPs for all databases used for training are available as interactive plots which can be viewed in a web browser, which can be obtained in the supporting information.

late with large errors. Interestingly, KL-divergences for these types of bond distances suggested that they were underrepresented in the reference sets, see Figure 3.7. The moieties corresponding to large errors change based on the different databases used to train the NN model, see Figure S13. Interactive plots are available in the supporting information.

3.5 Discussion and Conclusions

The prominence of ML has raised concerns regarding the 'interpretability' of the models conceived[166, 213]. This awareness also increases for complex models because a rational relationship between initial data used for training and resulting prediction becomes less transparent. Therefore, it is important to develop quantifiable and intuitive tests for how ML models "work" and how trustworthy predictions by them are[214]. One recently proposed procedure is "post-hoc" interpretation for which the practitioner analyzes a trained model with the aim to understand what the model has learned from the data without changing the underlying model[163, 165]. Here, post-hoc interpretability techniques were used to investigate the effect of different features of the database on predicting a chemical property (tautomerization energy). The selected features are considered important for the construction of robust quantum chemical databases for ML. In the present case this implied the analysis of individual features of several databases to quantify how these modify the prediction of a chemical property on an unseen set of examples using statistics and visualization techniques. With sufficient information from such an analysis it is expected that it will be possible to identify which features of the training databases are essential for good performance on a given task, making a rational design/enhancement of databases for training ML models for a given task possible.

The present work aimed at quantifying and analyzing the suitability of NNs trained on five different reference data sets (QM9, PC9, ANI-1E, ANI-1, and ANI-1x) to predict the tautomerization energies of molecules contained in Tautobase. It was found that depending on which characteristics are considered, the predicted MAEs can behave very differently and can, in part, be related to geometrical and/or chemical properties encoded in the databases. Such analyses attempt to digress from "black box" applications of ML methods and move towards "interpretable ML". Hence, one of the questions is "what features need to be present and covered in a training database for application to a concrete chemical question". In the present case the databases to choose from were QM9, PC9, ANI-1E, ANI-1, and ANI-1x' and the application was computation of the gas phase tautomerization energy.

The results indicate that the exploration of chemical space is essential for meaningful results. The coverage of chemical space can be quantified by the chemical diversity expressed as the number of amons on the database (see Figure 3.9). The energy prediction improves when the overlap of the number of amons in the training set and the Tautobase increases. If the number of amons in the chemical database does not cover all the amons on the target set, addition of non-equilibrium structures to the training set can improve the results (see Section 3.4.4). Results obtained with ANI-1 and ANI-1E indicate that including non-equilibrium structures generated from normal mode sampling does not yield clearly improved performance on the quantity of interest. However, using a complementary technique such as active learning can substantially improve the results, as was found for ANI-1x. It is interesting to note that normal mode sampling was also found to be insufficient for generating sufficiently reliable, full-dimensional NN-based near-equilibrium potential energy surfaces for harmonic and anharmonic normal modes.[215]

Another determinant property is the number of heavy atoms in molecules covered in the database (Section 3.4.1). Not surprisingly, better results are obtained for the range covered by the database, and if a sufficient number of samples is available, e.g. when considering the performance depending on the number of heavy atoms in SetLE9. Outside that range, the energy prediction quality decreases with the number of atoms for most databases. One of the training databases (ANI-1x) shows good results because the non-equilibrium structures help in predicting the energies. The present work finds that the different chemical environments need to be sufficiently covered by the reference database because functional diversity is key to assure good results.

To summarize the essential findings: the structural composition of the data sets used for training the NNs (QM9, PC9 and ANI-1E) and the data set to which the trained models were applied to (Tautobase) can be compared through the Kullback-Leibler divergence (Section 3.4.2). The overlap between these distributions already provides an indication how suitable a particular reference data set will be for application to the target task. In other words: the KL divergence can be used for the rational design of databases for NN models. It will be of interest to extend this to angles and dihedrals for a comprehensive exploration of the structural overlap. With respect to the molecular geometries it was also shown that the minimum energy structure used to evaluate a trained NN is essential for maintaining its performance.Using the TMAP algorithm it was possible to identify regions of chemical space that are poorly covered by the trained models. Combination of the KL-divergence and analysis of chemical coverage through TMAP is found to be an excellent aid to assess suitability and to further improve databases for specific tasks (here tautomerization energy) as those considered in the present work, which included QM9, PC9, ANI-1E, ANI-1 and ANI-1x.

In conclusion, the present work demonstrates that ML-trained models on five different reference databases and applied to one specific task (tautomerization energy) perform with a MAE ranging from 1.7 kcal/mol to 4.6 kcal/mol. The best performing reference database (ANI-1x with 5 M structures) performs on average by 1 kcal/mol better than PC9 which contains about two orders of magnitude fewer reference structures (≈ 85 K). On the other hand, PC9 is chemically more diverse by a factor of 5 (as judged from the number of amons) compared with the ANI family of databases. This indicates that lack in chemical diversity can be compensated for by increased number of non-equilibrium structures. However, the scaling of these two properties is very different. Together with quantitative descriptors, such as the KL divergence, the present results and analyses suggest that a rational approach to database generation for specific tasks may be possible. The present work also lays the groundwork for exploring the prediction of tautomer ratios in solution.[216]

3.6 Supporting Information

The supporting information for the results of this chapter can be found at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00363 or at: https://github.com/LIVazquezS/SI_PhD_Thesis/blob/main/SI_Chapter3.pdf.

Chapter 4

Uncertainty Quantification

> Uncertainty is a personal matter; it is not the uncertainty but your uncertainty.
>
> Dennis Lindley

This chapter discusses using different uncertainty quantification (UQ) techniques to explore chemical space. The first technique, Deep Evidential Regression (DER), is based on the Bayesian probability theory. Therefore, it is assumed that the data follows a Normal distribution while a Normal Inverse Gamma distribution represents the prior distribution. This formulation is convenient because, after a small modification on the output layer of a predefined NN model, it is possible to predict the uncertainty on the prediction at the same time as the main property (i.e. Energy). The model is tested through different metrics to quantify its calibration, the quality of its predictions, and whether prediction error and the predicted uncertainty can be correlated. The observed variance provides insight into the data quality used for training. Additionally, the influence of the chemical space covered by the training data set was studied using a biased database. The results clarify that noise and redundancy complicate property prediction for molecules. In addition to DER, we tested the Regression Prior (RP) Model technique, in which an ensemble of models creates a multivariate Gaussian distribution that is used to parameterise (distillate) a Normal Wishart distribution. This model is an alternative to creating a bridge between ensemble and single models for UQ through the use of knowledge distribution. Although the promise of this model is appealing, in practice, it does not provide adequate results. The main reason behind this is that the training procedure leads to numerical instabilities that are a result of theoretical limitations. Additionally, the proposed data distribution is not flexible enough to characterize the complexity of the data used.

Part of the results presented in this chapter have been previously published in Chem. Sci., 2022, 13, 13068-13084.

4.1 Deep Evidential Regression

4.1.1 Introduction

Undoubtedly machine learning (ML) models are becoming part of the standard computational/theoretical chemistry toolbox. This is because it is possible to develop highly accurate trained models in an efficient manner. In chemistry, such ML models are used in various branches ranging from the study of reactive processes,[217, 218] sampling equilibrium states,[219], the generation of accurate force fields,[96, 100, 220–222], to the generation and exploration of chemical space.[56, 75, 223] Nowadays, an extensive range of robust and complex models can be found.[119, 159, 188, 224, 225] The quality of these models is only limited by the quality and quantity of the data used for training.[62, 96] For the most part, however, the focus was on obtaining more extensive and complex databases as an extrapolation from applications in computer science. Therefore, it is believed that more significant amounts of data will beat the best algorithms.[27]

On the other hand, it has been found that even the best model can be tricked by poor data quality.[226–229] For example, in malware detection it was found that ML-based models can fail if the training data does not contain the event the model had been designed for.[226, 228] The notion of underperforming models trained on low-quality data ("garbage in-garbage out") can be traced back to Charles Babbage.[230] The ML community is starting to notice the importance of data quality used for training and the relevance to balance amount of data (“big data”) versus quality of data. From other fields in Science, it is known that using biased and low-quality data in ML can result in catastrophic outcomes[231] such as discrimination towards minorities,[232] reduction in patient survival, and the loss of billions of dollars.[233] As a result of these findings, the concept of "smart data" emerged[234–236] which describes data sets that contain validated, well-defined and meaningful information that can be processed.[235] However, specifically for chemical applications, an important additional consideration concerns the type of data that is required for predicting a particular target property.

Considering that data generation for training quantum ML models implies the use of considerable amounts of computational power[160, 161, 237] which increases the carbon footprint and makes the use of ML difficult for researchers without sufficient resources, it is essential to optimize the full workflow from conception to a trained model. With this in mind, the concept of smart data is of paramount importance for conceiving future ML models in chemistry. This necessity has been considered in previous reviews about ML in chemistry[62, 96]; however, it is still poorly understood how the choice of training data influences the prediction quality of a trained machine-learned model. One such effort quantitatively assessed the impact of different commonly used quantum chemical databases on predicting specific chemical properties.[88] The results showed that the predictions from the ML model are heavily affected by data redundancy and noise implicit in the generation of the training dataset.

Identifying missing/redundant information in chemical databases is a challenging but necessary step to ensure the best performance of ML models. In transfer learning from a lower level of quantum chemical treatment (e.g. Møller-Plesset second order theory - MP2) to the higher coupled cluster with singles, doubles and perturbative triples (CCSD(T)) it has been found for the H-transfer barrier height in malonaldehyde (MA) that it is the selection of geometries included in TL rather than the number of additional points that leads to a quantitatively correct model.[162] This has been further confirmed by computing tunneling splittings for MA from quantum instanton calculations.[238] It is also likely that depending on the chemical target quantity of interest the best database differs from the content of a more generic chemical database. Under such circumstances, uncertainty quantification (UQ) on the prediction provides valuable information on how prediction quality depends on the underlying database used for training the statistical model.

For chemical applications, ensemble methods which involve the training and evaluation of several independently trained statistical models to obtain the quantities of interest (average and variance for an observation) have been used.[239, 240] Despite their widespread use their disadvantage is the high computational cost they incur. An alternative to this are methods based on Gaussian process regression[241, 242]. However, these are limited by the size of the database that can be used. As ML models become more prevalent in different fields, new and efficient techniques for UQ have emerged which are potentially useful in chemical applications as well. These include Bayesian NNs[243] and single deterministic networks[244] with good prospects to be used in chemistry. One challenge for Bayesian NNs is that they need to be able to predict probability distributions over network parameters. This can become computationally intractable for NNs with a large number of parameters and data.[243] On the other hand, single deterministic networks are of particular interest because they are computationally cheaper given that these models need to train and evaluate only one model allowing to predict the variance for forecasting using a single deterministic model.

Some methods for UQ based on single deterministic networks have been proposed, among them regression prior networks[29], mean variance estimation[245], or Deep Evidential Regression (DER).[30] This last method has been recently applied in molecular discovery and inference for virtual screening.[31] In a recent benchmark study[245] four different UQ approaches were tested on a range of datasets. However, none of the methods tested performed best on all tasks. Part of this finding may be related to the notion that 2D networks were applied to the 3D problem of molecular structure which implies that such methods do not describe the system adequately and are not suitable to for uncertainty prediction.[245] Finally, it was concluded that UQ is a challenging task which can be highly specific for the problem at hand. However, high dimensional NNs together with random forest or mean variance estimation (which is a type of single deterministic networks) were among the best-performing approaches.

The aim of the present study is twofold. First a model for uncertainty prediction and quantification rooted in deep evidential regression is implemented as a final layer in a message passing NN based on the PhysNet architecture. This model is referred to as PhysNet-DER. Starting from the QM9 dataset a variety of metrics for hyperparameter optimization are tested quantitatively. Secondly, the trained model is used to address two concrete chemical questions at a molecular level to highlight the value of UQ in practical applications. They include characterization of a biased database and the prediction of tautomerization energies. Both applications pose different challenges to the trained model and associated uncertainty quantification in that details of chemical bonding encoded in the data set used for training directly impacts the quality of the predictions. Finally, the results are discussed in a broader context.

4.1.2 Methods

As a regression model, PhysNet[32] was selected for the present purpose. PhysNet was implemented within the PyTorch framework [147] to make it compatible with modern GPU architectures and in line with community developments. The original architecture of PhysNet was modified to output the energy and three extra parameters required for the representation of the uncertainty (Figure 4.1). Following earlier work,[30] it is assumed that the targets to predict (here energies E_i for samples i) are drawn from an independent and identically distributed (i.i.d) Gaussian distribution with unknown mean (μ) and variance (σ^2) for which probabilistic estimates are desired:

$$(E_1, \ldots, E_N) \approx \mathcal{N}(\mu, \sigma^2)$$

For modeling the unknown energy distribution, a prior distribution is placed on the unknown mean (μ) and variance (σ^2). Following the assumption that the values are drawn from a Gaussian distribution, the mean can be represented by a Gaussian distribution and the variance as an Inverse-Gamma distribution

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \nu^{-1}), \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$$

where $\Gamma(\cdot)$ is the gamma function, $\gamma \in \mathbb{R}$, $\nu > 0$, $\alpha > 1$ and $\beta > 0$.

The desired posterior distribution has the form:

$$q(\mu, \sigma^2) = p(\mu, \sigma^2 | E_1, \ldots, E_N).$$

where p indicates a generic distribution. Following the chosen representations for mean and variance, it is assumed that the posterior distribution can be factorized as $q(\mu, \sigma^2) = q(\mu)q(\sigma^2)$. Consequently, the joint higher-order, evidential distribution is represented as a Normal-Inverse Gamma distribution (Figure 4.1) with four parameters ($\mathbf{m} = \{\gamma, \nu, \alpha, \beta\}$) that represent a distribution over the mean and the variance.

$$p(\mu, \sigma^2 | \gamma, \nu, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \text{Exp}\left(-\frac{2\beta + \nu(\gamma - \mu)^2}{2\sigma^2}\right) \tag{4.1}$$

The four parameters that represent the Normal-Inverse Gamma distribution are the output of the final layer of the trained PhysNet model (Figure 4.1) and the total predicted energy for a molecule composed of N atoms is obtained by summation of the atomic energy contributions E_i:

$$E = \sum_{i=i}^{N} E_i \tag{4.2}$$

In a similar fashion, the values for the three parameters ($\nu, \alpha,$ and β) that describe the distribution of the variance for a molecule composed of N atoms are obtained by summation of the atomic contributions and are then passed to a softplus activation function to fulfill the conditions given for the distribution ($\gamma \in \mathbb{R}$ and ν, α, $\beta > 0$)

$$\begin{aligned} \alpha &= \log\left(1 + \exp\left(\sum_{i=i}^{N} \alpha_i\right)\right) + 1 \\ \beta &= \log\left(1 + \exp\left(\sum_{i=i}^{N} \beta_i\right)\right) \\ \nu &= \log\left(1 + \exp\left(\sum_{i=i}^{N} \nu_i\right)\right) \end{aligned} \tag{4.3}$$

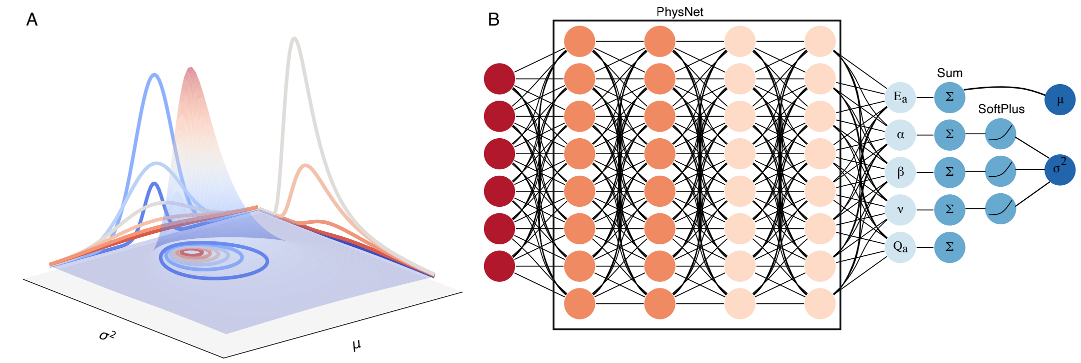

Figure 4.1: Modified PhysNet for uncertainty quantification. **A** Schematic 3D representation of the Negative Inverse Gamma distribution as a function of the mean (μ) and the variance (σ^2) (See Equation 4.1). **B** The modified architecture of PhysNet for the addition of the 'evidential' layer. The input layer receives atomic positions, atomic numbers, charges, and energies. In the next step, those values are passed to the regular architecture of PhysNet. The final layer is modified to output five values (E_a, Q_a, α, β, and ν) per each atom in a molecule. In the next step, the values of the outputs are summed by each molecule. Then, the three extra parameters are passed to a SoftPlus activation function (See Equation 4.3). The final output of the model are the values that characterize the Normal Inverse Gamma distribution. The mean value for the prediction (Equation 4.4) corresponds to the energy of the predicted molecule, and the parameters to determine the variance of the predicted energy which can be obtained using Equations 4.5 and 4.6.

Finally, the expected mean (Equation 4.4), and the aleatory (Equation 4.5) and epistemic (Equation 4.6) uncertainty of predictions can be calculated as:

$$\mathbb{E}[\mu] = \gamma \tag{4.4}$$

$$\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1} \tag{4.5}$$

$$Var[\mu] = \frac{\beta}{\nu(\alpha - 1)} \tag{4.6}$$

Including the new parameters in the output of the neural network changes the loss function of the model. The new loss function consists of a dual-objective loss $\mathcal{L}(x)$ with two terms: the first term maximizes model fitting and the second penalizes incorrect predictions according to

$$\mathcal{L}(x) = \mathcal{L}^{\mathrm{NLL}}(x) + \lambda(\mathcal{L}^{\mathrm{R}}(x) - \varepsilon) \tag{4.7}$$

In equation 4.7, the first term corresponds to the negative log-likelihood (NLL) of the model evidence that can be represented as a Student-t distribution (Equation 4.8)

$$\begin{aligned}\mathcal{L}^{\mathrm{NLL}}(x) = \frac{1}{2}\log\left(\frac{\pi}{\nu}\right) - \alpha\log(\Omega) + (\alpha + \frac{1}{2})\log((x-\gamma)^2\nu + \Omega) \\ + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + \frac{1}{2})}\right)\end{aligned} \tag{4.8}$$

where $\Omega = 2\beta(1+\nu)$ and x is the value predicted by the neural network.[30]. The second term in Equation 4.7, $\mathcal{L}^{\mathrm{R}}(x)$, corresponds to a regularizer that minimizes the evidence for incorrect predictions (Equation 4.9).

$$\mathcal{L}^{\mathrm{R}}(x) = |x - \gamma| \cdot (2\nu + \alpha) \tag{4.9}$$

The hyperparameter λ controls the influence of uncertainty inflation on the model fit and can be calibrated to obtain more confident predictions. For $\lambda = 0$, the model is overconfident. i.e. results are less likely to be correct. Alternatively, for $\lambda > 0$, the variance is inflated, resulting in underconfident predictions.

The neural network architecture was that of standard PhysNet, with 5 modules consisting of 2 residual atomic modules and 3 residual interaction modules. Finally, the result is pooled into one residual output module. The number of radial basis functions was kept at 64, and the dimensionality of the feature space was 128. Electrostatic and dispersion corrections were not used for the training to keep the model as simple as possible. All other parameters were identical to the standard version of PhysNet[32], unless mentioned otherwise.

For training, a batch size of 32 and a learning rate of 0.001 were used. An exponential learning rate scheduler with a decay factor of 0.1 every 1000 steps and the ADAM optimizer[146] with a weight decay of 0.1 were employed. An exponential moving average for all the parameters was used to prevent overfitting. A validation step was performed every five epochs.

Hyperparameter Optimization

The hyperparameter λ in equation 4.7 was optimized by training a range of models with different values of λ, using a portion of the QM9 dataset consisting of 31250 structures: 25000 structures for training, 3125 for validation and the remaining 3125 for testing. The splitting of the selected molecules of QM9 was performed randomly. The top panel of Figure S7 shows that the energy distributions from the training and test sets overlap closely which demonstrates that the dataset used for training is representative of the overall distribution of energies. Models were trained for 1000 epochs and the values for λ were 0.01, 0.1, 0.2, 0.4, 0.5, 0.75, 1.0, 1.5, and 2.0. The calibration of the NN models is required to assure that the computed uncertainties can be related with the obtained errors on the prediction. It should be mentioned that although this procedure is computationally expensive, it only needs to be done once.

Metrics for Model Assessment and Classification

In order to compare the performance/quality of the trained models, suitable metrics are required. These metrics are used to select the best value for the hyperparameter λ. Different metrics that have been reported in the literature[246–248] were evaluated.

The first metric considered is the Root Mean Variance (RMV) defined as:

$$\mathrm{RMV}(j) = \sqrt{\frac{1}{|B_j|}\sum_{t\in B_j}\sigma_t^2} \tag{4.10}$$

Here, σ_t^2 is the variance in the $j-$th bin B_j. For the construction of the bins B_j the data is first ranked with respect to the variance and then split into bins $\{B_j\}_{j=1}^{N}$ of size N which is adjustable and the effect of changing it on RMV is assessed when discussing the results.

The next metric was the empirical Root Mean Squared Error (RMSE):

$$\mathrm{RMSE}(j) = \sqrt{\frac{1}{|B_j|}\sum_{t\in B_j}(y_i - \hat{y}_t)^2} \tag{4.11}$$

where y_i is the i-th prediction and $\hat{y}_t$ is the average value of the prediction in a bin B_j. Using equations 4.10 and 4.11, the Expected Normalized Calibration Error (ENCE):

$$\mathrm{ENCE} = \frac{1}{M}\sum_{j=1}^{M}\frac{|\mathrm{RMV}(j) - \mathrm{RMSE}(j)|}{\mathrm{RMV}(j)} \tag{4.12}$$

can be obtained. Additionally, it is possible to quantify the dispersion of the predicted uncertainties for which the Coefficient of Variation (C_v) is

$$C_v = \frac{1}{\mu_\sigma}\sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\sigma_i - \mu_\sigma)^2} \tag{4.13}$$

In equation 4.13, μ_σ is the mean predicted standard deviation and σ_t is the predicted standard deviation for M samples.

The last metric used for the characterization of the predicted variance of the tested models is the 'sharpness'

$$\mathrm{sha} = \frac{1}{N}\sum_{i=1}^{N}var(F_n) \tag{4.14}$$

In equation 4.14, the value $var(F_n)$ corresponds to the variance of the random variable with cumulative distribution function F at point n.[247] The purpose of this metric is to measure how close the predicted values of the uncertainty are to a single value.[249]

In addition to the above metrics, calibration diagrams were constructed with the help of the uncertainty toolbox suite.[250] Calibration diagrams report the frequency of correctly predicted values in each interval relative to the predicted fraction of points in that interval.[247, 251] Another interpretation of the calibration diagram is to quantify the 'honesty' of a model by displaying the true probability in which a random variable is observed below a given quantile; if a model is calibrated this probability should be equal to the expected probability in that quantile.[250]

The results obtained for the test dataset were then classified into four different categories following the procedure described in Kahle and Zipoli.[252] For the present purpose, $\varepsilon^* = \mathrm{MSE}$ (mean squared error) and $\sigma^* = \mathrm{MV}$ (mean variance), and the following classes were distinguished:

- True Positive (TP): $\varepsilon_i > \varepsilon^*$ and $\sigma_i > \sigma^*$. The NN identifies a molecule with a large error through a large variance. In this case, it is possible to add training samples with relevant chemical information to improve the prediction of the identified TP. Alternatively, additional samples from perturbed structures for a particular molecule could be added to the increase chemical diversity.

- False Positive (FP): $\varepsilon_i < \varepsilon^*$ and $\sigma_i > \sigma^*$ in which case the NN identifies a molecule as a high-error point but the prediction is correct. In this case, the model is underconfident about its prediction.

- True Negative (TN): $\varepsilon_i < \varepsilon^*$ and $\sigma_i < \sigma^*$. Here the model recognizes that a correct prediction is made with a small value for variance. For such molecules the model has sufficient information to predict them adequately by assigning a small variance. Therefore, the model does not require extra chemical information for an adequate prediction.

- False Negative (FN): $\varepsilon_i > \varepsilon^*$ and $\sigma_i < \sigma^*$. The model is confident about its prediction for this molecule but it actually performs poorly on it. One possible explanation for this behaviour is that molecules in this category are rare[253] in the training set. The model recognizes them with a small variance but because there is not sufficient information the target property (here energy) can not be predicted correctly.

In the above classifications, i refers to a particular molecule considered for the evaluation. The classification relies on the important assumption that the MSE and the MV are comparable in magnitude which implies that the variance predicted by the model is a meaningful approximation to the error in the prediction. A second desired requirement is to assure the validity of the classification procedure and that the obtained variance is meaningful is that $\mathrm{MSE} > \mathrm{MV}$. This requirement is a consequence of the bias-variance decomposition of the squared error[145]

$$\begin{aligned}\mathbb{E}(\mathrm{MSE}) &= \mathbb{E}[(y(x) - \mu(x))^2|_{x=x_0}] \\ &= \underbrace{\sigma^2}_{\text{Irreducible Error}} + \underbrace{[\mathbb{E}\mu(x_0) - y(x_0)]^2}_{Bias^2} + \underbrace{\mathbb{E}[\mu(x_0) - \mathbb{E}\mu(x_0)]^2}_{\text{Variance}}\end{aligned} \tag{4.15}$$

Equation 4.15 states that the expected value ($\mathbb{E}$) of the MSE consists of three terms: the irreducible error, the bias, and the variance. Therefore, the MSE will always be smaller than the variance except for the case that $\mu(x) = y$ for which those quantities are equal.[254].

As a measure of the overall performance of the model, the *accuracy* is determined as[63]:

$$\mathrm{ACC} = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}} + N_{\mathrm{TN}} + N_{\mathrm{FP}}} \tag{4.16}$$

In equation 4.16, N_{TP}, N_{TN}, N_{FP}, and N_{FN} refers to the number of true positive, true negative, false positive, and false negative samples, respectively. Additionally, it is possible to compute the true positive rate (R_{TP}) or *sensitivity* as:

$$R_{\mathrm{TP}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \tag{4.17}$$

As a complement to equation 4.17, the true positive predictive value (P_{TP}) or *precision* is

$$P_{\mathrm{TP}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \tag{4.18}$$

Model Performance for Tautomerization

As a final test, the performance of the evidential model was evaluated using a subset of the Tautobase[87], a public database containing 1680 pairs. Previously, those molecules were calculated at the level of theory of the QM9 database.[88, 255] For the purpose of the present work, only molecules that contain less than nine heavy atoms were included. Three neural networks with λ values of 0.2, 0.4, and 0.75 were trained with the QM9 database. The QM9 database was filtered to remove molecules containing fluorine and those that did not pass the geometry consistency check. The size of final database size was 110 426 molecules. That number was split on 80 % for training, 10% for validation and 10% for testing. The three models were trained for 500 epochs with the same parameters as for the hyperparameter optimization.

4.1.3 Results

In this section the calibration of the network is analyzed and its performance for different choices of the hyperparameter is assessed. Then, an artificial bias experiment is carried out and finally, the model is applied to the tautomerization data set. Before detailing these results, a typical learning curve for the model is shown in Figure S1. As expected, the root mean squared error obtained for the test set decreases with increasing number of samples. For the mean variance, see Figure S2, it is found that its magnitude reduces up to a certain size of the training set after which it increases again. This observation is further discussed in “Discussion and Conclusions”.

Calibration of the Neural Network

The selection of the best value for the hyperparameter λ can be related to the calibration of the neural network model. Ideally, a calibrated regression model should fulfill the condition[246] that

$$\forall\sigma : \mathbb{E}_{x,y}[(\mu(x) - y)^2|_{\sigma(x)^2=\sigma^2}] = \sigma^2$$

where $\mathbb{E}$ is the expected value for the squared difference of the predicted mean evaluated at x minus the observed value y. In other words: the squared error for a prediction can be directly related to the variance predicted by the model.[246]

Figure 4.2 compares the root mean squared error with the root mean variance for a number of bins ($N = 100$) and shows that the correlation between RMSE and RMV can change between different intervals. Analyses were also carried out for different numbers N of bins and the effect on RMV was found to be negligible, see Figure S3. Additionally, the slope of the data can be used as an indicator as to whether the model over- or underestimates the error in the prediction. A slope closer to 1 indicates that the model is well-calibrated. Consequently, the predicted variance can be used as an indicator of the error with respect to the value to be predicted. The results in Figure 4.2 also show that smaller values of $\lambda = (0.01, 0.2, 0.4)$ result in increased slopes of the RMSE versus RMV curve, i.e. leads to less well-calibrated models, resulting in a model that is overconfident in its predictions. Results that are more consistent with a slope of 1 are obtained for $\lambda = 1$. However, for all trained models it is apparent that RMSE

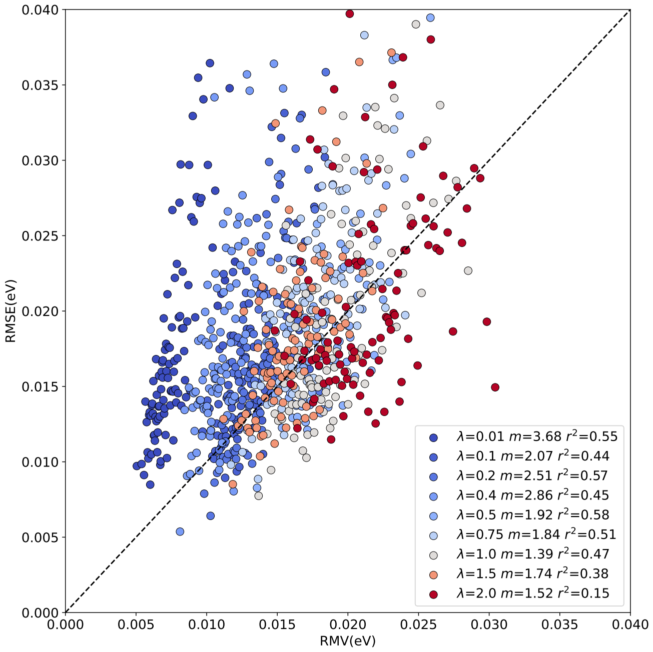

Figure 4.2: Empirical root mean squared error compared with the root mean variance of the evidential model trained on 25000 structures from the QM9 database. The values were divided in 100 bins ranked with respect to the predicted variance, 25 bins with 32 samples and 75 with 31 samples were considered. The value of λ together with the slope (m) from a linear regression analysis and the Pearson correlation coefficient (r^2) are given in the legend.

and RMV are not related by a “simple” linear relationship as is sometimes assumed in statistical modeling.

In previous studies,[247] the dispersion of the predicted standard deviation was considered as a measure of the quality of a regression model. Hence a wider distribution of the predicted standard deviation by the model is desired. To remove the influence of pronounced outliers, Figure 4.3A shows the distributions up to 99% of the predicted variance. It is clear that the center of the distribution, and its width, depend on λ. Larger values of the hyperparameter lead to wider distributions. However, the displacement of the center of mass of the distribution indicates that the standard deviation will be consistently overestimated. Also, $p(\sigma)$ is not Gaussian but rather resembles the inverse gamma distribution that was used as prior for the variance.

Predicted standard deviations from machine learned models must follow some characteristics that help to assess the quality of model predictions.[247] Among those characteristics, it is expected that the distribution of the predicted variance is narrow, i.e. will be ’sharp’. This has two objectives, the first is that the model returns uncertainties that are as tight as possible to a specific value.[249] With this property the model gains confidence on its prediction. The second goal of a 'sharp' model is that it is able to capture the 'trueness'[256], i.e. the distance between the true value and the mean of the predictions, on the forecast. Another desired characteristic is that $p(\sigma)$ is disperse and does not return a constant value for the uncertainty which would make the model likely to fail for predictions on molecules outside the training data and compromise its generalizability.

The previously described characteristics of the distribution of uncertainties are related to the value of the hyperparameter λ in the loss function (Equation 4.7) because, as can be seen in Figure S6, the MSE by percentile is independent on the choice of λ. Therefore, the model should be calibrated by selecting a value of the hyperparameter that fulfills the desired characteristics for the distribution of uncertainties.

Figure 4.3A shows that the spread of the distribution of standard deviations increases with increasing λ. However, the second desired feature for those distribution - sharpness - decreases with increasing λ to become almost constant for $\lambda \geq 0.75$. In consequence of this contradictory behaviour, it is necessary to find a value of λ that yields an accurate estimation of the uncertainty but it does not return a distribution of uncertainties but rather a constant value for each case. It is important to notice that both characteristics, sharpness and width of the distribution, are equally important and one of them should not be sacrificed in favour of the other.[247] In other words: a calibrated model is characterized by uncertainty distributions with a certain sharpness and a certain width.

A deeper understanding of the difference between the error of a predicted value and the predicted variance can be obtained through the ENCE (Equation 4.12) as described in the methods section. This metric is similar to the expected calibration error used in classification[247]. The ENCE quantifies the probability that the model incorrectly predicts the uncertainty of the prediction made. Figure 4.3B reports the values of ENCE (blue line) and shows that, typically, smaller values for ENCE are expected for increasing hyperparameter λ. For $\lambda = 0.4$, the value of ENCE increases as opposite of the expected trend because the predicted value of the RMSE is larger than the value for RMV for most of the considered bins. However, it is clear that for $\lambda \geq 0.5$, the ENCE is almost constant - which indicates that, on average, the model has a low probability to make incorrect predictions.

As a complement to the ENCE metric, the coefficient of variation (C_{v}) was also com-

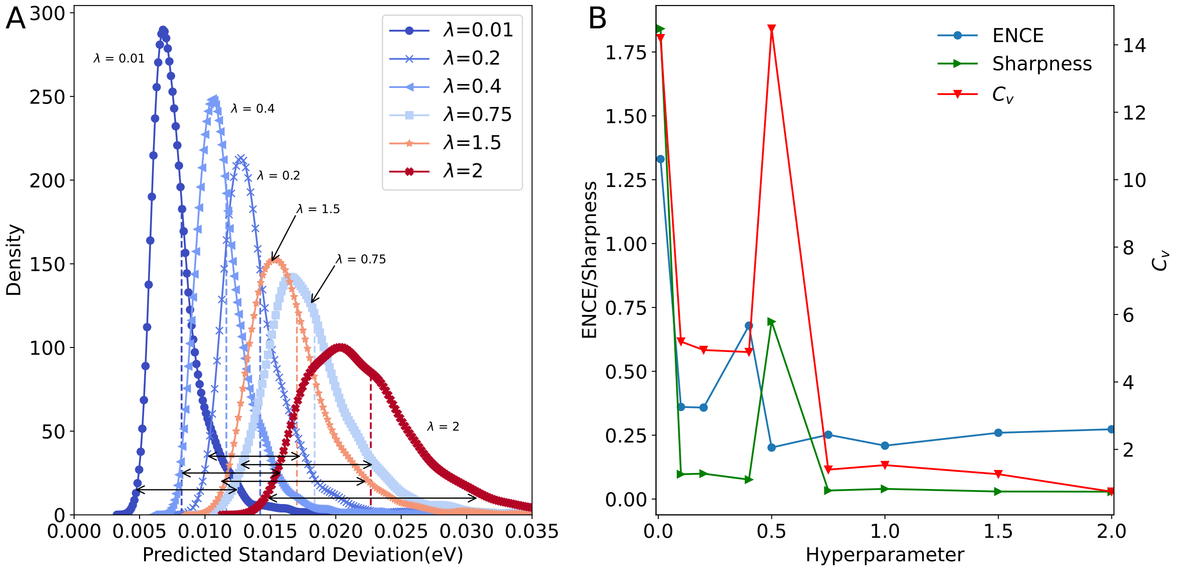

Figure 4.3: Metrics for the distribution of predicted variance. **A** Kernel density estimate of standard deviation($\sigma = \sqrt{\text{Var}}$) for different values of hyperparameter λ. Values up to the 99% percentile of the variance were considered. The internal arrows show the 'width' of the distributions. Dotted lines inside the distribution report their sharpness. Not all distributions are shown for clarity. **B** Evolution of the Expected Normalized Calibration Error (ENCE), sharpness, and the Coefficient of Variation (C_v) depending on λ.

puted (red trace in Figure 4.3B). This metric is considered to be less informative because the dispersion of the prediction depends on the validation/test data distribution[247, 257]. However, it is useful to characterize the spread of standard deviations because it is desired that the predicted uncertainties are spread and therefore cover systems outside the training data which help to generalize the model and make it transferable to molecules outside the training set. Comparing the results from Figure 4.3A and the values for C_{v} in Figure 4.3B, it is found that the largest dispersion is obtained for small values of λ. This indicates that the standard deviations for all predictions are concentrated in a small range of values for values in the 95th percentile of the distribution. For $\lambda \geq 0.75$ both ENCE and C_{v} values do not show pronounced variation. It should be noted that the distributions in Figure 4.3A are restricted to the 99% quantile of the data; on the other hand, the values for C_{v} covered the whole range of data. If the complete range of data is analyzed, it is possible to arrive at wrong conclusions. Figure 4.3B shows that for $\lambda = 0.5$, the C_{v} value is large which suggests a flat distribution (Figure S4), however it should be noticed that this behaviour arises primarily due to pronounced outliers that impact the averages used for the calculation. However, 95% of the distribution is concentrated around a small range of variances as shown in Figure 4.3A. Nevertheless, if only 95% of the data is studied, it is found that $\lambda \geq 0.5$ yields increased C_V (see Figure S5).

As shown in Figure 4.3A, the center of mass of $P(\sigma)$ displaces to larger σ with increasing λ. A more detailed analysis of the difference between MSE and MV for different percentiles of the variance was performed (Figure S6). Following the bias-variance decomposition of the squared error (Equation 4.15), the bias of the model can be quantified as a function of the different values of λ. Figure S6 shows that the MSE is constant regardless of the value of the hyperparameter λ or the percentile of the variance. On the other hand, the variance increases as a function of λ but it is constant regarding the value of the percentile with the exception of $\lambda = 1$. Thus, the MV is larger than the MSE which is counter-intuitive in view of the bias-variance decomposition of the squared error. Finally, it is clear that the difference between MSE and MV decreases as the value of λ increases. This indicates that the assumed posterior distribution does not correctly describe the data and, as a consequence, it can not adequately capture the variance of a prediction. In other words, a better "guess" of the posterior will improve the predicted variance.

A common method to judge whether a model is well-calibrated is by considering the calibration curves described in the methods section. The results in Figure 4.4 show that, as λ increases, the model is closer to the diagonal which indicates perfect calibration. The best calibrated models are obtained for small values of λ ($\lambda = 0.1$ and $\lambda = 0.2$). Calibration curves help to evaluate the 'honesty' of the model predictions. Previously,[31] calibration curves were employed to select a suitable value for λ using the SchNet architecture[258] for QM9. These results largely agree with what is found here with $\lambda = 0.1$ and $\lambda = 0.2$ as the best values. Although calibration curves are extensively used in the literature to assess the quality of uncertainty predictions by ML models, they also have weaknesses that complicate their use. For example, it was reported[246] that perfect calibration is possible for a model even if the output values are independent of the observed error. Furthermore, it was noticed[246] that calibration curves work adequately when the uncertainty prediction is degenerate (i.e. all the output distributions have the same variance) which is not the desired behavior. In addition to this, it was found that the shape of these curves can be misleading because there are percentiles for which the model under- or overestimates the uncertainty. Then the calibration curves need to be complemented with additional metrics for putting their interpretation in perspective. Here, the analysis of calibration curves was complemented by using the miscalibration area (the area between the calibration curve and the diagonal representing perfect calibration). Using this metric, it is clear that λ values of 0.75 have

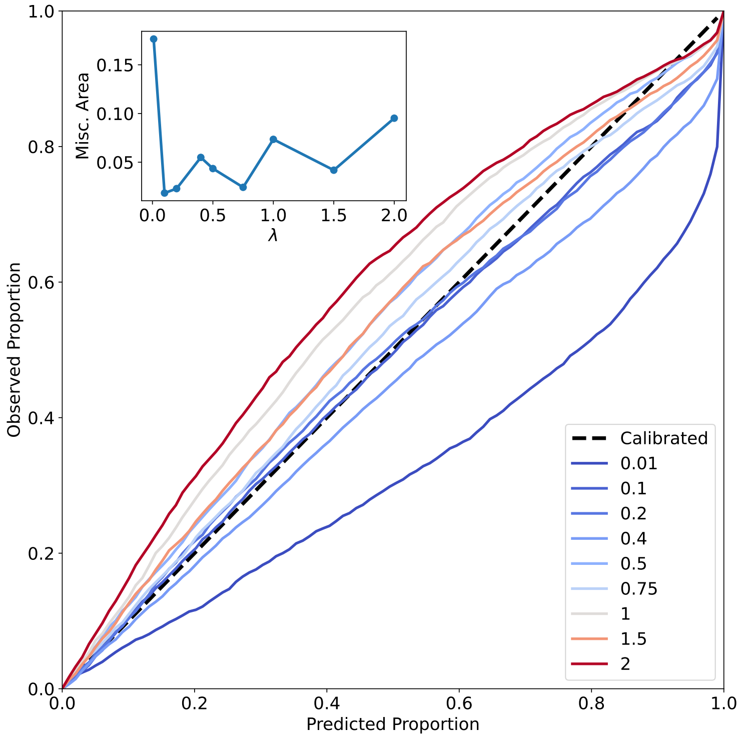

Figure 4.4: Calibration curves with respect to the hyperparameter λ. The x-axis shows the predicted probability to obtain the correct value for the error in a given percentile, the y-axis shows the true probability. The trend line shows the behavior of a perfectly calibrated model. Inside the plot, the area between the curve and the trend line, also called Miscalibration Area, is shown as a function of the hyperparameter λ. A smaller miscalibration area indicates a better model.

a performance as good as $\lambda = 0.1$ and $\lambda = 0.2$.

Classification of Predictions

The effect of bias in the training set for PhysNet-type models was previously found to negatively impact prediction capabilities across chemical space.[88] In the context of uncertainty quantification, it is also of interest to understand how the predicted variance can be related to the error in the prediction for an individual prediction. For this, the relationship between the predicted variance and the error of prediction was studied following a classification scheme, see methods section. To this end, the subset of QM9 used for hyperparameter optimization was considered. Then the molecules in the test set were evaluated with the models trained with different values of the hyperparameter λ.

For all the tested models, the largest percentage of molecules ($\approx 80\%$) was found to be True Negatives (TN), see Figure 4.5A. This indicates that the model recognizes for

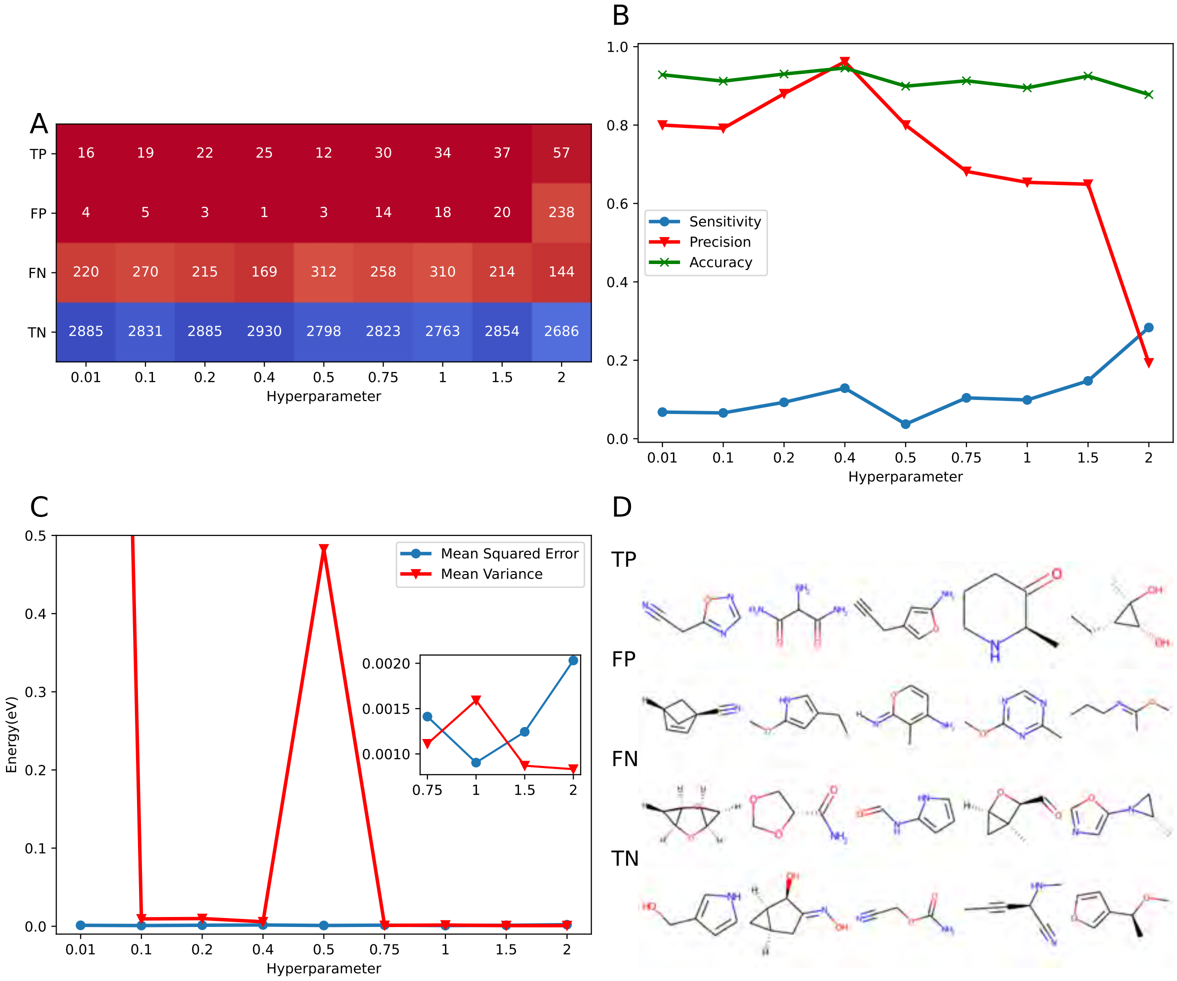

Figure 4.5: Results of the classification procedure. **A** Confusion matrix with respect to the value of the hyperparameter λ, inside each panel is the number of molecules that belong to the defined categories. The abbreviations refer to TP: True Positive, FP: False Positive, TN: True Negative, and FN: False Negative for information in how those categories are defined consult the methods section. **B** Accuracy (green, Equation 4.16), sensitivity (blue, Equation 4.17), and precision (red, Equation 4.18) depending on λ. **C** The MSE and MV for the full set of molecules as a function of λ. The Mean Variance for $\lambda = 0.01$ is not shown for clarity. The inset of the plot shows the behavior for $\lambda \geq 0.75$. **D** Chemical structures of the top 5 most common molecules in each of the four classes.

most of the samples that there is sufficient information for a correct prediction. On the other hand, molecules classified as True Positives (TP) correspond to samples for which predictions are difficult. Hence, these molecules lie outside the training distribution because they are associated with large prediction errors and the model is ’aware’ of this. As expected, the number of TP and FP increases with increasing λ. This is a consequence of the inflation of uncertainty by making the model less confident about its prediction which results in misclassification of molecules because - as described before - the error in the prediction is independent on the value of λ, see Figure S6. Finally, the number of False Negative (FN) samples in the data is approximately independent on λ. As described before, the molecules in this category contain information on the boundary of the training distribution which compromises the model’s prediction capability. The constant number of FN is indicative of a systematic problem that can only be corrected by providing additional samples of similar molecules. The distribution of FP and FN was further analyzed in Figure S7. The results indicate that the categories distribute uniformly over the energies sampled. It is also observed that false negatives (i.e. "underconfident") tend to be more present at smaller total energy (~ -65 eV) whereas false positives ("overconfident") are more common for larger total energy (~ -80 eV). Furthermore, the number of FPs decreases rapidly with decreasing value of the hyperparameter λ, whereas for FNs this number is rather insensitive to λ, see also Figure 4.5A.

A summary of the relationship between the four classifications in term of model accuracy, sensitivity, and precision is given in Figure 4.5B. In all cases the accuracy of the model is appropriate, since the largest part ($\approx 90\%$) of the studied samples are correctly predicted (i.e. TN) and the variance reflects the prediction error. On the other hand, the precision of the model is also high ($\approx 80\%$) but starts to decrease as λ increases. In the present context, precision is a measure for the model’s capability to recognize ‘problematic’ cases which also correspond to a real deficiency in the model which can be assessed by comparing the prediction with the true value and the predicted variance. It is expected that as the model becomes more underconfident, the precision decreases as there are more molecules misclassified due to inflation of the uncertainty. Conversely, sensitivity describes how many of the molecules that present a problem in the prediction are identified by the model. Here, the sensitivity increases for $\lambda > 0.5$: as the model becomes less confident, the probability to detect samples that are truly problematic increases. It should, however, also be pointed out that the numerical values for (ϵ^*, σ^*) to define the different categories will impact on how the classifications impact model characteristics such as “precision” or “sensitivity”.

The MV and MSE for the complete set of samples as a function of λ are provided in Figure 4.5C. It is found that with the exception of $\lambda = 0.01$ and $\lambda = 0.5$, MV and MSE are comparable, which is a desired characteristic of the model. However, since it is additionally desirable that MV<MSE the variance obtained by the model accounts for the variance term in equation 4.15. Therefore, the difference between MSE and MV is a constant value that corresponds to the combination of the bias of the model and the irreducible error. The advantage of this definition is that the variance can be mainly attributed to the data used for training. This provides a rational basis for further improvement of the training data. It is noted that the condition MV<MSE is only fulfilled for $\lambda = 0.75$ and $\lambda > 1.5$. A summary with the values of all the metrics tested for calibration is given in Table S1.

Figure 4.5D and Figures S8 to S11 present concrete molecules from each of the four categories. Although the molecules used in the training, validation and test sets were kept constant for the different models, the molecules identified as outliers differed for each value of λ. However, it is instructive to identify molecules that appear more frequently in the various tests. These chemical structures are studied in more detail on the following sections with the aim to identify systematic errors and sampling problems and how they can be corrected.

Artificial Bias Experiment

To provide a more chemically motivated analysis of predicted energies and associated variances, a model was trained using the first 25k molecules of QM9. The question addressed is whether predicted energies and variances for molecules not used in the training of the model are more likely to be true positives than for molecules with little coverage in the training set. Since the structures in QM9 were derived from graph enumeration, the order of the molecules in the database already biases certain chemical motifs, such as rings, chains, branched molecules and other features.

Figure 4.6A reports the Tree MAP (TMAP) projection[86] of the entire QM9 database (pink) and the first 25k molecules (blue). TMAP is a dimensionality reduction technique with good locality-preserving properties for high dimensional data such as molecular fingerprints. Analysis of the projection suggests that, as a general structural bias, the first 25k molecules over-represent aromatic heterocyclic, 5- and 6- membered rings, and structures with multiple substituted heteroatoms with regards to the relative probability of other structures also present in QM9.

For training the NN, as described in the methods section, 31500 structures were randomly split (train/validation/test of 0.8/0.1/0.1) and a model with $\lambda = 0.4$ was trained to make predictions on the test set. A TMAP projection of the test and train compounds is shown in Figure 4.6C. The connectivity of the different tree branches on the TMAP provides information about the local similarity of the molecules where dense regions of the map correspond to clusters of high similarity. The average degree i.e. number of edges between one molecule and its neighbors, for the TNs in the test set - which was the majority class ($\approx 90\%$ of the test samples) - was 2.0 compared with classes FN (169 molecules), TP (25 molecules), and FP (1 molecule) which had average degrees of 1.7, 1.3, 1.0. The lower connectivity for FP compared with TN indicates that "good predictions for the right reason" are more likely if coverage of particular structural and/or chemical motifs is better. Furthermore, it is observed that FPs have a low connectivity which indicates that these molecules are 'rare' in the training set. On the other hand, the different sample sizes of the four classes need to be kept in mind when generalizing such conclusions.

The TMAP projection of the test set in Figure 4.6B shows the chemical similarity between specific molecules seen during training or testing. In general, molecules identified as TPs contained common scaffolds seen during training in combination with unusual substituents. For example, the moiety of imidazole (a five-membered 1,3-C_3N_2 ring) was a common fragment in the training set and lies in the biased region of chemical space depicted in Figure 4.6A. Common true positives contained this imidazole scaffold inside uncommon fused three ring systems. When the model makes predictions for compounds close in chemical space to molecules of which it has seen diverse examples in the training set, the estimates of variance appear to be more reliable.

Figure 4.6D reports three examples of false positives (i.e. molecules with high error and low predicted variance) in the test set. The molecules in the training set are labelled as i, iii and v, whereas those used for prediction from the test set were ii, iv and vi. The pair (i/ii) consists of a diazepane core that goes through a double bond migration. Although the rest of the structure is conserved for i and ii, the error in the prediction for molecule ii (test) is ≈ 0.1 eV, but the predicted variance is the same for molecules i and ii. A possible explanation is that the model recognizes that i and ii are similar which leads to assigning a small variance to ii. However, this contrasts with the energy difference between molecules i and ii which is ≈ 0.5 eV.

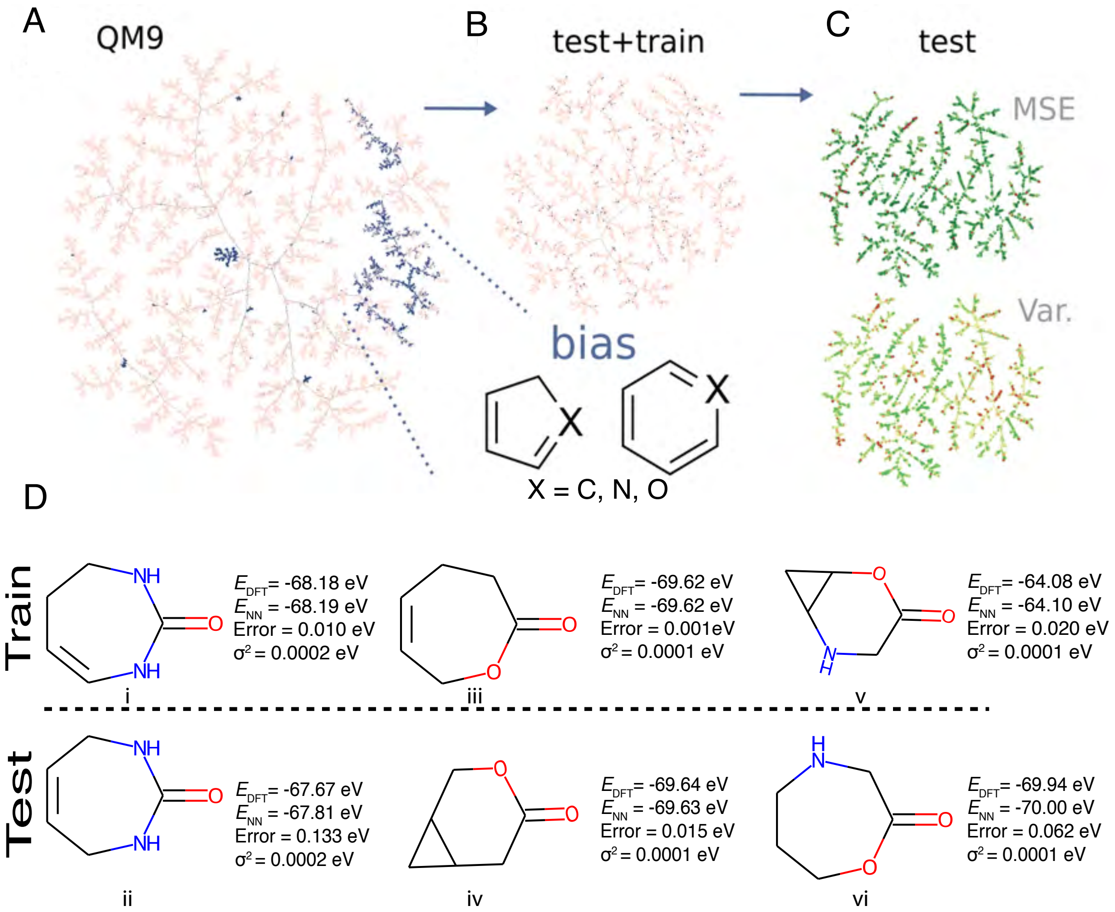

Figure 4.6: Artificial bias experiment. **A** TMAP of the QM9 database. In blue the structures used for training, the inset shows that the selected part of the database bias the data towards specific chemistry, in this case, aromatic 6- and 5-membered heterocyclic rings scaffolds. In pink, the rest of the structures on QM9. **B** TMAP of the reduced dataset. In pink the structures used for training and validation and in blue the selected random compounds used for test. **C** TMAP of the test set. On top TMAP, for the MSE and down the corresponding for variance. The colormaps which span from the minimum value (green) to $1\,\sigma$ (red). **D** Pairs of similar molecules (i/ii), (iii/iv), and (v/vi) for which one molecule was in the training set (top) and the related molecule was in the test set (bottom) with reference, prediction and difference energies displayed together with associated variance.

Pair (iii/iv) involves an oxepane ring with a carbonyl (iii) which is in the training set and the prediction is for an oxabicycloheptane (iv). In this case the model predicts the energy with an error of 0.015 eV. Hence, for pair (iii/iv) the information that the model has from molecule iii, in addition to the significant presence of bicycles in the training set, makes it easier to predict the energy for molecule iv. Finally, pair (v/vi) is opposite to (iii/iv): training on an Oxa-azabicycloheptane for predicting an Oxazepane. The error for this prediction is considerably higher (~ 0.06 eV). This shows that it is easier for the NN to predict bicycles than seven-membered rings and reflects the fact that there are more bicycles in the training set than seven-membered rings. An intriguing aspect of the totality of molecules shown in Figure 4.6D is that they all have the same number of heavy atoms, and that they share multiple structural and bonding motifs. This may be the reason why the model assigns a small variance to all of them because the NN is primed to make best use of structural information at the training stage. However, additional tests are required to further generalize this.

Similarly, cases where a ring was expanded or contracted by a single atom between molecules in the training and test set commonly resulted in similar failure modes due to over-confidence. This observation is particularly interesting because it suggests that the model might be overconfident when predicting compounds it has seen sparse but highly similar examples of during training. Uncertainty quantification, in this conception, is effective at predicting in-distribution errors, however, out-of-distribution errors are not as easily quantified by this model.

Tautomerization Set

As a concrete chemical application of how uncertainty quantification can be used, the prediction of energy of tautomer pairs was considered. Tautomerization is a form of reversible isomerization involving the rearrangement of a charged leaving group within a molecule.[167] The structures of the molecules involved in a tautomeric pair (A/B) only differ little which makes this an ideal application for the present developments. For the study of tautomeric pairs, three NN models with different values of $\lambda = 0.2, 0.4, 0.75$ were trained with QM9 database as described on the methods section. The test molecules considered come from the Tautobase database.[255] For the purpose of this work, only molecules with less than nine heavy atoms (C, N and O) were tested. A total of 442 pairs (884 molecules) was evaluated.

The training of PhysNet involves learning of the Atomic Embeddings (AtE) and the centers and widths of the Radial Basis Functions (RBF). These features encode the chemical environment around each atom and therefore contain the "chemical information" about a molecule. This opens the possibility to further analyze the potential relationship contained in the learned parameters to the information about the chemical space contained in the training dataset and how it compares with the chemical space of the test molecules that are the target for prediction. Hence, for the following the mean distances between each of the tested molecules and the molecules in the training set of the database for $\langle \mathrm{AtE} \rangle$ and $\langle \mathrm{RBF} \rangle$ were determined according to the procedure described in Section

I. Figure 4.7 shows the results for the relationship between the mean distance of the AtE and RBF, the error, variance and number of atoms for the molecules in the tautobase.

The bottom row of Figure 4.7A (panels i to v) report $\langle \mathrm{AtE} \rangle$ and $\langle \mathrm{RBF} \rangle$, the prediction errors and associated variances sorted by the number of heavy atoms $N = 3$ to 9 together with the distribution $P(N)$. The dependence of $\langle \mathrm{AtE} \rangle$ and $\langle \mathrm{RBF} \rangle$ on N shows that with decreasing number of heavy atoms the mean distance with respect to the molecules with the same number of atoms increases (Figure 4.7A i and ii). Additionally, the violin plots in Figure 4.7A i and ii show that the mean distance values are more spread as the number of atoms increases. One explanation for these results is that the available chemical space to explore increases with N which is also reflected in the number of samples with a given number of heavy atoms in the training dataset; consequently, the distance between the molecules with a low number of atoms increases. In other words, a larger molecule explores chemical space more extensively in terms of chemical environments, atom types, bonding patterns and other characteristics of chemical space. The relationship between error and the number of atoms illustrates how the smaller mean distance in RBF and AtE leads to a smaller error. Furthermore, the number of outliers also scales with the size of the molecules. Comparing error and variance by the number of heavy atoms, it is clear that for up to 5 atoms they behave similarly (Figure 4.7A iii and iv). From Figure 4.7A iii, it is clear that the error distribution shifts with increasing number of atoms in the molecule. For the center of mass of the predicted variance distribution (Figure 4.7A iv) is at a high value and progressively decreases until 5 heavy atoms to increase again. It should be noted that the number of outliers for error and variance increases with the number of heavy atoms which affects the displacement on the center of mass. Finally, the spread of error and variance by the number of atoms (Figure 4.7A iii and iv) presents similar shapes up to 8 heavy atoms. For molecules with 8 and 9 atoms, the variance is more spread out whereas the error distribution is more compact.

Panels vi, vii, x, and xi in Figure 4.7A show that variance and error are similarly distributed depending on $\langle \mathrm{AtE} \rangle$ and $\langle \mathrm{RBF} \rangle$, respectively. For the entire range of $\langle \mathrm{AtE} \rangle$ and $\langle \mathrm{RBF} \rangle$ low variance (< 0.0002 eV) and low prediction errors (< 0.25 eV) are found. Increased variance (~ 0.0005 eV) is associated with both, larger $\langle \mathrm{AtE} \rangle$ and $\langle \mathrm{RBF} \rangle$ whereas larger prediction errors (> 1.0 eV) are found for intermediate to large $1.0 \leq \langle \mathrm{RBF} \rangle \leq 1.5$. This similarity is also reflected in a near-linear relationship between $\langle \mathrm{AtE} \rangle$ and $\langle \mathrm{RBF} \rangle$ reported in panel xiii of Figure 4.7A.

Prediction error and variance are less well correlated for the evaluated molecules from tautobase, see panel viii of Figure 4.7A. This can already be anticipated when comparing panels i and ii. With increasing N, the position of the maximum error shifts monotonously to larger values whereas the variance is higher for $N = 3$, decreases until $N = 6$, after which it increases again. Hence, for tautobase and QM9 as the reference data, base error and variance are not necessarily correlated.

To gain a better understanding of the prediction performance of QM9 for molecules in the Tautobase from the point of view of feature space, polar plots considering extreme cases were constructed, see supporting information for technical details. Figure 4.7B shows the case for the molecule (center) with the largest average distance in RBF and AtE for molecules with the same number of atoms used for training for this representation; only the ten closest neighbours are shown. Although the molecule is relatively simple, no structure in the training set contains sufficient and appropriate information for a correct prediction. Despite abundant information about similar chemical environments but with different spatial arrangements, combination with different functional groups or different bonding arrangements, potentially conflicting information in the training set leads to uncertainties in the prediction. A second example, that of the molecule with largest variance and largest distance in RBF, is shown in Figure 4.7C. As for molecule ii in Figure 4.6D this case also highlights how seemingly small changes in bonding pattern, functional groups and atom arrangements can lead to large errors. However, in this case the abundant and similar structural information in the training set leads to a large predicted variance. In other words, "redundancy" in the training set can lead to vulnerabilities in the trained model as was previously found for predictions based on training with the ANI-1 database compared with the much smaller ANI-1E set: despite its larger size, predictions based on ANI-1 were less accurate than those based on ANI-1E.[88]

As a final example of the relationship between error and variance, the chemical structures for a set of molecules with low error but high variance is highlighted in Figure 4.7D and shows that heterocyclic rings and bicycles are well covered in the training set. An interesting aspect is that molecules with a nitro-group (–NO_2) appear with high variance and low error. This effect can be rationalized by considering the design of the GDB-17 Database[49] which is the source of the QM9 set: for GDB-17 aliphatic nitro groups were excluded, but aromatic nitro groups were retained. Therefore, the trained model will have similar information based on structural considerations but the quality of the data in view of a molecules' energetics is low which leads to significant variance.

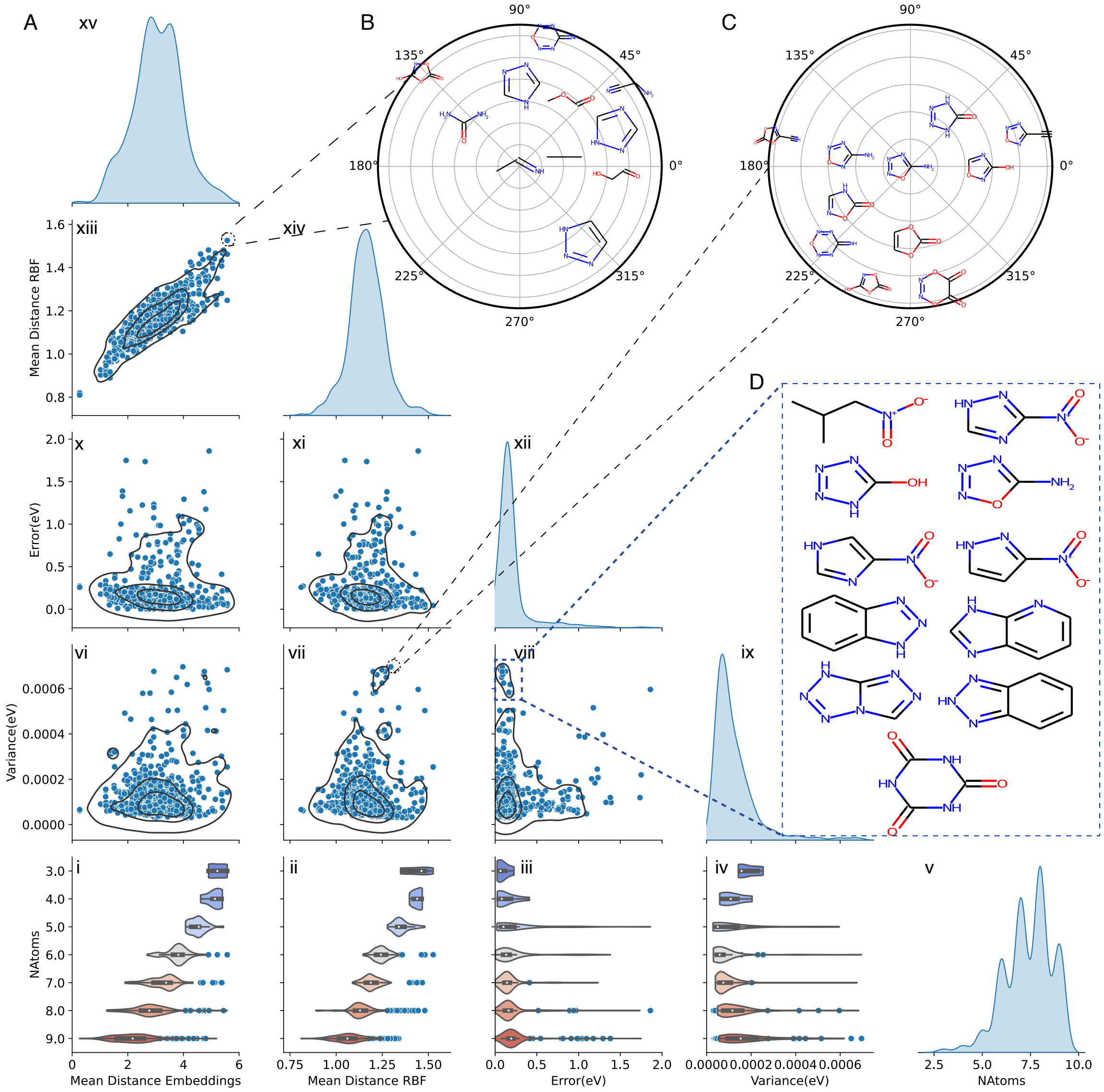

Figure 4.7: **A** Overview of the comparison between different results for the evaluation of molecules on the tautobase for $\lambda = 0.75$ up to the 95th percentile. The diagonal of the figure shows the kernel density estimate of the considered properties (Mean Distance Embeddings, Mean Distance RBF, Error (in eV), Variance (in eV) and Number of Atoms). For each of the panels a correlation plot between the variable and a 2D kernel density estimate is shown. On the last row, violin plots for the different considered properties with respect to the number of atoms is shown. Similar plots for $\lambda = 0.2$ and $\lambda = 0.4$ can be found in the Supporting Information. **B** Radial plot of the ten closest molecules of the training set on feature space for the molecule in tautobase with the largest distance in embedding and RBF space. **C** Radial plot of the ten closest molecules for the molecule in tautobase with the largest predicted variance and the largest distance in RBF space. **D** Examples of molecules with large predicted variance and small error. Enlarged views of panels B and C are provided in Figures S14 and S16.

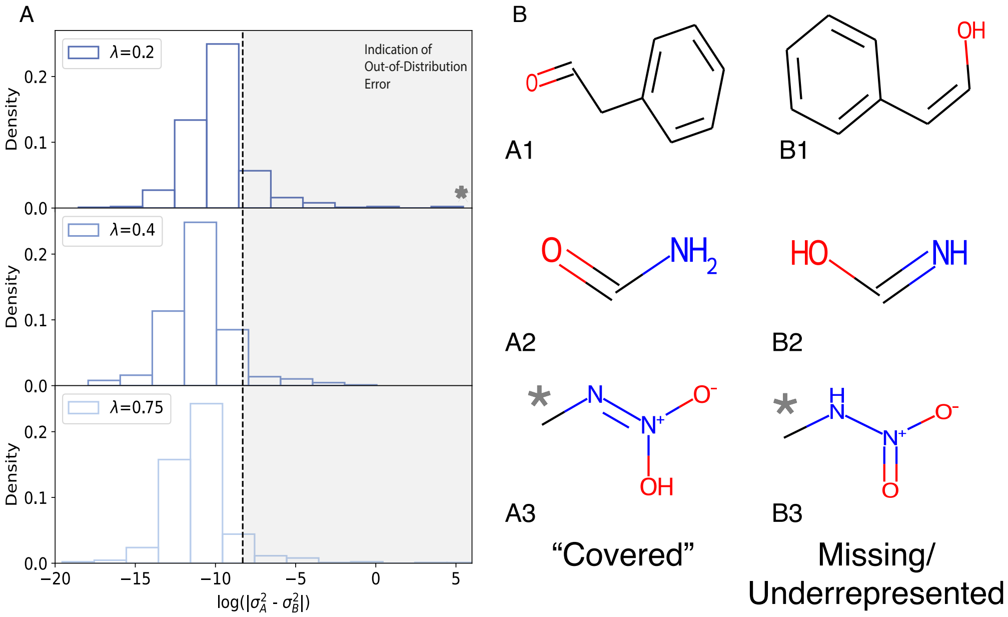

Figure 4.8: **A** The log distribution of differences in predicted variance between tautomer pairs, A (low variance) and B (high variance). **B** Tautomer pairs (A/B) containing chemical groups, nitro and vinyl alcohols, outside the training set (B1-3) are easily identified. The imine group in B2 was present in only one molecule in the training set. Numerical values for energies and variances are summarized in Table 4.1.

Finally, it is of interest to analyze tautomer pairs (A/B) for which the difference in the predicted variance is particularly large. Figure 4.8A reports the distribution $p(\sigma_A^2 - \sigma_B^2)$ for trained models with different values of the hyperparameter λ. First, it is found that the distribution of variance differences depends on the value of λ. Therefore, particularly prominent outliers can be avoided by careful evaluation of the predictions. Secondly, large differences (star in Figure 4.8A) in the variances can occur and indicate that the trained models are particularly uncertain in their prediction. To illustrate this, three tautomer pairs were identified and are analyzed in more detail in the following. For molecules B1 to B3 it is found that their functional groups are not present or are poorly represented in QM9. These include the N=O nitro group in an aliphatic chain (B3), vinyl alcohol (B1), and hydroxyl imine (B2, only one representative in QM9). Furthermore, the pair (A3/B3) is zwitterionic.

As is shown in Figure 4.8B the chemical motifs and functional groups in A1 to A3 are covered by QM9 whereas those in their tautomeric twins (B1 to B3) are not. For molecule B1 (vinyl alcohol) examples are entirely absent in QM9 and the presence of hydroxyl groups bound to sp^2(aromatic) carbons is not sufficient for a reliable prediction

Table 4.1: Reference energy (E_{DFT}), predicted energy (E_{NN}) and variance (σ^2) for selected molecules in Figure 4.8. All values are in eV.

Molecule	E_{DFT}	E_{NN}			σ^2		
λ		0.2000	0.4000	0.7500	0.2000	0.4000	0.7500
A1	-79.8900	-79.6800	-79.6900	-79.6800	**0.0018**	0.0002	0.0002
Δ		**0.2100**	0.2000	**0.2100**			
B1	-79.5900	-79.2800	-79.4400	-79.3600	**0.0249**	0.0002	0.0002
Δ		**0.3100**	0.1500	0.2300			
A2	-23.5900	-23.5200	-23.5200	-23.5200	**0.0016**	0.0012	0.0002
Δ		**0.0700**	**0.0700**	**0.0700**			
B2	-23.0200	-22.7500	-22.8800	-22.9100	**0.0019**	0.0011	0.0007
Δ		**0.2700**	0.1400	0.1100			
A3	-30.8600	-32.6700	-32.5800	-32.2000	0.0046	0.0004	**0.2243**
Δ		**1.8100**	1.7200	1.3400			
B3	-31.6300	-32.0200	-32.0800	-32.6900	**229.6200**	0.0035	0.0033
Δ		0.3900	0.4500	**1.0600**			

for B1. It is also noted that the difference Δ between the target energy (E_{DFT}) and the predictions (E_{NN}) are largely independent on λ for A1 but differ by a factor of two for B1. This is also observed for the pair (A2/B2) for which the uncertainties are more comparable than for (A1/B1).

Finally, the pair (A3/B3) poses additional challenges. First, the variance for one value of λ for B3 is very large and for A3 one of the variances is also unusually large, given that similar examples to A3 are part of the training set. Secondly, although A3 is better represented in the training set, the difference between target value and prediction is larger than 1 eV for all models. These observations are explained by the fact that (A3/B3) are both zwitterionic and the uncertainty associated with B3 may in part be related to the fact that QM9 only contains few examples of sp^2 NO bonds except for a small number of heterocyclic rings which are chemically dissimilar compounds compared with B3. Furthermore, for B3 some of the atom-atom separations ("bond lengths") are poorly covered by QM9. For the N–N distance, the QM9 database contains the range from 1.2 Å to 1.4 Å (see Figure S17) whereas N–N in B3 is 1.383 Å which is a low probability region for $p(r_{\mathrm{NN}})$. This is also the case for compound A3 although $p(r_{\mathrm{NN}})$ has a local maximum at the corresponding N–N separation. In conclusion, the majority of prediction problems in Figure 4.8B can be related to origins in the underlying chemistry. Interestingly, even a careful analysis of the performance of a trained model on the training set (see compound A3) may provide insight into coverage and potential limitations when making predictions from the trained model.

4.1.4 Discussion and Conclusions

The present work introduces uncertainty quantification for the prediction of total energies and variances for molecules based on a trained atomistic neural network. The approach is generic and it is expected that it can be generalized to other NN-architectures and observables.

With respect to computational effort it is noted that the current approach requires training of several independent models for a range of values for the hyperparameter λ. However, the uncertainty on a prediction can be obtained from evaluating a single model. This is an advantage compared to ensemble models which require the evaluation of all trained models to obtain an estimate of the uncertainty. For ensemble-based approaches the statistical error of a prediction $\sim 1/\sqrt{N}$ whereas for DER considered here this is not the case. Rather, a number of models needs to be trained for calibration but as demonstrated here, N 10 is a meaningful estimate for this. On the other hand, Bayesian methods rapidly become impractical for larger data sets as already mentioned in the Introduction. One possible way to avoid training for a range of $\lambda-$values is to use recalibration methods.[250, 259] However, such methods are quite new and still need to be validated by different metrics. Finally, the results obtained here can be used as a starting point for model training on other databases but it remains to be seen if the calibration results are transferable to other databases.

Data completeness and quality directly impact the forecasting capabilities of statistical models. Although quantum chemical models are trained, for example, on total energies of a set of molecules, it is not evident how to select the best suitable training set for most accurately predicting energy differences between related compounds, such as structural isomers as demonstrated in this work. PhysNet-DER is a step towards the design of validated, well-defined databases containing meaningful information ("smart data").[234–236] In this process one also anticipates that targeted databases will become available for specific applications in chemistry, such as tautomerization energies, hydration energies, or HOMO-LUMO gaps, to name a few. Also, the findings from the present work will be useful to be employed together with established methods like Gaussian Process Approximation.[242]

A large part of the present work was concerned with the impact of redundant/missing information in the databases used to train a model on the prediction of specific properties in chemical space. The results confirm that redundancies can impact heavily the prediction of a property and its variance. However, it is still necessary to systematically identify and remove conflicting information while retaining training quality. In this regard, the combination of unsupervised machine learning methods[72, 260–262] with the approach introduced here will hopefully allow to design workflows to broadly explore chemical space at low computational cost. Another point that requires attention is the underlying assumption in many similar applications that the predicted property can be represented as a normal (Gaussian) distribution. The present and earlier studies[155] indicate that this assumption is only valid approximately.

It was noted in Figure S2 that the average predicted variance for a hold-out set of molecules decreased with increasing training set size until a certain point. Beyond that, models trained on the most extensive training corpus predicted higher variance. This is consistent with the expectation that as new molecules are introduced to the training set, the probability of adding previously unseen information is initially large, but decreases as the training set grows. This is indicative of the law of diminishing returns.[263] The artificial bias experiment carried out here suggests that the model may become sensitive to redundant information which leads to overconfident estimates of variance for over-represented chemical motifs at the expense of being under-confident for motifs with fewer training examples. The observation that larger training sets can introduce higher uncertainties is compelling and highlights the need for a deeper understanding of the role of bias when evaluating atomistic neural networks for predictions made across chemical space.

Distances in the embedding space (AtE/RBF) of the neural network were studied to visualize and analyze the proximity between molecules in the training and test set, see Figures S14 and S16. This allowed to assess how similar molecules can influence the prediction by making the model less confident. On the other hand, it was also possible to recognize molecules for which insufficient information was available in the database for a prediction. In other words, analysis of the embedding space also hinted towards the role of similar information on model degradation. It is of interest to note that analysis of the embedding space was previously done for uncertainty determination.[239, 245] As used in the present work, distances in embedding space provide a qualitative picture for what information influences a prediction. This can be used in a more targeted fashion for model improvement but more systematic studies for this natural next step are required.

Some of the essential findings of the present work concern the notion that single metrics are not particularly meaningful to judge the calibration of a trained model. Exploration and development of meaningful metrics will benefit evidence-based inference.[264] Also, it is not always true that error and variance are directly related which is counter typical expectations in statistical learning. It is also demonstrated that mean variance and mean squared error can behave in counter-intuitive ways which points towards deficiencies in the assumed posterior distribution.

As found here, uncertainty quantification is essential and reveals that the nature and coverage of the training set used for model construction plays an important role when applied to specific chemical tasks. For example, it is demonstrated for tautomerization energies that classification of predictions can be used to identify problematic cases at the prediction stage. Furthermore, it was found that similar information in low quantity returns low uncertainties but high errors, whereas similar information in large quantities results in small errors but high predicted uncertainties. A notable example of this is the nitro group in the training database, which is not present for aliphatic chains but for aromatic rings. Thus, for a balanced ML-based model for chemical exploration an equilibrium between the quantity and the quality of data in the database is required. The information from UQ can be used in the future to build targeted and evidence-based datasets for a broad range of chemical observables based on active learning strategies and for constructing robust high-dimensional potential energy surfaces of molecules.

4.1.5 Supporting Information

Supporting information associated with this chapter can be found at: https://doi.org/10.1039/D2SC04056E or at: https://github.com/LIVazquezS/SI_PhD_Thesis/blob/main/SI_Chapter4.pdf

4.2 Regression Prior Networks

4.2.1 Introduction

It is known that ensembles of NN models are a powerful method for the prediction of uncertainty quantification. This method is based on the separate training of several models with different initial conditions that later are averaged to obtain a prediction mean value and its standard deviation. In chemistry, these are usually applied through the popular method 'query-by-committee'[265]. Unfortunately, the use of ensembles is limited by the high computational cost associated. Additionally, it has been found that ensemble methods could lead to overconfident models that underestimate the uncertainty of prediction[252]. On the other hand, single-pass uncertainty prediction reduces the computational cost by only needing to train one model and making assumptions about the underlying training distribution. Nevertheless, single-pass uncertainty models have problems too. In particular, problems related to DER will be discussed in Chapter 6.

A convenient method to mix ensemble and single-pass NN is the Regression Prior (RP) Network [29]. RP combines ensemble methods with a single-pass uncertainty prediction by distilling several neural network knowledge into a model representing a multidimensional Gaussian distribution (Figure 4.9). In this case, the computational effort is only done once during training by training several models, while at the inference moment, the time is considerably reduced because only one model is required to be evaluated.

An additional advantage that RP provides is the capability to separate the contributions to the uncertainty. In general, the uncertainty in a prediction can be decomposed into two main contributions[266]. The first one is called *aleatoric*, also known as statistical or data, and refers to the notion of randomness. Then, the variability in the prediction is a consequence of the inherently random effects. It is important to notice that this can not be reduced because it is caused by noise in the training data or limitations of the model. The second source of uncertainty is called *epistemic*, also known as systematic or knowledge, and refers to the lack of knowledge by the model. Therefore, epistemic uncertainty can be reduced by adding more information during the training procedure. This last one is usually the target for active learning procedures.

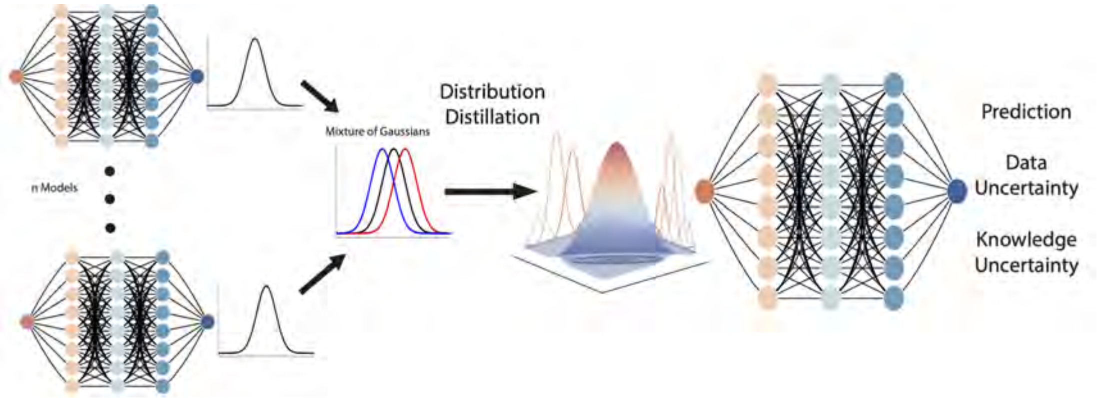

Figure 4.9: **Regression Prior Network.** The scheme represents the process of training. Initially, a set of n NN models where each of them returns a 1D Gaussian distribution as output. The outputs of the n NN models are mixed to create a teaching model by a mixture of Gaussians. A distribution distillation from the Gaussian mixture model parameterizes a new model based on the high dimensional Normal Wisard distribution. The model obtained is then used to predict the quantity of interest, the data uncertainty (a.k.a. aleatoric), and the knowledge uncertainty (a.k.a. epistemic).

4.2.2 Methods

Basic Theory Let us assume that each individual model trained will return a *probabilistic regression model*. This means that the model instead of returning a point estimate of the prediction, the model parameterises a distribution of the property of interest (i.e. Energy), $p(E|\mathbf{x}, \theta)$, which in place is dependent on the parameters of the NN model and the input vectors. Usually, the chosen return is the normal distribution ($\mathcal{N}$) as follows:

$$p(E|\mathbf{x}, \theta) = \mathcal{N}(E|\boldsymbol{\mu}, \boldsymbol{\Lambda}), \quad \{\boldsymbol{\mu}, \boldsymbol{\Lambda}\} = \mathbf{f}(\mathbf{x}; \theta) \tag{4.19}$$

Here, the normal distribution of the energy is parameterized with a mean $\boldsymbol{\mu}$ vector and a precision matrix, $\boldsymbol{\Lambda}$. The precision matrix is related to the covariance matrix, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$, which is typically used to parameterise multidimensional normal distributions.

Next, we train n probabilistic regression models, each of them returning Gaussian distribution parameters. Mixing the results of those models creates an ensemble of NN models that would parameterise a multivariate normal distribution (MVN).

$$\text{MVN} = \{p(E|\mathbf{x}, \theta^{(i)}\}_{i=1}^{n} \tag{4.20}$$

This ensemble can be interpreted as a set of draws from a higher-order implicit distribution over normal distributions[29]. By using the Bayes theorem, the aim is to find a distribution that can emulate the ensemble of Equation 4.20 by explicitly parameterizing a higher-order distribution over the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$[29]. The selected prior is the Normal Wishard ($\mathcal{NW}$) distribution, which is a prior to the MVN[267]. The $\mathcal{NW}$ distribution is a compound distribution of a normal distribution over the mean and a Wishard distribution over the precision.

$$\mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{m}, \boldsymbol{L}, \kappa, \nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, \kappa\boldsymbol{\Lambda})\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{L}, \nu) \tag{4.21}$$

Here $\boldsymbol{m}$ is the prior mean, $\boldsymbol{L}$ is the inverse of the positive-definite prior scatter matrix, κ and ν are the strengths of the belief in each prior. Complementary, the multidimensional Gaussian distribution is defined as:

$$\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, \kappa\boldsymbol{\Lambda}) = \left(\frac{\kappa}{2\pi}\right)^{D/2} |\boldsymbol{\Lambda}|^{1/2} \exp\left(-\frac{\kappa}{2}(\boldsymbol{\mu}-\boldsymbol{m})^{\top}\boldsymbol{\Lambda}(\boldsymbol{\mu}-\boldsymbol{m})\right) \tag{4.22}$$

While the Wishard distribution[1] is defined as:

$$\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{L}, \nu) = \frac{|\boldsymbol{\Lambda}|^{\frac{\nu-K-1}{2}}}{2^{\frac{\nu K}{2}}\Gamma_K\left(\frac{\nu}{2}\right)|\boldsymbol{L}|^{\frac{\nu}{2}}} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{L}^{-1}\right)\right) \tag{4.23}$$

Where $\Gamma_K(x)$ is the multivariate gamma function:

$$\Gamma_K(x) = \pi^{\frac{K(K-1)}{4}} \prod_{i=1}^{K} \Gamma\left(\frac{x+(1-i)}{2}\right)$$

and K are the degrees of freedom[267].

By using the distribution of equation 4.21, an RP model parametrizes the $\mathcal{NW}$ distribution over the mean and precision of normal output distributions as follows[29]:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{m}, \boldsymbol{L}, \kappa, \nu), \quad \{\boldsymbol{m}, \boldsymbol{L}, \kappa, \nu\} = \boldsymbol{\Omega} = f(\boldsymbol{x}; \boldsymbol{\Theta}) \tag{4.24}$$

The parameters in the set $\boldsymbol{\Omega}$ are predicted by the neural network model. Hence, we calculate the posterior predictive as:

$$\begin{aligned} p(E|\boldsymbol{x}, \boldsymbol{\Theta}) &= \iint p(E|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\Theta})d\boldsymbol{x} & (4.25)\\ &= \int \mathcal{N}(E|\boldsymbol{\mu}, \boldsymbol{\Lambda})\mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\Omega})d\boldsymbol{\Omega} & (4.26)\\ &= \mathcal{T}\left(E\middle|\boldsymbol{m}, \frac{\kappa+1}{\kappa(\nu-K+1)}, \nu-K+1\right) & (4.27) \end{aligned}$$

[1] The Wishard distribution is the generalization of the Gamma distribution to positive definite matrices[267]

The results of the integral of Equation 4.26 is the multivariate $\mathcal{T}$-Student distribution defined as:

$$\mathcal{T}(E|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+K}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{\kappa}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left(1 + \frac{1}{\nu}(E-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(E-\boldsymbol{\mu})\right)^{-\frac{(\nu+K)}{2}} \quad ,\nu \geq 0; \tag{4.28}$$

Once the posterior distribution is known, passing to the training step is possible. Malinin *et al.*[29] suggest using a double task loss function ($\mathcal{L}$) in which an in-domain and an out-of-distribution data are used. Although this looks convenient and has a sound theoretical base, in the case of chemical problems, it is complicated to make that distinction. Further discussion about this topic can be found in Chapter 6. To overcome this problem, the solution is to parameterize the model from an ensemble of distributions such as the one in Equation 4.20 by minimizing the Kullback-Leibler (KL) divergence between the model and the ensemble as:

$$\mathcal{L}_{\text{EnD}} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M}\left(\int p(\boldsymbol{y}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(\boldsymbol{m})}) \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(\boldsymbol{m})})}{p(\boldsymbol{y}|\boldsymbol{x}^{(i)}, \phi)}\right) d\boldsymbol{x}^{(i)}\right) \tag{4.29}$$

This process is called ensemble distillation (EnD). Equation 4.29 has problems keeping the diversity of the ensemble, more so in cases where the distributions in the ensemble can be spread. As a consequence of this loss of diversity, the resulting model will have poor performance in terms of prediction and uncertainty estimation. Therefore, an alternative is to consider that the adjusted model does not represent a single distribution but an ensemble of them [268]. This strategy helps to overcome the problem of diversity. However, it is hard to reproduce the ensemble diversity exactly without incurring on a high computational cost. The alternative proposed by Malinin *et al.* consists of modelling the ensemble's behaviour by averaging the ensemble's outputs and using those results to train the model distribution. This method is called ensemble distribution distillation (EnD^2)[269].

For the construction of the loss function of EnD^2, we start by assuming an ensemble of models such as the one described in equation 4.20 in which each model returns a value of the mean and precision of a normal distribution refer as $\mathcal{D}_{train}$. Then, an empirical distribution over the mean and the precision can be defined as:

$$\hat{p}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{x}) = \left\{\{\boldsymbol{\mu}^{(mi)}, \boldsymbol{\Lambda}^{(mi)}\}_{m=1}^{M}, \boldsymbol{x}^{(i)}\right\}_{i=1}^{N} = \mathcal{D}_{train} \tag{4.30}$$

The distillation of the model is accomplished by minimizing the negative log-likelihood of the mean and precision of the ensemble under the $\mathcal{NW}$ prior. This is equivalent to minimising the Kullback-Leibler divergence between the model and the empirical distribution constructed in Equation 4.30.

$$\mathcal{L}_{\mathrm{EnD}^2} = \int \hat{p}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x}) \log \left(\frac{\hat{p}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x})}{p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x}; \phi)} \right) d\boldsymbol{x} \tag{4.31}$$

The optimization of Equation 4.31 is numerically challenging, so it is required to use a temperature-annealing trick to make this process easier. Then, the ensemble is reduced to its mean as:

$$\boldsymbol{\mu}_T^{(mi)} = \frac{2}{T+1}\boldsymbol{\mu}^{(mi)} + \frac{T-1}{T+1}\bar{\boldsymbol{\mu}}^{(i)}, \quad \bar{\boldsymbol{\mu}}^{(i)} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\mu}^{(mi)} \tag{4.32}$$

$$\boldsymbol{\Lambda}_T^{-1(mi)} = \frac{2}{T+1}\boldsymbol{\Lambda}^{-1(mi)} + \frac{T-1}{T+1}\bar{\boldsymbol{\Lambda}}^{-1(i)}, \quad \bar{\boldsymbol{\Lambda}}^{-1(i)} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\Lambda}^{-1(mi)} \tag{4.33}$$

The new mean and precision matrix, scaled by the temperature (T), are then substituted on Equation 4.31. To avoid scaling the gradients by the temperature, the loss function is divided by the temperature value. In practice, Malinin *et al.*[269] suggest using a linear temperature scheduler in which the temperature is initially high for the model to match the mean. The temperature is annealed down to 1 in the second step, following a linear decay. Note that the temperature is unitless as it only represents a rescaling of the loss function.

An advantage of RP with respect to other models for uncertainty quantification is the possibility of quantifying the different components of the uncertainty. This could be derived by obtaining the mutual information between the parameters of the output distribution, $\{\boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ for the $\mathcal{NW}$ distribution, and the target values for a complete derivation the reader is referred to [29]. In this work, we use the law of total variance to derive the data and knowledge uncertainty then,

$$\mathrm{Var}_{p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x}, \boldsymbol{\theta})}(\boldsymbol{\mu}) = \mathrm{Var}_{p(\boldsymbol{y}|\mathbf{x}, \theta)}(\boldsymbol{y}) - \mathbb{E}_{p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x}, \boldsymbol{\theta})}(\boldsymbol{\Lambda}^{-1}) \tag{4.34}$$

In equation 4.34 the term on the left hand side ($Var_{p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x}, \boldsymbol{\theta})}(\boldsymbol{\mu})$) is the epistemic uncertainty. The right hand side are the total uncertainty ($Var_{p(\boldsymbol{y}|\mathbf{x}, \theta)}(\boldsymbol{y})$) and the expected data uncertainty ($\mathbb{E}_{p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{x}, \boldsymbol{\theta})}(\boldsymbol{\Lambda}^{-1})$). These terms are defined as follows:

$$\mathrm{Var}_{p(\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{x},\boldsymbol{\theta})}(\boldsymbol{\mu}) = \frac{1}{\kappa(\nu - K - 1)}\boldsymbol{L}^{-1} \tag{4.35}$$

$$\mathbb{E}_{p(\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{x},\boldsymbol{\theta})}(\boldsymbol{\Lambda}^{-1}) = \frac{1}{\nu - K - 1}\boldsymbol{L}^{-1} \tag{4.36}$$

$$\mathrm{Var}_{p(\boldsymbol{y}|\mathbf{x},\theta)}(\boldsymbol{y}) = \frac{1+\kappa}{\kappa(\nu - K - 1)}\boldsymbol{L}^{-1} \tag{4.37}$$

Notice that these results are in line with the corresponding values for DER (Equations 4.5 and 4.6). Additionally, similar results are obtained if the generalization of DER to multiple dimensions [270] is considered.

Setting Up the Experiments The regression prior network method was implemented on top of the PhysNet NN model[32]. The output was modified to return energy and the parameters required by either a Normal distribution (Equation 4.19) or the Normal Wishart distribution (Equation 4.21). To create the ensemble (Equation 4.30), five models were trained such that each model returned the parameters of a Gaussian distribution. Each single model was trained for 100 epochs using the parameters of the standard version of PhysNet. The ensemble is then used to parameterise the $\mathcal{NIW}$ distribution using a temperature value of 10 for 1000 epochs. The dataset used for training was the QM9 database[189] that was filtered to remove molecules that did not pass the geometry consistency check. As before, the total number of samples was split into 80 % for training, 10% for validation, and 10% for testing. The RP model was evaluated according to the metrics used for DER and described in section 4.1.

4.2.3 Results

The evaluated system indicates an inferior predictive power of the RP model with an MAE of 0.47 kcal/mol and an RMSE of 1.36 kcal/mol (Figure 4.10). This error is three times the error obtained with the vanilla version of PhysNet for the same dataset. However, it is slightly larger than PhysNet DER with $\lambda = 0.2$, which has an MAE and RMSE of 0.36 and 0.87 kcal/mol, respectively. DER with a different value of $\lambda = 0.4$ returns an MAE of 0.30 kcal/mol and an RMSE of 0.75 kcal/mol. From the plot in Figure 4.10, it is noticed that the variance in the prediction increases at the centre of the energy range (-80 to -40 eV) while regions with lower or higher energy show small variance values.

An important aspect of applying the RP model is that the variance obtained with it can be related to the error. As mentioned before, the variance returned by the model must

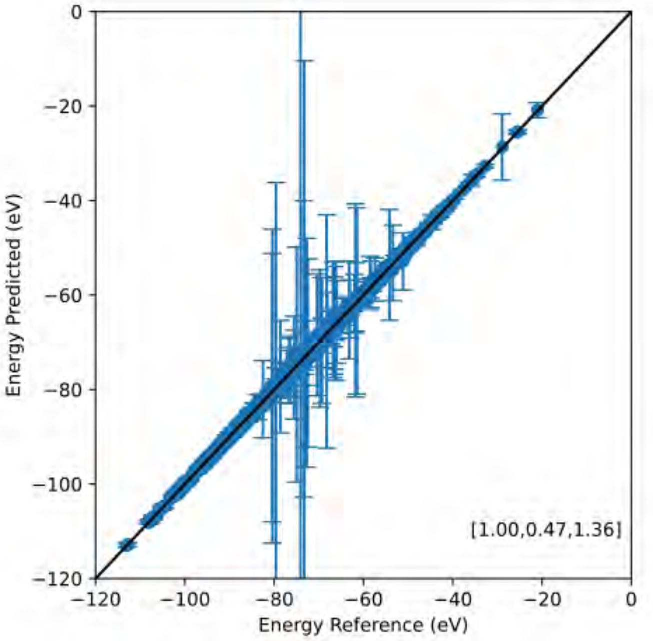

Figure 4.10: Scattering plot of the results for the RP method with the QM9 dataset on the test set. The error bars show the predicted variance. The insight shows the r^2 Pearson coefficient, mean absolute error (kcal/mol) and the root-mean-squared error (kcal/mol), respectively.

be directly related to the squared error[246]; if a model returns a value that can fulfil that condition, it is said is *calibrated*. Several tests can be performed to judge if a model is calibrated. As in the previous section, here, the calibration of the model was judged by plotting the empirical RMSE with respect to the Root Mean-Variance (RMV) of the model. Results in Figure 4.11A show a linear relationship between RMSE and RMV with a Pearson correlation coefficient close to 1 with small dispersion. The results for the correlation coefficient are better than the results obtained with DER models (See Figure 4.2). Nevertheless, the slope of the linear relationship between RMSE and RMV has a value close to zero. The reason behind that is that the scale of the RMV is one order of magnitude larger than the values of RMSE. This indicates that the variance values are overestimating the prediction error.

The distribution of the squared error and the variance highlights the difference between the scales of the two quantities, as shown in Figure 4.11B. The distribution of the squared error in the predictions obtained with the RP model is very sharp, with a peak centred close to zero, and has small tails that do not extend for very large values. On the contrary, the variance distribution is similar to an inverse Gamma distribution with a large value α and a small value β displaced to be centred around 0.2 eV. The decomposition of the variance in data and knowledge uncertainty shows the origin of the

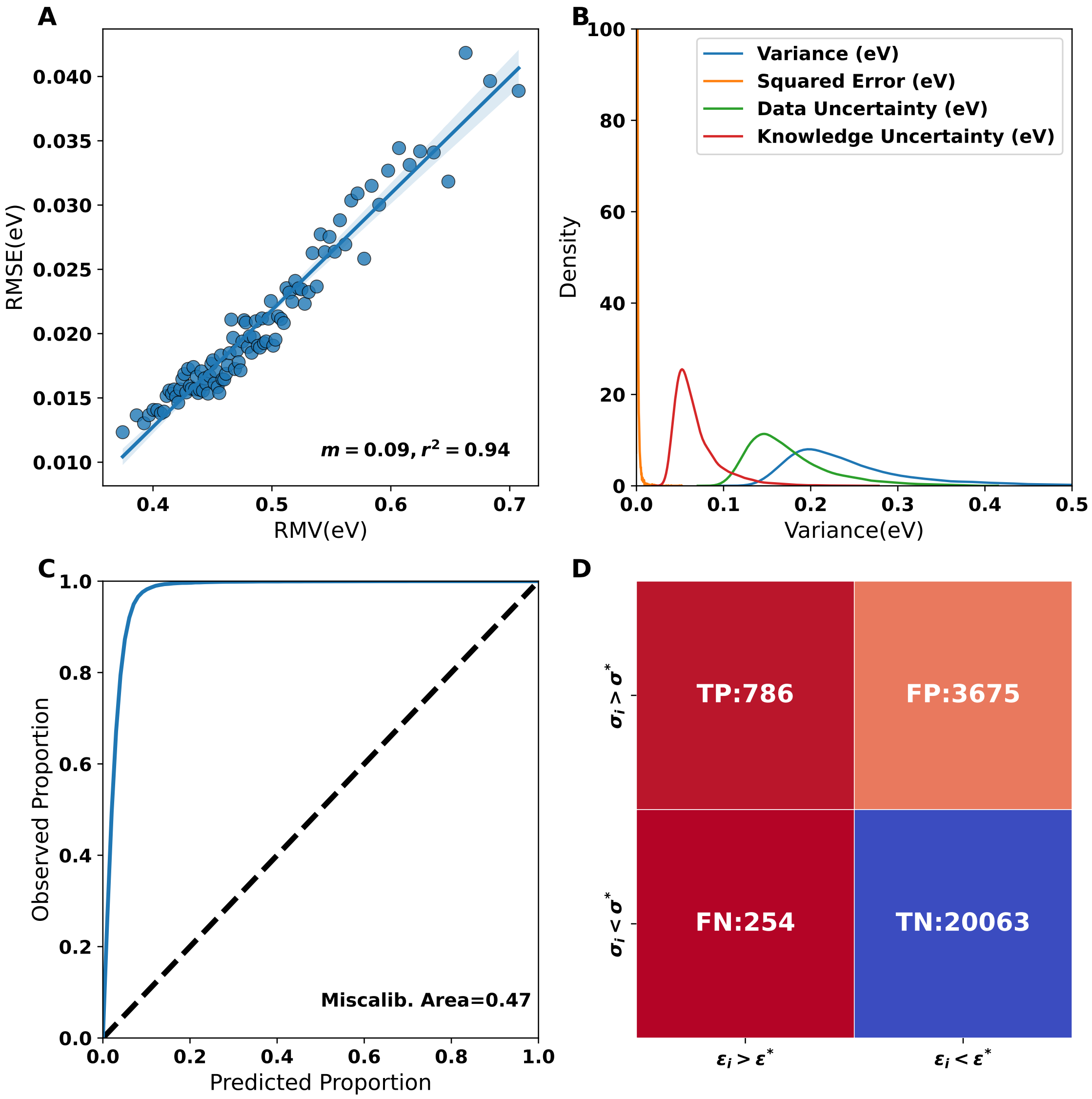

Figure 4.11: Calibration of the uncertainty predicted by the Regression Prior Method. Panel A shows the empirical Root-Mean Square Error (RMSE) compared with the Root Mean Variance (RMV) of the RP model training in QM9. The values were divided into 100 bins ranked with respect to the predicted variance. The slope (m) and Pearson correlation coefficient (r^2) are described inside the graph. Panel B displays the kernel density estimate of variance distribution, squared error, data and knowledge uncertainty for the RP model with values up to 95 %. Panel C depicts a calibration curve for the RP model. The x-axis shows the probability of obtaining the correct error value for the error in a given percentile, while the y-axis shows the true probability. The trend line shows a perfected correlated model. The value of the miscalibration area is given in the plot. Panel D displays the confusion matrix for the points of the test set. Inside each panel is the number of molecules that belong to the defined categories. The abbreviations refer to TP: True Positive, FP: False Positive, TN: True Negative, and FN: False Negative. For information on how those categories are defined, consult section 4.1.2.

problems in the prediction. Then, the knowledge uncertainty term related to the model is large and resembles again an inverse gamma distribution shifted to 0.07 eV. This can be interpreted as quantifying the limitations of using normal distributions to model the data. However, this value is also overestimated. The other uncertainty determined corresponding to the data uncertainty shows larger values. The distribution has a centre of mass around 0.15 eV but more spread with large tails. This indicates that the model is not confident about the amount of data that was provided. Previously, it was found that the database used for training does not have a large chemical diversity, and the large value of data uncertainty can be related to this lack of information in the training dataset.

The last test performed to assess the calibration of the RP model is by using a calibration curve (Figure 4.11 C), which helps evaluate the 'honesty' of the model predictions. Calibration curves have several limitations that were previously discussed in section 4.2.1. In this case, the calibration curve immediately shows that the model performs poorly with a step increase of the observed values in a small percentile of the predicted values. Complementary to the results, the miscalibration area was computed, and the value 0.47 indicates that the model does not return a meaningful variance prediction. Although the model showed that the variance prediction is insensitive to the value of the error, it can be instructive to analyze the results of the classification process, which separates the prediction of the molecules in the defined classes in section 4.1.2. The results are shown in Figure 4.11 D. It is clear that most of the samples are classified as true negatives (i.e. small error and small variance). These results need to be taken with care because the values of the variance are spread over a large range. On the other hand, the values of false positives (i.e. small error and large variance) are considerably large, which is another evidence of the underconfidence of the model.

From the previous analysis, it is noticed that the capability to relate the error in prediction with the obtained variance from the RP model is very poor, and the model is underconfident by returning large variances for a considerable number of samples. Nevertheless, the RP model can still be useful for detecting outliers, which will be the goal of the next chapter. In that case, it is interesting to evaluate if, among the predicted structures with large variance also, structures with large error can be found. To this end, the outlier detection plot of Figure 4.12 is used. The diagram shows the probability of finding the molecules with the largest errors among those labelled by the model with large variance. Conversely, the inverse relationship also holds. The results indicate a good performance in detecting large outliers with a probability of 92% to find the 25 largest structures among the 1000 structures with the largest variance. After that, a

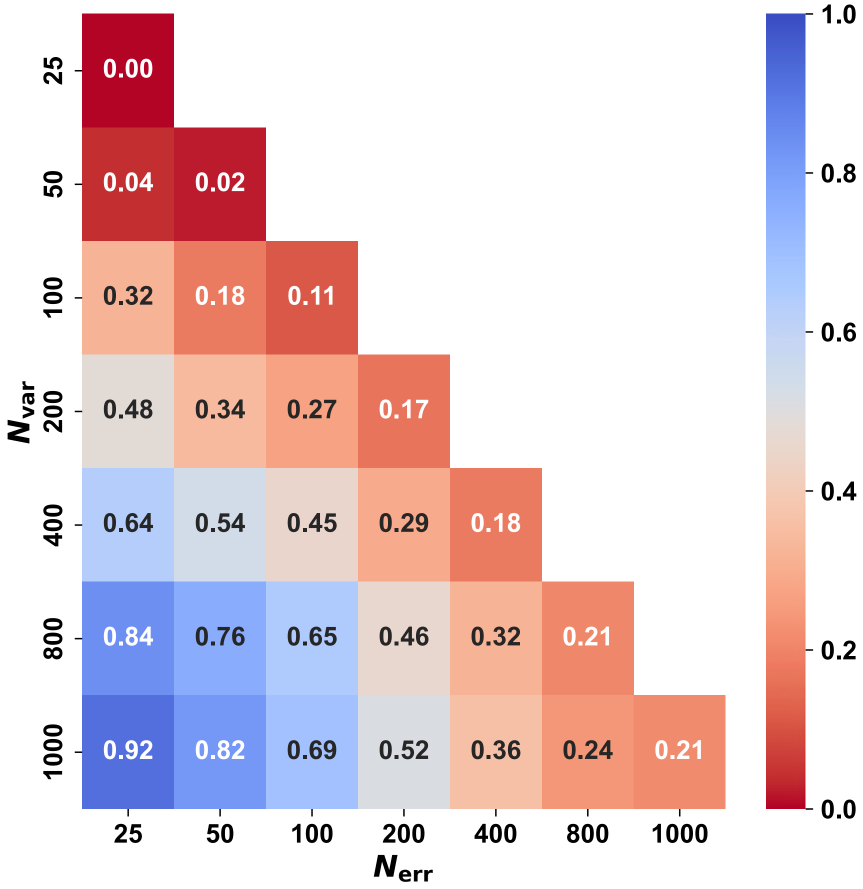

Figure 4.12: Outlier capability detection of the RP model. It is evaluated if a number N_{error} can be found in the N_{var} with the highest variance. A value of one means that all the structures with the largest errors can be found among those with the largest variance. The inverse relationship, the structures with large variance among those with large variance, also holds.

constant decay of the probability is observed. As expected in the direct case (i.e. the same number of structures with large error and large variance), the probability is zero, which indicates the lack of direct correlation between error and variance. However, this slowly increases to a maximum of 21%, meaning that around 210 of the 1000 molecules identified with large variance correspond to large errors.

4.2.4 Outlook

As a summary of this subchapter, it was found that despite the promising formulation of the RP model and the solid theoretical formulation of it, the model performs averagely for the prediction of the mean quantity (i.e. energy) and poorly for the variance. Some possible explanations for this behaviour are i) the limitations of assuming a Gaussian distribution to model the underlying data distribution ii) problems with the optimization procedure. The first limitation is associated with the data used for training. By using a normal distribution, the optimization is forced to reduce the predictive variance to improve the prediction of the mean value. This problem was previously observed for DER[271]. In consequence, it could be argued that the MVN model (Equation 4.20) used to fit the $\mathcal{NW}$ distribution (Equation 4.24) is not informative or that the data used to train the model can not be adequately described by a $\mathcal{NW}$ distribution.

The second problem associated with the training of RP is that the minimization of the loss function is numerically challenging, and the use of the temperature-annealing trick might not be the best solution to deal with it[2]. We tested different temperature schedulers to deal with this issue, resulting in the poorest results. Nevertheless, at the core of these problems are theoretical limitations. It is known that the training of multidimensional Gaussian models is complicated because the prediction of covariance matrices is numerically unstable[272–274]. Another problem found particularly for knowledge distillation techniques is that minimizing the KL divergence between student and teacher distribution leads to asymmetric gradients as a consequence of the asymmetry of the KL divergence (i.e. $KL(p||q) \neq KL(q||p)$)[275]. An alternative to deal with the asymmetry is to decouple the KL divergence into an MSE loss function and a cross-entropy term, which are weighted to avoid one having larger importance than the other[275]. Complementary, the use of a symmetric divergence such as the Jensen-Shannon divergence[276, 277] could be interesting to test. A simpler alternative is modifying the loss function to use only the MSE part or doing a sequential distillation[278].

Despite the problems related to training and the statistical assumptions, the RP model still has a good capability for outlier detection that corresponds with a linear relationship between RMSE and RMV. It would be of interest to take advantage of the current capabilities of the model by using it in combination with a *posthoc* uncertainty recalibration strategy[241, 249], which assures that the returned uncertainty matches the error prediction.

[2] We try to contact the authors to discuss other possible optimization strategies. Unfortunately, we have not engaged in discussion with them yet.

Chapter 5

Morphing of Potential Energy Surfaces

> It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with the experiment, it's wrong.
>
> Richard Feynman

This chapter presents the results of the application of a modified version of the morphing technique to a potential energy surface (PES). Here, a given PES is stretched or compressed by a linear transformation of the coordinates and the energy based on physically motivated constraints. The process is aimed to match highly accurate spectroscopic experimental data of so-called Feshbach resonances with theoretical scattering calculations. The technique is applied on the system of He-H_2^+ for which a PES at the highest level of theory (FCI), despite relativistic and quantum electrodynamic effects, is available. Additionally, PES at other levels of theory (MRCI and MP2) were tested. Our findings indicate an improvement in the position and intensities of the energy distributions of the Feshbach resonances. Remarkably, it was found that even the PES constructed with the FCI method showed small changes with respect to the experimental results. Additionally, it was found that the changes in the PES were more significant for the long-range region of the PES. The findings of this chapter set the stages for more comprehensive methods involving diverse spectroscopic information to obtain a better agreement with experimental results.

The results presented in this chapter have been previously published in Sci. Adv. in press

****Karl P. Horn collaborated in this work as joint first author*

5.1 Introduction

The potential energy surface (PES) representing the total energy of a molecule is a fundamental concept for characterizing the dynamics both in the gas and condensed phase[33, 34]. With high-quality PESs, the computation of experimental observables becomes possible with predictive power at a quantitative level. On the other hand, while essential measurable observables such as reaction cross sections, thermal rates, or relaxation times directly depend on it, the PES itself cannot be observed. This raises the question of how to obtain the most accurate PES for a given system. From an electronic structure perspective, it is known that within the Born-Oppenheimer approximation and neglecting relativistic and quantum electrodynamic corrections[279] full configuration interaction (FCI) calculations with large basis sets provide the highest quality for the total energies of a molecule. However, the unfavourable scaling of FCI with the number of electrons and basis functions prevents its routine use for constructing full-dimensional PESs for any molecule consisting of more than a few light atoms. Alternatively, one may approach the question from an experimentalist's perspective and argue that the "most accurate PES" is the one that best describes physical observations. Such an approach has been developed for diatomic molecules: the rotational Rydberg-Klein-Rees (RKR) method solves the "inversion problem" of obtaining the potential energy curve given spectroscopic information.[280] Rotational RKR has also been applied to triatomic van der Waals complexes[281, 282] but cannot be extended to molecules of arbitrary size. Indeed, solving the "inverse problem", i.e., determining the PES given experimental observables and an evolution equation from which these observables are calculated has in general turned out to be very difficult in chemical physics.[283] This concerns both the choice of observables as well as the actual inversion procedure.

An alternative that is not particularly sensitive to the dimensionality of the problem is to reshape the PES which was first done by trial-and-error[284, 285] and eventually lead to "morphing" PESs.[37] This method exploits the topological relationship between a reference and a target PES. Provided that calculations with the reference PES yield *qualitatively* correct observables $\mathcal{O}_{\rm calc}$ when compared with experimental observations $\mathcal{O}_{\rm exp}$, the squared difference $\mathcal{L} = |\mathcal{O}_{\rm calc} - \mathcal{O}_{\rm exp}|^2$ can be used to reshape the PES through linear or non-linear coordinate transformations ("morphing").[37] It capitalizes on the correct overall topology of the reference PES and transmutes it into a new PES by stretching or compressing internal coordinates and the energy scale, akin to stretching and bending a piece of rubber. Alternatives for reshaping PESs are machine learning-based methods such as $\Delta-$ML[286], transfer learning[287, 288], or differential trajectory reweighting.[289] Morphing has been applied successfully to problems in spectroscopy,[290] state-to-state reaction cross sections,[291] and reaction dynamics[292] for systems with up to 6 atoms[293]. Near-quantitative, full-dimensional reference PESs from electronic structure calculations have, however, so far rarely been available for direct comparison. For scattering experiments with He–H_2^+ such a PES is now available.[294] On the other hand, for weakly interacting triatomic van der Waals complexes accurate PESs were determined from fitting well depths and positions of radial minima to parametrized functions.[295] However, these studies relied heavily on explicitly available long-range information of the intermolecular interactions. The present work approaches the problem from a broader perspective, formulates and solves it as a machine learning-based task, and applies it to recently measured scattering data covering a wide range of intermolecular energies.

The He–H_2^+ molecular complex is an ideal proxy for the present work owing to the fact that the PES can be calculated rigorously at the highest level of quantum chemistry (FCI). The complex is also interesting in itself, and the current status of experimental and computational spectroscopy and reaction dynamics has recently been reviewed.[296] He–H_2^+, which is isoelectronic to H_3, is stable in its electronic ground state and features a rich reaction dynamics and spectroscopy. Experimentally, near- dissociation states[297, 298] and the low-resolution spectroscopy were reported for both, He–H_2^+ and He–D_2^+.[299] Assignments of the vibrational bands were possible by comparing with bound state calculations utilizing a FCI PES.[294] Only recently, it was possible to estimate the dissociation energy of ~ 1800 cm^{-1} from spectroscopic measurements.[299] This compares with earlier bound state calculations using the FCI PES predicting a value of $D_0 = 1784$ cm^{-1}.[294] This value was confirmed from a subsequent focal point analysis resulting in $D_0 = 1789(4)$ cm^{-1} for para-H_2^+.[300] Furthermore, a range of reactive collision experiments was carried out which yielded total and differential cross sections, depending on the vibrational state of the diatomic,[296] but with marked differences between observed and computed results. In particular, computationally predicted sharp reactive scattering resonances have not been found experimentally as of now.[296] Finally, the role of nonadiabatic couplings is of considerable current interest as a means to clarify the role of geometric phase in reaction outcomes and as a source of friction in the formation of the He–H_2^+ complex in the early universe. This provides additional impetus for a comprehensive characterization of this seemingly “simple” system.

The present work uses all very high quality experimentally measured Feshbach resonances for He–H_2^+[36] to morph potential energy surfaces. Feshbach(-Fano) resonances arise if a bound molecular state on a potential energy surface of a closed channel couples to scattering states in an open channel. [301, 302] The recoil translational energy is determined from measurements which are expected to probe large spatial areas of a PES and the underlying intermolecular interactions.[302] The redistribution of energy due to the Feshbach resonances has recently been mapped out comprehensively for He–H_2^+ and Ne–H_2^+ with coincidence velocity map imaging of electrons and cations, yielding very favourable agreement between theory and experiment.[36] In these experiments, the ionic molecular complexes are generated at separations of up to 10 a_0 between the rare gas atom and the molecular ion, confirming that the experiment indeed probes a large spatial extent of the PES, including its long-range part.

Here, morphing is applied to initial PESs ranging from essentially exact FCI (apart from non-Born-Oppenheimer, relativistic, quantum electrodynamic and remaining basis set effects) to medium- and lower-level methods, that is, Multi-Reference Configuration Interaction including the Davidson correction (MRCI+Q) and second-order Møller-Plesset perturbation theory (MP2). This allows us to determine the sensitivity of the PES and information content in the experimental observables about local and global features of the PES and to assess the performance of lower-level methods (e.g. MP2) compared with FCI. We found that starting from a PES of sufficiently high quality, the changes introduced by morphing can be related to parts of the PES that are probed by the experiments. At the same time, additional experimental observables, probing primarily the bound region for He interacting with H_2^+, will be required for morphing at the lower levels of quantum chemical theory.

5.2 Results

The three PESs considered in the present work, in decreasing order of rigour, were determined at the FCI, MRCI+Q, and MP2 levels of theory, using Jacobi coordinates R (distance between the centre of mass of the H_2^+ and He), r (distance between the two hydrogen atoms), and θ (the angle between the two vectors $\vec{\mathbf{R}}$ and $\vec{\mathbf{r}}$), see Figure 5.1A. To set the stage, scattering calculations with the FCI PES are considered which give quite satisfactory results when compared with the experimental data (Figure 5.2 A and Table 5.1). The measured kinetic energy distributions feature a series of peaks which reflect the rovibrational quenching associated with the decay of the Feshbach resonances.[36] On average, the positions of the peak maxima are reproduced to within 10.8 cm^{-1} whereas the maximum intensities, I_{max}, of $P(E)$ differ by 20.9 arb. u. (blue squares in Figure 5.2A).

Next, morphing is applied to all three PESs, including the FCI PES. The FCI PES has been validated with respect to experiment[36, 297–299] and therefore can serve as a suitable proxy for changes required for PESs at the MRCI+Q and MP2 levels of theory. Two morphing strategies were considered (Figure 5.1B): For Morphing M1, the total energy was decomposed into one-body ($\mathcal{V}_i^{(1)}$), two-body ($\mathcal{V}_i^{(2)}$) and three-body ($\mathcal{V}^{(3)}$) contributions,

$$V(R,r,\theta) = \mathcal{V}_{\rm He}^{(1)}+\mathcal{V}_{\rm H}^{(1)}+\mathcal{V}_{\rm H+}^{(1)}+\mathcal{V}_{\rm HeH}^{(2)}(r_{\rm HeH})+\mathcal{V}_{\rm HeH+}^{(2)}(r_{\rm HeH+})+\mathcal{V}_{\rm H_2^+}^{(2)}(r_{\rm H_2^+})+\mathcal{V}^{(3)}(R,r,\theta)\,, \tag{5.1}$$

and the morphing transformation was applied only to $\mathcal{V}^{(3)}(R,r,\theta)$. It should be noted that $\mathcal{V}^{(3)}(R,r,\theta)$ is defined as the difference between the total energy and the one- and two-body terms without implying a physical origin of the three-body contribution, such as an Axilrod-Teller interaction. Approach M1 is motivated by the assumption that all diatomic potentials $\mathcal{V}_i^{(2)}$ are of high quality so that changes are only required in the higher-order correction three-body term. In the second approach, called "M2", the PES is globally modified, including the two-body contributions. In other words, for M1 and M2 the morphing transformation (Eq. 5.4) is applied to $\mathcal{V}^{(3)}(R,r,\theta)$ and to $V(R,r,\theta)$, respectively. The reduction of the total loss and the associated parameter values are reported in Figures S1 and S2.

Morphing M1 applied to the FCI PES leaves most of the peak positions unchanged, see filled vs. open blue symbols in Figure 5.2D, but improves the peak heights considerably (by 30 %) as demonstrated in Figure 5.2E and Table 5.1 (rightmost column). These improvements are accommodated by reshaping the underlying PES as shown in Figure 5.3A: In the long-range ($R > 3.0$ a$_0$), the anisotropy of the morphed FCI-PES is somewhat decreased due to reshaping the PES around $\theta = 90°$ (T-shaped geometry) and D_e is decreased by ~ 50 cm^{-1}. One-dimensional cuts along the $r_{\rm HH}$ and R coordinates for a given angle θ show that changes in the PES become more substantial for larger $r_{\rm HH}$ with small changes in the depth of the potential wells but maintaining the overall shape (Figures S3 and S4). The changes with respect to R are noticeable for $R < 3.0$ a$_0$ with distortions of the energy contours at different angles θ, but maintaining the overall shape of the curves. For increasing R the changes are negligible compared with the original PES, reflecting the accurate treatment of the long-range interaction (Figure S3). 2D projections of the combined changes of $r_{\rm HH}$ and R at different angles show that the most pronounced modifications in the shape of the PES concern regions for $r_{\rm HH}$ larger than the equilibrium geometry of H_2^+ (Figures S5A ,S6A and S7A).

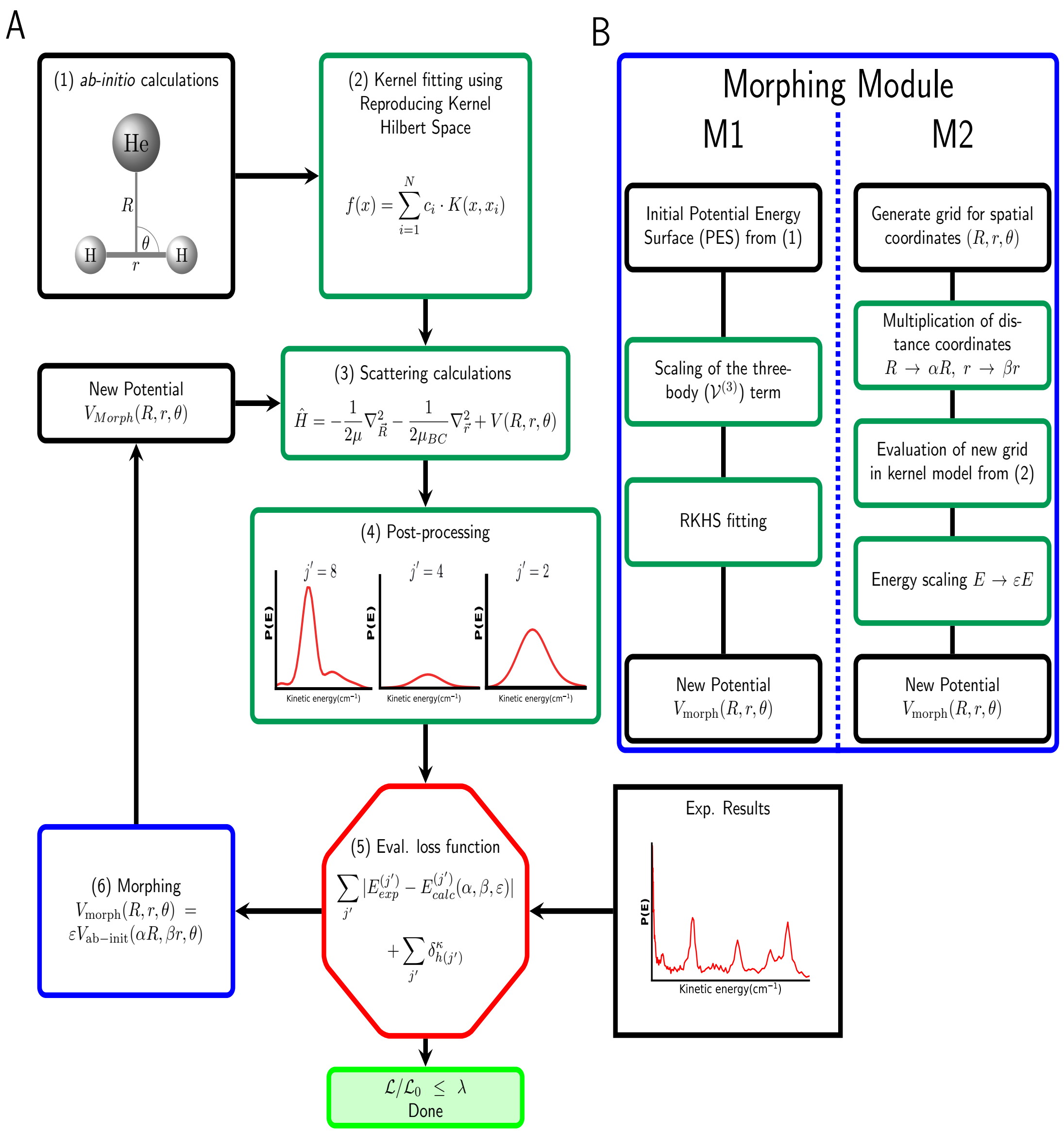
A
(1) ab-initio calculations
He
R
θ
H
H
r
(2) Kernel fitting using Reproducing Kernel Hilbert Space
$f(x) = \sum_{i=1}^{N} c_i \cdot K(x, x_i)$
New Potential
$V_{Morph}(R, r, \theta)$
(3) Scattering calculations
$\hat{H} = -\frac{1}{2\mu}\nabla^2_{\vec{R}} - \frac{1}{2\mu_{BC}}\nabla^2_{\vec{r}} + V(R, r, \theta)$
(4) Post-processing
$j' = 8$
$j' = 4$
$j' = 2$
P(E)
Kinetic energy(cm^{-1})
(6) Morphing
$V_{\rm morph}(R, r, \theta) = \varepsilon V_{\rm ab-init}(\alpha R, \beta r, \theta)$
(5) Eval. loss function
$\sum_{j'} |E^{(j')}_{exp} - E^{(j')}_{calc}(\alpha, \beta, \varepsilon)|$
$+ \sum_{j'} \delta^{\kappa}_{h(j')}$
Exp. Results
P(E)
Kinetic energy(cm^{-1})
$\mathcal{L}/\mathcal{L}_0 \leq \lambda$
Done
B
Morphing Module
M1
M2
Initial Potential Energy Surface (PES) from (1)
Scaling of the three-body ($\mathcal{V}^{(3)}$) term
RKHS fitting
New Potential
$V_{\rm morph}(R, r, \theta)$
Generate grid for spatial coordinates (R, r, θ)
Multiplication of distance coordinates
$R \to \alpha R,\ r \to \beta r$
Evaluation of new grid in kernel model from (2)
Energy scaling $E \to \varepsilon E$
New Potential
$V_{\rm morph}(R, r, \theta)$

Figure 5.1: **Morphing of *ab initio* potentials based on experimental data.** General flowchart of the morphing procedure (A): Module (1) implements the calculation of *ab-initio* points for the system under study, the triatomic HeH_2^+ with the definition of the respective coordinates indicated. Module (2) represents the fitting of the points obtained from the previous step using the Reproducing Kernel Hilbert Space Method, with the functional form used to approximate the given PES. Module (3) corresponds to the scattering calculations performed with the potential obtained in module (2), calculating the eigenstates of the Hamiltonian. Module (4) post-processes the results of the scattering calculations to yield $P(E)$ with examples for three values of j' displayed. Module (5) evaluates the loss function Eq. 5.5for morphing, comparing the experimental values of the energy distributions with the results of the scattering calculations. Module (6) carries out the actual morphing procedure, as explained in panel B. Morphing results in a new potential, and the procedure continues until the value of the loss function in module (5) does not improve further. The termination conditions are $\mathcal{L}/\mathcal{L}_0 \leq \lambda_{\rm M1} = 0.3$ or $\mathcal{L}/\mathcal{L}_0 \leq \lambda_{\rm M2} = 0.4$ for M1 and M2, respectively where $\mathcal{L}_0$ is the loss function of the unmorphed energy distribution, see Figure S1. Panel B: Morphing module (6) for procedures M1 (3-body) and M2 (global).

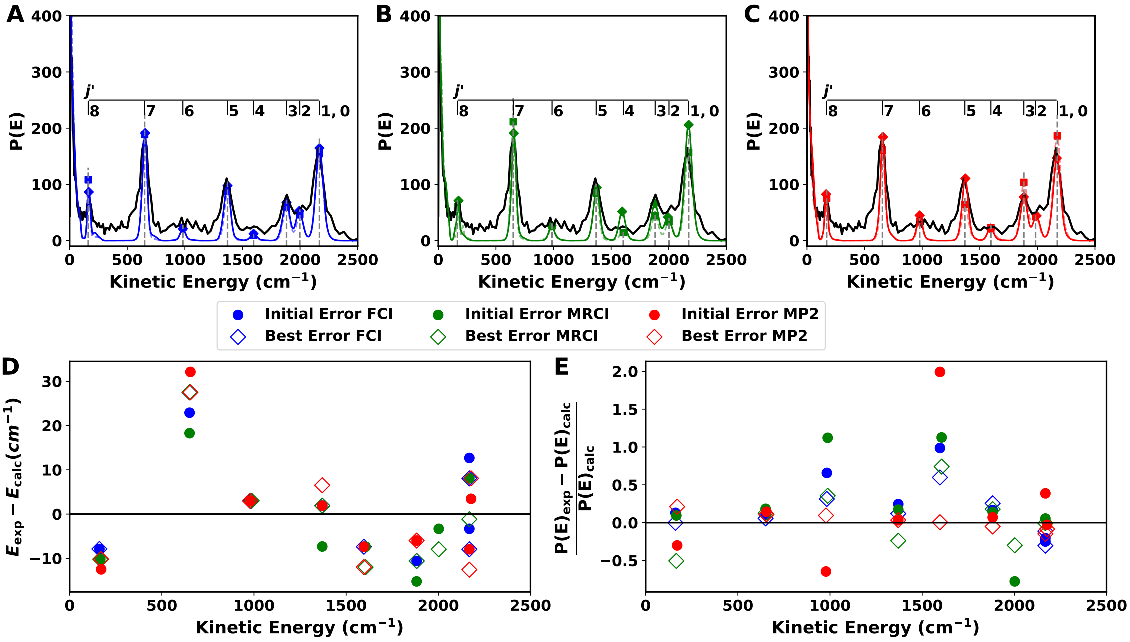

Figure 5.2: **Comparison of calculated energy distributions $P(E)$ from unmorphed and morphing M1 PESs with experimental results.** $P(E)$ obtained from experiment (black, data taken from Ref. 36) and full coupled channels calculations using the unmorphed and M1-morphed PESs for FCI (A), MRCI+Q (B), and MP2 (C). Computed results for the initial (blue, green, red dashed) and best (blue, green, red solid) morphed PESs are reported, with the residuals for the peak positions ($E_{\rm exp} - E_{\rm calc}$) and fraction of error in the peak heights ($\frac{P(E)_{\rm exp} - P(E)_{\rm calc}}{P(E)_{\rm calc}}$) for each PES shown in Panels D and E. The statistical measures for all models are summarized in Table 5.1. The experimental uncertainties are 3.5 cm^{-1} for the peak positions and $\sim$ 10 % for peak heights.

FCI calculations of entire PESs with sufficiently large basis sets are only feasible for few-electron systems. For larger systems, quantum chemical methods such as Møller-Plesset perturbation theory, multi-reference configuration interaction or coupled-cluster-based techniques need to be used instead. As reported in the two rightmost columns of Table 5.1, the initial MRCI+Q and MP2 PESs reproduce experimental peak positions within 10.3 and 13.1 cm^{-1} compared with 10.8 cm^{-1} from the FCI PES and for the peak intensities the RMSEs are 23.9 and 22.4 compared with 20.9 a.u. from using the highest level of electronic structure theory. On the other hand, the dissociation energy is smaller by more than 10% compared with the FCI PES due to partial neglect of correlation energy in the MRCI+Q and MP2 methods. This confirms that Feshbach resonances are not particularly informative with regards to features of the PES around the minimum energy structure ($R \sim 3.0$ a$_0$), although the wavefunctions sample this region extensively, see Figure 5.6. In other words, although an important characteristic of a PES such as the stabilization energy of the complex differs by 10 % or more, the energies and intensities measured in collision experiments are matched within similar

Surface	D_e (cm^{-1})	R_e/a$_0$	r_e/a$_0$	RMSE(E)	RMSE(I)
				(cm^{-1})	(arb. u.)
FCI Initial	2818.9	2.97	2.07	10.8	20.9
FCI Morphed (M1)	2772.0	2.95	2.07	11.9	13.7
FCI Morphed (M2)	2819.1	2.99	2.07	10.8	13.8
MRCI+Q Initial	2557.3	2.98	2.07	10.3	23.9
MRCI+Q Morphed (M1)	3414.7	2.98	2.08	12.2	21.9
MRCI+Q Morphed (M2)	2557.0	3.00	2.03	8.9	17.6
MP2 Initial	2494.0	2.99	2.07	13.1	22.4
MP2 Morphed (M1)	1685.6	2.93	2.12	12.8	10.9
MP2 Morphed (M2)	2492.8	2.97	1.74	10.0	11.8
MP2 Morphed (PES-to-PES)	2502.3	2.98	2.06	13.0(7)	22.9

Table 5.1: Dissociation energies (D_e in cm^{-1}) for He+H$_2^+$, coordinates for the minimum energy structures, R_e and r_e, and root mean squared errors (RMSE) for the peak positions and heights of the kinetic energy spectra for all initial and morphed PESs using both M1 and M2 methods. In all cases, the equilibrium geometry is linear He–H$_2^+$, i.e. $\theta = 0$ or $\theta = 180°$.

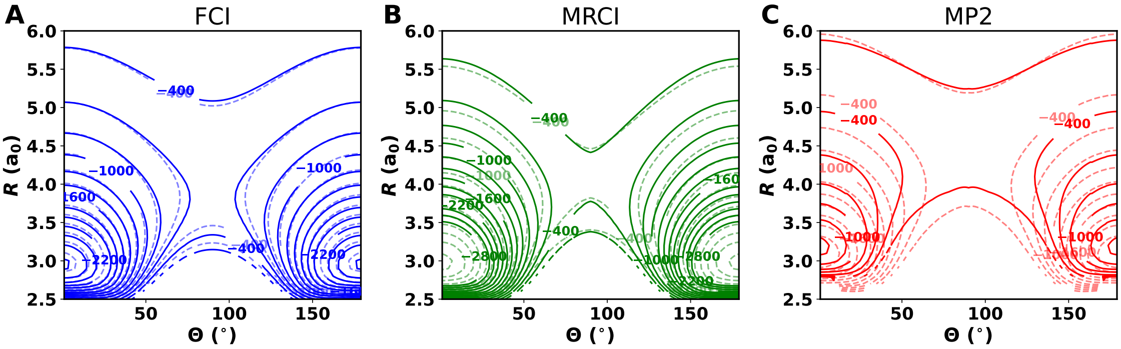

Figure 5.3: **Comparison between unmorphed and morphed M1 PESs.** Projections of the PESs for $r_{\mathrm{HH}} = 2.0$ a$_0$ for the three methods studied here. Isocontours for unmorphed PESs (FCI (blue), MRCI+Q (green) and MP2 (red) from left to right) are shown as dashed lines, whereas the M1-morphed PESs are solid lines. The zero of energy is set by the value at $r = 2.0$ a$_0$ and $R = \infty$. Energies are in cm^{-1}.

bounds.

Morphing M1 applied to the MRCI+Q and MP2 PESs supports this observation. The loss function evaluated in module (5) of the optimization, see Figure 5.1, decreased by 74% and 88% for the two cases, with improvements in the intensities by up to 50% for the MP2 PES, see Table 5.1 (rightmost column). However, the resulting PESs are clearly unphysical, with pronounced distortions in particular for the MP2 PES, see Figure 5.3C and dissociation energies either increased by 40 % for MRCI+Q or decreased by 30 % for MP2, respectively. Low-resolution experiments[299] provide an estimate for the dissociation energy $D_0 \sim 1800$ cm^{-1}, compared with $D_0 = 1794$ cm^{-1} from bound state calculations on the initial FCI PES[294] which features a well depth of $D_e \sim 2820$ cm^{-1}. This value of D_e serves as a reference for the remainder of the present work.

The percentage changes of the parameters $[\alpha, \beta, \varepsilon]$ scaling (R, r, V) provide further information about the transformation induced by morphing the initial PESs. For the FCI PES they are $(-0.6, -3.6, 0.0)\%$ compared with $(-0.6, 11.6, 1.0)\%$ for the MRCI+Q and $(0.3, -9.7, 0.1)\%$ for the MP2 PES. The most notable changes concern the H_2^+ vibrational coordinate $r_{\rm HH}$ for MRCI+Q $(+12.0\%)$ and MP2 (-10.0%). Such large changes are problematic since the many-body expansion used for morphing M1, cf. Eq. (5.1), relies on the quality of the two-body contributions, i.e., the H_2^+ and HeH$^+$ potential energy curves. However, MP2 underestimates the experimentally determined dissociation energy of the HeH$^+$ two-body interaction by 285 cm^{-1} (Figure S9) and accounts for an overall error of ~ 500 cm^{-1} in D_e for He–H_2^+. On the other hand, the two-body term for H_2^+ agrees to within 3 cm^{-1} between the three methods with remaining differences compared with experiment primarily due to neglect of non-Born-Oppenheimer contributions (Figure S10), relativistic corrections, quantum electrodynamic effects and remaining basis set incompleteness. To summarize: while M1-morphing improves the match between experimentally measured and calculated observables, it modifies the PES for the lower-level methods in an unphysical way. This is attributed to the fact that M1-morphing operates on the three-body term only and can thus not compensate for inaccuracies in the two-body contributions to the overall PES. In contrast, for FCI the changes for all characteristics of the morphed PES are moderate, underscoring the accuracy of both, the initial and morphed PESs from FCI calculations.

To reduce the dependence of the morphed PESs on the quality of the two-body contributions, morphing M2 was carried out. M2-morphing acts *globally* and independently on each of the internal degrees of freedom, see Figure 5.1. This makes M2 less prone to overcompensatory effects as observed for M1-morphing. For the MRCI+Q PES the improvement in the observables amounts to ≈ 14 % for the peak positions and ≈ 26 % for the peak heights. At the same time the changes in the PES are moderate, see Figure 5.4B, and the dissociation energy does not change (Table 5.1) although the energy scaling parameter, ε was allowed to vary. Similarly, for MP2, the RMSE for the positions and heights of the peaks improve by about 22 % and 47 %, respectively. Contrary to M1, morphing M2 does not substantially modify the well depth as reflected by the value of D_e, see Table 5.1. For FCI, morphing changes D_e by 0.2 cm^{-1} which is plausible as increasing the basis set from aug-cc-pv4z to aug-cc-pv5z changes D_e

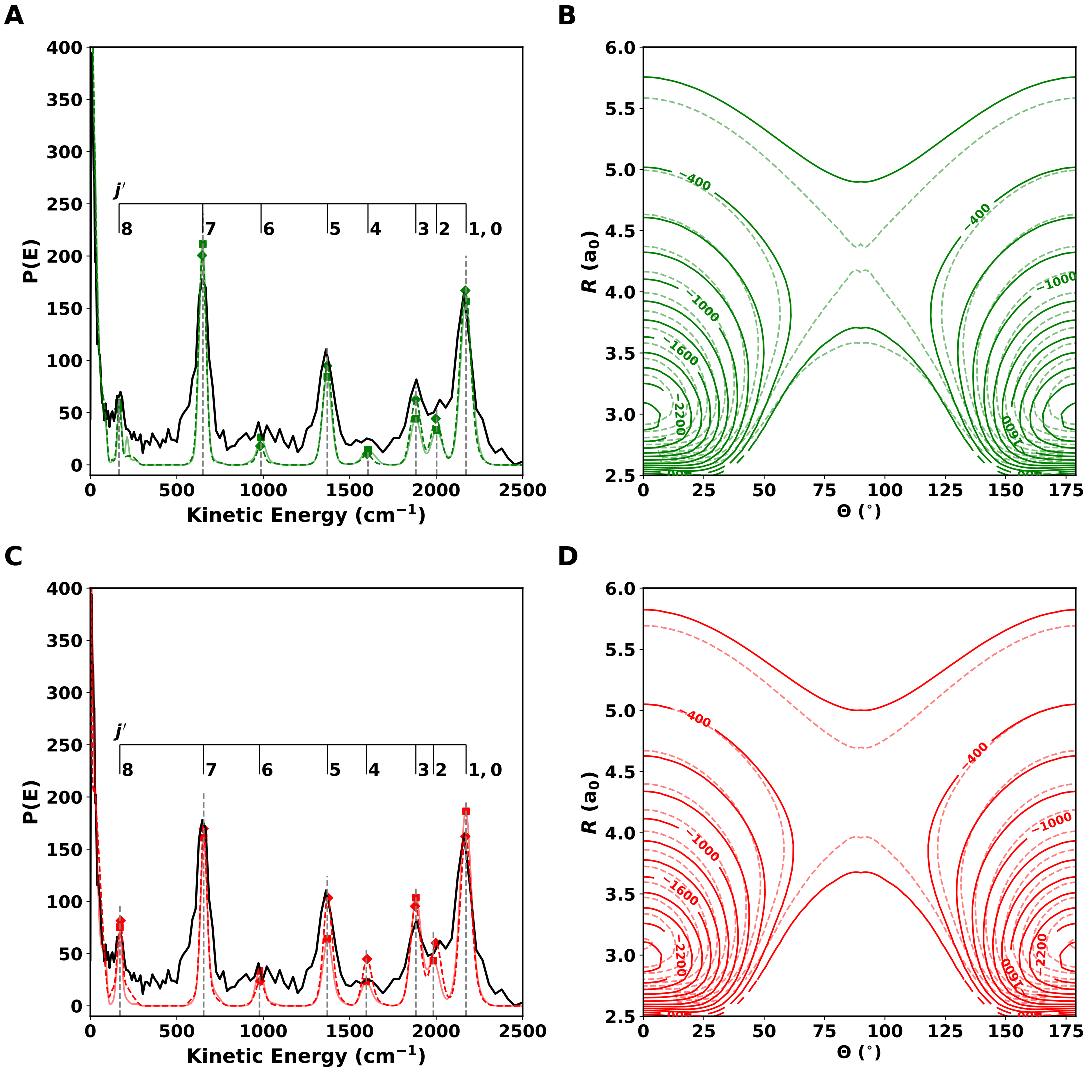

Figure 5.4: **Results of Morphing method M2.** Distributions $P(E)$ (panels A,C) obtained from experiment (black, data taken from Ref. 36) and full coupled channels calculations using the unmorphed (dashed lines) and M2-morphed (solid lines) PESs (B,D) for MRCI+Q (A,B), and MP2 (C,D). The RMSE for the peak positions and heights are reported in Table 5.1. The projections of the PES (B,D) are shown for $r = r_e$ (see Table 5.1) with the zero of energy set for the $r-$value considered and $R = \infty$. Energies are in cm^{-1}. The changes in the PES suggest that the observables are primarily sensitive to the long-range part and the repulsive wall of the PES.

by 5 cm^{-1} (Figure S11) and expected smaller changes when further increasing to the aug-cc-pv6z basis. This is confirmed for MRCI+Q calculations for which D_e changes by 2.5 cm^{-1} between aug-cc-pv5z and aug-cc-pv6z bases, see Figure S12.

For the optimal morphing parameters, M2 applied to the MRCI+Q PES yields an enlargement of R by ~ 1 % whereas r_{HH} is reduced by 1.9% and ε remains unaffected. The reduction in r_{HH} leads to a small increase in the height of the barrier between the two wells of the potential (Figure 5.4B) and a corresponding increase in the energy of the transition state, as observed in the minimum energy path (MEP), see Figure S13, for the angular coordinate. This effect is compensated by a positive displacement of the values of R (Figure S14) for the MEP. On the other hand, for the MP2 surface, the morphing parameters are ($+0.6$, $+19.0$, -0.04) %. The large positive value for β results in a displacement of the H_2^+ bond length to a shorter equilibrium value (Figure S15 and S16). For the R coordinate, the values are also reduced while the barrier height remains unchanged (Figure S14). As for M1, in the MP2 and MRCI+Q PESs the largest changes are observed in the r_{HH} coordinate. However, in the M2 method, scaling of the global PES results in a better performance for the calculation of the observable and a better physical description.

Finally, morphing one PES into another one can probe the flexibility of the morphing transformation as a whole. To this end, the MP2 PES was morphed to best compare with the FCI PES in a least squares sense according to method M2, i.e., by finding parameters $(\alpha, \beta, \varepsilon)$ that minimize $(V_{\mathrm{FCI}}(R, r, \theta) - \varepsilon V_{\mathrm{MP2}}(\alpha R, \beta r, \theta))^2$ without specifically weighting low- or high-energy regions in the fit. In this case, no experimental data was used in the refinement. Rather, the performance of the morphed PES was tested *a posteriori*. This optimization procedure reduces the RMSE between the FCI and unmorphed vs. morphed PES by about 30% (from 138 cm^{-1} to 87 cm^{-1}, see Figure S17). The changes in the topology of the surface in Figure 5.5C indicate that the morphed MP2 PES is "pulled towards" the FCI PES: Consider, for example, the isocontours for -400 cm^{-1} for which the original MP2 isocontour (blue) is far away from the FCI target contour (red), whereas the morphed PES (grey) is deformed towards the grey target isocontour. Closer inspection reveals this to occur for all the other isocontours in Figure 5.5C as well. The barrier separating the [He–HH]$^+$ and [HH–He]$^+$ minima is reduced, which is also seen in the minimum energy path (see Figure S18).

The results of the scattering calculations performed with the surface from the PES-to-PES morphing procedure (Figure 5.5A) are overall slightly inferior to those obtained

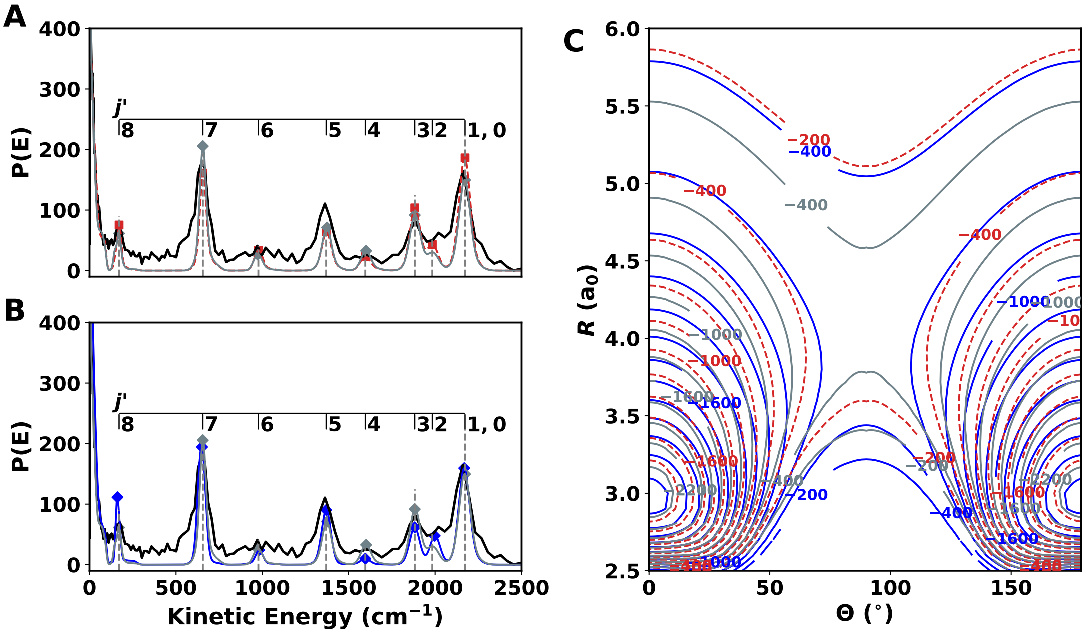

Figure 5.5: **PES-to-PES Morphing.** Panel A: Cross-sections obtained from experiments (black, data taken from Ref. 36) and scattering calculations on the unmorphed MP2 (dashed light red) and the morphed (grey) PESs for M2 PES-to-PES morphing procedure with the FCI PES as target. Panel B: Same as Panel A but comparing the best morphed PES (grey) to the unmorphed FCI surface (solid blue). Panel C: 2D projections of the PES for $r = 2.0$ a$_0$ for unmorphed FCI (solid blue), unmorphed MP2 (dashed light red) and best-morphed PES (grey). The zero of energy is set to the value of the PES at $r_{\rm HH} = 2.0$ a$_0$ and $R = \infty$. Energies are in cm^{-1}. All data points are equally weighted; the performance of the morphing transformation may be changed by differentially weighting attractive and repulsive regions of the PES.

from the initial FCI and MP2 PESs, when compared with the experimental data: a negligible increase of the RMSE for the peak positions ($< 1\%$) and intensities (2.2 %) is found. Moreover, the fact that the morphing transformation increases the well depth by merely 10 cm^{-1} indicates that a morphing transformation operating only on distances and the energy is not sufficiently flexible to accommodate global changes between topologies as different as FCI vs. MP2. Some further improvement might be obtained by more heavily weighting data points in the attractive region compared with the repulsive well which was, however, not considered in the present work.

The results indicate that at all levels of theory improvements in describing the experimental observables are possible. At the same time morphing applied in the fashion done here provides a stringent test to probe the quality of an initial PES at a quantitative level - with higher initial levels of theory, the changes that need to be accommodated decrease and specific deficiencies of a particular quantum chemical approach can be unveiled.

5.3 Discussion and Outlook

Given that essentially exact quantum calculations are possible for the He–H_2^+ complex,[36, 294, 299] the present results highlight what can and cannot be learned about molecular PESs — the central concept in classical and quantum molecular dynamics — from accurate and rather comprehensive experimental data based on Feshbach resonances. One hallmark of such quantum scattering resonances is the large spatial extent of the PES which the resonance wavefunction probes (Figure 5.6 and discussion in SI). In this regard, the kinetic energy spectrum obtained from the decay of the Feshbach resonances differs from spectroscopic observables, typically involving bound states sensitive to smaller spatial regions of the PES.[37]

In addition to the actual changes of the PES, a comparison of the two morphing procedures employed provides insight into the relationship between the PES, the information provided by specific observables, and how this information can be used to improve an initial PES. First, the much better performance of morphing the global interaction energy instead of restricting to the three-body contributions reveals the importance of corrections already at the level of two-body interactions. Moreover, the physically meaningful changes to the PES identified by the global morphing concern essentially the anisotropy in the long range. To this end, comparatively small changes of the PESs result in notable improvements in the agreement between calculated and measured observables. This is in line with the expectation that Feshbach resonance wavefunctions mainly probe the anisotropy of the PES in the long-range. Both observations taken together suggest extending the morphing transformation to include higher order terms (e.g. $\alpha r \rightarrow \alpha_1 r + \alpha_2 r^2 + \cdots$) or non-linear terms (akin to a neural network activation function) in the coordinate transformation. Including the angular degree of freedom θ in the morphing transformation as well yields further improvements, see Figures 5.8 to S21 and discussion in the SI.

The present work provides information about the behaviour of molecular PESs from lower (MP2) to very high (FCI) levels under morphing. It would also be interesting to characterize the effect of using different basis sets in the quantum chemical calculations. As an example, MRCI+Q calculations using the aug-cc-pV5z and aug-cc-pV6z basis sets changes the interaction energy between He and H_2^+ by 2.5 cm^{-1}, see Figure S12, compared with a well depth $D_e = 2557$ cm^{-1}. Hence for the basis sets used in the present work the effect is expected to be small. However, if smaller basis sets need to be used, as will be the case for larger systems, the effect will be considerably larger.

It is valuable to juxtapose the present effort to improve molecular PESs using experimental data with earlier work on van der Waals complexes between rare gas atoms and diatomic molecules. This approach was based on heavily parametrized functions including detailed expressions for the long-range part of the intermolecular interactions in which primarily the well depths and positions of the minima of the radial strength functions and the steepness of the repulsive wall were allowed to vary.[295, 303] Such a strategy was successful in fine-tuning PESs but also relied on an appreciable amount of detailed information: for example, more than 20 parameters are required to define the long range interaction between the rare gas and the diatomic molecule. In addition, uncertainties in the parameter values provided information about their sensitivity to experimental observables.

Contrary to this, the present work adopts a more holistic approach that also scales well to larger systems by deforming the entire PES to embed experimental observables. No particular physical meaning is then attributed to the morphing parameters and reporting uncertainties on them is of less immediate interest also because it is evident that multiple valid and meaningful solutions to the problem exist in general. The technique capitalizes on the fact that high-dimensional, global PESs can now be computed at sufficiently high levels of quantum chemistry[288] and obtaining a flexible machine learning-based rendering either from (reproducing) kernel representations or from neural networks is feasible.[100, 304] The approach followed here can be easily scaled to larger systems whereas the earlier “fine-tuning” strategies are typically limited to small systems.

At a fundamental level, the present findings raise the question how much and what experimental data is required to completely characterize a molecular PES. Indeed, the present work proposes several PESs with comparable average performance on the scattering observables, even though the shapes and local characteristics of the PESs differ greatly, illustrating that the information contained in the Feshbach resonances is not sufficient to uniquely define the PES. In particular, information on the bound-state region is missing. One possible way to answer the question which combination of observables is suited to completely characterize the dynamics of a molecular system has been developed in quantum information science and is referred to as quantum process tomography.[305] This has to be distinguished from the "Tomography of Feshbach resonance states" [36] which referred to the simultaneous measurement of multiple reaction products. Quantum process tomography goes substantially further by providing a mathematical prescription to completely characterize a quantum dynamical process. It has been adapted to molecular systems for example in the context of ultrafast spectroscopy. It is, however, still an open question how to adapt it to two- or many-body processes such as molecular scattering. In future work, quantum process tomography could be applied to the quest of uniquely defining a PES by making use of the mapping between the real-space representation of the molecular Hamiltonian and qubits.[306] This should allow for a systematic approach to identify the most important measurements which would then provide additional data for morphing PES.

5.4 Methods

5.4.1 Potential Energy Surfaces

For the present work, three PESs were employed. Full-dimensional PESs for He–H_2^+ were previously determined at the FCI/aug-cc-pV5Z and MRCI+Q/aug-cc-pV6Z levels of theory, respectively.[294] The reference data was represented as a reproducing kernel Hilbert space (RKHS)[38, 97] which provides a highly accurate interpolation and allows to encode the leading order long-range behaviour for large separations. In addition, a third PES using the same underlying grid for determining reference energies at the MP2/aug-cc-pV5Z level and also represented as a RKHS, was constructed for the present work. These calculations were carried out using the MOLPRO suite of codes.[307] All PESs are represented as a sum of diatomic potential energy curves together with an explicit three-body interaction. The complete many-body expansion for the He–H_2^+ system is given in Eq. (5.1), where distances $r_i \in \{r_{\mathrm{HeH}}, r_{\mathrm{HeH^+}}, r_{\mathrm{H}_2^+}\}$ in the two-body terms $\mathcal{V}_i^{(2)}$ are the distances between the respective atoms, whereas for the three-body term $\mathcal{V}^{(3)}(R, r, \theta)$ the coordinate r is the H_2^+ separation $r_{\mathrm{H}_2^+}$, R the distance between He and the centre of mass of the diatomic, and θ the angle between the two distance vectors $\vec{\mathbf{r}}$ and $\vec{\mathbf{R}}$. Finally, $\mathcal{V}^{(1)}$ corresponds to the respective atomic energies. The energies $\mathcal{V}_i^{(1)}$ and $\mathcal{V}_i^{(2)}$ were also determined at the respective level of theory from electronic structure calculations and the contributions $\mathcal{V}_i^{(2)}$ were fitted to analytical expressions described in Ref. 294. The fitting parameters for the FCI and MRCI levels of theory were published before and those for the MP2 level of theory are provided in the supporting information. Combining all this information, the three-body contribution $\mathcal{V}^{(3)}(R, r, \theta)$ was obtained on the grid used in the electronic structure calculations for $V(R, r, \theta)$ and represented as a RKHS.

5.4.2 Scattering Calculations

Integral scattering cross sections and scattering wave functions for He-H_2^+, resulting from a spatially distributed input wave packet, were evaluated using a home-written coupled-channels collision simulation based on the renormalized Numerov method.[308, 309] Details on these calculations have been given in earlier work [36] and only the salient features are presented here. The wavepacket simulations use Jacobi coordinates with $\vec{\mathbf{r}}$ the vector between the hydrogen atoms, $\vec{\mathbf{R}}$ the vector from the dihydrogen centre of mass to the helium atom and θ the angle between the two vectors. With $R = |\vec{\mathbf{R}}|$ and $r = |\vec{\mathbf{r}}|$, the total Hamiltonian is then

$$H_{\text{tot}} = -\frac{\hbar^2}{2\mu_{\text{cmplx}}}\nabla^2_{\vec{\mathbf{R}}} - \frac{\hbar^2}{2\mu_{\text{diat}}}\nabla^2_{\vec{\mathbf{r}}} + V(R, r, \theta) \ , \tag{5.2}$$

where μ_{cmplx} is the reduced mass of the three-body complex, μ_{diat} the reduced mass of the dihydrogen molecule, and $V(R, r, \theta)$ the three-dimensional PES. The total wavefunction of the system $\Psi(\vec{\mathbf{R}}, \vec{\mathbf{r}})$ is written as a product of $R-$, $r-$, and angularly dependent terms,

$$\Psi^{JMvj\ell}(\vec{\mathbf{R}}, \vec{\mathbf{r}}) \propto \sum_{v'j'\ell'} G^{Jvj\ell}_{v'j'\ell'}(R)\chi_{\text{diat},v'j'}(r) \sum_{m_j=-j}^{j} \sum_{m_\ell=-\ell}^{\ell} C^{JM}_{m_j m_\ell} Y_{\ell,m_\ell}(\theta_R, \varphi_R) Y_{j,m_j}(\theta_r, \varphi_r) \ , \tag{5.3}$$

see Ref.36 for more detail. Channels consist of tuples of quantum numbers v, j, and ℓ, corresponding to diatomic vibration, rotation and orbital angular momentum, respectively. In Eq. (5.3), $\chi_{\text{diat},v,j}(r)$ designates the rovibrational eigenstates of the molecule. Starting from a given entrance channel, the Schrödinger equation is solved numerically to obtain the radial wave functions $G(R)$ for the exit channel with quantum numbers (v', j', ℓ') connected with the entrance channel (v, j, ℓ). The total angular momentum, $\vec{\mathbf{J}}_{\text{tot}} = \vec{\mathbf{j}} + \vec{\mathbf{L}}$ obtained from coupling diatomic and orbital rotation, and parity are conserved under the Hamiltonian (5.2).

In the experiments, the He–H_2^+ complex (plus a leaving electron) is formed by Penning ionization (He*+H_2), and the scattering calculations considered in the present work describe the half-collision on the He–H_2^+ PES. The initial wavepacket $\phi(R)$ along the $R-$coordinate is approximated by Gaussian distributions centered around $R \approx 8\, a_0$.[36] The experiment prepares the input wavepacket with $j_{\text{wp}} = 0, 1$ for para- and ortho-H_2^+, respectively. However, as the system is prepared in a superposition of $J-$states, individual simulations need to be carried out for each possible value of J and partial wave ℓ. Then, the integral cross section is calculated as a weighted sum over the individual contributions for a given collision energy $E_{\text{col}}/\text{k}_\text{B} \approx 2.5$ K. The $J-$weights, which were calculated separately,[310] are shown in Figure S22. Experimentally, the initial state is prepared "in situ" whereby Penning ionization generates the He–H_2^+ complex. Thus, the initial state of the present quantum wavepacket simulations is in fact the result of an incoherent decay of a population of He*–H_2 complexes which leaves one with an unknown normalization. The experimentally observed quantity is a probability distribution $P(E)$ which is dimensionless. Here, the computed intensities are scaled such as to best reproduce the experimentally measured ones.

Evaluation of the collision cross section due to the spatially distributed input wavepacket can be accomplished by expanding $\phi(R)$ in a basis of eigenfunctions of H_{tot}. To this end, the time-independent Schrödinger equation was solved on a discretized interval of 1002 energies ranging from $100\ \text{cm}^{-1}$ below to $100\ \text{cm}^{-1}$ above the dissociation threshold of the given entrance channel. Because full coupled-channel calculations are computationally demanding, the considered set of initial wavepacket quantum numbers J and ℓ was limited to $(\ell/J) \in \{(0/0), (1/1), (2/2), (3/3), (4/4)\}$ for para- and $(\ell/J) \in \{(0/1), (1/1,2), (2/1,2,3)\ , (3/2,3,4), (4/3,4,5)\}$ for ortho-dihydrogen, respectively. For each coupled channel calculation a converged basis set of diatomic rotational states up to $j_{\text{max}} = 19$ and diatomic vibrational states up to $v_{\text{max}} = 5$ was used.

Solving the Schrödinger equation in this fashion allows for calculating the channel-resolved integral cross section for each energy in the discretized interval. For a given output channel, the eigenenergy $E_{v'j'\ell'} = E_{\text{int},v'j'\ell'} + E_{\text{kin},v',j',\ell'}$ can be decomposed into its internal and kinetic parts, respectively. By generating a histogram for all output channels (v',j',ℓ'), the cross-section can be expressed as a function of kinetic energy, which can be compared with the experimental results. Next, the kinetic energy histogram is convoluted using a Gaussian envelope to account for the finite resolution in the experiments.[36] Before convolution, and as shown in Figure S23, the computed peaks are sharp in E_{kin} which is a signature of Feshbach resonances. It should be noted that experimental peaks are clearly distinguishable and energetically match the theoretical predictions. However, the peak shapes and heights can vary, dependent on the histogram resolution and convolution width. In this work, only single initial vibrational excitations ($v = 1$) were considered, in order to exploit the experimental resolution of separate j' peaks in the cross-section as a function of kinetic energy [311].

5.4.3 Morphing

The morphing transformation considered here is

$$V_{\mathrm{morphed}}(R, r, \theta) = \varepsilon V_{\mathrm{ab-initio}}(\alpha R, \beta r, \theta)\ . \tag{5.4}$$

In Eq. (5.4), the three parameters $(\alpha, \beta, \varepsilon)$ are used for energy- (ε) and geometry-(α, β) related scalings. For the purpose of this work, the angle θ was not modified. The morphing procedure described further below optimizes the values of $(\alpha, \beta, \varepsilon)$ such that the difference between observed and computed features of the resonances is minimized. Application of such a procedure modifies local features (e.g. slope, curvature) of the PES but maintains its global shape.

For morphing M1 and M2 the refinement with respect to experimental values is formulated as an optimization problem with a loss function,

$$\mathcal{L} = \min_{\alpha,\beta,\varepsilon} \left[w_E \sum_{j'} |E^{(j')}_{\mathrm{exp}} - E^{(j')}_{\mathrm{calc}}(\alpha, \beta, \varepsilon)| + w_h \sum_{j'} \delta^{\kappa}_{h(j')} \right]\ , \tag{5.5}$$

to be minimized. Here, $E^{(j')}$ is the kinetic energy of each cross-section corresponding to an exit-channel j', and $\delta^{\kappa}_{h(j')}$ accounts for the difference in the peak heights between experimental and calculated values:

$$\delta^{\kappa}_{h(j')} = \begin{cases} (\Delta h(j') - h_{\mathrm{noise}})^{\kappa}, & (\Delta h(j') - h_{\mathrm{noise}})^{\kappa} > 0\ , \\ 0, & (\Delta h(j') - h_{\mathrm{noise}})^{\kappa} \leq 0\ , \end{cases} \tag{5.6}$$

where, $\delta_h(j')$ is *regularized* by subtracting $h_{\mathrm{noise}} = 10.0$ to avoid fitting experimental noise. By design, only values $\delta_h(j') > 0$ contribute to the error. Here $\Delta h(j') = |h^{(j')}_{\mathrm{exp}} - \gamma h^{(j')}_{\mathrm{calc}}(\alpha, \beta, \varepsilon)|$, where $h^{(j')}$ is the peak height of the cross section corresponding to an exit-channel j'. The parameter γ is recalculated after each iteration to best match the experiment by performing an additional 1d minimization over the squared difference in peaks heights. The weights $w_E = 1\,(\mathrm{cm}^{-1})^{-1}$ and $w_h = 1$ ensure that all terms and the total loss $\mathcal{L}$ in Eq. (5.5) are dimensionless and can be used to bias the fit to better reproducing certain observables than others which was, however, not done here.

The workflow to perform the optimization of Eq. (5.5) is shown schematically in Figure 5.1. In the first step, *ab initio* points of the PES are used to generate a RKHS kernel. Depending on the morphing procedure chosen, a new RKHS needs to be generated (for M1) or the existing kernel will be reused (for M2). All kernels are constructed and evaluated using the "fast" method.[38] The obtained PES is passed to the scattering code to perform the wavepacket propagation. Next, the resulting cross-sections are processed and then compared with the available experimental data. If the difference between experimental and calculated values matches a given tolerance the cycle finishes; otherwise, the PES is modified by three parameters as described in Eq. (5.4) following the chosen morphing approach. The values of the parameters α, β and ε were obtained by a non-linear optimization using the NLopt package[312]. For further details about the optimization procedure, see the SI.

5.5 Additional results

Complementary to the presented results, here we discuss the quality of the results of morphing M1 for the methods MRCI and MP2 and the wavefunction changes in the resonances. Finally, a first attempt to perform morphing of the potential energy surface while modifying the angular term is discussed.

5.5.1 Morphing M1 for the MRCI and MP2 PESs

Multi-Reference CI: Figure 5.2B compares the cross sections from experiments with the results from computations with PESs before and after morphing M1 for the MRCI+Q PES. Overall, the RMSE for the energies changes from 10.3 to 12.2 cm^{-1}, whereas the intensities improve from an RMSE of 23.9 to 21.9 arb. u. The results indicate that M1 has the most pronounced impact on intermediate values of j' (i.e. $j' = 4, 5$); see Figures 5.2D and E. Changes in the peak energies do not show a clear trend. The largest improvements are observed for $j' = 5$ and for $j' = [0, 1]$. Errors for peaks with $j' = 8$ and $j' = 6$ do not reduce using M1. The remaining peaks showed an increase in the error after applying M1. For the peak intensity, again, the largest improvement is observed for the $j' = [0, 1]$ peak. For most other peaks, with the exception of $j' = 5$ and $j' = 8$, there is clearly an improvement in the intensities.

The initial and morphed MRCI PESs are compared in Figure 5.3B. In this case, morphing increases the anisotropy at long-range compared to the initial PES. However, changes are more pronounced than for the FCI PES. One-dimensional cuts along the r_{HH} and R coordinates for given angle θ are provided in Figures S25 and S26. As for the FCI PES, the difference between the initial surface and the morphed surface is more pronounced as r_{HH} increases. The 1D cuts of the surface at different values of r_{HH} (Figure S26) show further evidence of the change in the depth of the potential well. The modifications of the energy curves with respect to the r_{HH} coordinate follow the same trend as the FCI surface.

MP2: The results for the lowest-quality surface (MP2) are shown in Figures 5.2C and 5.3C. The RMSE for the energies improves from 13.1 to 12.8 cm^{-1} whereas for the intensities, it changes from 22.4 to 10.9 arb. u. Particularly notable is the improvement in the intensities by more than a factor of two. Overall, the changes in the position of the energies and the intensities of the peaks for the calculated cross sections are more pronounced than for the FCI and MRCI+Q PESs. The energy position for peaks with large j' ($j' = 7$ and $j' = 8$) improve by ≈ 5 cm^{-1}. Another difference is that the shoulder of the peak at $j' = 8$ that appears for the two previously described surfaces is not visible for the MP2 surface. For the peaks with $j' = 4$ and $j' = 5$, the error with respect to the experimental spectra upon morphing increases slightly.

The original MP2 PES and its morphed variant for a H_2^+ separation of $r_{\mathrm{HH}} = 2.0$ a$_0$ are reported in Figure 5.3C. Because Møller-Plesset second-order theory is a single-reference method and makes further approximations, the changes in the topology of the PES are considerably larger than for the FCI and MRCI+Q PESs. Most of the isocontours are compressed compared with the initial MP2 surface, and the well depth is reduced from 2493 cm^{-1} to 1684 cm^{-1} (Table 5.1), see Figure S27. The one-dimensional cuts along the r_{HH} and R coordinates for given θ, see Figures S28 and S29, show that as r_{HH} increases the single-reference assumption of the method, leading to convergence problems for small R. As a consequence of the contraction of the potential wells, the barrier of the transition state at $\theta \approx 90^\circ$ is increased, which is further confirmed by the Minimum Energy Path (MEP) shown in Figure S30C. A more detailed analysis of the MEP (Figure S31C) reveals a small increase in the energy of the transition state along the angular coordinate θ. On the other hand, for the $R-$coordinate a non-physical barrier emerges at around 3.5 a_0.

5.5.2 Resonances under Morphing

The cross sections depend on the binding energy between He and H_2^+ as opposed to the relative kinetic energy of the two reactants show distinct peaks that are no longer separated by the final states (j') of the H_2^+ fragment but rather appear as one or several Feshbach Resonances per input J and ℓ at certain values of the binding energy. Both the energy at which a Feshbach Resonance appears, and the distribution of intensities in all exit channels, depend sensitively on the topography of the PES. In consequence, the effect of morphing on the PES can influence the number, energy and intensities of the Feshbach resonances. To illustrate this, it is instructive to consider projections of wave functions for particular resonances to characterize how changes in the PES, which lead to changes in the collision cross-section, are reflected in the radial and angular behaviour of the wave function.

Figure 5.6 shows the square of the $(v' = v)$ and $(j' = j)$ components of the resonance wave functions (first and third rows of panels) and corresponding resonances in the cross-section (second and fourth rows of panels) for the dominant ℓ and J contributions for para- and ortho-H_2^+ for all three unmorphed and morphed PESs, respectively. The number, position(s) and intensities of the spectroscopic features respond to morphing in a largely unpredictable way. As an example, the unmorphed and morphed PESs at the FCI level are considered for para-H_2^+ with $(\ell = 4, J = 4)$ (left column, rows 1 and 2 in Figure 5.6). Although M1 changes the topology of the morphed PES only in a minor fashion, the effect on the wavefunctions and resulting spectroscopic features is clearly visible. For the unmorphed FCI PES there is one resonance at –8.1 cm^{-1} which splits into two resonances at –2.1 cm^{-1} and –16.3 cm^{-1} of approximately equal height upon morphing the PES. Accordingly, the wavefunctions also differ, in particular in the long-range part, i.e. for large R. Similar observations were made for the wavefunctions on the MP2 PES, whereas for the MRCI PESs the changes in the wavefunctions are comparatively smaller.

Conversely, for ortho-H_2^+ the resonances of both FCI and MRCI PESs are affected in a comparable fashion and more noticeable changes to the resonance wave function are observed than for para-H_2^+. Whilst the resonance wave functions are shifted to larger R in the cases of FCI and MP2, the MRCI resonance wave function only experiences a small shift. Significantly, even though the anisotropy of the PESs only changes in a minor fashion under morphing, all three resonance wave functions respond owing to a change in the superposition of outgoing partial wave (quantum number ℓ'). For the FCI and MP2 PESs, angular/radial coupling is enhanced by morphing, which leads to the elongation of certain lobes in the wavefunctions along the $(R, \theta)-$direction for ortho-H_2^+–He. This contrasts with para-H_2^+–He for which unique assignments of the ro-vibrational quantum numbers are possible from conventional node-counting.

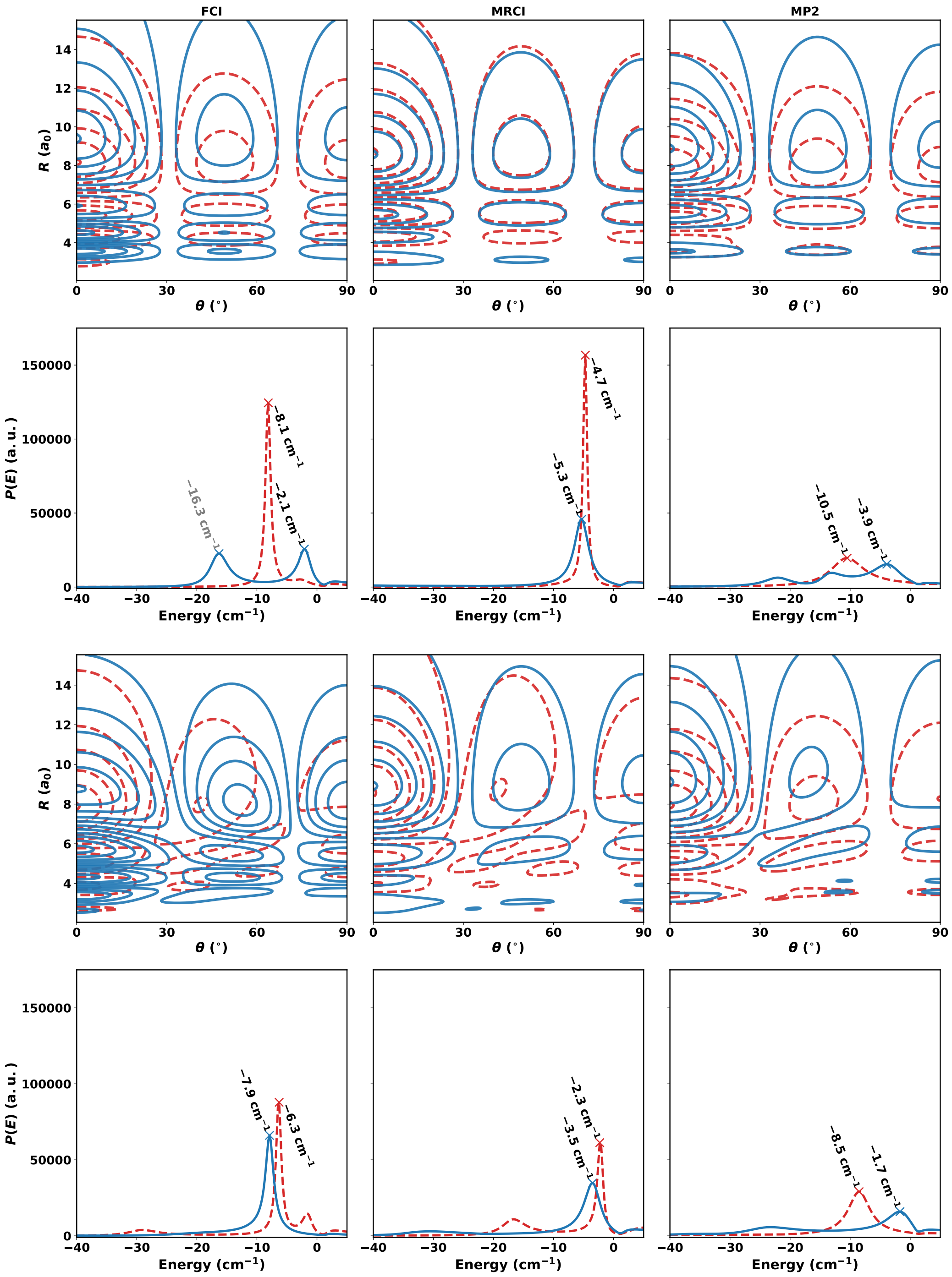

Figure 5.6: Comparison of the unmorphed (red, dotted) and morphed (blue) absolute value squared resonance wave functions in two dimensions (R, θ) in the case of para-H$_2^+$ $\ell = 4, J = 4$ (upper two rows) and ortho-H$_2^+$ $\ell = 4, J = 5$ (lower two rows) for resonance energies as marked and labelled in the corresponding cross sections are shown as a function of binding energy (second and fourth rows for para and ortho, respectively). The resonance wave functions have been scaled to have a maximal value of one, and the contours occur at 0.01, 0.1, 0.25, 0.5, 0.75 and 0.99.

5.6 Supporting information

Additional figures and supporting information for this work can be found at http://doi.org/10.1126/sciadv.adi6462 or at: https://github.com/LIVazquezS/SI_PhD_Thesis/blob/main/SI_Chapter5.pdf.

5.7 (Global) Angular Power Morphing of MP2

In addition to morphing distances and the energy scale, angular morphing was explored. Including the angular degree of freedom in morphing the PES is less straight forward, since any morphing transformation should conserve the underlying symmetry of the PES. Thus, straightforward linear scaling of the angle θ in the present case is not possible. Instead, the transformation needs to leave $\theta = 0, \pi/2$ (since the dihydrogen ion is homonuclear) and π invariant. One possible transformation which fulfills this requirement is

$$f_\eta(\theta) = \pi\left(1 - H\left(\frac{\theta}{\pi} - \frac{1}{2}\right)\right)\frac{1}{2}\left(\frac{2\theta}{\pi}\right)^\eta + \pi \cdot H\left(\frac{\theta}{\pi} - \frac{1}{2}\right)\left(1 - \frac{1}{2}\left(2 - \frac{2\theta}{\pi}\right)^\eta\right) . \tag{5.7}$$

Here, $H(x)$ is the Heaviside step function defined as:

$$H(x) = \begin{cases} 0 & If\ x < 0 \\ 0.5 & If\ x = 0 \\ 1 & If\ x > 0 \end{cases} \tag{5.8}$$

The first term in equation 5.7 is responsible for morphing angles $\theta < \pi/2$, whereas the second term is the mirror image around the $(\pi/2, \pi/2)$ point, responsible for morphing angles $\theta > \pi/2$ and $f_\eta(\pi/2) = \pi/2$ (See Figure 5.7).

The effect of including the angular coordinate in the morphing was explored by morphing the MP2 PES to the FCI PES as described in the main manuscript. The optimization procedure reduces the RMSE between FCI and the morphed PES by around 40 % from (138 cm^{-1} to 75 cm^{-1}). This compares with an improvement by 30 % (138 cm^{-1} to 87 cm^{-1}) without morphing the angle. Figure 5.8 shows the ensuing changes in the PES. It is found that the total loss improves by a factor of two compared to morphing without the angular degree of freedom, and for the PES-to-PES morphing, the RMSE improves by 25 % (Figure S17).

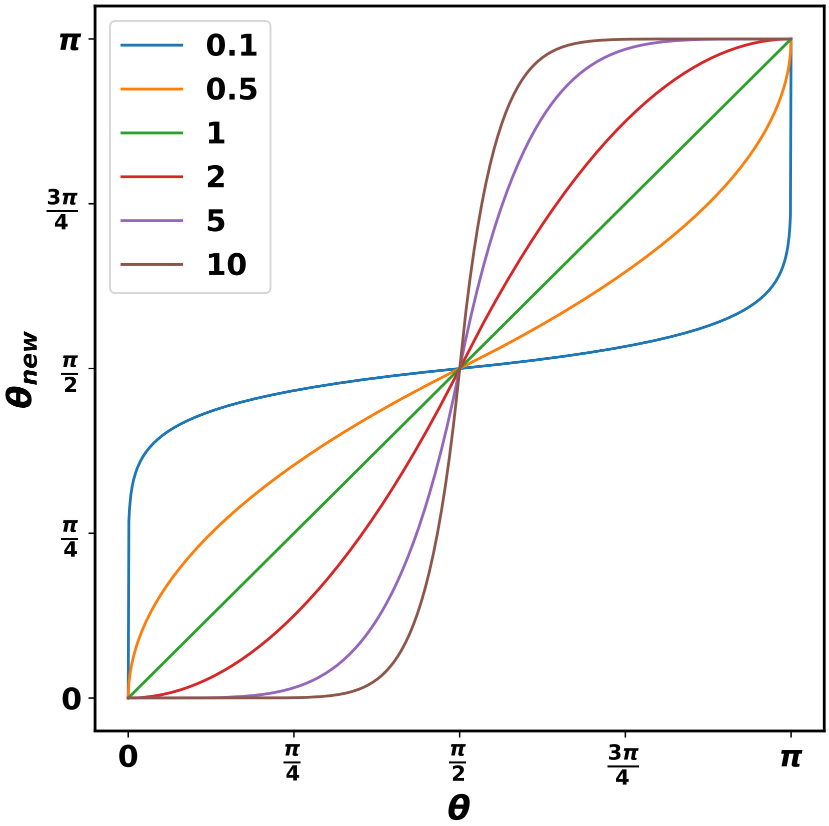

Figure 5.7: **Angular power morphing function** (see Eq. 5.7) used to transform the angular term of the PES. The transformation must keep the terms at $\theta = [0, \frac{\pi}{2}, \pi]$ invariant. The coloured lines corresponding to values of η ranging from 0.1 to 10 as indicated in the legend show the behaviour of the remapped angle as a function of the original in the domain $[0, \pi]$.

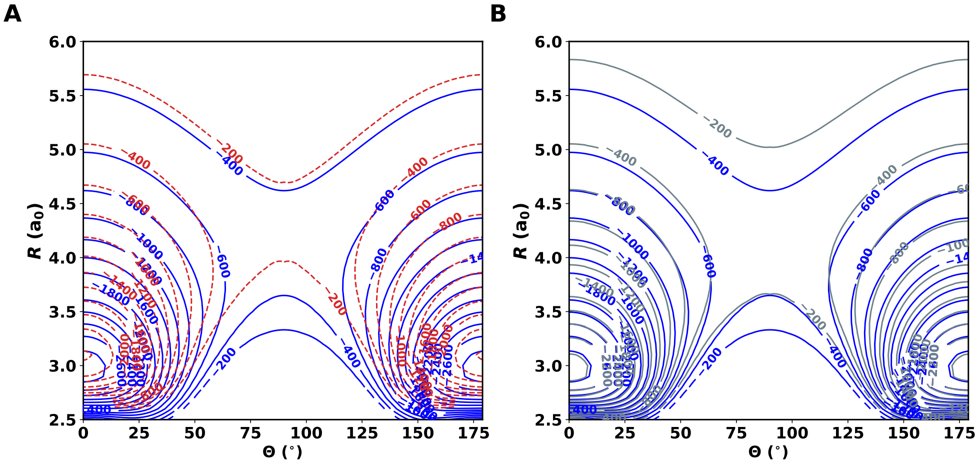

Figure 5.8: **Morphing PES-to-PES + Angular correction**. 2D projections of the PES for $r = 2.0$ a$_0$. Panel A shows the unmorphed FCI (solid blue) compared with unmorphed MP2 (dashed light red). Panel B compares unmorphed FCI (solid blue) and best-morphed PES (grey). The zero of energy is set to the value of the PES at $r_{\mathrm{HH}} = 2.0$ a$_0$ and $R = \infty$. Energies are in cm^{-1}. Note that all data points are equally weighted; the performance of the morphing transformation may be changed by differentially weighting attractive and repulsive regions of the PES.

Chapter 6

Using uncertainty to detect outliers in potential energy surfaces

There goes my hero. He's ordinary...

My Hero - Foo Fighters

This chapter evaluates the use of different uncertainty quantification techniques for detecting outliers (i.e. samples with large errors) in a reactive potential energy surface. In contrast with common approaches, here, the goal is to evaluate if the uncertainty predicted by the model is qualitatively related to the error in prediction. To fulfil this objective, three techniques were evaluated: the ensemble method, deep evidential regression (DER) and Gaussian mixtures model (GMM). Two new versions of DER are presented. The first one is based on modifying the loss function by using a Lipschitz regularizer (DER-L). The second version assumes that the data can be represented by a Normal Inverse Wishard distribution (DER-M). The system of study was the reaction of (*syn*)-Criegee to vinyl hydroxyperoxide. The generated PESs for the different models are evaluated by characterizing their stationary points. Additionally, its performance in simulation and the description of the reactive process. The capabilities of outlier detection were evaluated with different techniques. Furthermore, the relationship between structures with energy and variance is studied. The results show that ensemble models have the best performed for outlier detection. Among the DER versions, the introduced model DER-L has the best performance.

The results presented in this chapter will be submitted to npj Comput. Mater.

****Silvan Kässer collaborate in this work as joint first author*

6.1 Introduction

Computer simulations are an indispensable part of today's research and have become increasingly important in chemistry, physics, biology and materials science[313–316]. Commonly, molecular dynamics (MD) simulations involve the numerical integration of Newton's equations of motion, which requires the determination of potential energies and forces for a given atomic configuration.[96, 317] Ideally, those properties would be determined at the highest level of accuracy by solving the time-independent Schrödinger equation (SE). Unfortunately, this is only possible for small systems on a short time scale because the methods to solve the SE scale poorly with the system size and the method's accuracy. This limitation can be circumvented by using atomistic potentials that directly describe the relation between the atomic positions of a molecule and its potential energy through the mapping, $f : \{Z_i, \mathbf{r}_i\}_{i=1}^{N} \rightarrow E$, of the atomic charges (Z_i) and the atomic positions ($\mathbf{r}_i$) to the potential energy E [96]. Complementary, the atomic forces can be determined from the potential energy as its negative gradient ($F_i = -\nabla E$). The described mapping is known as a potential energy surface (PES).

Over the last decade, machine learning (ML) techniques such as neural networks (NNs) and kernel methods have been used to represent PESs. This originates from the methods' ability to *learn* relationships from provided data.[34] Therefore, it is possible to parametrize/learn the described mapping from a pool of reference *ab initio* calculations and eventually use it to describe the dynamics of a system of interest. Particularly, ML has been extensively used to represent PESs based on large, diverse, and high-quality electronic structure data.[318–323] While Machined Learned Potential Energy Surfaces (ML-PES), sometimes also called ML potentials[1] (MLP), reach unforeseen accuracies in the interpolation regime of the data set they are known to extrapolate poorly on unseen data due to their purely mathematical nature lacking any underlying functional form.[324, 325] Thus, ML-PESs crucially depend on the *globality* of the training data, which usually requires an iterative collection/extension of a data set.[34, 96, 326]

Nevertheless, constructing a global dataset that describes the dynamics of a chemical system is a complex task with challenges related to the quality and completeness of the training data and the inter- and extrapolation behaviour of the ML models. A way to tackle these critical aspects is through the use of uncertainty quantification with the primary goal of detecting outlier regions. Finding such outliers or outlier regions helps to increase the model's robustness and further improves its accuracy and reliability.

[1] Although in the literature it is common to find both names, the present work uses ML-PES to avoid confusion with multilayer perceptron also known as MLP.

Particularly for reactive PESs - one of the advantageous applications of ML-based PESs - quantitatively characterizing the confidence in predicted energies and/or forces for chemically interesting regions around the transition state(s) is very valuable. Such information can be used to distinguish well-covered regions from those that require additional training data. To facilitate the discussion, some working definitions must be given for error, uncertainty, and variance in this work. Here, the error is considered as the difference between the reference value of a property and the predicted value of that property with a given model. Complementary to this is the variance defined as the expected value for the square difference between the predicted value and the mean value of the model. Finally, uncertainty is considered as the degree of confidence in the prediction made by a given model. Uncertainty is related to the lack of knowledge or the model's limitations to describe a system.[266]

Currently, there are different approaches to quantifying the uncertainty in the prediction of an ML-PES. Those have been recently benchmarked on non-reactive systems[327]. Here the goal is to quantify uncertainty for a reactive system for which one of the Criegee Intermediates (CIs), *syn*-Criegee (CH_3CHOO), was used.

The manuscript is structured as follows. First, the methods, including data set generation, uncertainty quantification and analysis techniques, are described. Next, the performance of the PESs for computing geometrical and energetic properties is assessed. This is followed by the results on uncertainty quantification, outlier detection and an analysis of the relationship between molecular structure and errors/uncertainties. Finally, the findings are discussed in a broader context and conclusions are drawn.

6.2 Methods

This section describes the the *ab initio* reference data, the approaches to quantify uncertainty and further analyses. For the ensemble and deep evidential regression models, the variance is used for uncertainty quantification whereas for the Gaussian mixture model the negative log-likelihood (NLL) is used. In the text, "uncertainty" and "variance" are used synonymous, whereby a small variance value corresponds to a smaller uncertainty and a higher confidence in the prediction and *vice versa*. The models are characterized in terms of the Mean Squared Error (MSE), the Mean Absolute Error (MAE) and the Mean Variance (MV).

6.2.1 Data sets

The main ingredient for generating ML-PESs is reference electronic structure data to train the models on. Here, the H-transfer reaction from (*syn*)-Criegee to vinyl hydroxyperoxide (VHP) serves as a benchmark system (see Fig. 6.1) and reference data at the MP2/aug-cc-pVTZ level of theory is available from previous work.[328] From a total of 37399 structures covering the (*syn*)-Criegee → VHP reaction ~ 10 % are extracted semi-randomly (every 10th) and structures with very large energies (> 400 kcal/mol above the minimum) are excluded. A total of 3706 data points are used for obtaining a first-generation ML-PES (see the energy distribution in Fig. S1). Multiple rounds of diffusion Monte Carlo (DMC) simulations[329] and adaptive sampling[330] were run to detect *holes* and under-sampled regions. The resulting final data set contains a total of 4305 structures (see the energy distribution in Fig. S2) and is used to train different ML-PESs that are finally used for uncertainty prediction. It is important to note that the training data set is not considered to be comprehensive. If, e.g., a global PES for dissociation dynamics (i.e. formation of vinoxy radical, etc) is sought after, additional sampling would be required. Nevertheless, the small data set allows us to obtain different ML-based models and covers the relevant part of the configurational space of the reactive process of interest (H-transfer), and their ability to quantify uncertainty can be tested on an extensive test set. The (unseen) test set contains a total of 33402 structures covering the (*syn*)-Criegee → VHP reaction and the energy distribution is shown in Fig. S3.

6.2.2 Uncertainty Quantification

Ensembles The ensemble method based on the Query-by-committee[265] strategy is a frequently used and practical approach to uncertainty estimation. For this strategy, a "committee" of models is trained on the same data set. The uncertainty measure is obtained as the disagreement between the models (or within the committee/ensemble). If the predictions of the ensemble members agree closely, it can be assumed that the region on the PES is well described. For under-sampled regions, however, the predictions will diverge.[331] A commonly used uncertainty measure for the ensemble is the standard deviation given by[331]

$$\sigma_E = \sqrt{\frac{1}{\mathcal{N}} \sum_{n}^{\mathcal{N}} \left(\tilde{E}_n - \bar{E} \right)^2 }. \tag{6.1}$$

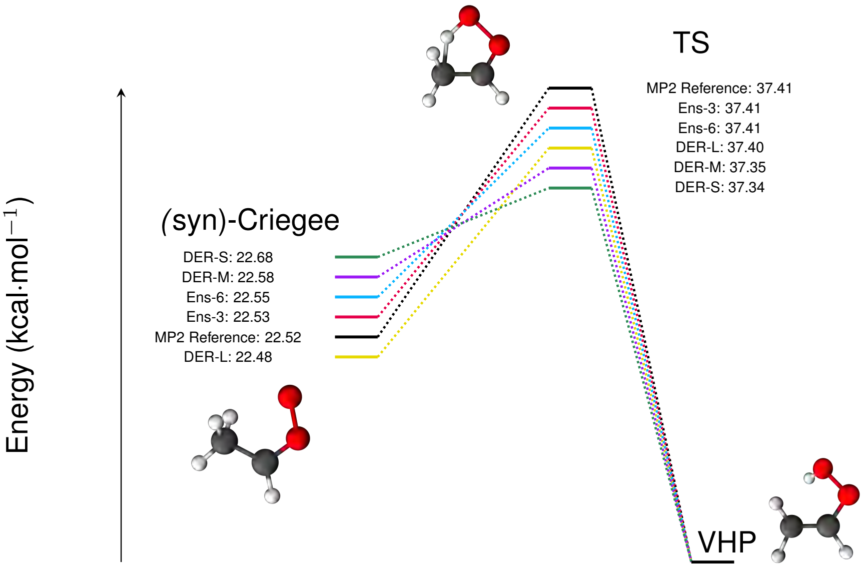

Figure 6.1: Characteristics of the stationary points of the PESs. The energy of the VHP minimum serves as a reference. The energy scale is exaggerated to better represent the differences between the methods.

Here, $\mathcal{N}$ corresponds to the number of committee models, $\tilde{E}_n$ is the energy predicted by committee model n and $\bar{E}$ is the ensemble average.

PhysNet[32] is chosen to learn a representation of the PES. A total of 6 models are trained to generate an ensemble. All models share the same architecture and hyperparameters. However, the random initialization prior to training and the splits of the training/validation data were altered (models 1/2, 3/4 and 5/6 were trained on exactly the same data). The 4305 data points were split into training/validation sets according to 80/20 %. The PhysNet models are trained on energies, forces and dipole moments according to the scheme outlined in Reference 32. Query-by-committee is performed with an ensemble of 6 models (PhysNet-6 or Ens-6) and 3 models (PhysNet-3 or Ens-3, models 1, 3, 5).

Deep Evidential Regression The present work employs a modified architecture[143] of PhysNet to predict energies and uncertainties based on Deep Evidential Regression (DER). DER assumes that the energies are Gaussian-distributed $P(E) = \mathcal{N}(\mu, \sigma^2)$. The prior distribution is a Normal-Inverse Gamma (NIG), described by four values (γ,

ν, α, β).[30] The total loss function $\mathcal{L}$ includes the Negative Log-Likelihood (NLL), $\mathcal{L}^{NLL}(x)$, which is regularized by the $\lambda-$scaled Mean Squared Error (MSE), $\mathcal{L}^{R}(x)$, that minimizes the evidence of incorrect predictions together with energies, forces, charges and dipole moments for all structures in the training set

$$\begin{aligned}\mathcal{L} = \mathcal{L}^{NLL}(E_{\text{ref}}, E_{\text{pred}}) + \lambda(\mathcal{L}^{R}(E_{\text{ref}}, E_{\text{pred}}) - \varepsilon) + W_F|F_{\text{ref}} - F_{\text{pred}}| \\ + W_Q\left|Q_{\text{ref}} - Q_{\text{pred}}\right| + W_D\left|D_{\text{ref}} - D_{\text{pred}}\right|.\end{aligned} \tag{6.2}$$

The NN is trained to minimize the difference between the NIG distribution and $p(E)$. The values of the hyperparameters were $W_F = 52.9177$ Å/eV, $W_Q = 14.3996\ e^{-1}$, and $W_D = 27.2113\ D^{-1}$, respectively,[32] and $\lambda = 0.15$ and $\varepsilon = 10^{-4}$ throughout. Notice that the forces and dipole moments were calculated as in the original version of PhysNet. In consequence, the variance of the forces can not be obtained because the derivative of the variance is the covariance matrix between energy and forces.[332] This model is referred to as DER-Simple or DER-S.

Modified Deep Evidential Regression The effectiveness in predicting uncertainties by DER-S has been recently questioned[271, 333]: Firstly, minimizing a loss function similar to Equation 6.2 is insufficient to uniquely determine the parameters of the NIG distribution because $\mathcal{L}^{NLL}(E_{\text{ref}}, E_{\text{pred}})$ is optimized independently of the data.[271] This leads to large uncertainty in poorly sampled regions. Secondly, it was shown that optimizing $\mathcal{L}^{NLL}(E_{\text{ref}}, E_{\text{pred}})$ is insufficient to obtain faithful predictions. Adding the term $\lambda(\mathcal{L}^{R}(E_{\text{ref}}, E_{\text{pred}}) - \varepsilon)$ as a regularizer addresses this problem but can lead to a gradient conflict between the two terms[333].

Two modifications to DER-S were considered. First, the multivariate generalization, DER-M, following the work of Meinert and Lavin[270] was implemented. In DER-M, the NIG is replaced by a Normal Inverse Wishart (NIW) distribution, which is the multidimensional generalization of the NIG distribution to predict a multidimensional distribution of energies (E) and charges (Q). The loss function for DER-M is

$$\begin{aligned}\mathcal{L} = &\log\left(\frac{\nu+1}{\nu-1}\right) - \nu\sum_j \ell_j + \frac{\nu+1}{2}\log\left(\det\left(\mathbf{L}\mathbf{L}^\top + \frac{1}{1+\nu}\mathbf{Y}\cdot\mathbf{Y}^\top\right)\right) + \\ &W_F\left|F_{\text{pred}} - F_{\text{ref}}\right| + W_D\left|D_{\text{pred}} - D_{\text{ref}}\right|\end{aligned} \tag{6.3}$$

where $\mathbf{Y} = [E_{\text{ref}}, Q_{\text{ref}}]^\top - [\mu_0, \mu_1]^\top$. μ_0 is the predicted energy (E_{pred}) and μ_1 the respective predicted total charge (Q_{pred}). Then, the model output will contain six values: the objective values ($E_{\text{pred}}, Q_{\text{pred}}$), the corresponding parameters of the covariance matrix $\mathbf{L}$, $\vec{l} = diag(\mathbf{L})$, and a parameter ν. The outputs of the model will be transformed to become the parameters of the multidimensional evidential distribution. Details on the construction of the $\mathbf{L}$ matrix, boundaries of ν and the uncertainty are given in the SI.

For the second modified architecture, a Lipschitz-modified loss function $\mathcal{L}^{Lips}$ was used[333] as a complementary regularization to the NLL loss

$$\begin{aligned}\mathcal{L} = \mathcal{L}^{NLL}(E_{\text{ref}}, E_{\text{pred}}) + \lambda(\mathcal{L}^{R}(E_{\text{ref}}, E_{\text{pred}}) - \varepsilon) + \mathcal{L}^{Lips.}(E_{\text{ref}}, E_{\text{pred}}) \\ +W_F\left|F_{\text{ref}} - F_{\text{pred}}\right| + W_Q\left|Q_{\text{ref}} - Q_{\text{pred}}\right| + W_D\left|D_{\text{ref}} - D_{\text{pred}}\right|\end{aligned} \tag{6.4}$$

Here, $\mathcal{L}^{Lips.}(E_{\text{ref}}, E_{\text{pred}})$ is defined as

$$\mathcal{L}^{Lips.}(E_{\text{ref}}, E_{\text{pred}}) = \begin{cases} (E_{\text{ref}} - E_{\text{pred}})^2 & \text{If } \lambda^2 < U_{\nu,\alpha} \\ 2\sqrt{U_{\nu,\alpha}}|E_{\text{ref}} - E_{\text{pred}}| - U_{\nu,\alpha} & \text{If } \lambda^2 \geq U_{\nu,\alpha} \end{cases} \tag{6.5}$$

where $\lambda^2 = (E_{\text{ref}} - E_{\text{pred}})^2$ and $U_{\alpha,\nu}$ are the derivatives of $\mathcal{L}^{NLL}$ with respect to each variable

$$\begin{cases} U_\nu = \frac{\beta(\nu+1)}{\alpha\nu} \\ U_\alpha = \frac{2\beta(1+\nu)}{\nu}[\exp(\Psi(\alpha + 1/2) - \Psi(\alpha)) - 1 \end{cases} \tag{6.6}$$

and $\Psi(\cdot)$ is the digamma function. This model is referred to as DER-L. For training DER-M and DER-L, the weights for forces, dipoles and charges were the same as in DER-S.

Gaussian Mixtures Models A third alternative to quantify the uncertainty is the so-called Gaussian Mixture Model, GMM. This method is convenient for representing - typically - multimodal distributions in terms of a combination of simpler distributions, such as multidimensional Gaussians[113]

$$\mathcal{N}(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)\right) \tag{6.7}$$

Here, μ_i is a N-dimensional mean vector and Σ_i is the $N \times N$-dimensional covariance matrix. The distribution of data, here the distribution of molecular features, x, given parameters θ can be represented as a weighted sum of N-Gaussians:

$$p(x|\theta) = \sum_{i=1}^{N} \omega_i \mathcal{N}(x|\mu_i, \Sigma_i) \tag{6.8}$$

with mixing coefficients ω_i obeying[136] $\sum_{i=1}^{N} \omega_i = 1$ and $0 \leq \omega_i \leq 1$. The ω_i coefficients are the prior probability for the ith-component.

Following the work of Zhu *et al.*[334], the parameters of Equation 6.8 ($\theta = \{\omega_i, \mu_i, \Sigma_i\}$) to construct the Gaussian mixture model (GMM) were obtained from the molecular features of the last layer of a trained PhysNet model. The distribution of molecular features from the training set is used to acquire the values of θ. The initial μ_i values were determined from k-means clustering. To each Gaussian i in the GMM model a covariance matrix Σ_i is assigned. The number of Gaussian functions required was determined by using the Bayesian Information Criterion (BIC) and was $N = 37$. Finally, the fitted model was evaluated by using the negative log-likelihood (NLL) of the molecular feature vector as:

$$NLL(p(x|X)) = -\ln\left(\sum_{i=1}^{N} \omega_i \mathcal{N}(x|\mu_i, \Sigma_i)\right) \tag{6.9}$$

Here, $p(x|X)$ is the conditional probability of a molecular feature vector x with respect to the distribution of feature vectors in the training dataset X. The value of NLL is used as a measure of the uncertainty prediction, whereby smaller NLL-values indicate good agreement. The "detour" involving the feature vectors is a disadvantage over the other methods studied here because it is not possible to directly relate the predicted energy with the corresponding uncertainty.

6.2.3 Analysis

Outlier detection. In this work, outliers are detected by considering whether a number N_{error} can be found in the N_{var} with the highest variance (or NLL in the case of GMM). Therefore, the accuracy for detecting outliers is defined as:

$$Acc = \frac{n(N_{\text{error}} \cap N_{\text{var}})}{N_{\text{var}}} \tag{6.10}$$

Here, $n(\cdot)$ is the cardinality of the intersection between the set of samples with the largest errors and the set with the largest variances. Complementary to this, a classification analysis of prediction over error and predicted variance was performed; details can be found on the SI.

Inside-Outside distribution The definition of inside-outside distribution is a controversial topic in the ML literature. Here, the natural definition of statistical learning theory is used:[335] Assume a training data distribution $p_{\text{train}}(x)$ and a testing distribution $q_{\text{test}}(x)$; a point x_i is defined as out-of-distribution if[336]

$$q_{\text{test}}(x_i) \neq p_{\text{train}}(x_i).$$

To assess whether a given molecular structure is inside or outside a given distribution, a rank is considered. First, all 28 intermolecular distances for *syn*-Criegee were computed. These distances were classified into "bonded" and "non-bonded" separations as follows: if the distance is smaller than the mean of the van der Waals radii of the two atoms involved plus 20%, the value is considered "bonded"; otherwise it is non-bonded. The van der Waals radii used here[337] were H: 1.10 Å, C: 1.70 Å, and O: 1.52 Å. Next, the 28 distances were determined for all structures in the training data set to determine $p_{\mathrm{bond}}(r)$ and $p_{\mathrm{no-bond}}(r)$. Using those distributions, it was possible to test a given distance of the samples in the testing dataset to be inside ($Q_{5\%}(r) < r_i < Q_{95\%}(r)$) or outside (otherwise) the distribution $p(r)$. Here $Q_{5\%}(r)$ and $Q_{95\%}(r)$ are the 5 % and 95 % quantile of $p(r)$. Using this criterion the contribution $\chi_j(r_i)$ of distance r_i for structure j is

$$\chi_j(r_i) = \begin{cases} 1 & r_i \in p_{\mathrm{bond}}(r) \\ 0.5 & r_i \in p_{\mathrm{no-bond}}(r) \\ 0 & r_i \notin [p_{\mathrm{bond}}(r) \cap p_{\mathrm{no-bond}}(r)] \end{cases} \tag{6.11}$$

From this, $rank_j$ for sample j is determined according to

$$rank_j = \sum_i^R \chi_j(r_i) \tag{6.12}$$

where $R = 28$ is the total number of distances.

6.3 Results

6.3.1 Characterization of the Trained PESs

The performance of all trained models is assessed on a hold-out test set and the MAEs and RMSEs on energies and forces are given in Table S1. While most models reach similar MAE(E) ≤ 1.0 kcal/mol, the performance on the forces deserves more attention and is provided below. An essential requirement of an ML-PES is to adequately describe geometries and relative energies of particular structures, including the minima and transition states, Figure 6.1. It is found that all models considered perform adequately to predict energies of stationary points with errors of < 0.1 kcal/mol. However, it is noticeable that most models, except for DER-L, overestimate the energy for the *(syn)*-Criegee conformation, while the transition state is underestimated for all except the ensembles.

The errors for the *syn*-Criegee structure are 0.01, 0.03, 0.16, -0.04, and 0.06 kcal/mol for Ens-3, Ens-6, DER-S, DER-L, and DER-M compared with errors lower than 0.01 kcal/mol for the TS using ensembles, and -0.07,-0.01 and 0.06 kcal/mol with DER-S, DER-L and DER-M, respectively. The smaller error of Ens-3 compared with Ens-6 is counter-intuitive and may be a consequence of random noise in the prediction caused by, e.g., parameter initialization, convergence of the loss function, or numerical inaccuracies[338, 339].

Complementary to the energy of the equilibrium structures, the Root Mean Squared Displacement (RMSD) between optimized geometries from the trained NN models and at the MP2 level were compared, see Figure S4. Generally, the deviations between the obtained geometries and the reference structures are very small. However, some differences between the tested models can be highlighted. First, it is noticed that models that use DER have an RMSD two or three orders of magnitude larger than ensembles. Additionally, it is observed that the geometry of the TS is predicted with more accuracy than the *(syn)*-Criegee or VHP conformations. For the DER models, the geometries obtained with DER-S are the most accurate by approximately two orders of magnitude compared to the ones produced with their counterparts. On the other hand, structures obtained with DER-M have the largest RMSD among the models tested here. The last of the DER models tested, DER-L, produces constant RMSD for the different molecules. Finally, the results obtained with GMM are of a slightly lower quality than the ones obtained with the ensemble model. This is expected because the GMM model is based on one of the ensemble members.

Another quantity that can be used to characterize a PES are the harmonic frequencies of the stationary points obtained from the Hessian matrix ($H = \partial^2 E/\partial \boldsymbol{r}^2$). The results (Figure S5) indicate that the best performers are the ensemble models and GMM with a MAE one order of magnitude lower than the DER models. Regarding the DER models, the best performer is DER-L, followed by DER-S and DER-M. In the case of the *(syn)*-Criegee molecule, DER-L has errors on the harmonic frequencies between -50 cm^{-1} and 50 cm^{-1}, most of the frequencies below 1500 cm^{-1} were underestimated while those above 2000 cm^{-1} (XH stretch) were overestimated. Conversely, for the same molecule, DER-S underestimates most frequencies, showing the largest errors for the vibrations at larger frequencies. The worst performing model for *(syn)*-Criegee, DER-M, shows a large overestimated value at around 500 cm^{-1} and a large underestimated value at high frequencies. The harmonic frequencies for the TS and for VHP follow similar trends. It is interesting to note that the large errors in the harmonic frequencies are also observed for the forces (See Table S1); in general, DER models have an MAE(F) one order of magnitude larger than the other three models evaluated here. This is a direct consequence and a limitation of the assumed normal distribution for the energies. The forces and Hessians are derivatives of the energy expression and the associated errors are $\propto \frac{\text{Error}^2_{\text{Ener.}}}{\sigma^2}$ and $\propto \frac{\text{Error}^3_{\text{Ener.}} - \sigma^2}{\sigma^4}$, respectively. Hence, the DER models have an inferior performance for forces and harmonic frequencies.

6.3.2 Calculations and Simulations with the PESs

Next, the performance of the different PESs for reactive MD simulations is assessed. For this, the minimum energy and minimum dynamic paths (MEP, MDP) are assessed and finite-temperature molecular dynamics simulations were carried out. The MEP describes the lowest energy path connecting reactants and products passing through the transition state. Complementary to the MEP, the MDP[340] provides information about the least-action reaction path in phase space.

Figure 6.2 A shows the MEP for the different models considered here. All MEPs are within less than 0.5 kcal/mol on each of the points sampled. Therefore, despite the differences in how errors are handled and their magnitude for each model, the MEP derived from the PESs are consistent with one another and nearly identical. The MDPs (see Figure 6.2C), initiated from the TS were determined with an excess energy of 10^{-4} kcal/mol. The TS structure is stabilized because it is a 5-membered ring and because little excess energy was used for the MDP. VHP is observed after 225 fs accompanied by pronounced oscillations in the potential energy primarily due to the highly excited OH-stretch. Overall, the time traces for potential energy (Figures 6.2C), one possible reaction coordinate $q = r_{\text{CH}} - r_{\text{OH}}$ (Figures 6.2D), and all atom-atom separations in Figure S6 are rather similar for the 6 models considered. Notable exceptions concern primarily DER-M (purple) for which the energy on the reactant side differs somewhat from the other five models. Along similar lines, the C1-H2 and C2-H3 separations deviate noticeably from the other 5 models; see Figure S6. On the product (VHP) side, the high-frequency oscillations with a period of $\sim$ 10 fs (see Figure 6.2C) correspond to a frequency of $\sim$ 3500 cm^{-1} characteristic of the OH-stretch vibration, whereas the low-frequency oscillation in Figure 6.2D is due to the azimuthal rotation of the -OH group.

Finally, NVE simulations with all six models were carried out; see the SI for details on these simulations. The simulations were run for 500 ps with a time step of 0.1 fs,

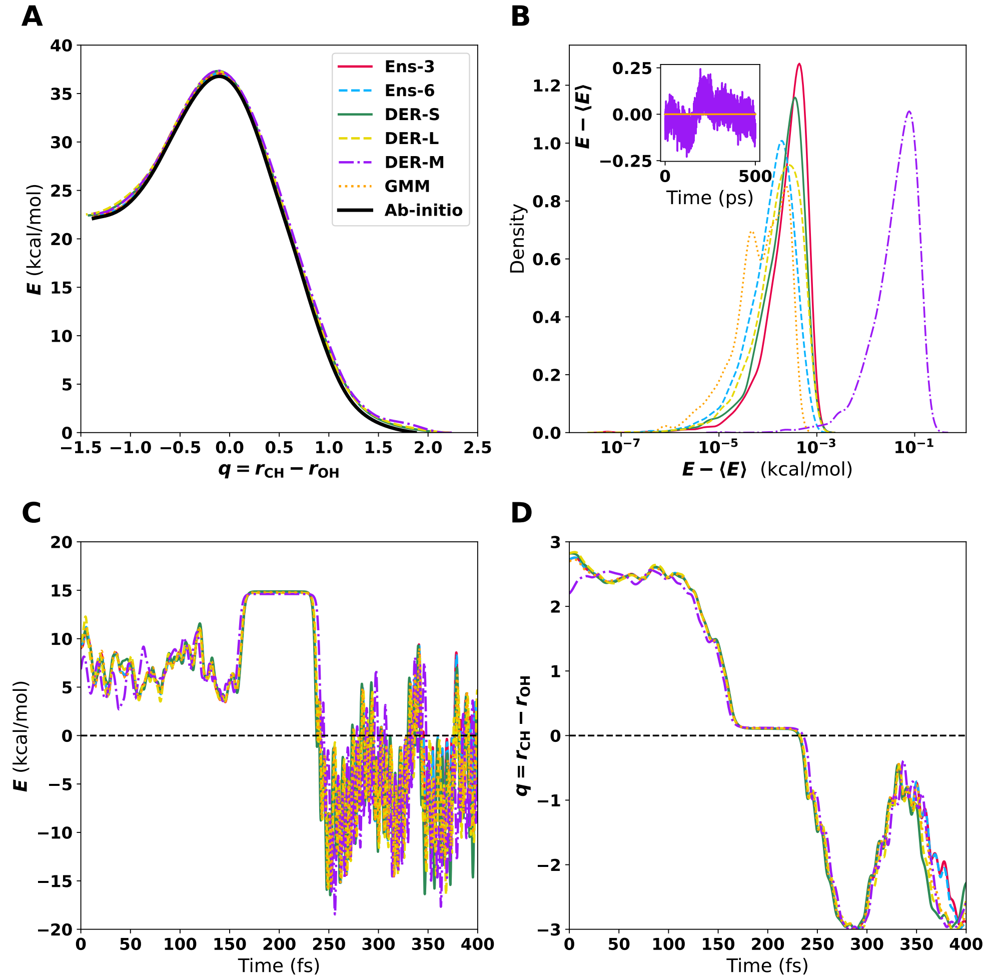

Figure 6.2: Behaviour during simulation of the different models. Panel A shows the Minimum energy path (MEP) from *syn*-Criegee to VHP for the different methods for UQ used in this work. The zero of energy is the corresponding value for the optimized structure of VHP. Panel B energy distribution for the different models during the simulation, note that the x-axis is on a logarithmic scale. Starting from the *(syn)*-Criegee, the system was simulated for 500 ps with a time step of 0.1 fs. On the insight, the time series of the energy. Panel C shows the variation of the energy for the Minimum Dynamic Path (MDP) of the different formulations of the ML-PESs starting from the optimized transition state Panel D shows the change in the defined reaction coordinate ($q = r_{\text{CH}} - r_{\text{OH}}$) with respect to the time for the MDP.

and energy is conserved to within ~ 2 kcal/mol, see Figure 6.2B. Importantly, no drift was found on this time scale for most of the models except DER-M.

6.3.3 Analysis of Error Distributions

Next, the errors, their magnitude and distributions for the trained models is analyzed in more detail. It is desirable that a model accurately predicts the energies across a wide range which points towards its extrapolation capabilities. The dataset considered contains structures for *(syn)*-Criegee, VHP, and the corresponding transition state. Residual plots were used to describe how the signed error $\Delta = E_{\text{Ref}} - E_{\text{Pred}}$, is distributed for energies between -700 and -300 kcal/mol.

Ensembles Figure 6.3 shows the performance of the ensembles. Noticeably, the error range is between -30 and 30 kcal/mol, with most errors near the centre (*i.e.* $\Delta = 0$). The region with the lowest energy ($E < -650$ kcal/mol) has higher accuracy with no noticeable outliers. The next region, between -650 and -500 kcal/mol, have the largest number of outliers broadly spread between positive and negative errors. For energies smaller than -500 kcal/mol range a small spread of the errors with few significant outliers is found. It can be noticed that the region with more outliers is close in energy to the transition state; therefore, the structures are expected to have larger deformation than the other regions. This is related to the fact that the training dataset was created to reproduce adequately the hydrogen transfer; nevertheless, side channels were not sampled.

The distributions of the squared error ($P((\Delta E)^2)$) and the variance ($P(\sigma^2)$) in Figure 6.3 are both rather sharp and centred around 0. Using a logarithmic scale further clarifies the structure of these distributions. The bimodal nature of $P((\Delta E)^2)$ and $P(\sigma^2)$ is the first distinctive feature. In addition, the predicted variance partially matches the squared error distribution (Figure 6.3 centre). The distribution agree closest near their centre. However, the height of the distribution is larger for $P(\sigma^2)$ than $P((\Delta E)^2)$. Furthermore, the tails of $P(\sigma^2)$ decay faster than for $P((\Delta E)^2)$. This is reflected in fewer samples labelled with large variance than the number of structures with large squared error.

Deep Evidential Regression. The results for the predictions of the DER models are displayed in Figure 6.4. For DER-S the errors are spread between -60 and 60

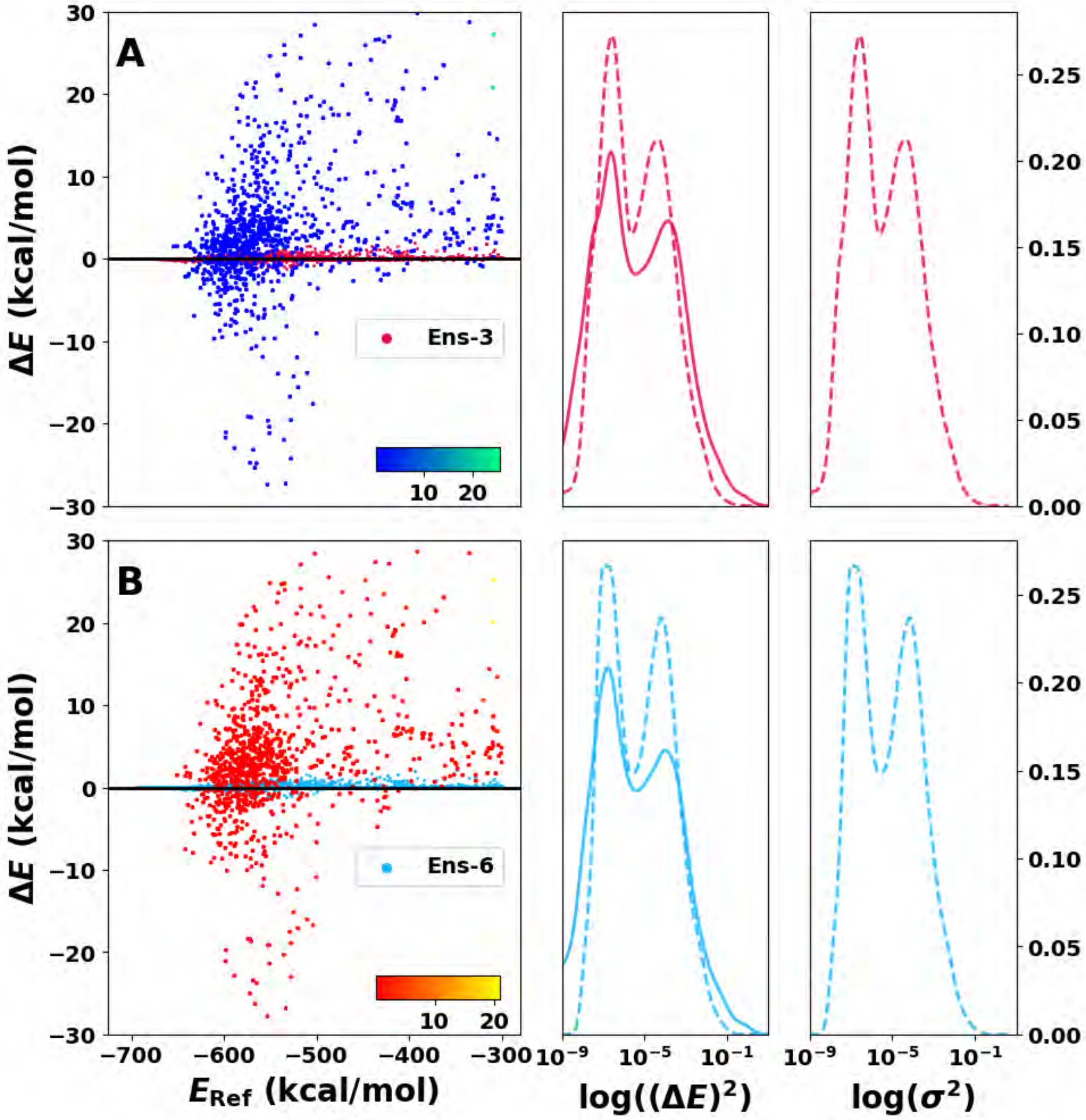

Figure 6.3: Performance of the Ens-3 and Ens-6 on the test set. Panels A and B on the left show residual plots of the error between reference and prediction. The 1000 energies with the largest variance are shaded with a different colour, and the corresponding colour bar represents the scale of the values. Squared error distribution (solid lines) and variance distributions (dotted lines) are shown in the centre next to panels A and B for comparison. Complementary to this is the variance distribution shown on the right of both panes. Notice that the x-axis on the centre and right are in logarithmic scale.

kcal/mol and the variances vary between 2×10^{-3} to 9×10^{-3} kcal/mol with a single sharp peak around 10^{-2} kcal/mol, i.e. the same uncertainty for nearly all predictions. This aligns with the previously discussed problems of DER[271] that reported models which improve the quality of the predictions by increasing their uncertainty. The small variances across the test set indicate that adding forces and dipole moments to the loss functions renders the model overconfident. One possible explanation for this is that terms depending on forces, charges and dipoles in Eq. 6.2 to DER-S act as extra regularizers to the evidence of incorrect predictions, akin to the $\mathcal{L}^R(x)$ term, during training of the NN. Hence, the variance predicted by DER-S loses its capability to detect outliers. Furthermore, DER-S tends to underestimate the energies with a larger population on the positive side of the ΔE. Finally, the squared error, centered around 10^0 is spread over a wide range from 10^{-4} to a few tens of kcal/mol.

Next, DER-L is considered (see Figure 6.4B) for which the error increases with the energy. Complementary, the variance is high for structures with positive ΔE (red points). The variance distribution is sharply peaked and centered around 10^{-3}, showing some overlap with the squared error distribution, whereas the distribution of squared error is unimodal and centered at 10^{-1} kcal/mol. However the tails are wide and extend to 10^2 kcal/mol. As for DER-S, the centre of mass of the variance distribution is between 1 or 2 orders of magnitude smaller than the corresponding distribution for the squared error, indicating that DER-L is overconfident about its predictions. It is also noted that DER-L is biased to identify predictions that overestimate the energy as outliers.

Finally, DER-M (Figure 6.4C) features a large dispersion of the predicted error around the energy range considered in this work. Predictions deteriorate quickly for low-energy configurations with almost no points near the diagonal. The squared error distribution is centered around 1 kcal/mol and extends from 10^{-2} to 10^2 kcal/mol with some overlap with the variance distribution. The variance distribution is bimodal, and its centre of mass is at $\sim 10^{-4}$, around four orders of magnitude smaller than the squared error distribution. Regarding the detection of outliers, it is noticed that samples that underestimate the energy display a large variance. On the technical side, it has been found that optimization of multidimensional Gaussian models, such as DER-M, can be numerically challenging because the NN-prediction of the covariance matrices can be numerically unstable.[272–274]

Differences between the three flavours of DER were noticeable. Firstly, DER-M performs worst on energy predictions with a poor quality of the underlying PES. On the

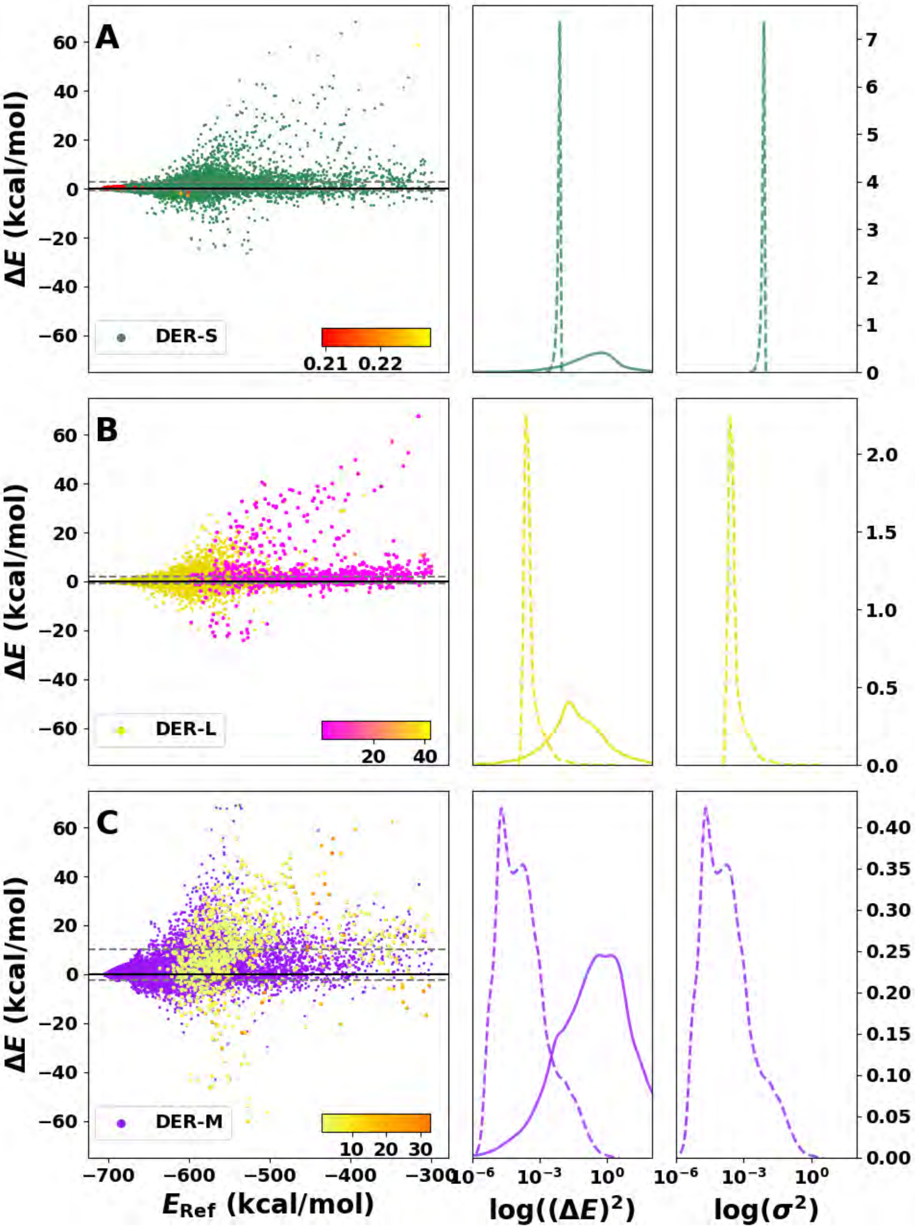

Figure 6.4: Performance of the different versions of PhysNet-DER through the range of energies of the test set. Panels A to C on the left show residual plots of the error between reference and inference for DER-S, DER-L, and DER-M, respectively. The 1000 points with the largest variance are shaded with a different colour, and the corresponding colour bar represents the scale of the values. Squared error distribution (solid lines) and variance distributions (dotted lines) are shown in the centre next to panels A B, and C for comparison. Complementary to this is the variance distribution shown on the right of both panes. Notice that the x-axis on the centre and right are in logarithmic scale.

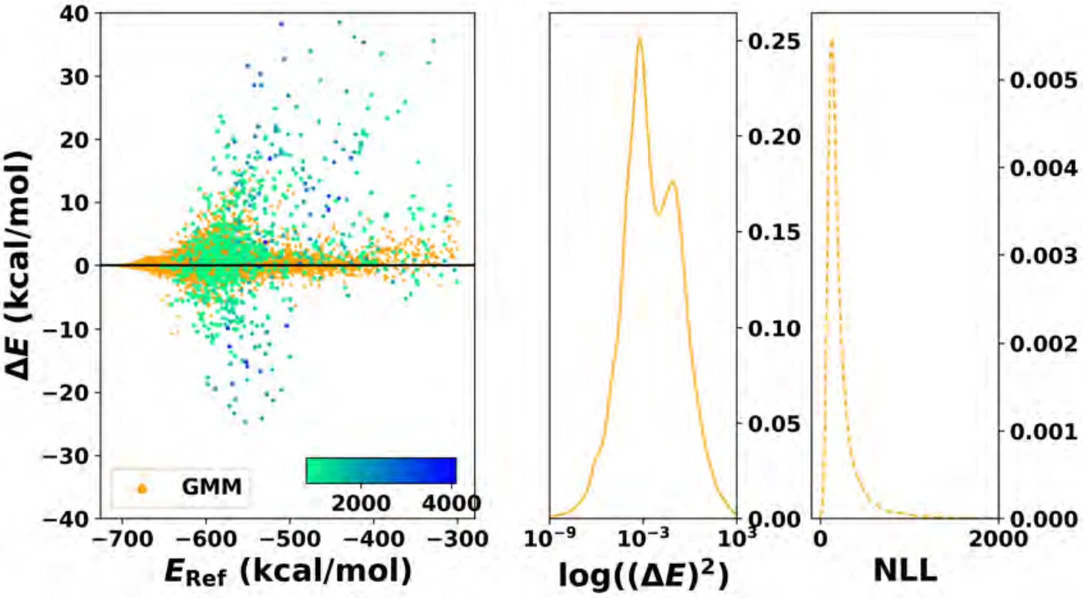

Figure 6.5: Performance of the PhysNet-GMM through the range of energies of the test set. A Residual plot of the error between reference and production is shown on the left. The 1000 points with the largest negative log-likelihood (NLL) value are shaded with a different colour, and the corresponding colour bar represents the scale of the values. The panel in the centre shows the squared error distribution. Note that the x-axis of the centre panel is in logarithmic scale for clarity. The panel on the right displays the distribution of the NLL, which is used to quantify the uncertainty.

other hand, DER-S and DER-L show a similar distribution of errors, see Figure 6.4. The variance distribution for DER-M is bimodal and considerably broader than for the other two models, which show a single sharp peak. The width of the variance distribution for DER-M increases the overlap with the $(\Delta E)^2$ distribution and, therefore, is more likely to identify outliers than the other two DER models. Unfortunately, the variance values predicted by DER-M underestimate the error by 2 to 3 orders of magnitude. From these results, DER-L is the best performer with the small MAE among the DER models and a medium quality for the variance estimation.

Gaussian Mixtures Models Finally, for the GMM (Figure 6.5) the dispersion of the error increases as the energy increases. Specifically, the largest errors occur for the highest energies. For the errors it is found that they are more evenly distributed in the over- ($\Delta E < 0$) and under-predicted ($\Delta E > 0$) regions. On the other hand, the squared error features a bimodal distribution centered at 10^{-3} with extended tails up to 10^3. As can be seen, the NLL is peaked at low values of NLL and decays rapidly for increasing NLL.

6.3.4 Outlier Detection

The focus of the present work is the detection of outliers. The error analysis carried out so far indicates that outlier detection is challenging. In this work, outlier detection capabilities of the models are evaluated using the accuracy metric defined in Equation 6.10 and the classification procedure described in the method section.

First, the number of structures with large variance was determined, and the magnitude of the error was assessed. Figure 6.6 shows the results for the 1000 structures with the largest predicted variance. The results indicate that as the number of structures with large errors sought increases, the probability of finding them among the top 1000 with large variance decreases. Overall, the best-performing model is Ens-6, closely followed by Ens-3 and GMM. The three DER models behave quite differently from one another. First, DER-S has a very poor performance that goes to practically null ability to detect outliers. Next, DER-L is very good at detecting extreme outliers, performing even better than Ens-3 for the first case. However, it decays quickly and is the second worst performer after DER-S. Finally, DER-M has an almost linear performance, meaning its capability predictions are constant, independent of the number of samples.

One interesting aspect of Figure 6.6 is that for the extreme cases (i.e. detecting the 25 samples with the largest error), four models (Ens-3, Ens-6, DER-L, and GMM) have a probability higher than 80% of detecting those extreme values. This trend continues for the ensemble models and GMM up to 200 samples. After this number, the accuracy decays for all of the models.This can be understood because the task at hand is harder to solve as the number of required samples to identify increases.

Next, a 2-dimensional analysis involving different numbers of structures with large errors and different numbers of high-variance structures was carried out. Figure 6.7 shows the probability of finding N_{err} structures with large error among the N_{var} structures with the large variance for each method. As an example, for Ens-3 the lower left corner reports a probability of 0.92 for finding the $N_{\text{err}} = 25$ structures with largest error among the $N_{\text{var}} = 1000$ structures with largest variance. Increasing N_{err} to 1000 reduces this probability to 0.52. This row corresponds to the data reported in Figure 6.6. More generally, the N_{var} can now be reduced from 1000 to 25, and the probability of finding corresponding large-error predictions is reported in the full triangle. Light and dark colours correspond to high and low probabilities, respectively. In practice one wants to keep N_{var} small and increase the probability to find a maximum of N_{err} structures. From this perspective the best-performing model is GMM.

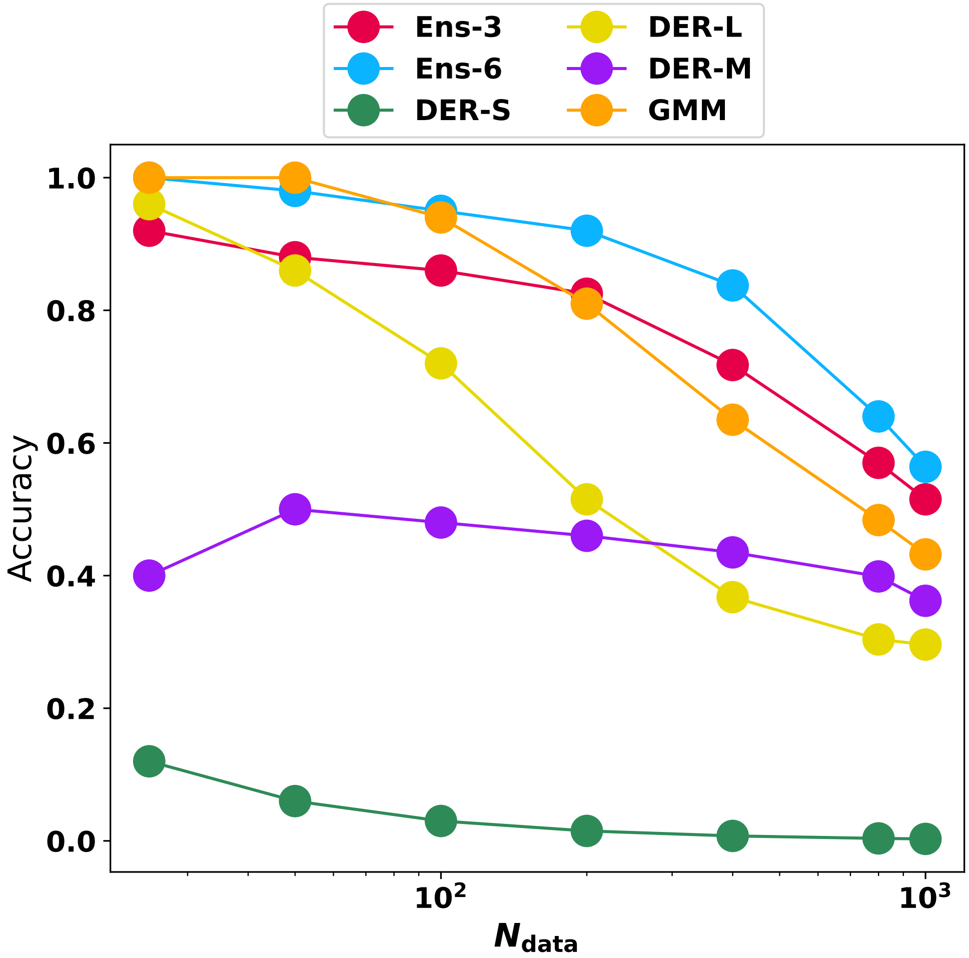

Figure 6.6: Reliability of outlier-detection for the different strategies: Given the 1000 structures that are predicted to have the highest errors/variance/uncertainty, it is evaluated whether they correspond to the structures that also have the highest errors from comparison to reference data for different $N_{\text{data}} = [25, 50, 100, 200, 400, 800, 1000]$. I.e. it is evaluated if the N_{data} structures with the actual highest errors are contained in the 1000 that are predicted to have high errors.

With Ens-3 as the reference, Ens-6 and GMM perform slightly better overall, whereas DER-L is comparable for small N_{err} and large N_{var}. As N_{var} decreases to 400 samples and below the reliability of DER-L drops drastically. DER-M performs inferior to DER-L for small N_{err} and large N_{var} but maintains a success rate of 0.2 to 0.4 for most values of N_{err} and N_{var}. Finally, DER-S has the smallest success rate throughout except for $N_{\text{err}} = N_{\text{var}} = 25$ for which it performs better than DER-L.

Complementary to the reliability analysis in Figures 6.6 and 6.7, the true positive rate (sensitivity or TPR, Eq. S3), that quantifies how many of the samples identified with a

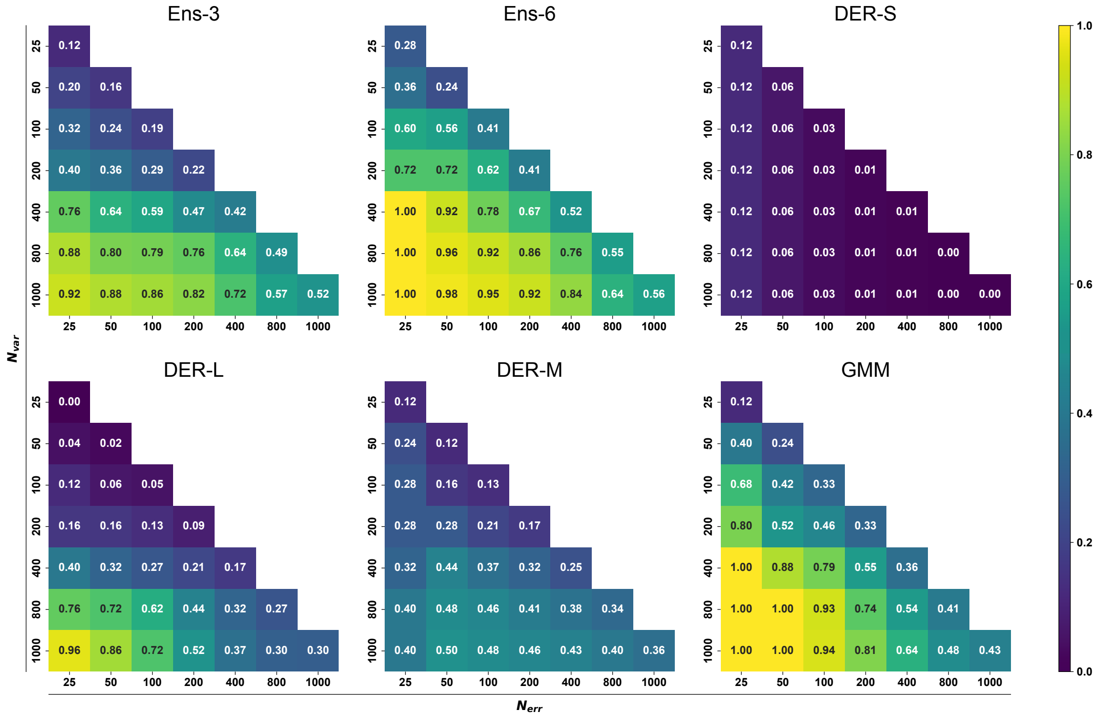

Figure 6.7: Reliability of outlier-detection for the different strategies: Given N structures with the highest error/variance, it is evaluated if they correspond to the N structures with the largest errors/variance. See Equation 6.10. The plot is coloured according to the accuracy. Exact values of the accuracy are given for each combination in white.

large variance also have a large error (c.f. true positives), and the positive predictive values (precision or PPV, Eq. S4) that measures how many of the samples with a large error are correctly labelled by the model were analyzed. This test was performed over different ranges of squared error and variance (or NLL for GMM), which can be used as confidence boundaries. Ideally, the model is expected to have large sensitivity and precision. Results for this analysis are reported in Figures S7-S12, which report a heatmap of TPR and PPV values using different thresholds for error or variance in the plot larger (desired) values are coloured blue while small values are shown in red. The results indicate that the model has a high sensitivity for Ens-6 and Ens-3 for all error ranges at low variance values (Figure S7 and S8). Conversely, PPV values are high at all variance ranges for a small error cutoff. It is also observed that the confidence range for Ens-6 (Figure S8) is larger than for Ens-3 (Figure S7). Results for the DER models also have large TPR values at small uncertainty values (Figures S9, S10 and S11). On the contrary, the PPV coverage is almost null for DER-S (Figure S9) and DER-L (Figure S10, while DER-M has high values for all variance ranges with a small error threshold (Figure S11). Note, however, that the scales for squared error and variance differ by 2 to 3 orders of magnitude. Hence, the magnitude of the MSE and MV need to be carefully inspected in addition to the color code. Lastly, the TPR for GMM shows a good performance over a large range of NLL values, which implies the model correctly assigns uncertainty to errors in a larger range of uncertainty (Figure S12). On the other hand, PPV values are obtained for large values of NLL but low squared error threshold (Figure S12)

.

Finally, two more metrics to quantify the reliability over the range of squared errors and variance were evaluated. The first is the “false positives rate” (FPR, Eq. S5), also known as “false alarm rate”, which measures how many of the samples identified with large variance do not correspond to a large error (i.e. false positives). Second is the false negative rate (FNR, Eq. S6) or miss rate, which measures how many samples not identified with a large variance correspond to a large error. For FPR and FNR small values (red) are desirable, whereas large values (blue) are undesirable. The results for both metrics are shown in Figures S13 to S18. For the ensemble models, $\mathrm{FPR} \sim 0$ over the range evaluated (Figures S13 and S14), indicating a low probability of misclassifying samples, i.e. suitable for outlier detection. Complementary, the FNR values are small for small variance values (Figures S13 and S14 left), while the probability of missing a sample with a large error increases with the variance. The results for DER models show low values of FPR except for very small values of variance (Figures S15,S16, and S17 left). Regarding the results for the FNR, large values are obtained except for very small values of variance (Figure S15, S16, and S17 right). As in the case of TPR and PPV, the difference in magnitude between variance and error gives way to misleading conclusions for DER-S. Finally, the GMM model has large values of FPR at low values of NLL (Figure S18 left) while the values of FNR are low in a large region but decay fast at large values of NLL (Figure S18 right). These results suggest that Ens-6 is the best model for detecting outliers with high TPR, and PPV complemented with a low FPR and FNR. On the contrary, the worst model is DER-S, which has a low probability of identifying outliers.

6.3.5 In- and Out of Distribution

A deeper understanding of the origin of the variances and the prediction error can be obtained by considering the distribution of structural features (atom distances) in the training and testing datasets, which can then be related to the predicted properties. Following the procedure described in Section 2.3, a score (the $rank$) for each molecule in the test set was calculated. The results in Figure 6.8 are combined with a histogram of the number of molecules with a given rank. The rank, see Equations 6.11 and 6.12, is interpreted as the degree to which a sample can be considered in or out of the distribution: a high $rank$ implies that more degrees of freedom (DOF) can be found in the training data. Thus, it is "in distribution" (ID), while a low $rank$ indicates that the sample has more DOFs farther away from the distribution and is "out of distribution" (OOD). The black histogram in Figure 6.8 shows that most samples are ID to some extent, with a most probable value $rank = 17$. Additionally, most of the samples have $rank > 14$, which can be related to good coverage of the interatomic distributions by the training dataset.

Figures 6.8A and B indicate that $rank$ and MSE or MV (coloured lines) are related. Similarly, the distribution of samples with given $rank$ also impacts MSE and MV, see black histograms. For the MSE (Figure 6.8A) all models except for DER-M behave similarly overall. Up to $\mathrm{rank} \sim 12$ the MSE varies between ~ 0 and ~ 100 kcal/mol and above the MSE decays monotonically well below 1 kcal/mol for all models except for DER-M. For DER-M the behaviour is not fundamentally different, but the magnitude of the MSE is considerably increased. The MV in Figure 6.8B reflects the behaviour of the MSE for DER-M, and the same is observed for Ens-3, Ens-6, and GMM. For DER-L, the decay of the MV with increasing rank is less pronounced, whereas for DER-S $\mathrm{MV} \sim 0.1$ throughout. One reason for the decay of MSE and MV with increasing $rank$ is the increased number of samples for given $rank$, $P(rank)$, see black histograms Figure S19. What distinguishes DER-M from the other five methods is the fact that the achievable MSE remains considerably larger for most rank-values.

The relationship between $rank$ and MSE/MV can also be considered individually for bonded and non-bonded separations, see Figure S20. Overall, the results from Figure 6.8A are replicated, but the relationship between $P(rank)$ and the MSE is yet more pronounced for bonded terms. For small sample sizes, the MSE is large and *vice versa*. Unexpectedly, for the non-bonded separations, the behaviour for all models except for DER-M differs: For the lowest ranks, which are sparsely populated, the MSE increases with increasing $P(rank)$ up to $rank = 6.5$, after which the MSE decreases monotonically. The MV, on the other hand, behaves as expected. It is noted that for DER-S both bonded and non-bonded separations yield an almost constant value for the MV irrespective of $P(rank)$.

The relationship between $rank$ and MEA/MV for bonded and non-bonded separations can also be analyzed in a 2-dimensional map. First, the average energy depending on bonded and non-bonded $rank$ is considered; see Figure 6.8C. This map can also be regarded as an abstract rendering of the PES. Low-energy structures correspond

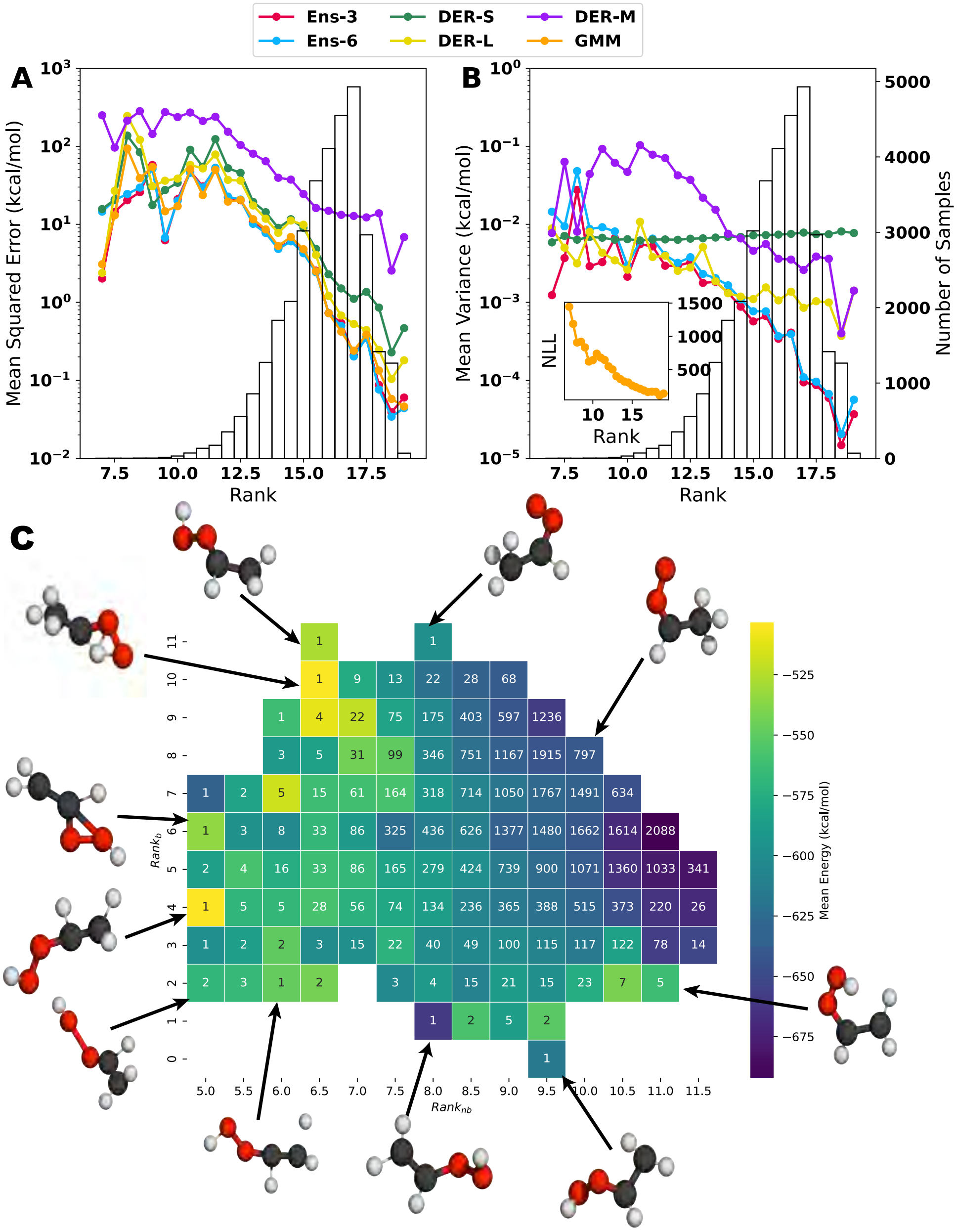

Figure 6.8: Evolution of the mean squared error (A) and the mean variance (B) concerning the rank of each structure in the test set. The bar plot (background) shows the number of structures with a particular rank. A large $rank-$value indicates that more degrees of freedom are covered by the training data and *vice versa*. The $y-$axis is displayed in logarithm scale to highlight the difference in the values of MSE or MV for the different rank values. Notice that for the Gaussian mixture model the negative log-likelihood is used to estimate the uncertainty. The insight on the right panel shows how the mean NLL changes concerning the defined rank. Panel C shows the 2d-map representation of the rank for bonded and non-bonded separations. Representative structures of different combinations are shown around the map.

to the *syn*-Criegee and VHP basins, followed by structures representative of the TS between the reactant and product and finally, higher-lying structures dominated by larger distortions. The majority of points (93 %, white numbers in Figure 6.8C) is for $8 \leq rank_{\text{nb}} \leq 11.5$ and $4 \leq rank_{\text{b}} \leq 9$. These structures cover an energy range from –700 to –300 kcal/mol with the lowest-energy structures featuring $rank_{\text{nb}} \geq 11.0$ and $rank_{\text{b}} \geq 5.0$. Hence, these are comparatively "open" structures, characteristic of an elongated molecule such as the one considered here. Examples for such structures are provided in Figure 6.8C.

Next, the MSE and MV are mapped onto this representation, see Figures S21 and S22. Hence, the map itself remains, but the coloration changes. For the MSE, darker colours indicate a low error, whereas lighter colours indicate higher errors. The regions for high MSE remain the same for all six models considered: $5.0 \leq rank_{\text{nb}} \leq 7.5$ and $2 \leq rank_{\text{b}} \leq 5$, i.e. What changes, however, is the *maximum* MSE which is 9 kcal/mol for Ens-3 and Ens-6 and increases up to 40 kcal/mol for DER-M.

For the MV, Ens-3 and Ens-6 are on the same scale and differ little. The largest variances for Ens-3 and Ens-6 are observed for similar ranks as for the MSE. On the other hand, DER-S, DER-M and DER-L are on rather different scales ranging from 10^{-3} (DER-S) to ~ 0.1 kcal/mol (DER-M and DER-L). DER-S returns a uniform value for all values of $rank_{\text{b}}$ and $rank_{\text{nb}}$. For DER-L, the MV is larger for $5.0 \leq rank_{\text{nb}} \leq 7.5$ and $0 \leq rank_{\text{b}} \leq 9$, while DER-M displays large values for a wider region ($rank_{\text{nb}} \leq 9.5$, $rank_{\text{b}} \leq 8$). Finally, the magnitude of NLL for GMM can not be directly compared with the other five models, but NLL is large for $rank_{\text{nb}} \leq 8$, $rank_{\text{b}} \leq 8$.

The analysis of the effect of inside and outside distribution degrees of freedom showed that a simple ranking such as the one presented here can highlight the effect of the differences between training and test distribution on the prediction and the uncertainty estimation. It must be mentioned that the $rank-$metric can be used as a proxy for how structure and error are related. However, further analysis is required to complement these results because averaging effects can play an important role. Yet, for improving reactive ML-PESs it is notable that samples with larger $rank$ feature lower average error and *vice versa*. It is also found that coverage of the non-bonded distances for predicting energies and uncertainties can be rather informative. This contrasts with the usual focus on sufficiently covering the range of chemical bonds when conceiving data sets for training ML-PESs.

6.4 Discussion and Conclusions

The present work analyzed in a quantitative fashion to what extent three different UQ-methods - ensembles, deep evidential regression, and Gaussian Mixture Models - are capable to detect outliers in samples from which full-dimensional reactive potential energy surfaces can be trained. The system investigated for this was one of the CIs *syn*-Criegee, CH_3CHOO.

From an electronic structure perspective, CIs are known to be challenging because they feature multi-reference effects.[341, 342] This can also be demonstrated from the present data and even be linked to the quality of the prediction and the MV. For this, molecular structures with the largest absolute errors (Figure 6.9A) and with the largest uncertainty (Figure 6.9B) for each of the models were determined. Generally, the largest errors arise either for deformed *(syn)*-Criegee or VHP structures, whereas structures with the largest variance are predominantly perturbed *(syn)*-Criegee structures except for GMM, which identifies one structure closer to the TS. Interestingly, none of the models assigns the largest uncertainty to the structure with the largest error. In all cases, the magnitude of the error is larger than the predicted variance. On the other hand, for structures with large variance, the errors are on the same scale for ensembles and DER-M, whereas they are almost constant for DER-S. Contrary to this, DER-L overestimates the uncertainty by one order of magnitude.

Structure #3429 (see Figures 6.9C and D) with the largest error is the same for four out of the six models. The remaining two models also show a large error for this structure, indicating that this structure is, in general, difficult to predict. Paradoxically, structure #3429 is predicted to have a large uncertainty for the models that do not identify it with the largest error (DER-M and DER-L), while the other four identify it with the smaller uncertainty. Structure #3986 is most difficult to predict with DER-M, while for the other models, it is better predicted with a difference between predictions of ≈ 50 kcal/mol. The GMM model assigns it a large uncertainty while the other models give it values in the same range as the predicted structure #3986. Lastly, structure #28980 features the largest error for DER-L but in the same magnitude as the other models except for DER-M. Regarding the uncertainty, Ens-6 identifies #28980 with a large uncertainty, while the other models attribute a small value to it. It is also found that Ens-3, Ens-6, DER-S, and GMM identify structures (e.g. #23366, #23550, #24576, #28980) that resemble those with the largest error; however, the error for these four structures is not large, see SI for a discussion.

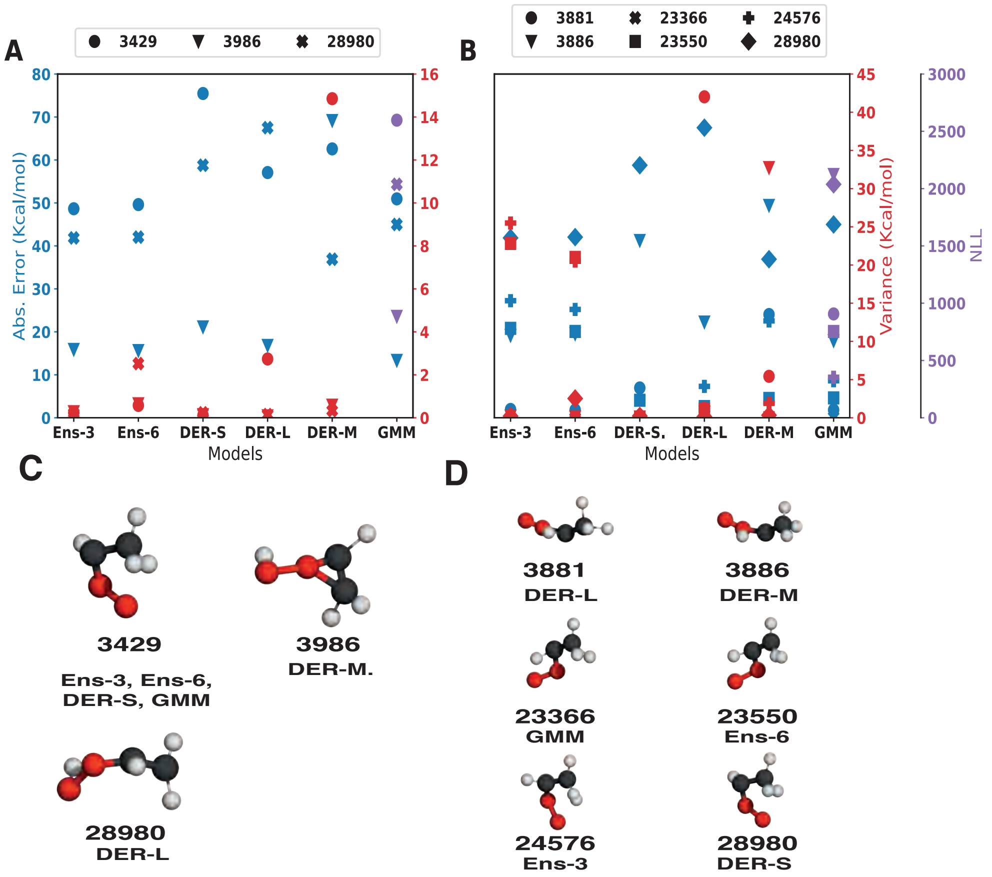

Figure 6.9: Extreme values in prediction. Panel A shows the values of the absolute error (blue) and variance (red) or NLL (purple) for each of the samples identified to have the largest error and its corresponding index. Molecular structures are shown in panel C with their corresponding index and the model for which the structure is identified to have the largest error. Panel B is similar to panel A but for the structures identified to have the largest variance. The corresponding structures are shown in panel D.

One possible reason for the difficulties to predict energies for particular geometrical arrangements concerns the multi-reference character of its electronic structure. To probe this, the T_1[343] and D_1[344] diagnostic coefficients were determined, see Table S5. All structures with large errors clearly display multi-reference character which are not captured from the single-reference MP2 reference data used in the present work. Interestingly, the uncertainty prediction of the models appears to be related to the MR effects as well (Table S6) because the molecules identified with large variance also have large values of T_1 and D_1 diagnostic. These findings are also consistent with earlier work on acetaldehyde.[345]

From the present analysis, ensemble models emerge as a viable route for outlier detection. The capability of the modified DER models are considerably improved over

DER-S, which is largely unsuitable for this task. On the other hand, DER-L is able to detect extreme cases with almost the same quality as the ensemble models thanks to the modifications of the loss function (c.f. Equation 6.4). However, this capability decays rapidly with the number of required samples N_{err}. Finally, DER-M has a constant probability of detecting outliers regardless of the number of samples considered. This is an interesting behaviour because it implies a strong correlation between the error in prediction and the variance. Unfortunately, the probability of detecting outliers for DER-M is ~ 40 % throughout. The remaining model, GMM, showed an intermediate performance between ensembles and DER. However, the NLL as the uncertainty measure is only qualitative and can not be used directly to estimate the error. Nevertheless, it performed well in detecting outliers with good reliability that decay at the same rate as ensemble models.

The fundamental insights gained from the present work are as follows. It is possible to carry out meaningful outlier detection for reactive PESs with the most successful approaches reaching 50 % detection quality for a pool of 1000 structures with the highest uncertainty. Two new formulations of the deep evidential regression method, DER-M and DER-L, were presented and evaluated. The most promising among the approaches tested here are ensemble methods and DER-L, and it is found that Ens-6 and GMM yield consistent results overall. All tested models generally can describe the PES with chemical accuracy (< 1 kcal/mol) on its stationary points. Potential future developments and improvements concern additional modifications to the loss function (scaled-by-variance[273], *post-hoc* recalibration of the uncertainty using isotonic regression[249] and using methods independent on the underlying statistics (e.g. Gaussian distribution of the data in DER) such as conformal prediction methods[346, 347]

6.5 Supporting Information

Supporting information associated with this chapter can be found at: https://github.com/LIVazquezS/SI_PhD_Thesis/blob/main/SI_Chapter6.pdf

Chapter 7

Enhancing chemical databases with information from conformational space

> If you want to have good ideas, you must have many ideas. Most of them will be wrong, and what you have to learn is which ones to throw away.
>
> Linus Pauling

This chapter presents a combination of the results of the previous ones for the study of how to improve the performance of a machine learning model trained on an unbalanced database by adding samples from conformational space. To this end, four artificially biased databases were constructed with the aim of exploring different chemical aspects, such as hybridization, oxidation in organic chemistry, chirality, and aromaticity. Conformers of a representative molecule of the target functionality were added to those biased databases. The effects of sampling temperature and the number of conformers added to the initial databases were evaluated. Alternatively, additions based on amons fragments or by structures with the largest error/uncertainty were also evaluated. The results show little improvement in performance by adding small fractions of molecules obtained at 300 K. Conversely, the addition of amons or based on uncertainty/error enhanced the predictions for the dataset with the largest difference with the training databases. All in all, the results indicate that adding conformations is beneficial if done following chemical criteria.

The results presented in this chapter will be submitted to J. Chem. Inf. Mod.

7.1 Introduction

Chemical space (CS) is the set of all possible molecules or materials[43]. In consequence, its size is extraordinarily large. It has been that the total number of possible substances[26, 46] is about 10^{200}. This large size makes the exploration of CS a big challenge for interested people but, at the same time, a necessary step for human development. In this regard, computational simulations have been consolidated as a powerful tool for this task. Currently, with the rise of machine learning (ML) methods, the obtention of high-quality predictions of chemical properties at a low computational cost has become easier than ever. Consequently, the exploration of CS has progressed in the direction of computational compound design[348, 349].

Nevertheless, for an ML method to perform adequately on different systems, it requires a large corpus of data that can cover as many situations as possible to be trained on. In chemistry, generating this reference data implies a high computational cost and the consequent amount of resources incurring on appreciable environmental costs [34] besides being limited by the size of the molecular systems of interest. In consequence, there is a need to obtain information from other sources that can help to explore CS with the use of ML models. In this regard, using information from conformational space represented by a potential energy surface (PES) represents a viable alternative. It has been proposed that the chemical information contained in a chemical bond and, consequently, in the conformational space provides valuable information that can help to study CS [350]. In particular, for ML methods, we previously found that that the exploration of chemical space can be improved by adding adequate information from the configurational space represented by the PES.

Although adding samples from conformational space is a convenient way to improve the ability of a model to explore CS, there is no clear guidance on how it should be done. Currently, this addition of samples is made by obtaining hundreds or thousands of conformers for a few molecules (i.e. QM7-X [351]) or for a large number of molecules (i.e. ANI-1[190]). However, this approach has the problem that creates redundancies in the data, and the prediction deteriorates as a consequence[88]. besides that, it is only possible if many computational resources are available. Data redundancy creates a problem well-known by the ML communities called dataset imbalance[352]. In chem-informatics, it has been efforts to deal with this problem[353–355] although limited to classification problems. Unfortunately, for atomistic machine learning and to the best of our knowledge, there is only one example of studies that tackle the question of chemical and conformational diversity for ML [356].

This work has a bifold aim. First, we would like to understand, from a chemical perspective, how a chemical database can be improved by adding samples from conformational space; therefore, aspects such as temperature and amount of samples will be evaluated. In addition, we aim to deal with dataset imbalance in a chemical dataset by explicitly biasing the initial dataset and then adding conformers to balance those initially biased datasets. The initial datasets are created to explore different chemical aspects and, therefore, are constructed with specific biases. As a difference to the study of Shenoy, *et al.*, 2023[356], we focused on specific chemical aspects of the databases while diversity is not extensively evaluated. The rest of the article is structured as follows. First, the construction of the artificial databases, data augmentation strategies, and ML method set-up are described in the methods section. Next, the results of the different aspects of the data augmentation are discussed. Finally, some conclusions of the different strategies evaluated are drawn.

7.2 Methods

7.2.1 Artificial Databases

The artificial databases were constructed using molecules extracted from the QM9 database[189], a subset of the GDB-17 chemical universe[49], comprised solely of molecules composed of carbon, nitrogen, oxygen, and fluorine elements. Each molecule in QM9 is limited to a maximum of nine heavy atoms. To ensure data quality, molecules failing the geometry consistency check were excluded from the dataset[189], resulting in a 'cleaned' version with 130,219 molecules, down from the initial 130,831.

The initial artificial databases were created using the *FragmentMatcher* tool within the RDKit software package[357]. This process involved considering the SMILES representations of molecules in QM9 for selection, alongside the generation of SMARTS patterns to identify functional groups of interest, with additional SMARTS patterns to exclude certain groups. Four sets were created to investigate various chemical trends. Table 7.1 and Figure 7.1 provide an overview of the artificial datasets created for this study.

The first set aimed to analyze changes in carbon atom hybridization (Figure 7.1 A). It consists of two subsets: one containing only molecules with single $\ce{C-C}$ bonds (sp^3), excluding double ($\ce{C=C}$, $\ce{C~-C}$, $\ce{C=N}$, $\ce{C=O}$, $\ce{N=N}$) and triple bonds (CC, CN), and another including molecules with $\ce{C=C}$ bonds (sp^2), while excluding triple bonds.

Table 7.1: Composition of the initial artificial datasets used in this work. The first column identifies the property that is wished to be inferred by the Neural Network model. The size of the subset column refers to the total number of molecules used for training, validation and testing.

Set	Composition of Subset	Target Molecules	Size of subset
1 Hybridation	Alkanes Alkanes + Alkenes	Alkynes	31250
2 Oxidation	Alcohols Alcohols + Aldehydes Alcohols + Aldehydes + Ketones	Carboxilic Acids	31250
3 Substituents	Primary Alcohols Secondary Alcohols	Tertiary Alcohols	10816 25695
4 Aromaticity	Alkenes + Cyclohexane	Aromatic rings of six atoms	15673

The goal is to predict $C{\equiv}C$ bonds (sp^1).

The second set aimed to examine changes in the oxidation state of organic molecules (Figure 7.1 B). This set was divided into subsets representing alcohols, aldehydes, ketones, and carboxylic acids. The ML method aimed to infer molecules of different oxidation states without explicitly including them in the training set. It must be mentioned that the QM9 dataset lacks carboxylic acids. Therefore, compounds for the target database were obtained from the PC9 database[51] and recalculated to the QM9 level of theory.

The third set aimed to explore the impact of substituents on molecule prediction (Figure 7.1C). Specifically, it assessed the model's ability to infer chirality from molecules lacking this property. Alcohols were chosen for this study as they can be differentiated based on the number of alkyl groups attached to the carbon in the α-position. The set was divided into two subsets: one comprising only primary alcohols ($RH_2C{-}OH$), and the other containing a mix of primary and secondary ($R_2HC{-}OH$) alcohols. The target compounds for this set were tertiary alcohols ($R_3C{-}OH$).

The final set aimed to ascertain whether an ML model could grasp the concept of aromaticity in chemistry. For this purpose, the dataset exclusively consisted of molecules containing cyclohexane and alkenes (Figure 7.1D). Alkenes from Set 1 were reused, and all compounds containing a cyclohexane ring were selected. The target dataset, in this case, comprised compounds with an aromatic ring containing six atoms.

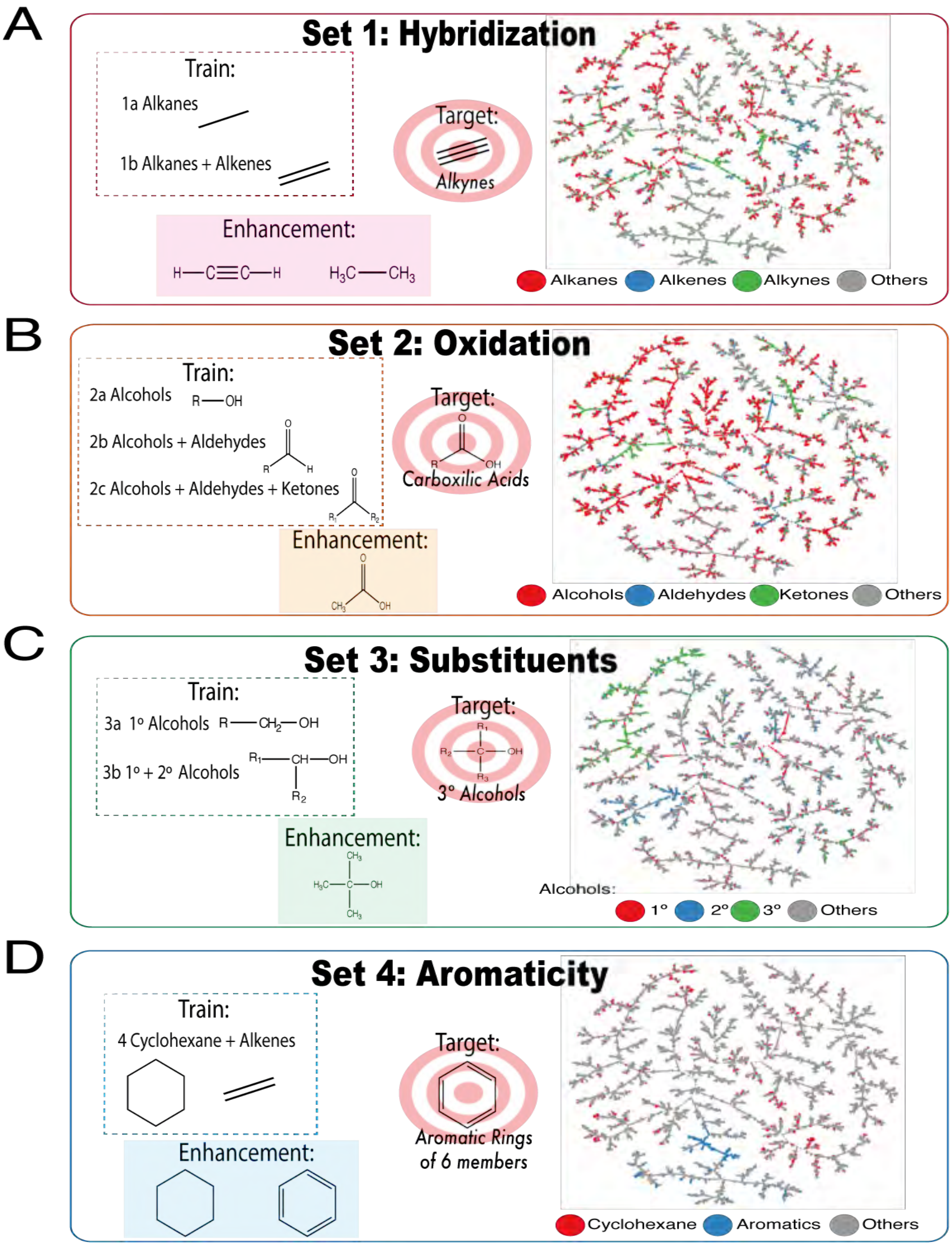

Figure 7.1: **Artificial Databases.** Summary of the constructed artificial databases used in this work. In each panel, the chemical structures of the training databases, together with the target structures and the molecules used for data enhancement. On the right side of each panel is the TMAP representation of the QM9 databases. The molecules with moieties of interest are highlighted if the sample does not present the fragment of interest is not coloured (grey). Panel A shows the molecules in the first set constituted by different hybridization of the C-C bond. Panel B shows different oxidation states of organic molecules; it is important to mention that QM9 does not have recognizable carboxylic acids. Panel C shows alcohol molecules with different numbers of substituents. Finally, panel D shows molecules with cyclohexane and aromatic rings with six atoms.

7.2.2 Machine Learning

The artificial databases were input into the PhysNet DER architecture [143] based on the Deep Evidential Regression[31] method for model training. Training occurred over 1000 epochs with a batch size of 32, and validation occurred every five. The hyperparameter λ, governing the neural network's confidence, was set to 0.2. Unless otherwise specified, all other parameters remained consistent with those detailed in Vazquez-Salazar *et al.*, 2021 [88]. A standard split of 8:1:1 for training, validation, and test sets was employed across all models. Each model underwent three training iterations with different starting seeds (28, 42, and 64). Initial model performance on the test set segment was assessed, with results presented in Table S1. Subsequently, the models were evaluated on the target databases outlined in Table 7.1.

7.2.3 Database Enrichment

Various strategies were employed to enhance the artificial databases. Initially, Normal Mode Sampling (NMS) was utilized to introduce samples from the conformation space. For this method, one or two representative molecules of the target functional group were selected for each artificial database (Figure 7.1) representing either the minimum or an extreme case of the functionality being studied. For instance, in Set 1, ethane and acetylene were selected as extreme examples of $\mathrm{C}-\mathrm{C}$ bonding (Figure 7.1A). Similar considerations were made for Set 4, where cyclohexane and benzene represented extreme cases of double bonds in a six-atom carbon ring (Figure 7.1 D). Set 2 featured formic acid, representing the minimum example of a carboxylic acid (Figure 7.1B). Lastly, Set 3 included conformations derived from *tert*-butanol, the minimum example of a tertiary alcohol (Figure 7.1C).

The selected molecules were then subjected to NMS to generate additional samples. Initially, the impact of temperature on sample generation was assessed by producing 1000 samples at different temperatures ($T = 300, 500, 1000, 2000$K), which were subsequently added to the biased databases. Subsequently, the temperature yielding the most significant reduction in mean absolute error on the target dataset was identified for sample generation, with the number of added samples determined by a percentage of the initial dataset size, ranging from 1% to 25%. The specific number of samples appended to each dataset is outlined in Table S2.

The second method involved enriching the database using Atoms-in-Molecule (amons) fragments[40] derived from molecules in the target database. Molecules ranging in size from 3 to 7 heavy atoms were generated for this purpose.

Finally, uncertainty prediction from the PhysNet model was leveraged to further enrich the databases. Following initial model evaluation on the target set of each database, molecules exhibiting the largest variance and error were identified. These molecules underwent NMS at 300K to obtain 100 samples per molecule, totalling 1000 samples for each database.

7.2.4 Normal Mode Sampling

Normal mode sampling is a proposed alternative to MD sampling to allow targeted sampling of relevant regions of a PES.[34, 188]. Starting from the vibrational normal modes vectors $\boldsymbol{Q} = \boldsymbol{q}_i$ obtained from harmonic analysis of a molecule in an equilibrium conformations $\boldsymbol{x}_{\mathbf{eq}}$. Then, random conformations are generated by displacing the coordinates at equilibrium by randomly scaled N_f normal mode coordinates by a factor defined as:

$$R_i = \pm\sqrt{\frac{3c_i N_a k_b T}{K_i}} \tag{7.1}$$

In equation 7.1, N_a is the number of atoms, k_b is the Boltzmann constant, K_i is the force constant for each of the N_f normal mode coordinates and c_i are pseudo-random numbers in the range of [0,1]. The sign in expression 7.1 is randomly defined by a Bernoulli distribution with $P = 0.5$. Finally, the value of T corresponds to the sampling temperature.

7.2.5 Amons Generation

Amons were generated from the SMILES representation of the different molecules in the target dataset. The SMILES representation is used to construct a molecular graph from which sub-graphs to a maximum number of atoms (excluding hydrogen) are generated. In this work, amons fragments containing between 3 and 7 atoms were generated with an in-house script. Once the structures were generated, they were tested for validity. The validity of the structures was tested by passing the generated structures to a geometry optimization step using MMFF94[198] as implemented in RDKit; samples that did not pass this step were discarded. Samples that passed the validity test were then checked for charge and multiplicity consistency. In this work, only molecules without charge and in basal state were considered. Lastly, the structures that passed the previous tests were passed to a geometry optimization procedure at the level of theory of QM9 using Gaussian16. Molecules that do not converge or present imaginary frequencies were removed. The number of molecules for each database is reported in the supporting information.

7.2.6 Electronic structure calculations

Single Point Energy Calculations For all the molecules generated with normal mode sampling, a single energy calculation was performed at the level of theory of the QM9 database using the Gaussian16 code[358].

Carboxylic Acids Given that the QM9 database does not have a recognizable carboxylic acid by the filtering method, molecules containing the moiety of carboxylic acid ($\mathrm{ROHC{=}O}$) in the PC9 database were considered. Following the same filtering procedure as for QM9, all the molecules containing the smarts string for carboxylic acid were obtained. The structures from PC9[51] were passed to Gaussian16 code[358] for geometry optimization and frequency calculation at the level of theory of QM9 (B3LYP/6-311G(2df,p)). It was checked that all the molecules correspond to a stationary point by assuring the absence of imaginary frequencies in the output.

7.3 Results and discussion

7.3.1 Effect of Sampling Temperature

Determining which region of conformational space provides valuable insights for enhancing predictions poses a primary challenge in data augmentation. Different temperatures influence the sampled space, prompting an initial test utilizing samples obtained at various temperatures. However, the energy distribution analysis revealed that samples from conformational space minimally reduced the disparity between training and test set energy distributions. Specifically, for Sets 1 and 2, added samples created peaks at high energies (> -40 eV), resulting in bimodal distributions (Figures S1 and S2), thereby complicating predictions with PhysNet DER because of the initial Normal assumption of the data. Set 3 exhibited new samples within the interval of -60 to -40 eV, again leading to a bimodal distribution (Figure S3). Set 4 showcased the most significant shift between training and target distributions (Figure S4). The addition of benzene samples in 4a resulted in a peak near the centre of the target distribution, aiding in reducing the disparity between the two distributions. Conversely, adding samples did not substantially disrupt the energy distribution in 4b.

Moving forward, we analyzed the Mean Absolute Error (MAE) changes of the target sets by models trained with the enhanced databases (Figure 7.2). The effect of temperature sampling varied across datasets, with different degrees of influence observed. Set 1 (Figure 7.2 A) showed a smaller MAE for 1b compared to 1a, indicating a positive impact from adding alkenes in 1b. Moreover, 1a demonstrated more significant sensitivity to additions, with broader error distributions than 1b (Figure S5). Regarding the effect of the added molecule, it is observed that adding samples from the acetylene molecule has a larger and more positive impact on the prediction than adding ethane samples. The temperature effect was predominantly negative for Set 1, with MAE generally increasing as the temperature rose.(Figure S5). The effect is considerably larger for 1a, which, regardless of the added molecule, has a larger MAE except for the subset improved with acetylene samples at the lowest temperature.

In Set 2 (Figure 7.2B), the consecutive addition of more oxidized compounds marginally improved the MAE with variations of ≈ 0.3 eV. Notably, the addition of formic acid, induced negligible changes in MAE, with consistent error distributions across temperatures (Figure S6). However, variations in MAE were more pronounced for 2a, increasing linearly with temperature, while 2b and 2c maintained relatively stable MAE.

Set 3 exhibited minimal changes in MAE (Figure 7.2C), with slight improvements observed for 3b compared to 3a. Notably, 3a showcased sensitivity to temperature, with an oscillatory MAE pattern. The error distributions (Figure S7) spread variedly without a clear trend, with the best performance for 3a observed at 2000 K. On the other hand, 3b has an almost constant value of MAE regardless of the sampling temperature.

Lastly, Set 4 demonstrated minor MAE variations regardless of the temperature (Figure 7.2 D). The addition of benzene samples slightly enhanced model performance, compacting the error distributions and shifting the centre of mass to ~ 0.4 eV (Figure S8). Conversely, adding cyclohexane did not significantly alter the distribution spread.

Complementary to the evaluation of the changes in the error of prediction, it is of interest to evaluate how many samples reduce the magnitude of the error with respect to the initial database. This change is quantified by the fraction of molecules that increase its error ($f_{\uparrow}$) defined as :

$$f_{\uparrow} = \frac{\sum_i n_i[|E_i^T| > |E_i^0|]}{n_{\text{total}}} \tag{7.2}$$

Here n_i is the number of molecules for which the condition that the absolute error of the sample i predicted by the dataset enhanced with samples obtained at temperature T,

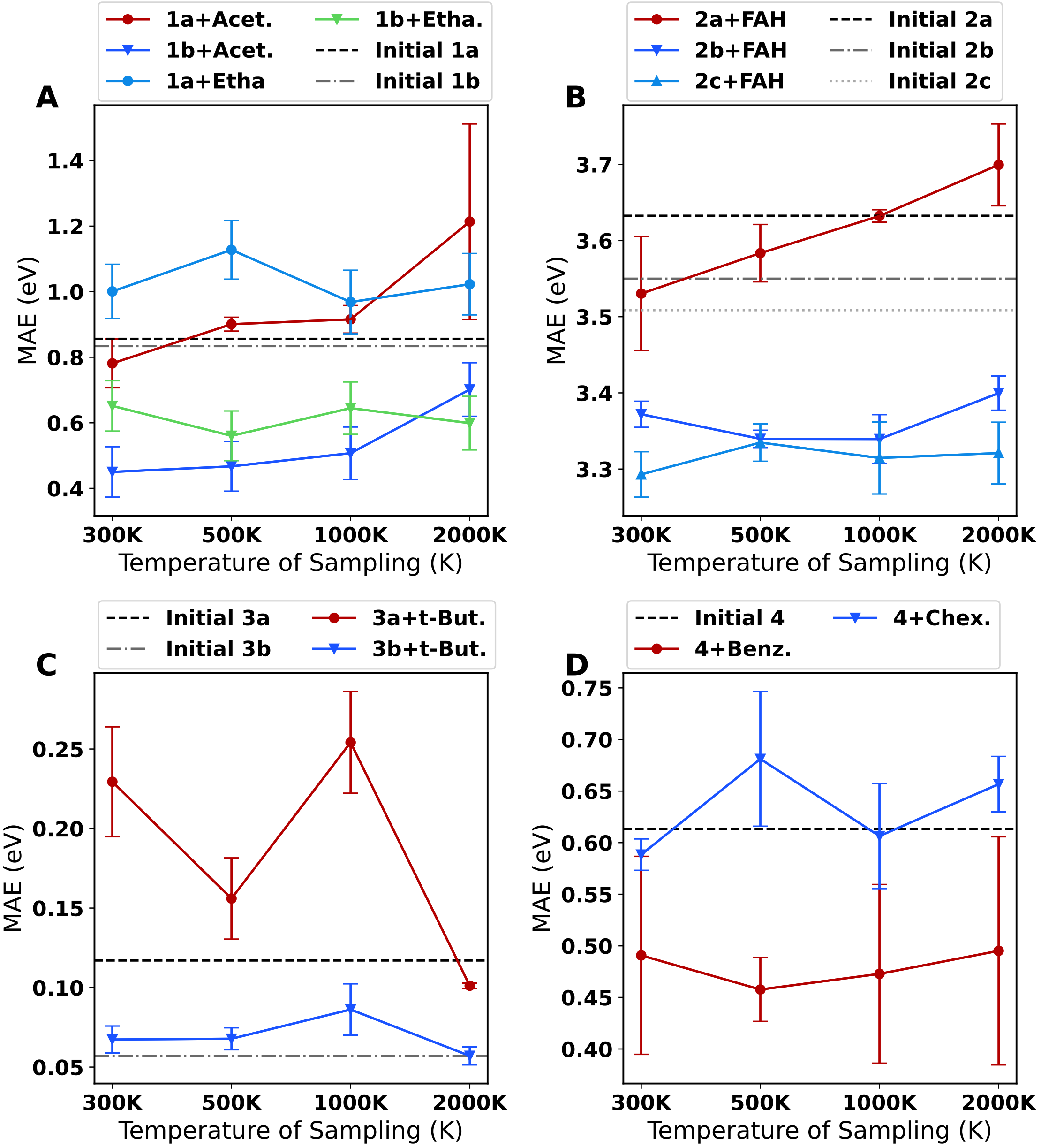

Figure 7.2: Change in the Mean Absolute Error (MAE) in the target dataset of the different databases with respect to the temperature used to obtain samples of a representative structure(s) using normal mode sampling. In all the cases, 1000 samples were added to the initial training dataset. The results show the mean over three models initialized with different seeds. The error bars represent the standard deviation of the MAE over the different values. In each of the panels, the performance of the model in the target dataset before the addition of the sample is shown in horizontal dotted lines.

E_i^T, increases in comparison with the error in the database without additional samples (E_i^0). Conversely, the fraction of molecules for which the absolute error decreases (i.e. $|E_i^T| < |E_i^0|$) is defined as:

$$f_\uparrow = 1 - f_\downarrow. \tag{7.3}$$

The results for the changes in the fraction of molecules that increase or decrease its error are shown in Figure 7.3. This analysis clarifies the effect of adding samples of the selected molecules obtained at different temperatures. As in the case of the MAE, the results are mixed, but the general trend seems negative for most of the datasets, with a large fraction of molecules increasing its error.

For Set 1, $f_\uparrow$ peaked at 2000 K for the datasets augmented with acetylene, while for ethane, the maximum was observed at 500 K. 1a-Acet shows a linear increase on $f_\uparrow$ to temperature rise. On the other hand,1a-Etha. shows oscillations between 0.6 and 0.8 for $f_\uparrow$. Notably, Set 1b exhibited stable behaviour, with $f_\downarrow$ consistently above 0.8, regardless of the temperature and the molecule used for augmentation. The distribution of changes in predicted energy ($\Delta E_{\mathrm{pred}} = E_0^{\mathrm{Pred}} - E_T^{\mathrm{Pred}}$) for set 1 is illustrated in Figure S9, complementing the analysis of $f_\uparrow$/ $f_\downarrow$. Notably, $P(\Delta E_{\mathrm{pred}})$ for 1b exhibits a shift towards positive values, while for 1a, it is centred at 0 (1a-Acet) or slightly shifted towards positive values (1a-Etha). Temperature influences the broadness of $P(\Delta E)$, and it is more evident in 1a-Acet's bimodal profile at 2000 K. Here, the first peak is near zero with a tail towards negative values, while the second peak is centred around -1.7 eV, indicating an increase in predicted energy ($E_T^{Pred} > E_0^{Pred}$) for samples with $\Delta E_{\mathrm{pred}} < 0$. Changes in energy prediction for 1a with ethane are less pronounced, with unimodal distributions centred at 0 with the exception of 300 K shifted slightly towards positive values of ΔE_{pred}. Overall, the initial models tend to overestimate predicted energy for samples in the target dataset. Adding samples from conformational space leads to reduced predicted energy for the best models, particularly at lower temperatures. However, at higher temperatures, the energy is overestimated again due to the inclusion of more disturbed samples. From a chemical standpoint, molecules that increase ΔE_{pred} predicted with 1a typically feature multiple triple bonds, while the same predicted in 1b tend to reduce ΔE_{pred}. Therefore, the effect of adding samples with $\mathrm{C{\equiv}C}$ functionalities for 1a is an increase on E_{pred}, whereas, in 1b, it decreases.

In further examination, we observe an intriguing trend in 2a, where the fraction $f_\downarrow$ increases with temperature (Figure 7.3B), contrary to the marginal increase in MAE shown in Figure 7.2. In further examination, we observe that the fraction $f_\downarrow$ increases

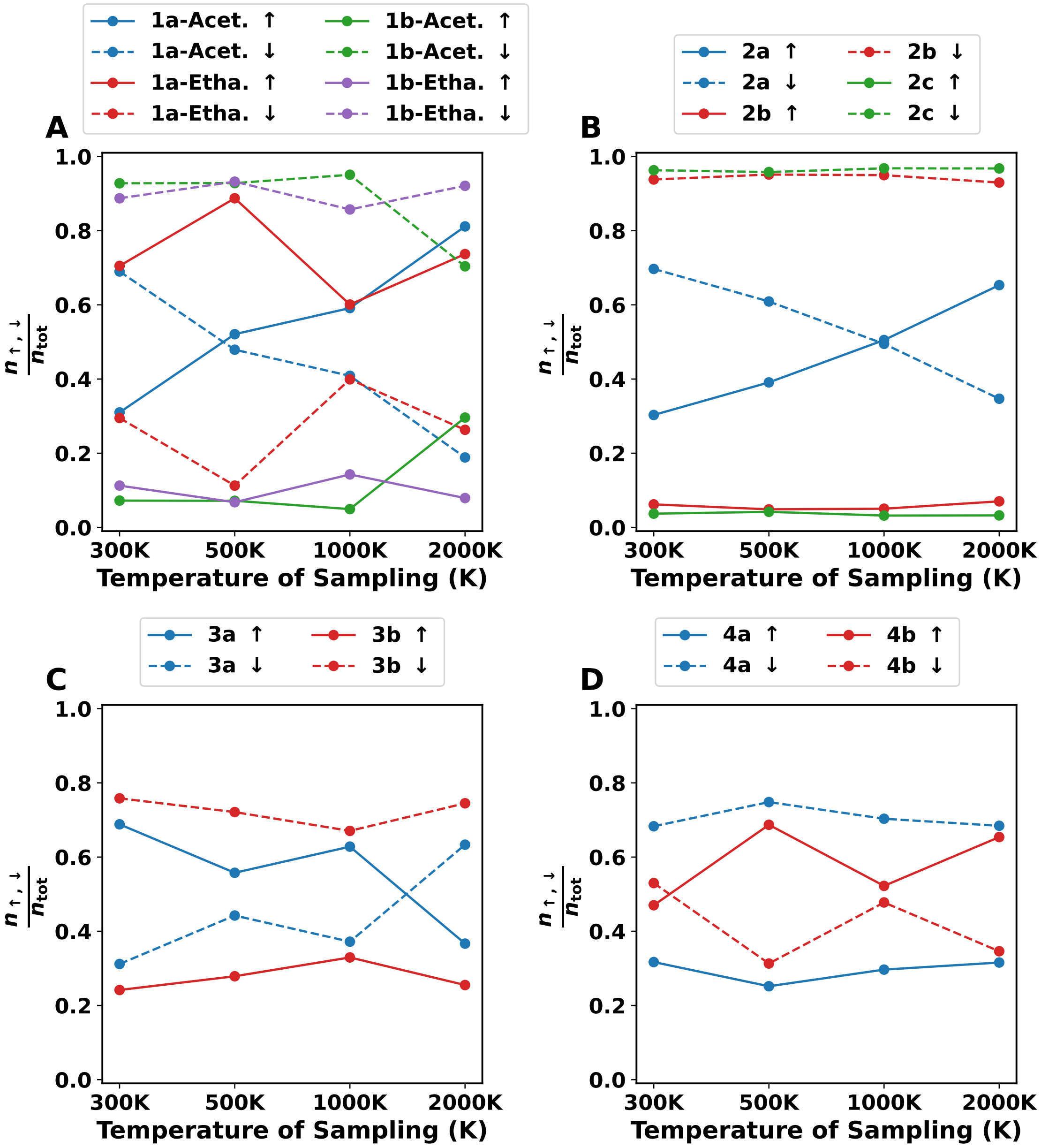

Figure 7.3: Fraction of samples for which the absolute error increases ($f_{\uparrow}$) or decreases ($f_{\downarrow}$) as a function of the sampling temperature for the different artificial datasets evaluated in this work.

with temperature (Figure 7.3B), in line with the marginal increase in MAE shown in Figure 7.2. Analysis of the $P(\Delta E_{\mathrm{pred}})$ distribution (Figure S10) reveals a reduction in E_{pred} for 2a at 300 and 500 K, with a slight increase at the other temperatures. Conversely, both subsets 2b and 2c demonstrate a $f_{\downarrow}$ value approaching 90%, reflected in a shift of the $P(\Delta E_{\mathrm{pred}})$ distribution towards positive values across all temperatures. Molecules exhibiting significant decreases and increases in predicted energy typically contain more heteroatoms (N, O) and carbonyl fragments.

Regarding $f_{\downarrow}$, 3a and 3b exhibit contrasting trends. In the case of 3a, $f_{\uparrow}$ oscillates around 60% for all temperatures except at 2000 K, where it drops to approximately 40%. Conversely, 3b maintains a stable value for $f_{\downarrow}$ at around 70% across different sampling temperatures (Figure 7.3 C). The distributions of ΔE_{pred} are sharply peaked around zero for all temperatures (Figure S11). However, for 3a, the dataset with the best performance, there is a slight displacement of the distribution towards positive values with large tails. Molecules with large negative values of ΔE_{pred} in 3 typically comprise complex structures with multiple rings, whereas those with large positive values exhibit simpler structures.

The findings for set 4a reveal an improvement in prediction for approximately 70% of the molecules in the target set, irrespective of the sampling temperature (Figure 7.3 D). Conversely, for 4b, the fraction $f_{\uparrow}$ oscillates between 40% to 60%, increasing with the sample temperature. The distribution of ΔE_{pred} highlights the opposing trends observed for 4a and 4b (Figure S12). In the case of 4a, the distribution shifts towards positive values with large tails extending up to 3 eV. Conversely, for 4b, the distributions shift towards a negative value of ΔE_{pred}. Additionally, the effect of temperature is evident in the width of the distribution and its tails, which grow with increasing temperature. Molecules with the largest negative value of ΔE_{pred} in 4a typically contain multiple heteroatoms organized in bicycles or feature the presence of the nitro group. Conversely, those that reduce their energy often consist of single aromatic rings.

In summary, the most favourable outcomes are achieved at low temperatures for most datasets, with the exception of set 3, which performs better at high temperatures. Set 2 proves to be the most challenging to predict, as there are no discernible changes in the error distributions after adding samples from conformational space. In contrast, set 1 exhibits the most significant changes in the mean absolute error (MAE), with 1b-Acet yielding the best results.

7.3.2 Effect of the Number of Added Samples

The subsequent aspect of our data augmentation exploration focuses on assessing the impact of the number of added samples to the initial database. After selecting the temperature (300 K) showing the most significant decrease in Mean Absolute Error (MAE), we added varying sample numbers. Similar to temperature assessment, we evaluated changes in energy distribution concerning the training and target sets (Figures S13 to

S16). Changes in the energy distributions are produced at different values than those covered by the target distributions. The analysis of the energy distributions revealed notable shifts for Sets 1 and 2, characterized by the emergence of secondary peaks at distinct intensities (Figure S13 and S14). Interestingly, the intensity of these peaks did not directly correlate with the number of added samples. Then, for 1a-Acet and 1b-Etha, the highest intensity is observed for 1% while for 1a-Etha, it is at 5% and for 1b-Acet, it is at 10%; a similar effect is noticed for set 2. Similarly, for set 3, the energy distribution has a new peak around -60 eV with different intensities (Figure S15). For 3a and 3b, the highest intensity of the new peak is for 1%. Energy distribution for set 4 enhanced with benzene also has a peak at around -60 eV that changes its intensity (Figure S16), reaching its maximum for the dataset enhanced with 1%. For set 4 enriched with cyclohexane, the peak on the energy distribution near -75eV modifies its intensity with the temperature, reaching a maximum at 1% of added samples.

Figure 7.4 illustrates the impact on the mean absolute error (MAE) of models trained with artificial databases and subsequently enhanced with varying sample sizes. Similar to previous observations, the effect is not consistent across all datasets and does not remain constant with the number of added samples to the training databases. For set 1, a notable discrepancy is observed between subsets 1a and 1b, with the MAE consistently smaller for 1a. Additionally, databases enriched with acetylene exhibit lower MAE values compared to those enriched with ethane. Remarkably, the MAE values for set 1 undergo minimal changes with an increase in the number of added samples. Furthermore, the MAE values remain more constant for 1b than for 1a, with a slight overall increase as the number of samples added increases. Analysis of the error distributions (Figure S17) reveals that distributions are more compact for 1b than for 1a, with a tendency to be closer to zero. Moreover, models trained with databases enhanced with acetylene demonstrate error distributions that are less spread out than those using ethane as the sampled molecule.

For set 2, the variations in MAE remain relatively minor, hovering around 0.4 eV, with the most significant fluctuations seen in 2a, followed by 2b, while 2c maintains a consistent value irrespective of the number of added samples. Consistently, the impact of incorporating more oxidized compounds is evident, with the most substantial MAE reduction observed in set 2c, followed by set 2b, and finally 2a. However, alterations in the error distributions across different subsets are marginal (Figure S18).

In the case of set 3, adding samples yields negative effects for both subsets, resulting in larger MAE values for all enhanced databases compared to the initial dataset, except for 3a with 1% augmentation. The MAE shows a slight yet continuous increase with the number of added samples, with a more pronounced effect observed for 3a than for 3b. This increase in MAE is further illustrated by shifts in the error distributions for set 3 (Figure S19), characterized by substantial displacements in the distribution's centre of mass for 3a and an expansion of the distribution tails for 3b.

Lastly, regarding set 4, the outcomes vary depending on whether benzene or cyclohexane is added, although the overall changes in MAE for both scenarios are approximately 0.1 eV. With the addition of benzene samples, a significant MAE reduction is observed between 1 and 5%, beyond which the improvement plateaus, suggesting that the impact of sample addition diminishes beyond the 5% threshold. Conversely, for cyclohexane, the MAE increases from 1 to 10%, with a slight decrease thereafter. Notable changes in the error distributions (Figure S20) are observed when benzene samples are added, characterized by a shift in the distribution centre towards zero and a more uniform distribution compared to the initial one. Contrarily, set 4+Cyclohexane shows no changes in shape but exhibits a slight increase in the distribution's centre of mass.

An analysis of $f_{\uparrow}$ (Equation 7.2) and $f_{\downarrow}$ (Equation 7.3) was conducted (see Figure 7.5). In set 1, notably, 1a exhibits larger values of $f_{\uparrow}$ compared to 1b, in line with results from the MAE values. Furthermore, datasets enriched with acetylene display larger values of $f_{\downarrow}$ compared to those augmented with ethane (Figure 7.5A). Regarding the impact of the number of added samples, an oscillatory pattern is observed across all datasets, except for 1b with acetylene, which maintains a consistent $f_{\downarrow}$ value regardless of the number of added molecules. Specifically, for 1a-Etha with 1% augmentation, $f_{\uparrow}$ is approximately 0.8, reaching nearly 90% at 5% augmentation, then declining to less than 60% at 10% augmentations, before rising again to 80% with the largest sample addition. Conversely, 1a-Acet displays larger values of $f_{\downarrow}$ oscillating between 70% and 30%. On the other hand, 1b shows more stable behaviour, independent of the number of added samples, with over 80% of samples reducing their error. While 1b-Acet maintains a constant $f_{\downarrow}$ value of around 90%, 1b-Etha fluctuates between 90% and 70% for all levels of addition.

Changes in the predicted energy (ΔE_{pred}) for set 1 (Figure S21) underscore variations induced by different models. Across all variants except 1a-Etha, there is an overall mean decrease in predicted energy (i.e., $\Delta E_{\text{pred}} > 0$). Notably, subset 1a-Acet initially exhibits positive ΔE_{pred} values after adding a few samples (i.e., 1% and 5%), shifting towards zero thereafter. Similarly, 1a-Etha consistently displays $\Delta E_{\text{pred}} < 0$ across

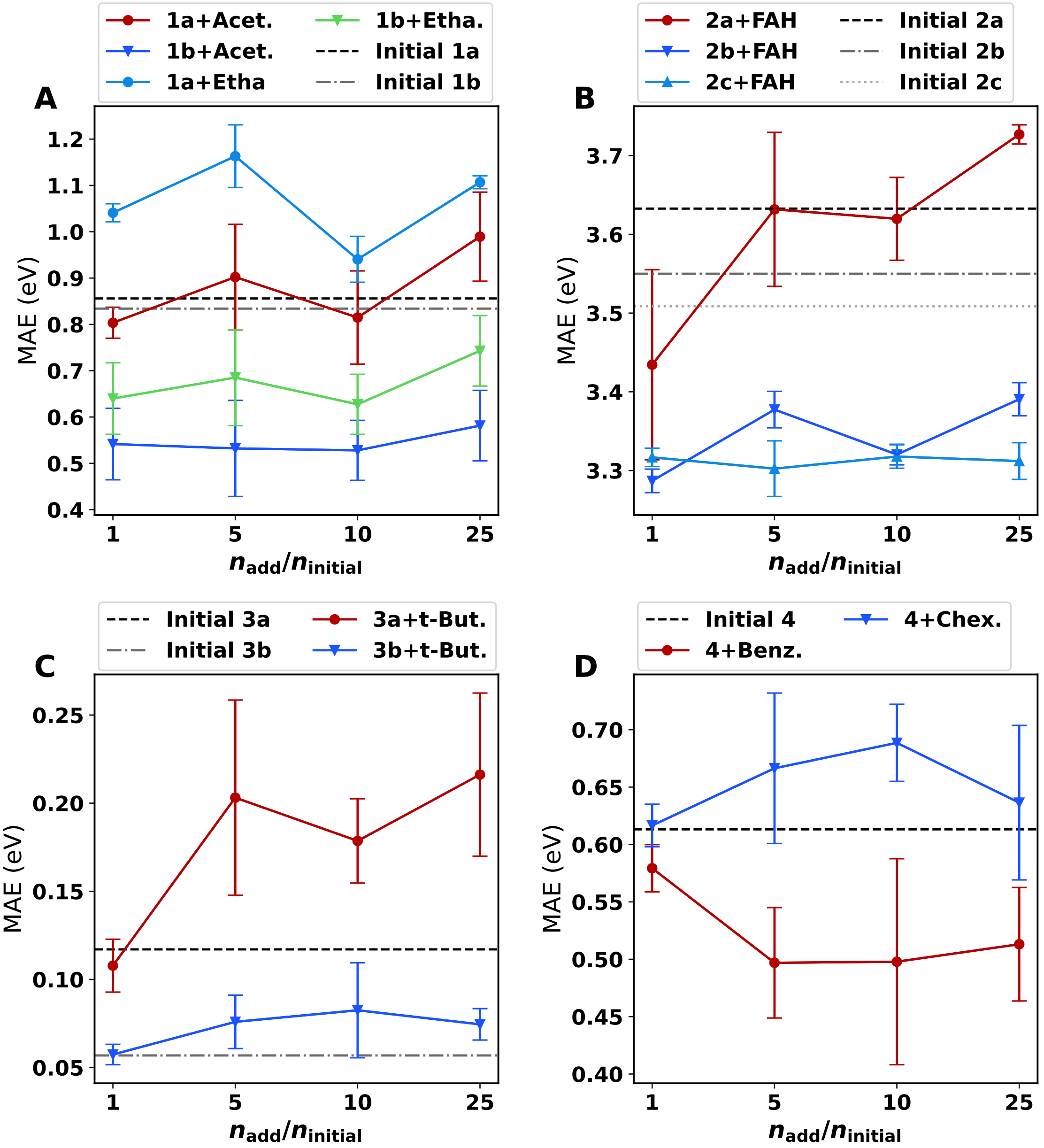

Figure 7.4: Change in the Mean Absolute Error (MAE) in the target dataset of the different databases to the number of samples of a representative structure added to the initial training dataset. In all the cases, samples were obtained using normal mode sampling with a temperature of 300 K. The x-axis in the graphs shows the percentage of added samples to the size of the training dataset. The results show the mean over three models initialized with different seeds. The error bars represent the standard deviation of the MAE over the different values. In each of the panels, the performance of the model in the target dataset before the addition of the sample is shown in horizontal dotted lines.

different augmentation levels. For 1b-Acet, there is a constant positive ΔE_{pred} centred at approximately 0.5 eV for all percentages tested. The case of 1b-Etha is particularly intriguing, with varying positions of ΔE_{pred} centre. Notably, at 5% augmentation, the distribution shows the largest shift with a centre at 0.7 eV, while at 25%, the centre shifts to negative values at approximately -0.2 eV. The chemical structures with significant decreases or increases in predicted energy lack a clear trend. In subsets 1a, decreased predicted energy is associated with structures featuring an oxazole ring or multiple triple bonds, while increased E_{pred} is observed for compounds with a $\mathrm{C{=}N{-}OH}$ moiety or multiple cyanide ($\mathrm{C{\equiv}N}$) fragments. Conversely, in 1b subsets, $\Delta E_{\text{pred}} > 0$ is observed for molecules with one carbon centre substituted by four $\mathrm{{-}CH_2{-}C{\equiv}CH}$ or the $\mathrm{C{=}N{-}OH}$ fragment, while negative values are seen for molecules with a $\mathrm{C{=}O}$ fragment or a formyl-acetamide fragment $\mathrm{O{=}C{-}NH{-}C{=}O}$.

Moving on to set 2, the findings align with the observations of the previous section on temperature effect. Specifically, 2a demonstrates an increase in the value of $f_{\uparrow}$ with the number of added samples, while 2b and 2c maintain constant values of $f_{\downarrow}$ exceeding 90% regardless of the sample size (see Figure 7.5B). The distributions of ΔE_{pred} generally shift towards positive values for most tested scenarios, except for 2a at low percentages (1% and 5%). Regarding chemical structures, they closely resemble those observed in the previous section, characterized by the presence of numerous heteroatoms (O, N) and $\mathrm{C{=}O}$ fragments.

Concerning set 3, a consistent opposite trend between 3a and 3b is evident (see Figure 7.5C). For 3a, there is a growth in $f_{\uparrow}$ with the number of added samples, whereas 3b maintains a high value ($> 70\%$) of $f_{\downarrow}$ regardless of database enrichment. The changes in ΔE_{pred} are illustrated in Figure S23. In 3a, the tails of $P(\Delta E_{\text{pred}})$ shift towards positive values, accompanied by an increase in the width of $P(\Delta E_{\text{pred}})$. These changes appear to align with the observed trend in energy distribution (see Figure S15) rather than the number of added samples. Conversely, subset 3b exhibits a $P(\Delta E_{\text{pred}})$ centred at 0 eV, with alterations primarily observed in the distribution's height. The structures of molecules displaying large ΔE_{pred} remain consistent with those observed in the previous section.

The last sets, 4-Benz and 4-Chex, exhibit opposing trends, with 4-Benz showing a $f_{\downarrow}$ value of approximately 60% (see Figure 7.5D). Contrariwise, 4-Chex demonstrates $f_{\uparrow}$ values close to 60%. The distributions of $P(\Delta E_{\text{pred}})$ for set 4 (see Figure S24) reveal contrasting outcomes for enhancement with benzene and cyclohexane. Benzene enrichment results in reduced energy predictions with positive values of ΔE_{pred}, cen-

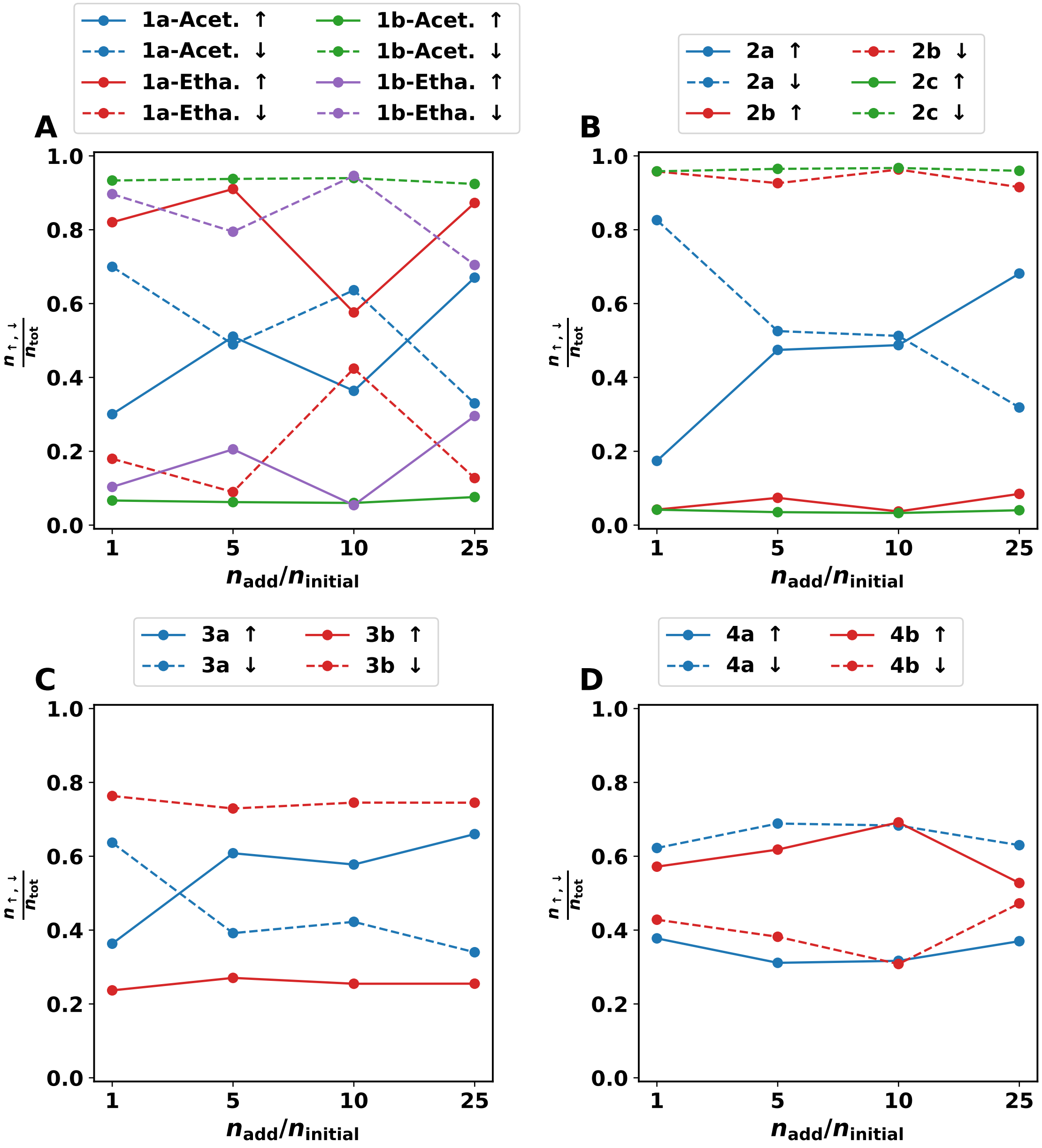

Figure 7.5: Fraction of samples for which the absolute error increases ($f_{\uparrow}$) or decreases ($f_{\downarrow}$) as a function of the fraction of added molecules ($n_{\mathrm{add}} n_{\mathrm{initial}}$) for the different artificial datasets evaluated in this work.

tring the distribution's mass at larger positive values for small addition percentages. In contrast, set 4 enriched with cyclohexane shows distributions centered at negative values, with the mass centering at more negative values for small addition percentages. Molecules exhibiting significant changes in ΔE_{pred} commonly feature fused rings with heteroatoms and nitro ($\mathrm{O} \overset{--}{-} \mathrm{N} \overset{--}{-} \mathrm{O}$) fragments.

In summary, this section highlights that the best results are achieved by adding a small fraction of samples to the initial databases. This finding highlights the importance of modifications in the energy distribution within the training dataset. Additionally, it was observed that adding more samples either harms or has no significant effect on prediction accuracy. Once again, set 2 emerges as the most challenging to predict, with marginal improvements, while set 1 undergoes the most significant changes.

7.3.3 Amons

The results of augmenting artificial databases with samples from the conformational space of a single molecule suggest that while there may be slight improvements in prediction performance, they are generally marginal. This could be attributed to the fact that adding only a simple functional group does not cover all the possible undersampled parts of chemical space present in the target database, e as evidenced by the energy distributions in Figures S1 to S4 and S13 to S16. Therefore, employing amons fragments[40], which systematically partition molecules to represent various regions of chemical space, could offer a more comprehensive approach to address undersampling. As in the previous sections, we will start by describing the changes in the distributions of energies of the datasets improved with amon fragments to obtain an overview of the changes in the initial training databases and their comparison with the target database.

Beginning with set 1, the energy distributions (see Figure S25) retain a similar shape to the initial ones. However, for 1a with amon sizes 4 and 5, the distributions become broader, exhibiting improved overlap with the target distribution. Similarly, 1b with amon sizes 4 and 6 also widens its distribution. Despite the described changes, the differences between target and enhanced distributions are maintained. Moving to set 2, the energy distribution (see Figure S26) shows increased overlap with the target distribution. However, a few peaks appear at low energy values (-100 to -80 eV), while the tails decay faster at higher energies (-60 to -40 eV), resulting in decreased overlap in both regions. Set 3 displays mixed behaviours (see Figure S27): 3a with amon size 5 shifts the distribution, losing coverage at low energies but gaining at high energies. At the same time, 3b shows minimal changes in the energy distribution. Finally, for set 4, changes in the height of the distribution are observed at low energy values. While the formation of a shoulder that varies in height with increasing amon size is observed at higher energy values (Figure S28).

In general, the effect of the addition of amons to the training set deteriorates the performance of the model on the prediction of the target dataset (Figure 7.6). Specifically, in set 1 (Figure 7.6A), minor differences between the outcomes of 1a and 1b are discernible, with 1a generally outperforming 1b. Concerning amon size, a substantial increase in MAE is observed for smaller amons (sizes 3 and 4). However, the MAE then steadily decreases, eventually reaching values lower than those of the initial database for the largest amon sizes (size 7). The distributions of MAE for 1a (see Figure S29) become broader with increasing amon size up to size six. Notably, for 1a enhanced with amons of size 7, the MAE distribution exhibits a prominent peak near 0 but extends into long tails, reaching up to 5 eV. On the other hand, for set 1b, the distribution also widens up to size 5. Sizes 6 and 7 display more concentrated distributions with peaks closer to 0, albeit accompanied by extensive tails. Still, the tails are more pronounced for size 6 compared to size 7.

Moving on to set 2, it has an interesting behaviour because the MAE, irrespective of amon size, is lower than the initial database. This is intriguing, considering that this dataset exhibited only marginal improvements in the previous section. Similar to previous cases, 2a displays the most significant variations in MAE, consistently decreasing it to a minimum at size 6. Conversely, sets 2b and 2c show a slight increase in MAE but still maintain a better MAE compared to the initial database. One plausible explanation for the enhanced prediction in set 2 with amons is its substantial divergence from the training datasets, originating from a different database (PC9). PC9 has a greater chemical diversity than QM9, the source of the initial training databases for this set. Hence, it's reasonable to expect that PC9 introduces different moieties not originally present in QM9, making the prediction tasks harder for the model. Notably, the distribution of MAE (see Figure S30) undergoes significant changes for 2a, featuring a broader profile. However, for size 6, the distribution exhibits a peak near 1 eV with extensive tails. In contrast, 2b and 2c demonstrate relatively minor changes in the distribution of MAE.

Moving forward to set 3, two distinct scenarios emerge yet again. For 3a, there is an observable increase in MAE, which then fluctuates as the amon size varies, while for 3b, there is negligible change compared to the initial value. The error distribution for set 3 (see Figure S31) undergoes significant alterations in shape, width, tails, and centre of mass for 3a, whereas 3b experiences minimal changes. Similarly, set 4 fails to exhibit improvement, displaying a larger MAE than the initial dataset for all amon sizes, with no notable changes in the error distribution shapes (see Figure S32).

The study of the fractions $f_\uparrow$ and $f_\downarrow$ is reported in Figure 7.7. For set 1, it was observed that the value of $f_\uparrow$ is around 80% for sizes 4 and 5 (Figure 7.7 A). Then, the value decays to 30 % following the same trend of the MAE in Figure 7.6. An analysis of the difference in the energy prediction ΔE_{ref} (Figure S33) shows that the distribution centre of mass shifts to positive values for small amon size for dataset 1a; the largest change is seen for size 5 with a distribution centred at ~ 5 eV. On the contrary, the distributions of large amons for 1a have a bimodal profile, with size 6 having the highest peak around 5 eV and size 7 at ~ 1 eV. Likewise, 1b has an unimodal distribution for amon size 4; at the same time, for other sizes, the distributions are bimodal. Again, $P(\Delta E)$ for 1b has the first peak near 1 eV, and it is the highest for 6 and 7. The second peak of $P(\Delta E)$ of 1b can be found around 5 eV, which is the highest for size 5. Molecules with the largest variations of ΔE_{pred} have some commonalities. In the case of samples with large positive values of ΔE_{pred}, the structures have multiple triple bond fragments and $\mathrm{C{\equiv}N}$ groups. On the contrary, those with large negative values have 5-member rings or the fragment $\mathrm{-\!\!-CH_2-C{\equiv}CH^{\cdot}}$

In Set 2, a substantial portion of molecules (> 70 %) experiences a reduction in error, a trend unaffected by the amon size (Figure 7.7B). The value of $f_\downarrow$ exhibits minor oscillations among subsets. Specifically, in 2a, $f_\downarrow$ rises initially and then declines with larger amon sizes. Conversely, 2b displays no clear trend, with slightly small $f_\downarrow$ for even sizes and slightly higher for odd sizes. Conversely, 2c demonstrates an opposite trend to 2a. Similar to prior cases, 2a showcases significant changes in predicted energy (Figure S34), while 2b and 2c remain relatively constant, with minor shifts in the distribution's center of mass. Molecules exhibiting significant changes in E_{pred} typically feature a considerable number of heteroatoms in their structures. Notably, across all datasets, the moiety $\mathrm{O-O-OH}$ exhibits the most negative ΔE_{pred} values, while molecules with large positive values often contain amine groups ($\mathrm{NR_3}$) or double carbon bonds ($\mathrm{C{=}C}$).

Moving to Set 3, two distinct behaviours regarding the values of $f_\uparrow$ are observed (Figure 7.7C). Specifically, 3a has a large value of $f_\uparrow$ that increases with amon size, reaching values above 80%. Conversely, 3b maintains a constant $f_\downarrow$ around 50%, suggesting a relatively stable ratio of improved and worsened molecules. Notably, the disparity in

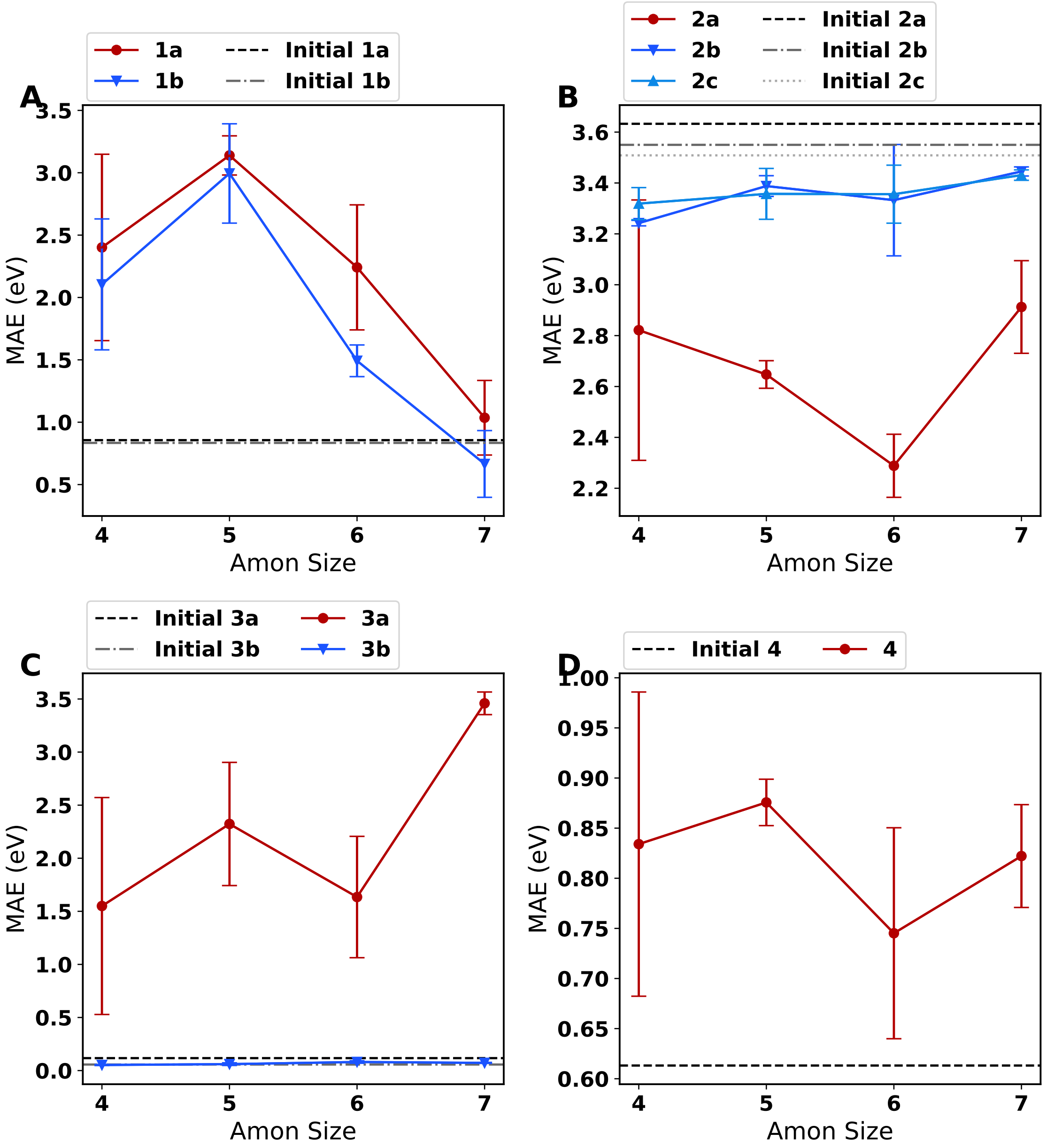

Figure 7.6: Change in the Mean Absolute Error (MAE) in the target dataset of the different databases with respect to the size of the added amons to the initial dataset. The results show the mean over three models initialized with different seeds. The error bars represent the standard deviation of the MAE over the different values. In each of the panels, the performance of the model in the target dataset before the addition of the sample is shown in horizontal dotted lines.

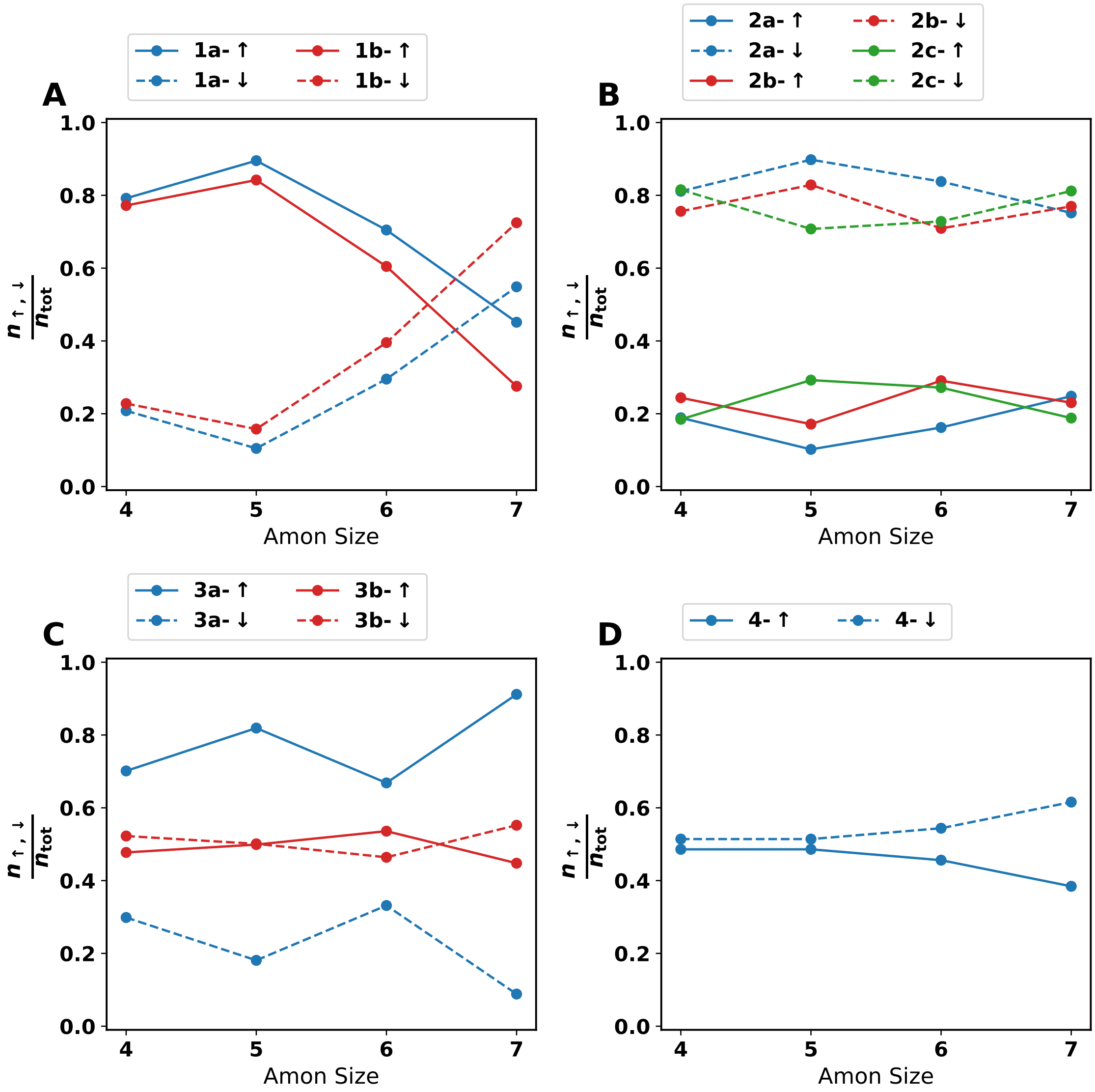

Figure 7.7: Fraction of samples for which the absolute error increases ($f_{\uparrow}$) or decreases ($f_{\downarrow}$) as a function of the amon size added to the initial databases for the different artificial datasets evaluated in this work.

changes becomes more apparent when examining the distributions of ΔE_{pred} (Figure S35). For 3a, $P(\Delta E_{\text{pred}})$ exhibits a bimodal shape, with one peak centred at 0 eV and another at 4 eV that intensifies with larger amon sizes. In contrast, the distribution for 3b features a sharp peak centred at zero, with variations primarily in intensity. The chemical structures with the most significant changes in ΔE_{pred} typically involve 5- or 4-membered rings for negative values and simpler structures for positive values.
For Set 4, variations in the fractions $f_{\uparrow}$ and $f_{\downarrow}$ are minimal, with $f_{\uparrow}$ consistently smaller than its counterpart (Figure 7.7D). The discrepancy between the fractions enlarges with the size of the added amon fragment. Changes in predicted energy are also negligible (Figure S36). For amon size 4, the centre of mass of $P(\Delta E_{\text{pred}})$ shifts to negative values, while for larger sizes, it shifts to positive values, with the shift becoming more pronounced as the amon size increases. Regarding chemical structures, those with significantly negative ΔE_{pred} values typically feature multiple nitrogen and oxygen atoms on aromatic rings, while those with notably positive values contain pyrimidine and triazine rings.

Overall, the results of this section indicate that the amon approach does not enhance energy prediction for the tested databases, except for Set 2. The amon size exhibits an inconsistent effect across the databases, with all datasets experiencing significant MAE changes in at least one subset. While Sets 1 and 2 show a positive impact, with Set 1 benefiting only from the largest amon size, Set 2 displays a positive effect across all sizes. Conversely, Sets 3 and 4 exhibit poor performance, characterized by substantial MAE increases.

7.3.4 Uncertainty/Error Guided Sampling

The final enhancement method examined in this study profits from the uncertainty predicted by the neural network model. Samples added to the training dataset were chosen from the target set based on their predicted variance. Complementary, structures with significant errors were selected from the target set for comparison. NMS was then applied to the selected structures to generate a total pool of 1000 samples.

Initially, the impact of adding samples was assessed by examining changes in energy distributions (Figures S37 to S40). In Set 1, there were minimal alterations in the energy distributions, except for a slight increase in peak height and the emergence of a small peak for 1a at high energy values (-50 to -30 eV). Likewise, Set 2 displayed only minor changes in peak height and the appearance of small peaks around -60 to -40 eV. Set 3 exhibited similar alterations, primarily in peak height, particularly noticeable between -60 and -40 eV. This trend persisted in Set 4, where the addition of samples with significant errors led to the emergence of a new peak around -60 eV in the energy distribution.

The MAE for databases augmented using two different metrics is depicted in Figure 7.8. Overall, augmenting samples based on uncertainty or error tends to degrade the model's performance, except for Set 2 when using uncertainty. An individual analysis starting with Set 1 exhibits an increase in MAE after sample addition in both subsets. Notably, the error-based addition yields smaller errors compared to uncertainty-based addition. For Set 1a, performance is slightly better than 1b when augmenting based on error, with broader error distributions observed for 1a (Figure S41). Contrarily, 1b demonstrates lower error but with a more spread-out distribution, albeit with a lower centre of mass. In Set 2, augmenting samples based on uncertainty leads to a lower MAE compared to error-based addition, with a difference of approximately 2 eV. Notably, subsets 2b and 2c exhibit MAE values close to the initial database, reflected in minimal changes in error distributions (Figure S42). Continuing with Set 3, error-based augmentation increases MAE by approximately 3 eV. In comparison, uncertainty-based augmentation increases MAE for 3b by around 2 eV, albeit with 3a showing a slightly smaller MAE than the initial value. Error-based augmentation notably shifts the distribution's centre of mass to approximately 4 eV, widening its spread. Similarly, uncertainty-based augmentation shifts the centre to 3 eV for 3b and increases the distribution's width (Figure S43). For Set 4, error-based augmentation results in a slightly larger MAE than uncertainty-based augmentation. The error distributions exhibit a shift in the centre of mass towards larger values, with increased tail and width (Figure S44).

The analysis of individual error changes was conducted using the fractions $f_{\downarrow}$ and $f_{\uparrow}$, as shown in Figure 7.9. Beginning with Set 1, it's evident that both subsets and augmentation methods yield higher values for $f_{\uparrow}$. Comparing the augmentation methods, $f_{\uparrow}$ tends to be slightly larger for uncertainty-based additions. Regarding subsets, 1b exhibits higher $f_{\downarrow}$ with error-based augmentation, while 1a shows the same trend with uncertainty-based augmentation (Figure 7.9). The distributions of ΔE_{pred} for both subsets and metrics are centred at positive values (Figure S45). For 1a, $P(\Delta E_{\mathrm{pred}})$ is centred at approximately 3.5 eV and 4.5 eV for error- and uncertainty-based additions, respectively. Moreover, the distribution's height is greater for uncertainty-based additions, with significantly larger tails compared to the other method. In the case of 1b, both methods yield a distribution centred at around 5 eV. An examination of chemical

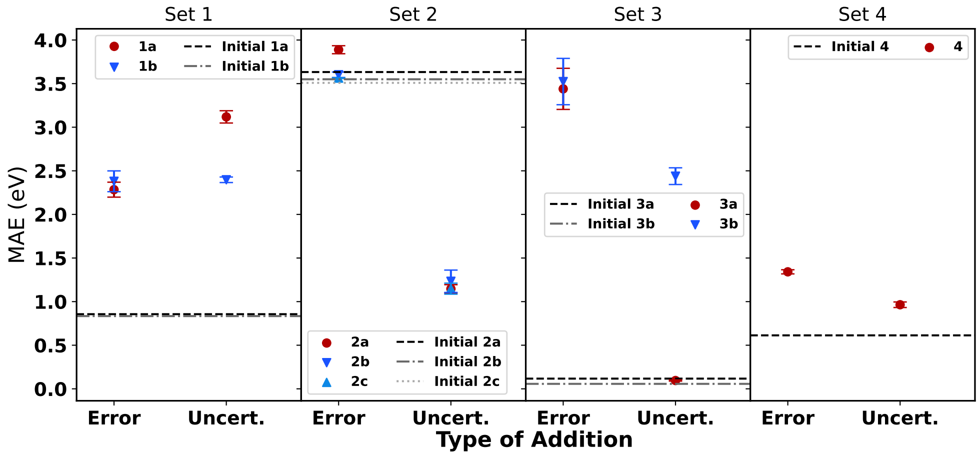

Figure 7.8: Changes in the Mean Absolute Error in the target dataset of the different databases with respect to the method of addition. In the case of uncertainty-based addition, the ten molecules with the largest uncertainty value in the initial evaluation were chosen to enhance the training database. For each of those molecules, 100 structures were generated using normal mode sampling at 300 K. The same procedure is repeated for the molecules with the largest error. The results show the mean over three models initialized with different seeds. The error bars represent the standard deviation of the MAE over the different values. In each of the panels, the performance of the model in the target dataset before the addition of the sample is shown in horizontal dotted lines.

structures reveals that molecules with the highest ΔE_{pred} values contain multiple $\mathrm{C{\equiv}C}$ fragments. Conversely, those with negative ΔE_{pred} values feature a ring system and a $\mathrm{C{=}O}$ group.

Moving to Set 2, the values of $f_{\uparrow}$ and $f_{\downarrow}$ exhibit distinct behaviours depending on the augmentation methods. With uncertainty-based addition, $f_{\downarrow}$ values surpass 80%, whereas for error-based addition, $f_{\downarrow}$ values fluctuate around 50% (Figure 7.9). As seen previously, significant disparities emerge in the distribution of ΔE_{pred} (Figure S46). For error-based addition, $P(\Delta E_{\text{pred}})$ forms sharp distributions centred at 0 for subsets 2b and 2c, while for subset 2a, it centres at a negative value. Conversely, with uncertainty-based addition, $P(\Delta E_{\text{pred}})$ shifts towards positive values with considerable width. The chemical structures undergoing the most significant changes commonly feature heterocycles with N and O or $\mathrm{—N—COOH}$ fragments.

For set 3, the analysis of $f_{\uparrow}$ and $f_{\downarrow}$ show interesting trends. Error-based addition results in a high $f_{\uparrow}$ for both subsets, while uncertainty-based enhancement yields a substantial fraction of $f_{\downarrow}$ for 3a and a minor fraction for 3b (Figure 7.9). Notably, there are considerable disparities in the predicted energy differences (Figure S47). For 3a, the error-based approach produces a broad distribution centred at approximately 3.5 eV,

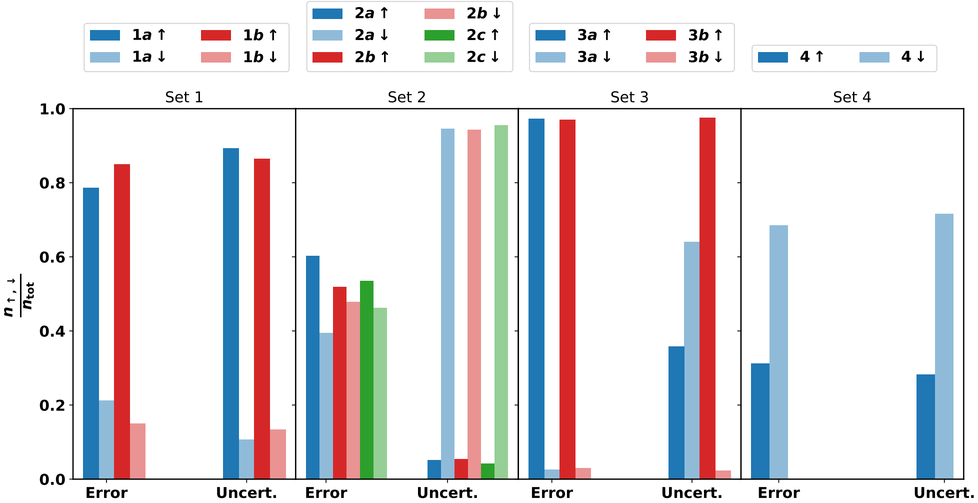

Figure 7.9: Fraction of samples for which the absolute error increases ($f_{\uparrow}$) or decreases ($f_{\downarrow}$) as a function of the strategy of addition used to select the initial samples for the different artificial datasets evaluated in this work.

whereas the uncertainty-based approach exhibits a sharp peak near 0 eV. In contrast, for 3b, both methods shift the distribution to positive ΔE_{pred} values with extensive tails decaying towards 0 eV. Chemical structures associated with positive ΔE_{pred} values feature multiple tertiary alcohols, while those associated with negative values entail multiple fused cycles.

Set 4 consistently shows $f_{\downarrow}$ values exceeding 60%, regardless of the augmentation method used (Figure 7.9). Furthermore, the examination of ΔE_{pred} (Figure S48) reveals similar shapes for both types of adddition, with a peak centred at 0 eV. However, the error-based addition results in a bimodal distribution, with a secondary peak at 3 eV. Structures with negative ΔE_{pred} values typically incorporate a ring with two nitrogen atoms and one oxygen atom, along with a $\mathrm{C{=}O}$ moiety. Conversely, structures with positive ΔE_{pred} values feature multiple nitrogen atoms within the ring.

In conclusion, the inclusion of samples from the conformational space, whether based on error or uncertainty, generally has a detrimental impact on most datasets. Specifically, adding samples based on error worsens prediction errors across all databases. Conversely, while uncertainty-based addition proves beneficial for set 2, it yields adverse outcomes for the remaining datasets.

7.4 Conclusions

This study investigated three methods for enhancing initially biased chemical databases. These databases were designed to cover various chemical aspects, including hybridization, oxidation, chirality, and aromaticity. The performance assessment of these methods focused on mean absolute error, the fraction of samples with increased/decreased absolute error in the target dataset, changes in E_{pred}, and the chemical structures of samples exhibiting significant changes in E_{pred}.

The first method assessed involved augmenting the datasets with samples generated using normal mode sampling of a representative molecule corresponding to the targeted chemical aspect. This analysis included examining the impact of sampling temperature and sample size. Generally, adding samples from a single molecule had minimal effects on sets 2, 3, and 4. However, the influence of temperature was found to slightly degrade prediction accuracy across most databases, with optimal results achieved at 300 K. Conversely, smaller sample sizes yielded better performance, suggesting that redundancy and highly disturbed structure addition adversely affect prediction quality.

The results of the first method indicate that adding a single moiety fails to fully address the distribution shift issue across different databases. Therefore, alternative methods were explored. The first involved utilizing the atoms-in-molecule fragments approach. While sets 1, 3, and 4 did not yield positive outcomes, set 2 exhibited significant improvement. This behaviour can be attributed to the fact that the target set of set 2 originates from a distinct dataset, thus encompassing various unexplored regions of chemical space by the initial database. Notably, subsets 2b, 2c, and 3b show no changes in the error regardless of the amon size. Conversely, set 1 displayed substantial differences, improving performance with larger amon sizes.

Another method assessed involved enhancing the database by utilizing samples with significant errors or variances, followed by normal mode sampling to generate samples from conformational space for each molecule. Unfortunately, this approach yielded negative results for sets 1, 3, and 4, resulting in increased MAE for these datasets. Intriguingly, employing samples selected based on uncertainty values yielded better performance compared to using molecules with large errors. Similarly to the amons method, set 2 demonstrated considerable improvement in MAE, which can be attributed to the unique construction of the target set for this dataset.

In general, it was shown that incorporating samples from conformational space could enhance property prediction, outperforming methods like amons fragments. Particularly, the generation of new samples must be performed at low temperatures to capture relevant regions of conformational space that help to improve prediction. Moreover, it was observed that a small number of samples can yield a significant impact compared to an excessive amount. Notably, when dealing with a target dataset from a different distribution, utilizing amons fragments or uncertainty-based sampling appears to be more effective. Future research should explore the influence of the sample size in uncertainty-based enhancement and the number of samples from conformational space per molecule. Additionally, analyzing changes in structural properties can further elucidate database modifications. This can be done by using the Kullback-Leibler divergence to evaluate changes in the structural properties of the databases (i.e. bond distances) in addition to the analysis of energy distributions made here.

7.5 Supporting Information

Supporting information related to this chapter can be found at: https://github.com/LIVazquezS/SI_PhD_Thesis/blob/main/SI_Chapter7.pdf

Notes

The author used Chat-GPT3.5 to improve the readability and conciseness of the text. However, the text was reviewed for errors.

Chapter 8

Conclusions

> The world little knows how many of the thoughts and theories which have passed through the mind of a scientific investigator have been crushed in silence and secrecy by his own severe criticism and adverse examination; that in the most successful instances not a tenth of the suggestions, the hopes, the wishes, the preliminary conclusions have been realized.
>
> Michael Faraday

Over the different chapters of this thesis, we explored different aspects of how chemistry insights can be gained by using machine learning algorithms. Those algorithms are a powerful option for exploring chemical and conformational space. However, ML models must be used carefully, and different aspects must be considered. This last chapter summarises the different pieces of work discussed here and provides some perspectives on future extensions for the various aspects of this work.

In Chapter 3, the influence of the composition of chemical databases was reviewed. It was found that the prediction of a chemical property (i.e. tautomerization energy) is strongly dependent on the chemical diversity of the training databases. The different aspects evaluated show that biases on chemical training sets can be identified. For example, $\mathrm{C{=}C}$ near heteroatoms (N, O) or azoles are harder to predict because they are not included in the training set. Additionally, it was found that an adequate addition of samples from conformational space compensates for the lack of chemical diversity. However, the number of conformers added to the chemical database must be balanced with the chemical diversity to maintain a good performance. Moreover, we introduce a quantitative measure of the deficiencies in chemical databases by means of the Kullback-Leibler (KL) divergence between distributions of bond lengths. Some aspects of this study that can be improved include extending the use of KL to other geometric properties of molecules like angles or dihedrals. From the results of our study, it was unclear if many-body terms have a larger importance than single 2-body terms. In addition, KL divergence over the different structural quantities of molecules can be used to construct chemical databases by minimising the difference between distributions of bonds, angles or dihedrals on test and training databases by generating synthetic data as recently done with Gaussian processes [359] or by the use of generative models[75]. Finally, using other metrics, such as Jensen-Shannon divergence[277], can be an alternative because it does not suffer from symmetry problems like KL.

As an alternative to the quantification and use of structural properties, uncertainty quantification (UQ) can be employed to construct chemical databases. Chapter 4 explores using models that can predict its own uncertainty with this end. Usual UQ approaches involve training several models and then calculating the mean and standard deviation of their predictions; this strategy is commonly known as ensembles. A disadvantage of using the ensemble method is the high computational cost of training the different models. Then, alternatives to quantify the uncertainty in the prediction were studied. In particular, a method based on Bayesian probabilities called Deep Evidential Regression was implemented on top of the PhysNet model. The new model was characterized through different tests. The calibration of the predicted uncertainty and the relationship between error and variance were evaluated. Although there is no linear correlation between error and uncertainty, it was found that the uncertainty can give insights into the data quality and biases of the databases. Additionally, it was quantified the ability of the model to label samples with large errors with a corresponding large variance. To obtain a better understanding of how the model predicts the values of variance or the lack of information that leads to large errors, the distances in the embedding space of the NN were used. This new analysis highlights that, as expected, the lack of information complicates an adequate prediction. At the same time, redundancy creates "confusion" in the model, leading to misclassifying samples with high errors with small variances. A second method for UQ that tries to create a bridge between ensemble models and single model prediction of uncertainty called Regression Prior Networks was also tested. However, its performance was poor because of technical limitations. The implemented methods provided several insights into the relationships of the chemical structures in training and test sets. However, the statistical assumptions made on its construction limit its prediction capability, as will be discussed next. Consequently, using distribution-free methods like conformal prediction [346, 347] are interesting alternatives to improve the quality of the uncertainties obtained.

Until now, the constructed models were created to reproduce results from *ab-initio* calculations. Nevertheless, it is interesting to create models that also reproduce experimental results. Chapter 5 explored this possibility for the system of He-H_2^+. Starting from high-quality potential energy surfaces (PES) generated with a Reproducing Kernel Hilbert Space model from high-level quantum chemistry calculation, a procedure to scale the coordinates and energy called morphing was used to match the results of scattering calculations and experimentally determined Fesbach resonances. The transformation of the coordinates was done by multiplying them by a scalar value, e.g. $V_{\mathrm{morphed}}(R, r, \theta) = \varepsilon V_{\mathrm{ab-initio}}(\alpha R, \beta r, \theta)$. The determination of the different scalar values was done by an iterative procedure on which a loss function of the position of the energy peaks and its intensities are minimized with respect to experimental values. The results obtained indicate that even the PES obtained at the highest level of theory needs to be scaled with respect to the experimental quantities to improve the results. In addition, it was found that the procedure is sensitive to the parts of the potential visited for the experimental measurements. Further, it was seen that the description of the PES needs to be done in a global form (i.e. avoiding a decomposition on n-body terms) to obtain physically sound results. Although the model returns adequate results, it is clear that the simple transformation used can not cover all the changes in a PES because the surface might require different adjustments on different regions. Consequently, a non-linear transformation is necessary to obtain better results. A possible model that can be used is the transformer architecture revised in Chapter 2 and the key behind the success of LLM. The selection of this model is because of the attention layer on it. The attention layer is able to capture local changes based on the context; for the case of PES, it is expected that some parts of the PES require changes while others do not. Complementary, those changes are expected to have dependencies between them, which here can be considered the 'context'. Then, the morphing of PES is formulated as follows for a triatomic molecule:

$$V_{\mathrm{morphed}}(R, r, \theta) = \mathrm{NN}(V_{\mathrm{surrogate}}(\mathrm{NN}(R, r, \theta))) \tag{8.1}$$

Here, NN is a transformer layer, and $V_{\mathrm{surrogate}}$ is the RKHS model that represents the surface. In principle, the model's parameters will be obtained by minimizing a simple MSE loss function between the values of energy obtained at a high level of theory in a defined grid with the energy obtained with the morphing model. Other probabilities are the use of polynomial transformations for each of the coordinates.

Continuing with the study of PES, the DER model introduced in Chapter 3 was tested on a reactive PES. The aim was to benchmark different uncertainty quantification methods for predicting outliers. To this end, two new formulations of DER were introduced.

Additionally, other quantities, such as the characteristics of the PES judged by the stationary points and harmonic frequencies, the adequate description of the reactive process evaluated with the minimum energy path and the minimum dynamic path, and the energy conservation in NVE simulations were evaluated. Finally, a connection between structural properties such as interatomic distances with the concept of inside-/outside-of-distribution and error and variance was made by introducing a heuristic metric based on the van der Waals radius. The results indicate that ensembles are the best performers for detecting outliers. Nevertheless, the results also show that the reactive process was accurately described for all models despite the fact they did not achieve the desired quality. In this chapter, the statistical limitations of DER were clearly noticed and became a problem for adequately predicting the quantities of interest. For example, forces which are the negative gradient of energy were found to have a large error, which is a consequence of the assumed Gaussian description of the energy. Then, the error in forces can be approximated as $\frac{Error^2_{\mathrm{energy}}}{\sigma^2}$. Alternative development avenues include using different loss functions, recalibration of the uncertainty, use of alternative optimizers to gradient descent and the mentioned conformal prediction models.

The last question treated in this thesis was how to take advantage of the information in conformation space to make better predictions in chemical space. This was done by constructing artificially biased datasets, which were then tried to be improved by adding conformers of minimum examples of a chemical functionality of interest. Methods of enhancement based on normal mode sampling, amons fragments, and the uncertainty/error of samples in the target set were tested. The results give us an idea of the complex interplay between conformational and chemical space. The addition of samples from a single molecule does not greatly impact some of the tested datasets, while others highly benefit from it. However, for those datasets that improve performance, adding samples from conformational space with normal mode sampling seems to return good results when done at 300 K for a small fraction of samples with respect to the total size of the training dataset. On the other hand, the other two methods of addition (amons and error/uncertainty-based addition) give good results only for cases where the target distribution is largely different from the training distribution. Several aspects of this work can be improved, the first is a similar analysis to the one performed in Chapter 3 of the KL divergence of bond distribution to quantify for which bonds larger differences lead to positive or negative improvements in performance. Next, testing adding more diverse examples to the database could be interesting. Lastly, mixing the amon method with normal mode sampling can be a possibility for evaluation.

Other aspects that can be explored in future works are the interpretability of the ML models, the evaluation of the globality of the PES generated by the ML model using either basin hopping[360] or minima hopping[361], the connection of information theory with the construction of the chemical databases, and the automatization of the complete cycle of construction of PES.

Bibliography

[1] V. Kepuska and G. Bohouta, "Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)", in 2018 ieee 8th annual computing and communication workshop and conference (ccwc) (IEEE, 2018), pp. 99–103.

[2] M. B. Hoy, "Alexa, siri, cortana, and more: an introduction to voice assistants", Med. Ref. Serv. Q. **37**, 81–88 (2018).

[3] C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: algorithms, business value, and innovation", ACM Trans. Manag. Inf. Syst. **6**, 1–19 (2015).

[4] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations", in Proceedings of the 10th acm conference on recommender systems (2016), pp. 191–198.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners", Advances in neural information processing systems **33**, 1877–1901 (2020).

[6] J. Manyika and S. Hsiao, "An overview of bard: an early experiment with generative ai", AI. Google Static Documents **2** (2023).

[7] K. Shailaja, B. Seetharamulu, and M. Jabbar, "Machine learning in healthcare: a review", in 2018 second international conference on electronics, communication and aerospace technology (iceca) (IEEE, 2018), pp. 910–914.

[8] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges", Sensors **23**, 4178 (2023).

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis", Med. Image Anal. **42**, 60–88 (2017).

[10] C. Park, C. C. Took, and J.-K. Seong, “Machine learning in biomedical engineering”, Biomed. Eng. Lett. **8**, 1–3 (2018).

[11] D. Guest, K. Cranmer, and D. Whiteson, “Deep learning and its application to lhc physics”, Annu. Rev. Nucl. Part. Sci. **68**, 161–181 (2018).

[12] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, “Machine learning and the physical sciences”, Rev. Mod. Phys. **91**, 045002 (2019).

[13] D. Baron, “Machine learning in astronomy: a practical overview”, arXiv preprint arXiv:1904.07248 (2019).

[14] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, “Deep learning for computational biology”, Mol. Syst. Biol. **12**, 878 (2016).

[15] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, “A guide to machine learning for biologists”, Nat. Rev. Mol. Cell Biol. **23**, 40–55 (2022).

[16] A. Agrawal and A. Choudhary, “Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science”, Apl Mater. **4**, 053208 (2016).

[17] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, “Machine learning unifies the modeling of materials and molecules”, Sci. Adv. **3**, e1701816 (2017).

[18] A. Aspuru-Guzik, R. Lindh, and M. Reiher, “The matter simulation (r) evolution”, ACS Cent. Sci. **4**, 144–152 (2018).

[19] D. Lu, H. Wang, M. Chen, L. Lin, R. Car, E. Weinan, W. Jia, and L. Zhang, “86 pflops deep potential molecular dynamics simulation of 100 million atoms with ab initio accuracy”, Comp. Phys. Comm. **259**, 107624 (2021).

[20] B. Kozinsky, A. Musaelian, A. Johansson, and S. Batzner, “Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size”, in Proceedings of the international conference for high performance computing, networking, storage and analysis, SC ’23 (2023).

[21] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al., “Scientific discovery in the age of artificial intelligence”, Nature **620**, 47–60 (2023).

[22] B. Cheng, G. Mazzola, C. J. Pickard, and M. Ceriotti, “Evidence for supercritical behaviour of high-pressure liquid hydrogen”, Nature **585**, 217–220 (2020).

[23] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, et al., “A deep learning approach to antibiotic discovery”, Cell **180**, 688–702 (2020).

[24] J. A. Pople, “Nobel lecture: quantum chemical models”, Rev. Mod. Phys. **71**, 1267 (1999).

[25] C. Puzzarini, J. Bloino, N. Tasinato, and V. Barone, “Accuracy and interpretability: the devil and the holy grail. new routes across old boundaries in computational spectroscopy”, Chem. Rev. **119**, 8131–8191 (2019).

[26] G. Restrepo, “Chemical space: limits, evolution and modelling of an object bigger than our universal library”, Digit. Discov. **1**, 568–585 (2022).

[27] P. Domingos, “A few useful things to know about machine learning”, Communications of the ACM **55**, 78–87 (2012).

[28] A. Tkatchenko, “Machine learning for chemical discovery”, Nat. Comm. **11**, 4125 (2020).

[29] A. Malinin, S. Chervontsev, I. Provilkov, and M. Gales, “Regression prior networks”, arXiv preprint arXiv:2006.11590 (2020).

[30] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, “Deep evidential regression”, in Advances in neural information processing systems, Vol. 33, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (2020), pp. 14927–14937.

[31] A. P. Soleimany, A. Amini, S. Goldman, D. Rus, S. N. Bhatia, and C. W. Coley, “Evidential deep learning for guided molecular property prediction and discovery”, ACS Cent. Sci. **7**, 1356–1367 (2021).

[32] O. T. Unke and M. Meuwly, “Physnet: a neural network for predicting energies, forces, dipole moments and partial charges”, J. Chem. Theory Comput. **15**, 3678–3693 (2019).

[33] D. J. Wales, “Exploring energy landscapes”, Annu. Rev. Phys. Chem. **69**, 401–425 (2018).

[34] S. Käser, L. I. Vazquez-Salazar, M. Meuwly, and K. Töpfer, “Neural network potentials for chemistry: concepts, applications and prospects”, Digit. Discov. **2**, 28–58 (2023).

[35] O. A. von Lilienfeld and K. Burke, “Retrospective on a decade of machine learning for chemical discovery”, Nat. Comm. **11**, 1–4 (2020).

[36] B. Margulis, K. P. Horn, D. M. Reich, M. Upadhyay, N. Kahn, A. Christianen, A. van der Avoird, G. C. Groenenboom, C. P. Koch, M. Meuwly, and E. Narevicius, “Tomography of feshbach resonance states”, Science **380**, 77–81 (2023).

[37]M. Meuwly and J. M. Hutson, "Morphing ab initio potentials: A systematic study of Ne–HF", J. Chem. Phys. **110**, 8338–8347 (1999).

[38]O. T. Unke and M. Meuwly, "Toolkit for the Construction of Reproducing Kernel-Based Representations of Data: Application to Multidimensional Potential Energy Surfaces", J. Chem. Inf. Model **57**, 1923–1931 (2017).

[39]J. L. Beck and K. M. Zuev, "Rare-event simulation", in *Handbook of uncertainty quantification*, edited by R. Ghanem, D. Higdon, and H. Owhadi (Springer International Publishing, Cham, 2016), pp. 1–26.

[40]B. Huang and O. A. von Lilienfeld, "Quantum machine learning using atom-in-molecule-based fragments selected on the fly", Nat. Chem., 1–7 (2020).

[41]J. M. Ribó, "Chirality: the backbone of chemistry as a natural science", Symmetry **12**, 1982 (2020).

[42]P. Kirkpatrick and C. Ellis, "Chemical space", Nature **432**, 823–824 (2004).

[43]C. W. Coley, "Defining and exploring chemical spaces", Trends Chem. **3**, 133–145 (2021).

[44]E. Whittaker, "Eddington's theory of the constants of nature", Math. Gaz. **29**, 137–144 (1945).

[45]M. M. Vopson, "Estimation of the information contained in the visible matter of the universe", AIP Adv. **11** (2021).

[46]A.-D. Gorse, "Diversity in medicinal chemistry space", Curr Top Med Chem **6**, 3–18 (2006).

[47]T. Fink, H. Bruggesser, and J.-L. Reymond, "Virtual exploration of the small-molecule chemical universe below 160 daltons", Angew. Chem. Int. Ed. **44**, 1504–1508 (2005).

[48]T. Fink and J.-L. Reymond, "Virtual exploration of the chemical universe up to 11 atoms of c, n, o, f: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery", J. Chem. Inf. Model. **47**, 342–353 (2007).

[49]L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17", J. Chem. Inf. Model. **52**, 2864–2875 (2012).

[50]J.-L. Reymond, "The chemical space project", Acc. Chem. Res. **48**, 722–730 (2015).

[51] M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy, and B. Da Mota, "Dataset's chemical diversity limits the generalizability of machine learning predictions", J. Cheminf. **11**, 69 (2019).

[52] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, et al., "Pubchem substance and compound databases", Nucl. Acid. Res. **44**, D1202–D1213 (2016).

[53] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, et al., "Pubchem in 2021: new data content and improved web interfaces", Nucl. Acid. Res. **49**, D1388–D1395 (2021).

[54] P. S. Gromski, A. B. Henson, J. M. Granda, and L. Cronin, "How to explore chemical space using algorithms and automation", Nat. Rev. Chem. **3**, 119–128 (2019).

[55] O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Exploring chemical compound space with quantum-based machine learning", Nat. Rev. Chem. **4**, 347–358 (2020).

[56] B. Huang and O. A. von Lilienfeld, "Ab initio machine learning in chemical compound space", Chem. Rev. **121**, 10001–10036 (2021).

[57] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln", Ann. Phys. **389**, 457–484 (1927).

[58] D. J. Tannor, *Introduction to quantum mechanics: a time-dependent perspective* (University Science Books, 2007).

[59] F. Jensen, *Introduction to computational chemistry* (John wiley & sons, 2017).

[60] E. Lewars, *Computational chemistry: introduction to the theory and applications of molecular and quantum mechanics*, 2nd (Springer, 2011).

[61] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation", Annu. Rev. Phys. Chem. **71**, 361–390 (2020).

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

[63] J. Watt, R. Borhani, and A. K. Katsaggelos, *Machine learning refined: foundations, algorithms, and applications* (Cambridge University Press, 2020).

[64] A. M. Turing, "Computability and λ-definability", J. Symb. Log. **2**, 153–163 (1937).

[65] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.", Psychol. Rev. **65**, 386 (1958).

[66] J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, “Applications of artificial intelligence for chemical inference. i. number of possible organic compounds. acyclic structures containing carbon, hydrogen, oxygen, and nitrogen”, J. Am. Chem. Soc. **91**, 2973–2976 (1969).

[67] J. Gasteiger, “Chemistry in times of artificial intelligence”, ChemPhysChem **21**, 2233–2242 (2020).

[68] J. Zupan and J. Gasteiger, *Neural networks in chemistry and drug design* (John Wiley & Sons, Inc., 1999).

[69] B. G. Sumpter and D. W. Noid, “Potential energy surfaces for macromolecules. a neural network technique”, Chem. Phys. Lett. **192**, 455–462 (1992).

[70] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, “Neural network models of potential energy surfaces”, J. Chem. Phys. **103**, 4129–4137 (1995).

[71] B. B. Goldman and W. P. Walters, “Chapter 8 machine learning in computational chemistry”, in, Vol. 2, edited by D. C. Spellmeyer, Annual Reports in Computational Chemistry (Elsevier, 2006), pp. 127–140.

[72] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, “Unsupervised learning methods for molecular simulation data”, Chem Rev **121**, 9722–9758 (2021).

[73] S. Gow, M. Niranjan, S. Kanza, and J. G. Frey, “A review of reinforcement learning in chemistry”, Digit. Discov. **1**, 551–567 (2022).

[74] J. P. Janet and H. J. Kulik, *Machine learning in chemistry*, Vol. 1 (American Chemical Society, 2020).

[75] D. Schwalbe-Koda and R. Gómez-Bombarelli, “Generative models for automatic chemical design”, in *Machine learning meets quantum physics* (Springer, 2020), pp. 445–467.

[76] D. M. Anstine and O. Isayev, “Generative models as an emerging paradigm in the chemical sciences”, J. Am. Chem. Soc. **145**, 8736–8750 (2023).

[77] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen, “Generative models for molecular discovery: recent advances and challenges”, Wiley Interdiscip. Rev. Comput. Mol. Sci. **12**, e1608 (2022).

[78] D. Fourches, E. Muratov, and A. Tropsha, “Trust, but verify: on the importance of chemical structure curation in cheminformatics and qsar modeling research”, J. Chem. Inf. Model. **50**, 1189 (2010).

[79] J. L. Medina-Franco, K. Martinez-Mayorga, E. Fernández-de Gortari, J. Kirchmair, and J. Bajorath, “Rationality over fashion and hype in drug design”, F1000Res. **10**, 397 (2021).

[80] R. Todeschini and V. Consonni, *Handbook of molecular descriptors* (John Wiley & Sons, 2008).

[81] H. Huo and M. Rupp, “Unified representation of molecules and crystals for machine learning”, Mach. Learn.: Sci. Technol. **3**, 045017 (2022).

[82] M. McGibbon, S. Shave, J. Dong, Y. Gao, D. R. Houston, J. Xie, Y. Yang, P. Schwaller, and V. Blay, “From intuition to ai: evolution of small molecule representations in drug discovery”, Briefings in bio **25**, bbad422 (2024).

[83] L. David, A. Thakkar, R. Mercado, and O. Engkvist, “Molecular representations in ai-driven drug discovery: a review and practical guide”, J. Cheminform. **12**, 1–22 (2020).

[84] B. Settles, *Active learning*, Vol. 6, Synthesis lectures on artificial intelligence and machine learning 1 (Morgan & Claypool Publishers, 2012), pp. 1–114.

[85] S. J. Pan and Q. Yang, “A survey on transfer learning”, IEEE Trans. Knowl. Data Eng. **22**, 1345–1359 (2009).

[86] D. Probst and J.-L. Reymond, “Visualization of very large high-dimensional data sets as minimum spanning trees”, J. Cheminf. **12**, 12 (2020).

[87] O. Wahl and T. Sander, “Tautobase: an open tautomer database”, J. Chem. Inf. Model. **60**, 1085–1089 (2020).

[88] L. I. Vazquez-Salazar, E. D. Boittier, O. T. Unke, and M. Meuwly, “Impact of the characteristics of quantum chemical databases on machine learning prediction of tautomerization energies”, J. Chem. Theory Comput. **17**, 4769–4785 (2021).

[89] W. Pronobis and K.-R. Müller, “Kernel methods for quantum chemistry”, in *Machine learning meets quantum physics* (Springer, 2020), pp. 25–36.

[90] M. Pinheiro Jr and P. O. Dral, “Kernel methods”, in *Quantum chemistry in the age of machine learning* (Elsevier, 2023), pp. 205–232.

[91] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, “Fast and accurate modeling of molecular atomization energies with machine learning”, Phys. Rev. Lett. **108**, 058301 (2012).

[92] B. Huang and O. A. von Lilienfeld, “Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity”, J. Chem. Phys. **145**, 161102 (2016).

[93] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, “Physics-inspired structural representations for molecules and materials”, Chem. Rev. **121**, 9759–9815 (2021).

[94] J. Mercer, “Xvi. functions of positive and negative type, and their connection the theory of integral equations”, Philos. Trans. R. Soc. A **209**, 415–446 (1909).

[95] M. Rupp, “Machine learning for quantum mechanics in a nutshell”, Int. J. Quantum Chem. **115**, 1058–1073 (2015).

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

[97] T.-S. Ho and H. Rabitz, “A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations”, J. Chem. Phys. **104**, 2584 (1996).

[98] G. B. Arfken, H. J. Weber, and F. E. Harris, *Mathematical methods for physicists: a comprehensive guide* (Academic press, 2011).

[99] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, Vol. 2, 3 (MIT press Cambridge, MA, 2006).

[100] D. Koner and M. Meuwly, “Permutationally invariant, reproducing kernel-based potential energy surfaces for polyatomic molecules: from formaldehyde to acetone”, J. Chem. Theory Comput. **16**, 5474–5484 (2020).

[101] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity”, Bull. Math. Biophys. **5**, 115–133 (1943).

[102] P. O. Dral, A. A. Kananenka, F. Ge, and B.-X. Xue, “Neural networks”, in *Quantum chemistry in the age of machine learning*, edited by P. O. Dral (Elsevier, 2023), pp. 183–204.

[103] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning”, Nature **521**, 436–444 (2015).

[104] G. Montavon, “Introduction to neural networks”, in *Machine learning meets quantum physics* (Springer, 2020), pp. 37–62.

[105] S. Prince, *Understanding deep learning* (MIT Press, 2023).

[106] G. Gybenko et al., “Approximation by superposition of sigmoidal functions”, Math. Control Signals Syst. **2**, 303–314 (1989).

[107] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators”, Neural Netw. **2**, 359–366 (1989).

[108] K. Hornik, “Approximation capabilities of multilayer feedforward networks”, Neural Netw. **4**, 251–257 (1991).

[109] R. Eldan and O. Shamir, “The power of depth for feedforward neural networks”, in Conference on learning theory (PMLR, 2016), pp. 907–940.

[110] N. Cohen, O. Sharir, and A. Shashua, “On the expressive power of deep learning: a tensor analysis”, in Conference on learning theory (PMLR, 2016), pp. 698–728.

[111] M. Telgarsky, “Benefits of depth in neural networks”, in Conference on learning theory (PMLR, 2016), pp. 1517–1539.

[112] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, “The expressive power of neural networks: a view from the width”, in Adv. neural. inf. process. syst. Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (2017), pp. 6231–6239.

[113] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

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

[115] C. C. Aggarwal, *Neural networks and deep learning: a textbook* (Springer, 2018).

[116] P. Veličković, “Everything is connected: graph neural networks”, Curr. Op. Struct. Biol. **79**, 102538 (2023).

[117] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry”, in Proceedings of the 34th international conference on machine learning, Vol. 70, edited by D. Precup and Y. W. Teh, Proceedings of Machine Learning Research (2017), pp. 1263–1272.

[118] C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola, and K. F. Jensen, “Convolutional embedding of attributed molecular graphs for physical property prediction”, J. Chem. Inf. Model. **57**, 1757–1772 (2017).

[119] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, “Schnet–a deep learning architecture for molecules and materials”, J. Chem. Phys. **148**, 241722 (2018).

[120] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Message passing neural networks”, in *Machine learning meets quantum physics* (Springer, 2020), pp. 199–214.

[121] A. Zee, *Group theory in a nutshell for physicists*, Vol. 17 (Princeton University Press, 2016).

[122] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, et al., "Graph neural networks for materials science and chemistry", Commun. Mater. **3**, 93 (2022).

[123] A. Duval, S. V. Mathis, C. K. Joshi, V. Schmidt, S. Miret, F. D. Malliaros, T. Cohen, P. Lio, Y. Bengio, and M. Bronstein, "A hitchhiker's guide to geometric gnns for 3d atomic systems", arXiv preprint arXiv:2312.07511 (2023).

[124] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data", IEEE Signal Process. Mag. **34**, 18–42 (2017).

[125] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: grids, groups, graphs, geodesics, and gauges", arXiv preprint arXiv:2104.13478 (2021).

[126] K. Atz, F. Grisoni, and G. Schneider, "Geometric deep learning on molecular representations", Nat. Mach. Intel. **3**, 1023–1032 (2021).

[127] J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs", in International conference on learning representations (2020).

[128] J. Gasteiger, F. Becker, and S. Günnemann, "Gemnet: universal directional graph neural networks for molecules", Advances in Neural Information Processing Systems **34**, 6790–6802 (2021).

[129] K. Schütt, O. Unke, and M. Gastegger, "Equivariant message passing for the prediction of tensorial properties and molecular spectra", in International conference on machine learning (PMLR, 2021), pp. 9377–9388.

[130] V. G. Satorras, E. Hoogeboom, and M. Welling, "$E(n)$ equivariant graph neural networks", in International conference on machine learning (PMLR, 2021), pp. 9323–9332.

[131] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "$E(3)$-equivariant graph neural networks for data-efficient and accurate interatomic potentials", Nat. Comm. **13**, 1–11 (2022).

[132] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, "Mace: higher order equivariant message passing neural networks for fast and accurate force fields", Advances in Neural Information Processing Systems **35**, 11423–11436 (2022).

[133] C. Joshi, "Transformers are graph neural networks", The Gradient **7** (2020).

[134] A. M. Bran and P. Schwaller, "Transformers and large language models for chemistry and drug discovery", arXiv preprint arXiv:2310.06083 (2023).

[135] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", Advances in neural information processing systems **30** (2017).

[136] C. M. Bishop and H. Bishop, *Deep learning: foundations and concepts* (Springer, 2024).

[137] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., "Highly accurate protein structure prediction with alphafold", Nature **596**, 583–589 (2021).

[138] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., "On the opportunities and risks of foundation models", arXiv preprint arXiv:2108.07258 (2021).

[139] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, W. J. Baldwin, N. Bernstein, et al., "A foundation model for atomistic materials chemistry", arXiv preprint arXiv:2401.00096 (2023).

[140] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning", arXiv preprint arXiv:2301.00234 (2022).

[141] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, and B. Smit, "Is gpt-3 all you need for low-data discovery in chemistry?", chemrXiv preprint chemrxiv-2023-fw8n4 (2023).

[142] Z. Xie, X. Evangelopoulos, Ö. H. Omar, A. Troisi, A. I. Cooper, and L. Chen, "Fine-tuning gpt-3 for machine learning electronic and functional properties of organic molecules", Chem. Sci. **15**, 500–510 (2024).

[143] L. I. Vazquez-Salazar, E. D. Boittier, and M. Meuwly, "Uncertainty quantification for predictions of atomistic neural networks", Chem. Sci. **13**, 13068–13084 (2022).

[144] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning", Ann. Data Sci., 1–26 (2020).

[145] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009).

[146] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization", arXiv preprint arXiv:1412.6980 (2014).

[147] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: an imperative style, high-performance deep learning library", in *Advances in neural information processing systems 32* (2019), pp. 8024–8035.

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

[149] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: composable transformations of Python+NumPy programs*, version 0.3.13, 2018.

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

[151] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces", Phys. Rev. Lett. **98**, 146401 (2007).

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

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

[154] 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, "Prediction errors of molecular machine learning models lower than hybrid dft error", J. Chem. Theory Comput. **13**, 5255–5264 (2017).

[155] O. T. Unke and M. Meuwly, "A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information", J. Chem. Phys. **148**, 241708 (2018).

[156] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, and M. Ceriotti, "Accurate molecular polarizabilities with coupled cluster theory and machine learning", Proc. Natl. Acad. Sci. USA **116**, 3401–3406 (2019).

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

[158] O. T. Unke, D. Koner, S. Patra, S. Käser, and M. Meuwly, “High-dimensional potential energy surfaces for molecular simulations: from empiricism to machine learning”, Mach. Learn.: Sci. Technol. **1**, 013001 (2020).

[159] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, “A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer”, Nat. Comm. **12**, 1–11 (2021).

[160] O. A. Von Lilienfeld, “Quantum machine learning in chemical compound space”, Angew. Chem. Int. Ed. **57**, 4164–4169 (2018).

[161] S. Heinen, M. Schwilk, G. F. von Rudorff, and O. A. von Lilienfeld, “Machine learning the computational cost of quantum chemistry”, Mach. Learn.: Sci. Technol. **1**, 025002 (2020).

[162] S. Käser, O. T. Unke, and M. Meuwly, “Reactive dynamics and spectroscopy of hydrogen transfer from neural network-based reactive potential energy surfaces”, New J. Phys. **22**, 055002 (2020).

[163] M. Du, N. Liu, and X. Hu, “Techniques for interpretable machine learning”, Commun. ACM **63**, 68–77 (2019).

[164] W. Samek and K.-R. Müller, “Towards explainable artificial intelligence”, in *Explainable ai: interpreting, explaining and visualizing deep learning* (Springer, 2019), pp. 5–22.

[165] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Definitions, methods, and applications in interpretable machine learning”, Proc. Natl. Acad. Sci. USA **116**, 22071–22080 (2019).

[166] R. Dybowski, “Interpretable machine learning as a tool for scientific discovery in chemistry”, New J Chem **44**, 20914–20920 (2020).

[167] A. Wilkinson and A. McNaught, “Iupac compendium of chemical terminology,(the" gold book")”, International Union of Pure and Applied Chemistry: Zürich, Switzerland (1997).

[168] E. D. Raczyńska, W. Kosińska, B. Ośmiałowski, and R. Gawinecki, “Tautomeric equilibria in relation to pi-electron delocalization”, Chem Rev **105**, 3561–3612 (2005).

[169] Y. C. Martin, “Let’s not forget tautomers”, J. Comput. Aided Mol. Des. **23**, 693–704 (2009).

[170] J. D. Watson and F. H. Crick, “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid”, Nature **171**, 737–738 (1953).

[171] M. K. Shukla and J. Leszczynski, “Tautomerism in nucleic acid bases and base pairs: a brief overview”, Wires Comput. Mol Sci. **3**, 637–649 (2013).

[172] V. Singh, B. I. Fedeles, and J. M. Essigmann, “Role of tautomerism in rna biochemistry”, RNA **21**, 1–13 (2015).

[173] S. Käser, O. T. Unke, and M. Meuwly, “Isomerization and decomposition reactions of acetaldehyde relevant to atmospheric processes from dynamics simulations on neural network-based potential energy surfaces”, J. Chem. Phys. **152**, 214304 (2020).

[174] M. Sitzmann, W.-D. Ihlenfeldt, and M. C. Nicklaus, “Tautomerism in large databases”, J. Comput. Aided Mol. Des. **24**, 521–551 (2010).

[175] J. R. Greenwood, D. Calkins, A. P. Sullivan, and J. C. Shelley, “Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution”, J. Comput. Aided Mol. Des. **24**, 591–604 (2010).

[176] P. Pospisil, P. Ballmer, L. Scapozza, and G. Folkers, “Tautomerism in computer-aided drug design”, J. Recept. Signal Transduct. **23**, 361–371 (2003).

[177] P. J. Taylor, G. van der Zwan, and L. Antonov, “Tautomerism: introduction, history, and recent developments in experimental and theoretical methods”, Tautomerism: methods and theories, 1–24 (2014).

[178] G. Fogarasi, “Studies on tautomerism: benchmark quantum chemical calculations on formamide and formamidine”, J. Mol. Struct **978**, 257–262 (2010).

[179] S. L. Baughcum, R. W. Duerst, W. F. Rowe, Z. Smith, and E. B. Wilson, “Microwave spectroscopic study of malonaldehyde (3-hydroxy-2-prop enal) .2. structure, dipole-moment, and tunneling”, J. Am. Chem. Soc. **103**, 6296–6303 (1981).

[180] D. W. Firth, K. Beyer, M. A. Dvorak, S. W. Reeve, A. Grushow, and K. R. Leopold, “Tunable far infrared spectroscopy of malonaldehyde”, J. Chem. Phys. **94**, 1812–1819 (1991).

[181] H.-H. Limbach, B. Wehrle, H. Zimmermann, R. D. Kendrick, and C. S. Yannoni, “Kinetic 15n-cpmas-nmr study of a double proton transfer in a crystalline malonaldehyde diimine derivative”, Angew. Chem. Int. Ed. **26**, 247–248 (1987).

[182] Y. Wang, B. J. Braams, J. M. Bowman, S. Carter, and D. P. Tew, “Full-dimensional quantum calculations of ground-state tunneling splitting of malonaldehyde using an accurate *ab initio* potential energy surface”, J. Chem. Phys. **128**, 224314 (2008).

[183] M. Schröder, F. Gatti, and H.-D. Meyer, "Theoretical studies of the tunneling splitting of malonaldehyde using the multiconfiguration time-dependent hartree approach", J. Chem. Phys. **134**, 234307 (2011).

[184] Y. Yang and M. Meuwly, "A generalized reactive force field for nonlinear hydrogen bonds: hydrogen dynamics and transfer in malonaldehyde", J. Chem. Phys. **133**, 064503 (2010).

[185] K. Karandashev, Z.-H. Xu, M. Meuwly, J. Vaníček, and J. O. Richardson, "Kinetic isotope effects and how to describe them", Struct. Dyn. **4**, 061501 (2017).

[186] D. K. Dhaked, L. Guasch, and M. C. Nicklaus, "Tautomer database: a comprehensive resource for tautomerism analyses", J. Chem. Inf. Model. **60**, 1090–1100 (2020).

[187] N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network", J. Chem. Phys. **148**, 241715 (2018).

[188] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost", Chem. Sci. **8**, 3192–3203 (2017).

[189] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules", Sci. Data **1**, 140022 (2014).

[190] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules", Sci. Data **4**, 3192–3203 (2017).

[191] J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev, and S. Tretiak, "The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules", Sci. Data **7**, 1–10 (2020).

[192] J.-D. Chai and M. Head-Gordon, "Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections", Phys. Chem. Chem. Phys. **10**, 6615–6620 (2008).

[193] A. D. Becke, "A new mixing of hartree–fock and local density-functional theories", J. Chem. Phys. **98**, 1372–1377 (1993).

[194] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields", J. Phys. Chem. **98**, 11623–11627 (1994).

[195] M. Nakata and T. Shimazaki, "Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry", J. Chem. Inf. Model. **57**, 1300–1308 (2017).

[196] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: sampling chemical space with active learning", J. Chem. Phys. **148**, 241733 (2018).

[197] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: an open chemical toolbox", J. Cheminf. **3**, 33 (2011).

[198] T. A. Halgren, "Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94", J. Comput. Chem. **17**, 490–519 (1996).

[199] M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, and P. J. Taylor, "The sampl2 blind prediction challenge: introduction and overview", J. Comput. Aided Mol. Des. **24**, 259–279 (2010).

[200] N. Foloppe and A. D. MacKerell Jr, "All-atom empirical force field for nucleic acids: i. parameter optimization based on small molecule and condensed phase macromolecular target data", J. Comput. Chem. **21**, 86–104 (2000).

[201] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field", J. Comput. Chem. **25**, 1157–1174 (2004).

[202] W. L. Jorgensen and J. Tirado-Rives, "The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin", J. Am. Chem. Soc. **110**, 1657–1666 (1988).

[203] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, and W. M. Skiff, "Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations", J. Am. Chem. Soc. **114**, 10024–10035 (1992).

[204] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, "Definition and testing of the gromos force-field versions 54a7 and 54b7", Eur. Biophys. J. **40**, 843–856 (2011).

[205] T. Hassinen and M. Peräkylä, "New energy terms for reduced protein models implemented in an off-lattice force field", J. Comput. Chem. **22**, 1229–1242 (2001).

[206] J. J. Stewart, "Optimization of parameters for semiempirical methods vi: more modifications to the nddo approximations and re-optimization of parameters", J. Mol. Model. **19**, 1–32 (2013).

[207] J. J. Stewart, "Mopac2016", Stewart Computational Chemistry: Colorado Springs, CO, USA (2016).

[208] M. J. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. Petersson, et al., "Gaussian 09, revision d. 01, gaussian", Inc.: Wallingford, CT (2009).

[209] U. Diwekar and A. David, “Probability density functions and kernel density estimation”, in *Bonus algorithm for large scale stochastic nonlinear programming problems* (Springer, 2015), pp. 27–34.

[210] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley Series in Telecommunications and Signal Processing (John Wiley & Sons, 2006).

[211] *Aqml: amons-based quantum machine learning for quantum chemistry*, https://github.com/binghuang2018/aqml, Accessed: 2020-10-15.

[212] V. T. Lim, D. F. Hahn, G. Tresadern, C. I. Bayly, and D. L. Mobley, “Benchmark assessment of molecular geometries and energies from small molecule force fields”, F1000Res. **9**, 1390 (2020).

[213] Z. C. Lipton, “The mythos of model interpretability.”, ACM Queue **16**, 31–57 (2018).

[214] K. T. Schütt, M. Gastegger, A. Tkatchenko, and K.-R. Müller, “Quantum-chemical insights from interpretable atomistic neural networks”, in *Explainable ai: interpreting, explaining and visualizing deep learning* (Springer, 2019), pp. 311–330.

[215] S. Käser, E. D. Boittier, M. Upadhyay, and M. Meuwly, “Transfer learning to ccsd (t): accurate anharmonic frequencies from machine learning models”, J. Chem. Theory Comput. **17**, 3687–3699 (2021).

[216] M. Wieder, J. Fass, and J. D. Chodera, “Fitting quantum machine learning potentials to experimental free energy data: predicting tautomer ratios in solution”, bioRxiv (2020).

[217] M. Meuwly, “Machine learning for chemical reactions”, Chem. Rev. **121**, 10218–10239 (2021).

[218] K. Töpfer, S. Käser, and M. Meuwly, “Double proton transfer in hydrated formic acid dimer: interplay of spatial symmetry and solvent-generated force on reactivity”, Phys. Chem. Chem. Phys. **24**, 13869–13882 (2022).

[219] F. Noé, S. Olsson, J. Köhler, and H. Wu, “Boltzmann generators: sampling equilibrium states of many-body systems with deep learning”, Science **365**, eaaw1147 (2019).

[220] S. Manzhos and T. Carrington Jr, “Neural network potential energy surfaces for small molecules and reactions”, Chem. Rev. **121**, 10187–10217 (2020).

[221] R. Conte, C. Qu, P. L. Houston, and J. M. Bowman, “Efficient generation of permutationally invariant potential energy surfaces for large molecules”, J. Chem. Theory Comput. **16**, 3264–3272 (2020).

[222] O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, D. Ahlin, M. Gastegger, L. M. Sandonas, A. Tkatchenko, et al., "Accurate machine learned quantum-mechanical force fields for biomolecular simulations", arXiv preprint arXiv:2205.08306 (2022).

[223] P. Ramos-Sánchez, J. N. Harvey, and J. A. Gámez, "An automated method for graph-based chemical space exploration and transition state finding", J. Comput. Chem. **44**, 27–42 (2023).

[224] X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith, and A. E. Roitberg, "Torchani: a free and open source pytorch-based deep learning implementation of the ani neural network potentials", J. Chem. Inf. Model. **60**, 3408–3415 (2020).

[225] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K.-R. Müller, "Spookynet: learning force fields with electronic degrees of freedom and nonlocal effects", Nat. Comm. **12**, 1–14 (2021).

[226] H. Sanders and J. Saxe, "Garbage in, garbage out: how purportedly great ml models can be screwed up by bad data", Proceedings of Blackhat **2017** (2017).

[227] M. F. Kilkenny and K. M. Robinson, "Data quality:"garbage in–garbage out"", Health Inf. Manag. J. **47**, 103–105 (2018).

[228] G. Canbek, "Gaining insights in datasets in the shade of "garbage in, garbage out" rationale: feature space distribution fitting", Wiley Interdiscip. Rev. Data Min. Knowl. Discov **12**, e1456 (2022).

[229] R. L. Tweedie, K. L. Mengersen, and J. A. Eccleston, "Garbage in, garbage out: can statisticians quantify the effects of poor data?", Chance **7**, 20–27 (1994).

[230] C. Babbage, *Passages from the life of a philosopher*, Cambridge Library Collection - Technology (Cambridge University Press, 2011).

[231] R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang, ""garbage in, garbage out" revisited: what do machine learning application papers report about human-labeled training data?", Quant. sci. stud. **2**, 795–827 (2021).

[232] J. C. Weyerer and P. F. Langer, "Garbage in, garbage out: the vicious cycle of ai-based discrimination in the public sector", in Proceedings of the 20th annual international conference on digital government research (2019), pp. 509–511.

[233] B. Saha and D. Srivastava, "Data quality: the other face of big data", in 2014 ieee 30th international conference on data engineering (IEEE, 2014), pp. 1294–1297.

[234] F. Iafrate, "A journey from big data to smart data", in *Digital enterprise design & management* (Springer, 2014), pp. 25–33.

[235] M. T. Baldassarre, I. Caballero, D. Caivano, B. Rivas Garcia, and M. Piattini, "From big data to smart data: a data quality perspective", in Proceedings of the 1st acm sigsoft international workshop on ensemble-based software engineering (2018), pp. 19–24.

[236] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, "Transforming big data into smart data: an insight on the use of the k-nearest neighbors algorithm to obtain quality data", Wiley Interdiscip. Rev. Data Min. Knowl. Discov **9**, e1289 (2019).

[237] S. Käser, D. Koner, A. S. Christensen, O. A. von Lilienfeld, and M. Meuwly, "Machine learning models of vibrating h2co: comparing reproducing kernels, fchl, and physnet", J. Phys. Chem. A **124**, 8853–8865 (2020).

[238] S. Käser, J. O. Richardson, and M. Meuwly, "Transfer learning for affordable and high-quality tunneling splittings from instanton calculations", J. Chem. Theory Comput. **18**, 6840–6850 (2022).

[239] J. P. Janet, C. Duan, T. Yang, A. Nandy, and H. J. Kulik, "A quantitative uncertainty metric controls error in neural network-driven chemical discovery", Chem. Sci. **10**, 7913–7922 (2019).

[240] P. Zheng, W. Yang, W. Wu, O. Isayev, and P. O. Dral, "Toward chemical accuracy in predicting enthalpies of formation with general-purpose data-driven methods", J. Phys. Chem. Lett. **13**, 3479–3491 (2022).

[241] F. Musil, M. J. Willatt, M. A. Langovoy, and M. Ceriotti, "Fast and accurate uncertainty estimation in chemical machine learning", J. Chem. Theory Comput. **15**, 906–915 (2019).

[242] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules", Chem Rev **121**, 10073–10141 (2021).

[243] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., "A survey of uncertainty in deep neural networks", Artif. Intell. Rev. **56**, 1513–1589 (2023).

[244] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al., "A review of uncertainty quantification in deep learning: techniques, applications and challenges", Inf Fusion. **76**, 243–297 (2021).

[245] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, and C. W. Coley, “Uncertainty quantification using neural networks for molecular property prediction”, J. Chem. Inf. Model. **60**, 3770–3780 (2020).

[246] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, “Evaluating and calibrating uncertainty prediction in regression tasks”, arXiv preprint arXiv:1905.11659 (2019).

[247] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, and Z. W. Ulissi, “Methods for comparing uncertainty quantifications for material property predictions”, Mach. Learn.: Sci. Technol. **1**, 025006 (2020).

[248] J. Busk, P. B. Jørgensen, A. Bhowmik, M. N. Schmidt, O. Winther, and T. Vegge, “Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks”, Mach. Learn.: Sci. Technol. **3**, 015012 (2021).

[249] V. Kuleshov, N. Fenner, and S. Ermon, “Accurate uncertainties for deep learning using calibrated regression”, in International conference on machine learning (PMLR, 2018), pp. 2796–2804.

[250] Y. Chung, I. Char, H. Guo, J. Schneider, and W. Neiswanger, “Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification”, arXiv preprint arXiv:2109.10254 (2021).

[251] P. Pernot, “The long road to calibrated prediction uncertainty in computational chemistry”, J. Chem. Phys. **156**, 114109 (2022).

[252] L. Kahle and F. Zipoli, “Quality of uncertainty estimates from neural network potential ensembles”, Phys. Rev. E **105**, 015311 (2022).

[253] K. Cheng, F. Calivá, R. Shah, M. Han, S. Majumdar, and V. Pedoia, “Addressing the false negative problem of deep learning mri reconstruction models by adversarial attacks and robust training”, in Medical imaging with deep learning (PMLR, 2020), pp. 121–135.

[254] M. J. Schervish and M. H. DeGroot, *Probability and statistics* (Pearson Education London, UK: 2014).

[255] L. I. Vazquez-Salazar and M. Meuwly, *Qtautobase: a quantum tautomerization database*, version 1.0, Apr. 2021.

[256] B. Ruscic, “Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and active thermochemical tables”, Int. J. Quantum Chem. **114**, 1097–1101 (2014).

[257] G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li, and W. H. Green, “Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction”, J. Chem. Inf. Model. **60**, 2697–2717 (2020).

[258]K. Schutt, P. Kessel, M. Gastegger, K. Nicoli, A. Tkatchenko, and K.-R. Müller, “Schnetpack: a deep learning toolbox for atomistic systems”, J. Chem. Theory Comput. **15**, 448–455 (2018).

[259]G. Palmer, S. Du, A. Politowicz, J. P. Emory, X. Yang, A. Gautam, G. Gupta, Z. Li, R. Jacobs, and D. Morgan, “Calibration after bootstrap for accurate uncertainty quantification in regression models”, Npj Comput. Mater. **8**, 1–9 (2022).

[260]P.-A. Cazade, W. Zheng, D. Prada-Gracia, G. Berezovska, F. Rao, C. Clementi, and M. Meuwly, “A comparative analysis of clustering algorithms: o2 migration in truncated hemoglobin i from transition networks”, J. Chem. Phys. **142**, 01B610_1 (2015).

[261]M. Ceriotti, “Unsupervised machine learning in atomistic simulations, between predictions and understanding”, J. Chem. Phys. **150**, 150901 (2019).

[262]G. Fonseca, I. Poltavsky, V. Vassilev-Galindo, and A. Tkatchenko, “Improving molecular force fields across configurational space by combining supervised and unsupervised machine learning”, J. Chem. Phys. **154**, 124102 (2021).

[263]A. V. Joshi, “Essential concepts in artificial intelligence and machine learning”, in *Machine learning and artificial intelligence* (Springer International Publishing, Cham, 2020), pp. 9–20.

[264]M. Naser and A. H. Alavi, “Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences”, Archit. Struct. and Const., 1–19 (2021).

[265]H. S. Seung, M. Opper, and H. Sompolinsky, “Query by committee”, in Proceedings of the fifth annual workshop on computational learning theory (1992), pp. 287–294.

[266]E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods”, Mach. Learn. **110**, 457–506 (2021).

[267]K. P. Murphy, *Machine learning: a probabilistic perspective* (MIT press, 2012).

[268]L. Tran, B. S. Veeling, K. Roth, J. Swiatkowski, J. V. Dillon, J. Snoek, S. Mandt, T. Salimans, S. Nowozin, and R. Jenatton, “Hydra: preserving ensemble diversity for model distillation”, arXiv preprint arXiv:2001.04694 (2020).

[269]A. Malinin, B. Mlodozeniec, and M. Gales, “Ensemble distribution distillation”, arXiv preprint arXiv:1905.00076 (2019).

[270]N. Meinert and A. Lavin, “Multivariate deep evidential regression”, arXiv preprint arXiv:2104.06135 (2021).

[271] N. Meinert, J. Gawlikowski, and A. Lavin, “The unreasonable effectiveness of deep evidential regression”, in Proceedings of the aaai conference on artificial intelligence, Vol. 37, 8 (2023), pp. 9134–9142.

[272] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks”, in Proceedings of the 34th international conference on machine learning, Vol. 70, edited by D. Precup and Y. W. Teh, Proceedings of Machine Learning Research (2017), pp. 1321–1330.

[273] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, “On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks”, in International conference on learning representations (Apr. 2022).

[274] D. Megerle, F. Otto, M. Volpp, and G. Neumann, *Stable optimization of gaussian likelihoods*, 2023.

[275] J. Cui, Z. Tian, Z. Zhong, X. Qi, B. Yu, and H. Zhang, “Decoupled kullback-leibler divergence loss”, arXiv preprint arXiv:2305.13948 (2023).

[276] J. Lin, “Divergence measures based on the shannon entropy”, IEEE Trans. Inf. Theory **37**, 145–151 (1991).

[277] F. Nielsen, “On the jensen–shannon symmetrization of distances relying on abstract means”, Entropy **21**, 485 (2019).

[278] T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun, “Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation”, arXiv preprint arXiv:2105.08919 (2021).

[279] J. Liu, E. J. Salumbides, U. Hollenstein, J. C. Koelemeij, K. S. Eikema, W. Ubachs, and F. Merkt, “Determination of the ionization and dissociation energies of the hydrogen molecule”, J. Chem. Phys. **130**, 174306 (2009).

[280] M. Child and D. Nesbitt, “RKR-based inversion of rotational progressions”, Chem. Phys. Lett. **149**, 404–410 (1988).

[281] D. J. Nesbitt, M. S. Child, and D. C. Clary, “Rydberg–Klein–Rees inversion of high resolution van der Waals infrared spectra: An intermolecular potential energy surface for Ar+ HF($\nu = 1$)”, J. Chem. Phys. **90**, 4855–4864 (1989).

[282] D. J. Nesbitt and M. S. Child, “Rotational-RKR inversion of intermolecular stretching potentials: Extension to linear hydrogen bonded complexes”, J. Chem. Phys. **98**, 478–486 (1993).

[283] L. Kurtz, H. Rabitz, and R. de Vivie-Riedle, “Optimal use of time-dependent probability density data to extract potential-energy surfaces”, Phys. Rev. A **65**, 032514 (2002).

[284] J. M. Bowman and B. Gazdy, “A simple method to adjust potential energy surfaces: Application to HCO”, J. Chem. Phys. **94**, 816–817 (1991).

[285] B. Gazdy and J. M. Bowman, “An adjusted global potential surface for HCN based on rigorous vibrational calculations”, J. Chem. Phys. **95**, 6309–6316 (1991).

[286] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, “Big data meets quantum chemistry approximations: the Δ-machine learning approach”, J. Chem. Theory Comput. **11**, 2087–2096 (2015).

[287] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, “Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning”, Nat. Commun. **10**, 1–8 (2019).

[288] S. Käser and M. Meuwly, “Transfer-learned potential energy surfaces: Toward microsecond-scale molecular dynamics simulations in the gas phase at CCSD (T) quality”, J. Chem. Phys. **158**, 214301 (2023).

[289] S. Thaler and J. Zavadlav, “Learning neural network potentials from experimental data via differentiable trajectory reweighting”, Nat. Comm. **12**, 6884 (2021).

[290] S. N. Yurchenko, L. Lodi, J. Tennyson, and A. V. Stolyarov, “Duo: a general program for calculating spectra of diatomic molecules”, Comp. Phys. Comm. **202**, 262–275 (2016).

[291] K. T. Lorenz, M. S. Westley, and D. W. Chandler, “Rotational state-to-state differential cross sections for the HCl–Ar collision system using velocity-mapped ion imaging”, Phys. Chem. Chem. Phys. **2**, 481–494 (2000).

[292] R. Vargas-Hernández, Y. Guan, D. Zhang, and R. Krems, “Bayesian optimization for the inverse scattering problem in quantum reaction dynamics”, New J. Phys. **21**, 022001 (2019).

[293] T. van Mourik, G. J. Harris, O. L. Polyansky, J. Tennyson, A. G. Császár, and P. J. Knowles, “Ab initio global potential, dipole, adiabatic, and relativistic correction surfaces for the HCN–HNC system”, J. Chem. Phys. **115**, 3706–3718 (2001).

[294] D. Koner, J. C. S. V. Veliz, A. van der Avoird, and M. Meuwly, “Near dissociation states for H_2^+–He on MRCI and FCI potential energy surfaces”, Phys. Chem. Chem. Phys. **21**, 24976–24983 (2019).

[295] J. M. Hutson, “Intermolecular forces from the spectroscopy of van der waals molecules”, Annu. Rev. Phys. Chem. **41**, 123–154 (1990).

[296] S. Adhikari, M. Baer, and N. Sathyamurthy, “HeH_2^+: structure and dynamics”, Intern. Rev. Phys. Chem. **41**, 49–93 (2022).

[297] A. Carrington, D. I. Gammie, A. M. Shaw, S. M. Taylor, and J. M. Hutson, “Observation of a microwave spectrum of the long-range He . . . H_2^+ complex”, Chem. Phys. Lett. **260**, 395–405 (1996).

[298] D. I. Gammie, J. C. Page, and A. M. Shaw, “Microwave and millimeter-wave spectrum of the He$\cdots$$H_2^+$ long-range complex”, J. Chem. Phys. **116**, 6072 (2002).

[299] O. Asvany, S. Schlemmer, A. van der Avoird, T. Szidarovszky, and A. G. Császár, “Vibrational spectroscopy of H_2He^+ and D_2He^+”, J. Mass Spectrom. **377**, 111423 (2021).

[300] D. Kedziera, G. Rauhut, and A. G. Császár, “Structure, energetics, and spectroscopy of the chromophores of HHe_n^+, $H_2He_n^+$, and He_n^+ clusters and their deuterated isotopologues”, Phys. Chem. Chem. Phys. **24**, 12176–12195 (2022).

[301] C. Chin, R. Grimm, P. Julienne, and E. Tiesinga, “Feshbach resonances in ultracold gases”, Rev. Mod. Phys. **82**, 1225 (2010).

[302] J. Pérez Ríos, *Introduction to cold and ultracold chemistry: atoms, molecules, ions and rydbergs* (Springer, 2020).

[303] R. J. Le Roy and J. M. Hutson, “Improved potential energy surfaces for the interaction of H2 with Ar, Kr, and Xe”, J. Chem. Phys. **86**, 837–853 (1987).

[304] S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, “Accurate global machine learning force fields for molecules with hundreds of atoms”, Sci. Adv. **9**, eadf0873 (2023).

[305] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information* (Cambridge University Press, 2010).

[306] P. J. Ollitrault, A. Miessen, and I. Tavernelli, “Molecular quantum dynamics: a quantum computing perspective”, Acc. Chem. Res. **54**, 4229–4238 (2021).

[307] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, et al., *Molpro, version 2019.2, a package of ab initio programs*, 2019.

[308] B. R. Johnson, “The renormalized numerov method applied to calculating bound states of the coupled-channel schroedinger equation”, J. Chem. Phys. **69**, 4678–4688 (1978).

[309] F. X. Gadéa, H. Berriche, O. Roncero, P. Villarreal, and G. D. Barrio, “Nonradiative lifetimes for LiH in the A state using adiabatic and diabatic schemes”, J. Chem. Phys. **107**, 10515–10522 (1997).

[310] M. Pawlak, Y. Shagam, A. Klein, E. Narevicius, and N. Moiseyev, “Adiabatic variational theory for cold atom–molecule collisions: application to a metastable helium atom colliding with ortho-and para-hydrogen molecules”, J. Phys. Chem. A **121**, 2194–2198 (2017).

[311] B. Margulis, J. Narevicius, and E. Narevicius, “Direct observation of a feshbach resonance by coincidence detection of ions and electrons in penning ionization collisions”, Nat. Comm. **11**, 3553 (2020).

[312] S. G. Johnson, *The nlopt nonlinear-optimization package*, http://github.com/stevengj/nlopt, Accessed: 2021-10-15.

[313] W. F. Van Gunsteren and H. J. Berendsen, “Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry”, Angew. Chem., Int. Ed. Engl. **29**, 992–1023 (1990).

[314] M. Ferrario, G. Ciccotti, and K. Binder, *Computer simulations in condensed matter: from materials to chemical biology*, Vol. 1 (Springer Science & Business Media, 2006).

[315] Q. Cui, “Perspective: quantum mechanical methods in biochemistry and biophysics”, J. Chem. Phys. **145**, 140901 (2016).

[316] D. Raabe, *Computational materials science: the simulation of materials microstructures and properties* (Wiley-vch, 1998).

[317] J. Behler, “Four generations of high-dimensional neural network potentials”, Chem. Rev. **121**, 10037–10072 (2021).

[318] Y. Wang, B. J. Braams, J. M. Bowman, S. Carter, and D. P. Tew, “Full-dimensional quantum calculations of ground-state tunneling splitting of malonaldehyde using an accurate ab initio potential energy surface”, J. Chem. Phys. **128** (2008).

[319] D. P. Tew and W. Mizukami, “Ab initio vibrational spectroscopy of cis-and trans-formic acid from a global potential energy surface”, J. Phys. Chem. A **120**, 9815–9828 (2016).

[320] N. M. Kidwell, H. Li, X. Wang, J. M. Bowman, and M. I. Lester, “Unimolecular dissociation dynamics of vibrationally activated ch_3choo criegee intermediates to oh radical products”, Nat. Chem. **8**, 509–514 (2016).

[321] J. Li, Z. Varga, D. G. Truhlar, and H. Guo, “Many-body permutationally invariant polynomial neural network potential energy surface for n_4”, J. Chem. Theory Comput. **16**, 4822–4832 (2020).

[322] M. O. Alves, V. C. Mota, J. P. Braga, A. J. Varandas, H. Guo, and B. R. Galvão, "High-accuracy dmbe potential energy surface for cno(a"4) and the rate coefficients for the c+ no reaction in the ^{2}a', ^{2}a", and ^{4}a" states", J. Chem. Phys. **159** (2023).

[323] K. P. Horn, L. I. Vazquez-Salazar, C. P. Koch, and M. Meuwly, "Improving potential energy surfaces using experimental feshbach resonance tomography", arXiv preprint arXiv:2309.16491 (2023).

[324] P. J. Haley and D. Soloway, "Extrapolation limitations of multilayer feedforward neural networks", in [proceedings 1992] ijcnn international joint conference on neural networks, Vol. 4 (IEEE, 1992), pp. 25–30.

[325] J. Behler, "Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations", Phys. Chem. Chem. Phys. **13**, 17930–17955 (2011).

[326] A. M. Tokita and J. Behler, "How to train a neural network potential", J. Chem. Phys. **159** (2023).

[327] A. R. Tan, S. Urata, S. Goldman, J. C. Dietschreit, and R. Gómez-Bombarelli, "Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles", Npj Comput. Mater. **9**, 225 (2023).

[328] M. Upadhyay and M. Meuwly, "Thermal and vibrationally activated decomposition of the *syn*-ch$_3$choo criegee intermediate", ACS Earth Space Chem. **5**, 3396–3406 (2021).

[329] I. Kosztin, B. Faber, and K. Schulten, "Introduction to the diffusion monte carlo method", Am. J. Phys. **64**, 633–644 (1996).

[330] G. Csányi, T. Albaret, M. Payne, and A. De Vita, ""learn on the fly": a hybrid classical and quantum-mechanical molecular dynamics simulation", Phys. Rev. Lett. **93**, 175503 (2004).

[331] M. Gastegger and P. Marquetand, "Molecular dynamics with neural network potentials", in *Machine learning meets quantum physics* (Springer, 2020), pp. 233–252.

[332] J. Klicpera, S. Giri, J. T. Margraf, and S. Günnemann, "Fast and uncertainty-aware directional message passing for non-equilibrium molecules", arXiv preprint arXiv:2011.14115 (2020).

[333] D. Oh and B. Shin, "Improving evidential deep learning via multi-task learning", in Proceedings of the aaai conference on artificial intelligence, Vol. 36, 7 (2022), pp. 7895–7903.

[334] A. Zhu, S. Batzner, A. Musaelian, and B. Kozinsky, "Fast uncertainty estimates in deep learning interatomic potentials", J. Chem. Phys. **158** (2023).

[335] V. Vapnik, *The nature of statistical learning theory* (Springer science & business media, 1999).

[336] S. Farquhar and Y. Gal, "What'out-of-distribution'is and is not", in Neurips ml safety workshop (2022).

[337] M. Mantina, A. C. Chamberlin, R. Valero, C. J. Cramer, and D. G. Truhlar, "Consistent van der waals radii for the whole main group", J. Phys. Chem. A **113**, 5806–5812 (2009).

[338] S. Käser and M. Meuwly, "Numerical accuracy matters: applications of machine learned potential energy surfaces", arXiv preprint arXiv:2311.17398 (2023).

[339] S. Goswami, S. Käser, R. J. Bemish, and M. Meuwly, "Effects of aleatoric and epistemic errors in reference data on the learnability and quality of nn-based potential energy surfaces", Art. Intel. Chem. **2**, 100033 (2024).

[340] O. T. Unke, S. Brickel, and M. Meuwly, "Sampling reactive regions in phase space by following the minimum dynamic path", J. Chem. Phys. **150** (2019).

[341] M. Upadhyay, K. Topfer, and M. Meuwly, "Molecular simulation for atmospheric reactions: non-equilibrium dynamics, roaming, and glycolaldehyde formation following photoinduced decomposition of syn-acetaldehyde oxide", J. Phys. Chem. Lett. **15**, 90–96 (2023).

[342] R. Dawes, B. Jiang, and H. Guo, "Uv absorption spectrum and photodissociation channels of the simplest criegee intermediate (ch_2oo)", J. Am. Chem. Soc. **137**, 50–53 (2015).

[343] T. J. Lee and P. R. Taylor, "A diagnostic for determining the quality of single-reference electron correlation methods", Int. J. Quantum Chem. **36**, 199–207 (1989).

[344] C. L. Janssen and I. M. Nielsen, "New diagnostics for coupled-cluster and møller–plesset perturbation theory", Chem. Phys. Lett. **290**, 423–430 (1998).

[345] S. Käser, O. T. Unke, and M. Meuwly, "Isomerization and decomposition reactions of acetaldehyde relevant to atmospheric processes from dynamics simulations on neural network-based potential energy surfaces", J. Chem. Phys. **152** (2020).

[346] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification", arXiv preprint arXiv:2107.07511 (2021).

[347] Y. Hu, J. Musielewicz, Z. W. Ulissi, and A. J. Medford, "Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials", Mach. Learn.: Sci. Technol. **3**, 045028 (2022).

[348] L. M. Sandonas, J. Hoja, B. G. Ernst, Á. Vázquez-Mayagoitia, R. A. DiStasio, and A. Tkatchenko, "“freedom of design” in chemical compound space: towards rational in silico design of molecules with targeted quantum-mechanical properties", Chem. Sci. **14**, 10702–10717 (2023).

[349] A. Fallani, L. Medrano Sandonas, and A. Tkatchenko, "Enabling inverse design in chemical compound space: mapping quantum properties to structures for small organic molecules", arXiv e-prints, arXiv–2309 (2023).

[350] S. Shaik, H. S. Rzepa, and R. Hoffmann, "One molecule, two atoms, three views, four bonds?", Angew. Chem. Int. Ed. **52**, 3020–3033 (2013).

[351] J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr, and A. Tkatchenko, "Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules", Sci. Data **8**, 43 (2021).

[352] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning* (Mit Press, 2008).

[353] P. Banerjee, F. O. Dehnbostel, and R. Preissner, "Prediction is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets", Front. Chem., 362 (2018).

[354] J. Hemmerich, E. Asilar, and G. F. Ecker, "Cover: conformational oversampling as data augmentation for molecules", J. Cheminf. **12**, 18 (2020).

[355] S. Korkmaz, "Deep learning-based imbalanced data classification for drug discovery", J. Chem. Inf. Model. **60**, 4180–4190 (2020).

[356] N. Shenoy, P. Tossou, E. Noutahi, H. Mary, D. Beaini, and J. Ding, "Role of structural and conformational diversity for machine learning potentials", in Neurips 2023 ai for science workshop (2023).

[357] G. Landrum et al., *Rdkit: a software suite for cheminformatics, computational chemistry, and predictive modeling*, 2013.

[358] M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, G. Petersson, H. Nakatsuji, et al., *Gaussian 16*, 2016.

[359] J. L. A. Gardner, Z. Faure Beaulieu, and V. L. Deringer, "Synthetic data enable experiments in atomistic machine learning", Digit. Discov. **2**, 651–662 (2023).

[360] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms", J. Phys. Chem. A **101**, 5111–5116 (1997).

[361] M. Krummenacher, M. Gubler, J. A. Finkler, H. Huber, M. Sommer-Jörgensen, and S. Goedecker, “Performing highly efficient minima hopping structure predictions using the atomic simulation environment (ase)”, SoftwareX **25**, 101632 (2024).

Luis Itza **Vazquez-Salazar**

PhD Student, MSc Chemistry

Johanniterstrasse 13, 4056 Basel, Basel-Stadt, Switzerland

☎ (+31) 640788878 | ✉ litzavazquezs@gmail.com | ✉ luisitza.vazquezsalazar@unibas.ch | ⌂ livazquezs.github.io | LIVazquezS | Luis Itza Vazquez-Salazar | 🎓 L.I.Vazquez-Salazar

Education

Faculty of Science, University of Basel (UniBas) *Basel, Switzerland*

PhD Chemistry *2019 - Present*

- **Thesis**: Inferring Chemistry from Data with Atomistic Machine Learning: Applications to Potential Energy Surfaces and Chemical Space
- **Supervisor:** Prof. Dr. Markus Meuwly
- **Second Supervisor:** Prof. Dr. Markus Lill
- **External Expert:** Prof. Dr. Alexandre Tkatchenko, University of Luxembourg.
- **Courses included**: Computational Quantum Mechanics Based Design of Matter, Modern Aspects of Bayesian Statistics, Molecular Modeling in Drug Design, A Practical Introduction to Data Science, Artificial Intelligence in Drug Discovery.

Faculty of Science and Engineering, University of Groningen (RUG) *Groningen, The Netherlands*

MSc. Chemistry *2017 - 2019*

- European Master in Theoretical Chemistry and Computational Modelling.
- **Courses included**: Advanced Quantum Mechanics, Mathematical Methods for Physicists, Molecular Quantum Mechanics, Classical Molecular Dynamics, Quantum Molecular Dynamics, Computational Quantum Chemistry.
- **Thesis**: MARTINI model of imidazolium-based Ionic liquids from nanostructure to liquid-liquid extractions. *Supervisors: Prof. Dr. Siewert-Jan Marrink and Prof.Dr. Shirin Faraji*

Faculty of Chemistry, National Autonomous University of Mexico (UNAM) *Mexico City, Mexico*

BSc (Hons) Chemistry *2011 - 2015*

- **Courses included**: Quantum Chemistry, Computational Chemistry, Molecular Modelling and Simulation, Molecular Symmetry, Statistical Mechanics, Chemical Kinetics, Numerical Methods and Supramolecular Chemistry
- **Thesis**: Quantum tunnelling description on chemical kinetics using analytical models. *Supervisor: Dr. Emilio Orgaz*

Work Experience

Faculty of Chemistry, UNAM *Mexico City, Mexico*

Lecturer *2017*

- I taught the following two courses:
  - **Fundamentals of Spectroscopy**: Where I conducted basic experiments on vibrations and waves.
  - **Structure of Matter**: A fully theoretical course introducing quantum mechanics to chemistry and engineering students.

Geophysics Institute, UNAM *Mexico City, Mexico*

Research Assistant *2015-2016*

- I worked as part of *Professor Emeritus Ismael Herrera* research group in the field of mathematical and computational modelling of multiscale phenomena applied to fluid dynamics. My work involved the development of mathematical models for multiphysics systems and the definition of boundary conditions to Quantum Mechanics/Molecular Mechanics techniques.

Faculty of Chemistry, UNAM *Mexico City, Mexico*

Teaching Assistant *2013-2016*

- I worked with Dr Alejandro Pisanty as an assistant in the Fundamentals of Spectroscopy course. My principal duty was the development of pedagogical strategies to obtain a better understanding of the topics reviewed in the course through the use of computational tools. Also, I collaborated in the design and evaluation of exams and homework.

Research Experience

Plant tissue culture Laboratory, Biology Institute, UNAM *Mexico City, Mexico*

Research Intern *2010*

Environmental engineering Laboratory, Engineering Institute, UNAM — *Mexico City, Mexico*
Research Intern — *2010 - 2011*

Department of Physics and Theoretical Chemistry, Faculty of Chemistry, UNAM — *Mexico City, Mexico*
Research Intern — *2013- 2016*

Molecular Dynamics Group, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen — *Groningen, The Netherlands*
Research Intern — *2017 - 2019*

Theoretical and Computational Chemistry Group, Department of Chemistry, Faculty of Science, KU Leuven — *Leuven, Belgium*
Research Intern — *May-July 2019*

Department of Chemistry, University of Cambridge — *Cambridge, United Kingdom*
Visiting PhD Student — *March 2023*

Honors & Awards

International

2017	**Scholarship**, Erasmus Mundus Scholarship to make postgraduate studies at University of Groningen	*The Netherlands*
2018	**Best Poster Award**, IUBMB/ IUPAB Advanced School and Workshop: "Protein-Protein and Protein-Membrane interaction: Experimental and Theoretical Approaches"	*Havana, Cuba*
2023	**Travel Grant**, Institute of Pure and Applied Mathematics, University of California-Los Angeles workshop on Learning and Emergence in Molecular Systems	*Los Angeles, CA, United States*
2023	**Participant**, 5th Merck Synthetic Challenge	*Berlin, Germany*
2024	**Fellowship**, Early Postdoc Mobility, Swiss National Science Foundation	*Bern, Switzerland*

Domestic

2010	**1st Place**, XVlll University Competition Fair of Sciences with the project "In vitro grow of Furcraea macdougalli, extinct specie in its natural area"	*Mexico City, Mexico*
2011	**2nd Place**, XII Mexican Young Competition of Water with the work " Production of safe water for human use through the use of nanofiltration membranes"	*Mexico City, Mexico*
2012	**1st place**, Experimental Exhibition of Physics, Faculty of Chemistry, UNAM with the work "Estimation of electron's charge"	*Mexico City, Mexico*
2012	**Honorific Mention**, Experimental Exhibition of Physics, Faculty of Chemistry, UNAM with the work "The CD as a diffraction net: an application of Bragg's Law"	*Mexico City, Mexico*
2013	**1st Place**, Experimental Exhibition of Physics, Faculty of Chemistry, UNAM with the work "Viscosimetry: A measurement of ethylene glycol viscosity"	*Mexico City, Mexico*
2012	**1st Place**, Demonstration Show with motive of the symposium : "Impact of coordination chemistry in 100 years" with the work "Coordination Dominoes"	*Mexico City, Mexico*

Scientific Production

Talks

European Conference of Theoretical Chemistry — *Perugia, Italy*
Presenter of MARTINI Coarse-Grained models of imidazolium-based ionic liquids: from nanostructural organization to liquid-liquid extractions — *2019*

Scientifica — *Zurich, Switzerland*
Presenter of 'Virtual reality in chemistry applications in teaching and research' — *2021*

MMSML Workshop Methods in Molecular Simulations and Machine Learning — *Barcelona, Spain*
Presenter of 'More data or better data? How the training data influences machine learned predictions in Chemistry' — *2022*

Theoretical Biology Group, Karlsruhe Institute of Technology *Karlsruhe, Germany*

Presenter of ’Uncertainty Quantification in Atomistic Neural Networks’ *2022*

Departamento de Fisica y Quimica Teorica, Facultad de Quimica, Universidad Nacional Autonoma de Mexico *Mexico City, Mexico*

Presenter of ’Hacia mejores bases de datos quimicas para el aprendizaje automatico atomistico’ *2023*

Department of Chemistry, University of Cambridge *Cambridge, United Kingdom*

Presenter of ’Uncertainty Quantification in Atomistic Neural Networks’ *2023*

Department of Physics & Astronomy, University College London *London, United Kingdom*

Presenter of ’Learning Potential Energy Surfaces: Morphing and Uncertainty quantification’ *2023*

Institute for Theoretical Physics, University of Heidelberg *Heidelberg, Germany*

Presenter of ’How to enhance chemical databases for atomistic machine learning?’ *2023*

Conference of the American Chemical Society *San Francisco, CA, USA*

Presenter of ’Towards better chemical databases for atomistic machine learning’ *2023*

Journal Articles

1. Vazquez-Salazar L.I., Selle M., De Vries A.H., Marrink S.J., & Telles de Souza P.C., Martini coarse-grained models of imidazolium-based ionic liquids: from nanostructural organization to liquid–liquid extraction, Green Chem., 2020, 22, 7376-7386 *(Selected as “hot article” by the editors)*
2. Vazquez-Salazar L.I., Boitter E., Unke O.T., & Meuwly M., Impact of the characteristics of quantum chemical databases on machine learning predictions of tautomerization energies, J. Chem. Theor. Comp., 2021, 17(8), 4769–4785.
3. Töpfer K., Pasti A., Das A., Salehi S.M., Vazquez-Salazar L.I., Rohrbach D., Feurer T., Hamm P., & Meuwly M., Structure, Organization, and Heterogeneity of Water-Containing Deep Eutectic Solvents, J. Am. Chem. Soc. 2022 144 (31), 14170-14180.
4. Vazquez-Salazar L.I.*, Boitter E., & Meuwly M., Uncertainty Quantification for Predictions of Atomistic Neural Networks, Chem. Sci., 2022, 13, 13068-13084. * Co-Corresponding author
5. Käser S, Vazquez-Salazar L.I., Meuwly M., & Töpfer K, Neural Network Potentials for Chemistry: Concepts, Applications and Prospects, Dig. Disc., 2023, 2, 28-58.
6. Fischer T.L., Bödecker M, Schweer S.M. ,Dupont J., Lepère V, Zehnacker-Rentien A., Suhm M.A., Schröder B., Henkes T., Andrada D.M., Balabin R.M., Singh H.K., Bhattacharyya H.P., Sarma M., Käser S., Töpfer K., Vazquez-Salazar L.I., *et al.*, The first HyDRA challenge for computational vibrational spectroscopy, 2023, Phys. Chem. Chem. Phys., 2023, 25, 22089-22102.
7. Song K., Käser S., Töpfer K., Vazquez-Salazar L.I. & Meuwly M., PhysNet Meets CHARMM: A Framework for Routine Machine Learning/ Molecular Mechanics Simulations, J. Chem. Phys., 2023, 159, 024125
8. Horn K, Vazquez-Salazar L.I.†, Koch C. & Meuwly M., Improving Potential Energy Surfaces Using Measured Feshbach Resonance States, 2023, (Preprint:arXiv:2309.16491). Sci. Adv. *In Press* † Equal contribution.

Articles in preparation

1. Vazquez-Salazar L.I.†, Käser S. & Meuwly M., Outlier-Detection for Reactive Machine Learning Potential Energy Surfaces, 2024, to be submitted to Nat. Comm.. † Equal contribution
2. Vazquez-Salazar L.I.* & Meuwly M., Enhancing chemical databases for atomistic machine learning by sampling conformational space, 2024, to be submitted to J. Chem. Inf. Mod.
3. Töpfer K., Vazquez-Salazar L.I.† & Meuwly M., *Asparagus*: A Toolkit for Automatic Construction of Machine-Learned Potential Energy Surfaces, 2024, to be submitted to Comp. Phys. Comm. † Equal contribution.

4. Vazquez-Salazar L.I., Töpfer K. & Meuwly M., *nscat*: A code to compute neutron and x-ray structure factors from atomistic simulations.
5. Vazquez-Salazar L.I. & Meuwly M., Regression Prior Networks for Uncertainty Quantification in Atomistic Neural Networks, 2024, to be submitted to Mach. Learn. Sci. Tech.
6. Vazquez-Salazar L.I., Tunnelling of heavy atoms in double well potentials using analytical models, To Be submitted to Chem. Phys.
7. Vazquez-Salazar L.I., Meuwly M. & Wales D., Exploring landscapes of Machine Learned potentials: a practical example with peptides.
8. Vazquez-Salazar L.I & Medina-Franco J.L., (In Spanish) Aprendizaje automático aplicado a química: ¿Evolución, revolución o moda? (English title: Machine learning applied to chemistry: Evolution, revolution or trend?)

Reviewer tasks

1. American Chemical Society Lab certificate reviewer
2. Reviewer for ICLR-2023 workshop: ML4Materials
3. Reviewer for Journal of Open Source Software