Vibrational Frequencies of Molecular Systems using High Dimensional Neural Network Potentials

Dissertation

for the award of the degree

“Doctor rerum naturalium"

of the Georg-August-Universität Göttingen

within the doctoral program Chemistry

of the Georg-August University School of Science (GAUSS)

submitted by

Dilshana Shanavas Rasheeda

from **Kerala, India**

Göttingen, **9.12.2022**

Thesis Committee

Supervisor:
Prof. Dr. Jörg Behler
Lehrstuhl für Theoretische Chemie II
Ruhr-Universität Bochum

Second Supervisor:
Prof. Dr. Martin Suhm
Institut für Physikalische Chemie
Georg-August-University Göttingen

Examination Board

Reviewer:
Prof. Dr. Jörg Behler
Lehrstuhl für Theoretische Chemie II
Ruhr-Universität Bochum

Second Reviewer:
Prof. Dr. Martin Suhm
Institut für Physikalische Chemie
Georg-August-University Göttingen

Further Members of the Examination Board
Prof. Dr. Ricardo Mata
Institut für Physikalische Chemie
Georg-August-University Göttingen

Jun.-Prof. Dr. Daniel Obenchain
Institut für Physikalische Chemie
Georg-August-University Göttingen

Dr. Tim Schäfer
Institut für Physikalische Chemie
Georg-August-University Göttingen

Jun.-Prof. Dr. Anna Krawczuk
Institut für Anorganische Chemie
Georg-August-University Göttingen

Date of the oral examination: 25.01.2023

Oath

I hereby declare that I have prepared this thesis all by myself, not used any sources or tools except for those explicitly stated, and marked all quotes, be it in literal or analogous form, accordingly.

I declare that this thesis has not, nor in excerpts, been submitted to this or any other university in the context of a failed examination.

Göttingen, **9.12.2022**

(**Dilshana Shanavas Rasheeda**)

Abstract

Computational Chemistry is an important field in chemistry which looks for solutions for several questions such as reaction mechanisms, design of experiments and understanding fundamental properties of molecules. For molecular systems, usually quantum chemical methods are used. These methods give highly accurate results albeit with high computational costs. They can be used to calculate several properties like energies, forces and also vibrational spectra. When it comes to vibrational computations, there are two limiting factors; the level of electronic structure methods and the level of vibrational treatment. A highly accurate Potential Energy Surface (PES) is needed to compute high quality vibrational frequencies. Machine Learning Potentials (MLPs) which have become increasingly popular in chemistry and material science can help in bridging the gap between accuracy and cost. Here, High-Dimensional Neural Network Potentials (HDNNPs) are the MLPs of choice used to construct an accurate Potential Energy Surface for use in vibrational spectroscopy.

The endeavour is to construct an HDNNP fitted to a computationally expensive Coupled Cluster method starting from a small molecule which is Formic Acid Monomer and then increase the system size to the Formic Acid Dimer. The constructed PES will be used to calculate vibrational frequencies at harmonic and anharmonic levels and also benchmark them against experimental and theoretical vibrational frequencies. The HDNNP has further benefits of accurately representing the energies and dynamics of the system at hand at low cost.

The work presents a methodology to construct a High-Dimensional Neural Network Potential for use in vibrational spectroscopy. The PES is constructed systematically with proper analysis and validation steps to reach a predefined threshold of 10 cm^{-1} for harmonic frequencies. Additionally, the HDNNP for Formic Acid Dimer is validated by computing anharmonic frequencies. Concurrently, it tests the capabilities of a High-Dimensional Neural Network in representing the fine details of the potential.

Contents

1 Introduction **13**

2 Theoretical Background **19**
2.1 Electronic Structure Methods 22
2.1.1 Hartree-Fock 23
2.1.2 Coupled Cluster 24
2.2 Explicitly Correlated Methods 25
2.3 Resolution of Identity (RI) Approximation 25
2.4 Basis Sets 25
2.5 High-Dimensional Neural Networks 27
2.5.1 Atom Centered Symmetry Functions 29
2.5.2 Training the High-Dimensional Neural Network 32
2.6 Molecular Dynamics 34
2.7 Theoretical Vibrational Spectroscopy 35
2.7.1 Harmonic Frequencies 35
2.7.2 Second-Order Vibrational Perturbation Theory 36
2.7.3 Variational Vibrational Computation 37

3 Computational Details **39**
3.1 Molpro 39
3.2 RuNNer 39
3.3 Construction of the HDNNP 40
3.4 Finite Difference 41
3.5 Molecular Dynamics Simulations 41

4 Results and Discussion **43**
4.1 Formic Acid Monomer 43
4.1.1 Initial HDNNP 44
4.1.2 Final HDNNP 50
4.1.3 Comparison with Tew Potential 53
4.2 Formic Acid Dimer 55
4.2.1 Construction of HDNNP 55
4.2.2 HDNNP1 57
4.2.3 Development of HDNNP2 64
4.2.3.1 Intermediate HDNNPs 69
4.2.4 HDNNP2 74
4.2.5 HDNNP3 80
4.2.6 Discussions on the Various HDNNP Iterations and their Frequencies 90
4.2.7 Construction of HDNNP with 3D Coupling Terms 98

5 Conclusion and Outlook **105**

Bibliography **109**

A Appendix **119**
A.1 HDNNP3-i 119
A.2 Analytical Harmonic Frequencies of FAD-HDNNP 122
A.3 RuNNer Settings 123
A.3.1 ACSFs 123
A.4 Harmonic Frequencies of Intermediate HDNNPs 125

Curriculum Vitae **133**

Glossary

ACE Atomic Cluster Expansion

ACSF Atom Centered Symmetry Function

BOA Born-Oppenheimer Approximation

CBS Complete Basis Set

CC Coupled Cluster

CCD Coupled Cluster Doubles

CCSD Coupled Cluster Singles and Doubles

CCSD(T) Coupled Cluster Singles, Doubles and Perturbative Triples

CCSDT Coupled Cluster Singles, Doubles and Triples

CI Configuration Interaction

DFT Density Functional Theory

GAP Gaussian Approximation Potential

GTO Gaussian Type Orbitals

HDNNP High-Dimensional Neural Network Potential

HF Hartree-Fock

LCAO Linear Combination of Atomic Orbitals

MCTDH Multiconfiguration Time-Dependent Hartree

MD Molecular Dynamics

MLPs Machine Learning Potentials

MP Møller-Plesset Perturbation Theory

NNP Neural Network Potential

PES Potential Energy Surface

RI Resolution of Identity

RMSD Root Mean Square Deviation

RMSE Root Mean Square Error

STO Slater Type Orbitals

VCI Vibrational Configuration Interaction

VPT2 Second-Order Vibrational Perturbation Theory

Chapter 1

Introduction

In the past few decades, human life has become increasingly digital. Even though, every new technological advancement has been welcomed with a sense of skepticism and hesitation, technology and in particular computers have become more and more prevalent in everyday life.

Computations have also become important in Chemistry, not only as a teaching aid but also as an active field of research. Computational Chemistry is a union of theoretical chemistry and computer programs to study chemical problems. Computational Chemistry is used to solve several types of chemical problems like molecular geometry, energies and reaction mechanisms, dynamics of a system, spectroscopy, drug design and physical properties in material chemistry.

Several tools are utilised in studying these problems. Molecular Mechanics considers the system as a classical object. The energies and geometries are obtained by considering the molecule as a ball and spring system. It utilises mathematical methods like steepest descent to find optimal geometries and uses empirical data to define energies as a function of bond distances, angles and dihedrals. Molecular Mechanics is fast and enables understanding of fairly large systems.

Ab-initio methods solve the Schrödinger equation. The Schrödinger equation cannot be solved exactly and ab-initio methods rely on approximations to solve for energy. Ab-initio calculations are slow but very accurate. Because of high computational costs, these cannot be used for large molecules.

Semi-empirical methods are based on Schrödinger equation but simplified extensively by using experimental data for approximations. These are slower than Molecular Mechanics but faster than wavefunction based methods.

Nowadays, one of the most widely used methods is Density Functional Theory (DFT). These utilise a functional of electron density to calculate energies instead of a wavefunction which is the case in ab-initio methods. These are faster than wavefunction based methods but slower than semi-empirical methods.

Another method used in Computational Chemistry is Molecular Dynamics which can be combined with any of the above described methods. It uses the time evolution of a system using Newton's laws of motion. The energies and forces can be defined using a forcefield. When the energy is defined by quantum mechanics, it is called ab-initio molecular dynamics. Ab-initio molecular dynamics can simulate chemical reactions.

There are several other tools or methods used in Computational Chemistry. One of the topics gaining much interest in the recent times is the use of Machine Learning in chemistry. Machine Learning is a subset of artificial intelligence. Artificial intelligence is when computers or machines are used to perceive and analyse and use information without human interference. Artificial intelligence is used from web search algorithms to self driving cars. One of the most active fields in artificial intelligence is Machine Learning.

Machine learning is developing methods that learn from data and then use what is learned to interpolate or extrapolate the data. The machine learning algorithm is able to predict behaviours in areas it is not trained using the underlying data. The machine learning algorithm relies on three aspects; the data, the model of learning and the output to be predicted [1]. In chemistry, machine learning potentials are used widely for many applications such as chemical reactions, molecular design and so on. [1–5].

To elucidate the use of MLPs in chemistry, one of the important concepts to introduce is the Potential Energy Surfaces (PES). A potential energy surface is a representation of the energy of a system as a function of the positions of nuclei and electrons. The potential energy surface is often a multi dimensional function of coordinates and it requires solving the Schrödinger equation for various atomic arrangements and is a difficult problem to solve. In addressing problems which require an extensive set of atomic arrangements, the potential energy surface becomes a difficult hurdle to overcome. This is the case of molecular dynamics where energies and forces are calculated on the fly for the propagation of the system. This also becomes a bottle neck for high level vibrational calculations. In these instances, it becomes useful to have a representation of the potential energy surface in a functional form. Analytical potentials are a solution to the problem but they involve coming up with a functional relation between the geometry and the energies and is only feasible for low dimensional molecules. Molecular Mechanics though useful in doing computations with very large systems lack in accuracy.

A clear solution to the above described problem is constructing a Machine Learning Potential. An earliest example for the use of MLPs in constructing a Potential Energy Surface is the work by Doren and et al. They used a feed forward Neural Network to study the adsorption of CO on Ni(III) [6]. The Potential Energy Surface of a system can be learned by an MLP algorithm by training it with a set of reference data using an electronic structure method. This provides a way to get a PES which has the accuracy of the quantum mechanical method used to construct the reference data. Over the years, various studies have used MLPs for constructing PESs. [7–12]

The first generation of MLPs was limited to small molecules [13, 14]. This is mainly due to the unavailability of descriptors which can handle large scale systems. This is addressed in the second generation of MLPs. This was solved by the introduction of a descriptor known as Atom Centered Symmetry Functions by Behler [15]. The scaling to large systems was also addressed when Behler and Parrinello [16] used element specific atomic neural networks to output atomic energies which can be summed up to give total energy of the system. The MLPs are known as High-Dimensional Neural Networks and will be used in this thesis. [7, 14–20]. Other types of MLPs in the second generation are Gaussian Approximation Potentials (GAPs) [21, 22], Spectral Neighbour Analysis Potentials (SNAPs) [23], Atomic Cluster Expansion (ACE) [24] and so on. These all use descriptors which are local. These do not include long range interactions such as electrostatics.

Including electrostatic interactions led to the third generation of MLPs. Here, the short range and long range energies are computed separately. The long range energy is expressed using environment dependent charges. An example for this is third generation HDNNPs [25, 26].The fourth generation of MLPs include long range electrostatic interactions depending on the atomic charges that relies on the whole structure of the system [27, 28]. The fourth generation High-Dimensional Neural Network Potential (4G-HDNNP) is an example of fourth generation MLPs which considers atomic electro-negativities as a function of local atomic environment. [29].

The MLPs provide a very flexible form for the Potential Energy Surface. It is possible that because of this the MLP may not fully replicate the correct form of the PES. Therefore, the selection of the training data and the validation step becomes crucial in constructing a PES. Even though, the fit is really good it is possible to have endured error compensation emphasizing the need to use the most useful validation technique. Usually RMSEs are used for validating the quality of a potential. But, since the RMSEs are printed for structures which are present in the reference data, it becomes important to propose another validation method to represent the quality of the potentially especially in domains like vibrational spectroscopy where the minute details of the PES become critical.

Here, it is proposed to use vibrational frequencies, both harmonic and anharmonic frequencies, to examine the efficacy of a potential. The goal is to benchmark vibrational frequencies. At the same time, they also give a necessary validation to the merit of the constructed potential. Neural Networks have been used to compute vibrational frequencies for small systems. [30, 31]. Several types of MLPs have been used for this. [32–34] Larger systems have been handled by second generation MLPs for large molecules, clusters and condensed systems. [26, 35–39] Though second generation MLPs have been designed for very large systems, here HDNNPs are used to test the limit of accuracy for moderate sized systems. The underlying data used Coupled Cluster theory to provide energies and with the help of HDNNPs, the intent is to produce high level vibrational frequencies based on a highly accurate electronic structure method. This is highly desirable as the quality of computed frequencies depend on both electronic structure method and vibrational treatments. The union of both these desirable features has been a big challenge and a highly accurate potential constructed by HDNNP could provide answer to the challenge. This is possible as a very accurate potential should mimic the behaviour of underlying level of theory. For this two systems are studied here: Formic Acid Monomer and Formic Acid Dimer.

Formic Acid Monomer has been studied extensively using experiments because of its importance in atmospheric chemistry [40] and interstellar chemistry [41–43]. Experimental IR and Raman spectroscopy have been studied for Formic Acid Monomer by gas phase and matrix isolation spectroscopy [44–53]. On the side of theory a few studies have been done for the vibrational spectra of Formic Acid Monomer. VSCF calculations have been performed on a Møller-Plesset PES to calculate overtones of the molecule [49]. Vibrational Perturbation Theory has been used on high level Coupled Cluster theory to calculate rotational and hyperfine parameters [54]. The trans rotamer of Formic Acid Monomer is energetically more stable than the cis rotamer and is the topic of study in the thesis. High level anharmonic calculations for Formic Acid Monomer on an accurate potential energy surface has been interesting for understanding its spectra as the fundamental O-H bend and the overtone of O-H torsion are in resonance and the assignment of the peaks have been a topic of active research for a long time [55, 56].

A recent work [57] uses a PES constructed from CCSD(T)-F12c/cc-pVTZ-F12 energies as a reference. The PES is constructed using LASSO-based regression model. Vibrational Configuration Interaction (VCI) vibrational frequencies are computed on the PES and benchmarked against experimental data for both the trans conformer for Formic Acid Monomer with an RMSD of 3 cm^{-1}. VCI frequencies are also calculated for the cis conformer. Another study constructed a PES based on CCSD(T)-F12a/aug-cc-pVTZ for both the conformers for Multiconfiguration Time-Dependent Hartree (MCTDH) vibrational calculations [58]. More recently a transfer learned potential was constructed for Formic Acid Monomer based on MP2 and then transfer learnt onto CCSD(T) with an aug-cc-pVTZ basis set [59]. Harmonic and VPT2 frequencies were computed on this surface.

Formic Acid Dimer is a doubly hydrogen bonded planar molecule. It is a very interesting and actively studied molecule for construction of PES for specifically spectroscopic use. The double proton transfer especially makes it a very interesting system for dynamic studies. The barrier for the double proton transfer [60] and the ground-state tunnelling-splitting [60, 61]has been of interest specifically. There has been a wealth of experimental studies for Formic Acid Dimer. Early works include thermal gas phase spectroscopy [62, 63]. Recent work includes jet-cooled infra red spectra and raman spectra in the finger print region of the monomer [64]. The intermolecular vibrational fundamental frequencies and many combination and overtone bands of Formic Acid Dimer have been studied in the gas phase with remarkable accuracy [65–67, 67–74]. A review [75] gives a very good overview of the experimental and theoretical work so far done in the case of vibrational spectroscopy of Formic Acid Dimer. Though several PESs are available and constructed for spectroscopic use in the case of Formic Acid Dimer, so far all these potentials are not able to describe harmonic frequencies accurately, reporting a maximum deviation above 20 cm^{-1} with respect to the harmonic frequencies from the reference ab-initio method.

One of the most widely used PES for vibrational studies is the Permutationally Invariant Potential (PIP) constructed by Qu and Bowman [60] based on 13475 structures using CCSD(T)-F12a/haTZ level of theory. VCI calculations were performed on this surface [76–78]. This PES was used in reduced dimensionality variational calculation [79]. In this work, the potential was found to have very good concurrence with experiments but two fundamental modes were majorly blue shifted and also the PES has some artefacts such as PES holes and this necessitates the requirement of another good quality PES for vibrational studies of Formic Acid Dimer. Another full dimensional PES has been recently published for Formic Acid Dimer and Formic Acid Monomer [59]. The underlying reference data is based on 26,000 MP2/aug-cc-pVTZ structures which are transfer learned based on 866 CCSD(T)/aug-cc-pVTZ energies to obtain a PES which is used to compute harmonic and anharmonic frequencies. These two PESs will be compared to the HDNNP for Formic Acid Dimer.

In this work, the goal is benchmarking of vibrational frequencies for both Formic Acid Monomer and Formic Acid Dimer as a validation step in assessing the quality of the potential. It also helps to understand where the potential needs improvement and to observe how it fares against experimental frequencies and the frequencies from available potentials. The objective is to construct an accurate potential for computing vibrational frequencies and such benchmarking the potential against the other potentials is necessary to assess the accuracy and the benefits in using this potential over others available.

The harmonic frequencies of Formic Acid Monomer is computed by constructing an HDNNP based on CCSD(T)-F12c/cc-pVTZ-F12 energies. Initially, a dataset obtained from David Tew [57] is used to construct an HDNNP which is analysed and validated to see the quality of the potential. This is performed by defining various validating parameters and systematically analysing and substantiating each step along the way. The final HDNNP is constructed by improving the underlying data through sampling critical regions of the PES. This HDNNP is then benchmarked with reference Coupled Cluster harmonic frequencies and the frequencies obtained from David Tew's analytical potential.

In the case of Formic Acid Dimer, the harmonic frequencies are benchmarked with the reference Coupled Cluster frequencies and harmonic frequencies obtained from the potential (QB16) of Qu and Bowman [60]. The reference data was initially obtained from Qu and Bowman used CCSD(T)-F12a/haTZ energies. The construction of an HDNNP for Formic Acid Dimer goes through various iterations of construction of a potential. The final HDNNP serves to do various vibrational studies in collaboration with Edit Matyus and Benjamin Schröder [80]. This gives opportunity to benchmark VPT2 frequencies from the HDNNP with Coupled Cluster VPT2 frequencies and experimental frequencies. The reduced dimensionality variational calculations are also done on the potential which leads to benchmarking with the variational frequencies from QB16 and also with experiments. The various qualities of the PES will be elucidated in this thesis.

In the case of both Formic Acid Monomer and Formic Acid Dimer, a procedure of developing a Machine Learning Potential for spectroscopic use will be underlined and the further step of quality control and substantiation will be proposed. The aim is to develop a sturdy full-dimensional potential of Coupled Cluster quality which can provide an opportunity to perform high-level vibrational calculations which are otherwise expensive and limited by the computational cost. The potential should not only give good values of desired quantities but must be a global surface which can be used in other applications such as Molecular Dynamics and representations of important regions of the PES.

Since the HDNNP is constructed with the goal of vibrational benchmarking along with RMSE and energetics, a goal of accuracy for the harmonic frequency is to be defined which is to be within a deviation of 10 cm^{-1} from ab-initio harmonic frequencies. In the case of Formic Acid Dimer, further to represent couplings properly, benchmarking with VPT2 is considered an additional parameter of quality control. And the final HDNNP serves to be used in high level computations such as variational methods.

Chapter 2

Theoretical Background

Computational chemistry uses concepts in quantum mechanics or classical mechanics to find solutions to chemical problems. It also uses certain approximations to predict observables and properties for chemical systems of varying sizes and problems of various kinds. When it comes to quantum chemistry, it mainly involves solving the Time Independent Schrödinger Equation. When this is done using methods without any experimental data, it is known as ab-initio methods. Computational Chemistry has numerous applications. It can be used to find out reaction mechanisms, to calculate dipole moments and polarizabilities, to produce vibrational, NMR and UV spectra and so on.

The Time Dependent Schrödinger Equation is a linear partial differential equation which describes the wavefunction of a system. It shows how the wavefunction of an isolated system evolves over time.

$$\hat{H}\psi(x,t) = i\hbar\frac{d}{dt}\psi(t) \tag{2.1}$$

Here, ψ is the wavefunction which encompasses all the properties of the system. It basically gives the *state* of the system. ψ is a function of coordinates of the system. If we know the state of a function at time t_0, it is possible to calculate the state of the system at a future time t.

Many applications in chemistry do not require the solution of the Time Dependent Schrödinger Equation. Instead, the Time Independent Schrödinger Equation is used.

$$H\psi = E\psi \tag{2.2}$$

Here, the hamiltionian H is given by,

$$H = \frac{-\hbar^2}{2m}\frac{d^2}{dq^2} + V(q) = \hat{T} + \hat{V} \tag{2.3}$$

The Hamiltonian basically gives the energy contributions from Kinetic Energy T

and Potential Energy V. This can be further broken down into the following for a system with N electrons and M nuclei.

$$\begin{aligned} H &= \hat{T}_e + \hat{T}_N + \hat{V}_{eN} + \hat{V}_{ee} + \hat{V}_{NN} \\ &= -\frac{1}{2}\sum_{i}^{N}\nabla_i^2 - \frac{1}{2}\sum_{A}^{M}\frac{1}{M_A}\nabla_A^2 - \sum_{i}^{N}\sum_{A}^{M}\frac{Z_A}{r_{iA}} + \sum_{i}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{r_{AB}} \end{aligned} \quad (2.4)$$

Here, $M_A = \frac{m_A}{m_e}$. Here, m_A is the mass of A atom and m_e is the mass of electron. $\hat{T_e}$ gives kinetic energy of electrons, $\hat{T_N}$ gives nuclear kinetic energy. $\hat{V_{eN}}$ is the nuclear-electron interaction energy, $\hat{V_{ee}}$ is the electron-electron repulsion energy and $\hat{V_{NN}}$ is the nuclear-nuclear repulsion energy contribution to the full potential energy of the system. Z_A gives the nuclear charge of an atom A. The Schrödinger equation in Eq. 2.2 can be solved for only very few systems. Hence, a few approximations have to be made to use it for solving chemical problems in molecular systems. One such approximation is Born-Oppenheimer Approximation (BOA) [81, 82].

The nuclei are much heavier than an electron. Therefore, their movement is much slower compared to an electron. This means that the electrons adjusts their position instantaneously when the nuclei move. Therefore, the movement of electrons and nuclei can be assumed to be independent of each other. This helps to approximate the wavefunction as follows.

$$\psi_{mol}(r_i, r_A) = \psi_{el}(r_i, R)\psi_{nuc}(r_A) \quad (2.5)$$

This helps to describe the electronic motion separately while substituting in Eq. 2.2. Thus,

$$\hat{H}_{el}\psi_{el}(r_i, R) = E_{el}(R)\psi_{el}(r_i, R) \quad (2.6)$$

Here, the electronic Hamiltonian is

$$\hat{H}_{el} = \hat{T}_e + \hat{V}_{eN} + \hat{V}_{ee} + \hat{V}_{NN} \quad (2.7)$$

Since the nuclei are considered stationary, the nuclear kinetic energy is assumed to be 0. Also since the nuclei are fixed, R is treated as a parameter. Therefore,E_{el} will depend on R parametrically. It is to be noted that $\psi_{nuc}(r_A)$ will define vibrational and rotational motion of the molecule. Here, the also the $\hat{V_{NN}}$ is considered as a constant.

The separation of a molecular Schrödinger equation into electronic and nuclear motion is known as the Born-Oppenheimer approximation. It is to be noted that the BO approximation can breakdown when the separation between electronic states is small. Basically, when there is a situation where the nuclear and electronic motion is coupled, the BO approximation does not hold true.

Also, the energy of a system with respect to the coordinates of the nuclei is called Potential Energy Surface(PES). The minima in this surface correspond to the equilibrium structures. The saddle points give rise to Transition State enabling us to map the profile of a reaction or other processes.

The variational principle [83] is another important concept used in quantum mechanics. Since an exact wavefunction for a system is impossible to construct, the energy obtained using an approximate wavefunction is always greater than the exact energy of the system.

With the help of these concepts, several methods have been used to solve quantum chemical problems. The earliest attempt was using Hartree-Fock method which uses the Slater-determinant to build a wavefunction from molecular orbitals. In Hartree-Fock method, it is also assumed that the electron moves in a mean field of all the other electrons. The difference between the Hartree-Fock energy and the exact energy of a system is called correlation energy. A number of methods called post-Hartree-Fock methods are devised to recover electron correlation which reduces deviation from experimental results.

2.1 Electronic Structure Methods

One of the important reasons to use quantum mechanical methods in chemistry is to determine the electronic structure of a system. For a given geometry, the wavefunction can be optimised and a probability distribution of the electrons can be obtained. Even the gradient can be determined according to the position of the nuclei in order to obtain the equilibrium structure. For this, often an electronic structure method needs to be chosen, eg: Hartree-Fock. And also a basis set need to be determined. Using these two, the energies or geometry of a system can be computed. There will be further discussions on basis sets. There are various electronic structure methods built on top of Hartree-Fock which correct for the electron correlation. Examples are Møller-Plesset Perturbation Theory (MP), Configuration Interaction (CI), Coupled Cluster (CC) etc. These are wavefunction based methods. Density Functional Theory (DFT) is also a quantum mechanical method to determine electronic structure. In DFT unlike, wavefunction based methods, the energies are dependent on electron density rather than the wavefunction. Another class of methods to obtain the electronic structure of a system is the semi-empirical methods. Often, Hartree-Fock and post Hartree-Fock methods are very expensive computationally. For example, the most costly part of Hartree-Fock algorithm is the two electron integrals. Semi-empirical methods approximate that using empirical data. As such they are built on Hartree-Fock but with serious approximations that reduces the computational cost considerably.

There are a few sources of error in ab-initio electronic structure calculations. One is from the incompleteness of the basis set which will be explained in a later section. Another is from incorrect treatment of electron correlation. All the post-Hartree-Fock methods are the attempt at recovering electron correlation energy. The correlation energy is defined as the difference between the exact energy of the system and the Hartree-Fock energy.

$$E_{corr} = E_{exact} - E_{HF} \tag{2.8}$$

There are two types of correlations, dynamic correlation and static correlation. Dynamic correlation arise from the instantaneous interaction of electrons due to their movement. Static correlation arises from the fact that in some cases the single-Slater-determinant HF wavefunction is a poor description of the electronic state. The wavefunction has significant contribution from multiple electronic states. To deal with this, multi-reference methods need to be used.

There are two factors that are desirable in a method other than being variational. One is size consistency. A method is considered to be size consistent if the dissociated parts of a molecule at infinite distance gives the same computed energy as the sum of computed energy of each of the parts. Another important characteristic is size extensivity. A method is considered size extensive if the computed energy of n non-interacting identical systems is n times the energy of one such system.

In this thesis, many calculations were performed using Coupled Cluster methods. Since Coupled Cluster is built on Hartree-Fock, the next section goes into an explanation of the Hartree-Fock method.

2.1.1 Hartree-Fock

Hartree-Fock (HF) [84–88] is an approximation method to determine wavefunction and energy of a many electron system. It finds an approximate solution to the Schrödinger equation.

A simple form of the wavefunction could be a product of molecular orbitals. This is called a Hartree Product.

$$\psi_{HP}(r_1, r_2 ... r_N) = \phi(r_1)\phi(r_2)..\phi(r_N) \tag{2.9}$$

Since, this does not fulfill the Pauli's principle, another method to construct the wavefunction using molecular orbitals had to be conceived. This should ensure that the wavefunction is anti-symmetric with respect to the interchange of any spin-spatial coordinate. This led to the wavefunction being represented as a Slater-determinant.

$$\psi_{HF}(r_1, r_2 ... r_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(r_1) & \psi_2(r_1) & \cdots & \cdots & \psi_N(r_1) \\ \psi_1(r_2) & \psi_2(r_2) & \cdots & \cdots & \psi_N(r_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \psi_1(r_N) & \psi_2(r_N) & \cdots & \cdots & \psi_N(r_N) \end{vmatrix} \tag{2.10}$$

Each term in the determinant is a spin orbital which means it has both spin and spatial components.

The expression for the Hartree-Fock energy comes out to be,

$$E_{HF} = \langle \psi_{HF} | \hat{H} | \psi_{HF} \rangle \tag{2.11}$$

This can be written in terms of one electron and two electron integrals,

$$E_{HF} = \sum_i \langle i | h | i \rangle + \frac{1}{2} \sum_{ij} [\langle ij | | ij \rangle - \langle ij | | ji \rangle] \tag{2.12}$$

The equation is solved for energy iteratively until certain predefined conditions are met. The Hartree-Fock method also utilises the variational theorem to compute energy. With a complete basis set, the best possible solution for the Hartree-Fock energy can be obtained. This is called the Hartree-Fock limit. Open-shell systems can also be handled using Unrestricted Hartree-Fock and Restricted Open shell Hartree-Fock methods. But, Hartree-Fock may not be used in systems where there are multiple significant electronic configurations contributing to the wavefunction.

2.1.2 Coupled Cluster

Coupled Cluster [89, 90] is a post-Hartree-Fock method used for many electron systems. Coupled Cluster is one of the most widely used post-Hartree-Fock method. Coupled Cluster is non-variational, size extensive and size consistent. It basically uses a cluster operator $\hat{T}$ on a Slater-determinant $|\phi_0\rangle$. The cluster operator $\hat{T}$ is an excitation operator. The wavefunction from Coupled Cluster is written as :

$$|\psi\rangle = e^{\hat{T}} |\phi_0\rangle \tag{2.13}$$

Here, $|\phi_0\rangle$ is the reference Hartree-Fock wavefunction in the form of a Slater-determinant and cluster operator $\hat{T}$ for an N electron system is defined as:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + ... + \hat{T}_N \tag{2.14}$$

Here, $\hat{T}_1$ indicates single excitations, $\hat{T}_2$ indicates double excitations and so on. Here these excitation operators are defined.

$$\hat{T}_1 \phi_0 = \sum_{i,a} t_i^a \phi_i^a \tag{2.15}$$

$$\hat{T}_2 \phi_0 = \sum_{\substack{i<j \\ a<b}} t_{ij}^{ab} \phi_{ij}^{ab} \tag{2.16}$$

Here, ϕ_i^a is the singly excited Slater-determinant with occupied spin-orbital u_i replaced by virtual spin-orbital u_a. $\hat{T}_1$ operator converts Slater-determinant ϕ_0 into a linear combination of all possible single excited Slater-determinants. ϕ_{ij}^{ab} is the doubly excited Slater-determinant with occupied spin-orbital u_i and u_j replaced by virtual spin-orbitals u_a and u_b respectively. The rest of the operators are also defined accordingly till $\hat{T}_N$ because N is the total number of electrons.

Now, $e^{\hat{T}}$ can be expanded using Taylor expansion.

$$e^{\hat{T}} = 1 + \hat{T} + \frac{\hat{T}^2}{2!} + \frac{\hat{T}^3}{3!} + ... \tag{2.17}$$

In order to use Coupled Cluster for calculations, an approximation is made. Only some parts of the cluster operator are included. For example when only the $\hat{T}_2$ operator is included, i.e, $\hat{T} = \hat{T}_2$ it is called CCD method (Coupled Cluster Doubles). When $\hat{T} = \hat{T}_1 + \hat{T}_2$, it is called CCSD (Coupled Cluster Singles and Doubles). CCSDT is Coupled Cluster Singles, Doubles and Triples with $\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3$. CCSD(T) is the most popular Coupled Cluster method which is Coupled Cluster Singles, Doubles and perturbative Triples. Here, the triple excitations are included as perturbation.

The Schrödinger equation in terms of Coupled Cluster wavefunction becomes:

$$\hat{H} e^{\hat{T}} |\phi_0\rangle = E e^{\hat{T}} |\phi_0\rangle \tag{2.18}$$

This can be solved as

$$E_{CC} = \langle\phi_0| \hat{H} e^{\hat{T}} |\phi_0\rangle \tag{2.19}$$

2.2 Explicitly Correlated Methods

The motivation behind the use of explicitly correlated methods is to solve the problem of slow convergence of the wavefunction. Electron correlation is dependent on the distance r_{12} between two electrons, 1 and 2. Therefore, including r_{12} directly in the definition of the wavefunction can describe the correlation energy more accurately. These methods which include the distance between two electrons explicitly in defining the wavefunction is called explicitly correlated methods.

There are various types of explicitly correlated methods [91–94]. CCSD(T)-F12 methods are widely used. In general F12 methods have an additional F12 doubles term on top of the double excitations term where a function of r_{12} (f_{12}) is included in the two electron integrals.

2.3 Resolution of Identity (RI) Approximation

The four orbital Coulumb integral term in a wavefunction is computationally expensive. The RI approximation [95, 96] involves representing pair products of atomic basis functions as a linear combination of an auxiliary basis functions. This helps to greatly simplify the four orbital integrals into three centre and two centre integrals. The RI approximation is also known as density fitting.

2.4 Basis Sets

Basis sets are sets of functions used to construct a wavefunction in quantum chemistry. Several types of basis sets can be used for this purpose. Plane wave basis set is often used in solid state chemistry and basis sets based on atomic orbitals is used for molecular systems.

A complete set of basis functions, i.e, an infinite number of basis functions, is required to represent the wavefunction exactly. With this and an exact representation of correlation energy would be required to find the exact energy of a system. An infinite basis set is not possible and a truncation is required to realistically use them. This gives rise to a basis-set truncation error. Therefore, choosing a small enough but accurate enough basis set for a given level of theory is one of the main challenges in any computational chemistry problem.

There are various types of basis sets generally used for molecular systems. Linear Combination of Atomic Orbitals (LCAO) are used to construct wavefunctions. One type is the Slater Type Orbitals (STO) which has an orbital exponent of the type $-e^{\zeta r}$. The calculation of two-electron integrals become computationally costly with STOs. And as a result a different kind of basis set is most commonly used. These are Gaussian Type Orbitals (GTOs).

GTOs were very helpful in reducing the computational cost of quantum chemical calculations as they handle the integrals efficiently. The exponents are of the form, $-e^{\zeta r^2}$. Product of two Gaussians give another Gaussian thus greatly simplifying two-electron integrals at multiple atomic centres. GTOs nevertheless does not represent orbitals at nuclei properly and as a result a larger basis set is often required to produce accurate results. Multiple GTOs are brought together to form contracted Gaussian functions whose linear combinations is used to represent orbitals.

A minimal basis set which is the simplest basis set uses a single Gaussian function to represent one spatial orbital. A split-valence basis set [97] was invented to improve accuracy. Here, the valence orbital is represented by multiple Gaussian functions and the core orbital is represented by a single Gaussian function. To account for distortions caused by nearby atoms, polarization functions [98] can be added to the basis set. Correlation consistent basis sets [99] are designed to converge systematically to Complete Basis Set (CBS) limit using extrapolation techniques. These are the most widely used basis sets. They include large polarization functions added. The basis sets used in this work will be discussed in the Computational Details section.

2.5 High-Dimensional Neural Networks

High-Dimensional Neural Network Potentials (HDNNP) were conceptualised by Behler and Parrinello [16, 19, 100]. These Neural Networks overcome the limitations of the previous Neural Network Potentials. First of all, the previous NNPs were limited to small scale systems. Secondly, they were not transferable with respect to system size. They also failed at being invariant with respect to rotation or translation of the system and the permutation of same atom types. All these have been addressed in the formulation of High-Dimensional Neural Network Potentials.

Here, in HDNNPs, for each atom there is an atomic Neural Network which outputs an energy E_i. Summing up the atomic energies give the energy of the system. It is to be noted that the atomic contributions to total energy is not an observable and hence not chemically relevant.

$$E_{total} = \sum_{i=1}^{N} E_i \tag{2.20}$$

The atomic Neural Networks are specific to the element. The atomic energy E_i strictly depends on the local environment of that particular atom which is defined by a cutoff radius R_c. R_c is chosen such that all the relevant atomic interactions are included.

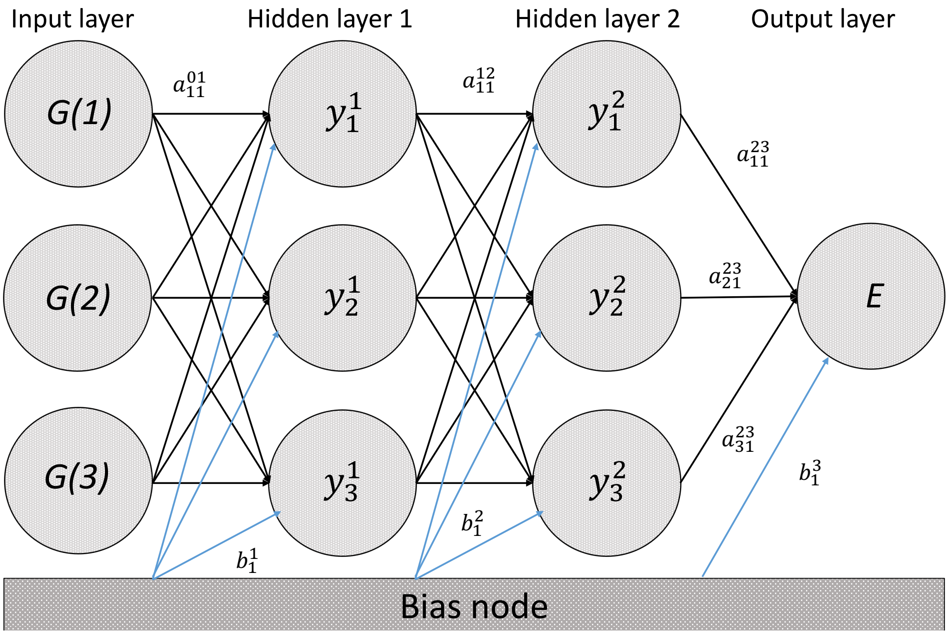


Figure 2.1: An example feed forward Neural Network which forms an atomic Neural Network in High-Dimensional Neural Networks. There are two hidden layers with 3 nodes each. The input is given as a symmetry function $G(i)$. The neurons are connected to each other by weight parameters a_{ij}^{kl}. The bias nodes connects each node with a bias parameter b_i^j

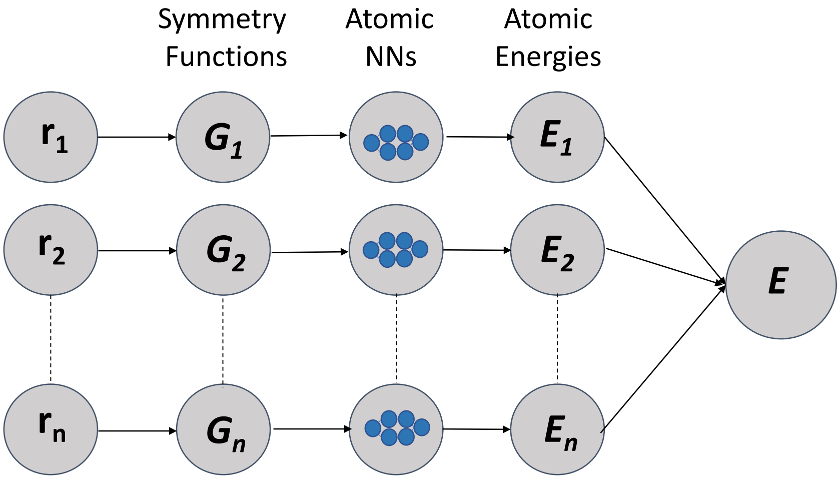


Figure 2.2: A High-Dimensional Neural Network comprising of several Atomic Neural Networks. The input is given as a vector of symmetry function G_i constructed from Cartesian coordinates r_i. Each Atomic Neural Network gives an atomic energy E_i which when summed up gives the full energy of the system E.

The general structure of a High-Dimensional Neural Network is as follows. For the feed-forward Neural Network for a particular atom, the input is a vector of symmetry function G_i. There are a number of layers with n number of nodes per layer. The layers in between the input and output layer are called hidden layers. Each node in a layer is connected to the previous one using a weight parameter which is what is optimised in the training process of the algorithm.

We can notate that a weight parameter a_{ij}^{kl} connects input layer k and l with the respective nodes i and j. Excepting the input nodes, all the nodes of a layer j are connected to the bias node by a bias weight b_i^j. The value of a node is calculated as

$$y_i^j = g_i^j(b_i^j + \sum_k a_{ki}^{j-1,j} \cdot y_k^{j-1}) \tag{2.21}$$

g_i^j is the activating function. Therefore, the value a neuron is calculated by multiplying the previous layer values of nodes with the weight parameter and adding them up along with the current bias weight and applying a function on them. The activation function is a non linear function which prevents the output energy value from being a linear combination of atomic coordinates. In HDNNP, the hyperbolic tangent is used as the activation function except for the output layer.

$$g(x) = tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.22}$$

For the output layer, $g(x) = x$ is used as the activation function which basically gives the sum of atomic energies while calculating the value of the output node.

If we imagine a feed forward Neural Network with two hidden layers with three nodes each, the atomic energy, i.e, the output of a single feed forward Neural Network is given as

$$E = g_1^3(b_1^3 + \sum_{k=1}^{3} a_{k1}^{23} \cdot g_k^2(b_k^2 + \sum_{l=1}^{3} a_{lk}^{12} \cdot g_l^1(b_l^1 + \sum_{i=1}^{3} a_{i1}^{01} \cdot G_i)) \tag{2.23}$$

2.5.1 Atom Centered Symmetry Functions

The input of each atomic Neural Network by a vector of symmetry functions which provide a fingerprint of the atomic environment. They are constructed out of the Cartesian coordinates of the atoms in the local environment.They ensure that the High-Dimensional Neural Networks are invariant with respect to rotation and translation of the system and permutation of same elements. These are called Atom Centered Symmetry Functions (ACSFs) [17]. The ACSFs depend on a cutoff radius R_c. R_c is used to define a cutoff function f_c. Only atomic environments up to R_c are contributing to the atomic energy E_i.

$$f_c(r_{ij}) = \begin{cases} 0.5[\cos(\frac{\pi r_{ij}}{Rc}) + 1], & \text{when } r_{ij} \leq R_c \\ 0, & \text{when } r_{ij} > R_c \end{cases} \tag{2.24}$$

Here, r_{ij} is the distance between atoms i and j. There are two types of symmetry functions, radial symmetry functions and angular symmetry functions. Radial symmetry functions encompass two body terms and angular symmetry functions encompass three body terms.

The radial symmetry function is defined as

$$G_i^{rad} = \sum_j e^{-\eta(r_{ij} - r_{shift})^2} \cdot f_c(r_{ij}) \tag{2.25}$$

Here, η gives the width of the Gaussian function and r_{shift} determines the centre of the Gaussian function. The fact that each symmetry function is multiplied by the cutoff function f_c ensures that the total symmetry function decays to zero at the cutoff radius. It also ensures that the double derivatives on the symmetry functions don't suddenly drop to zero at the cutoff distance and is hence differentiable.

The angular symmetry functions determine the orientation of atoms around a central

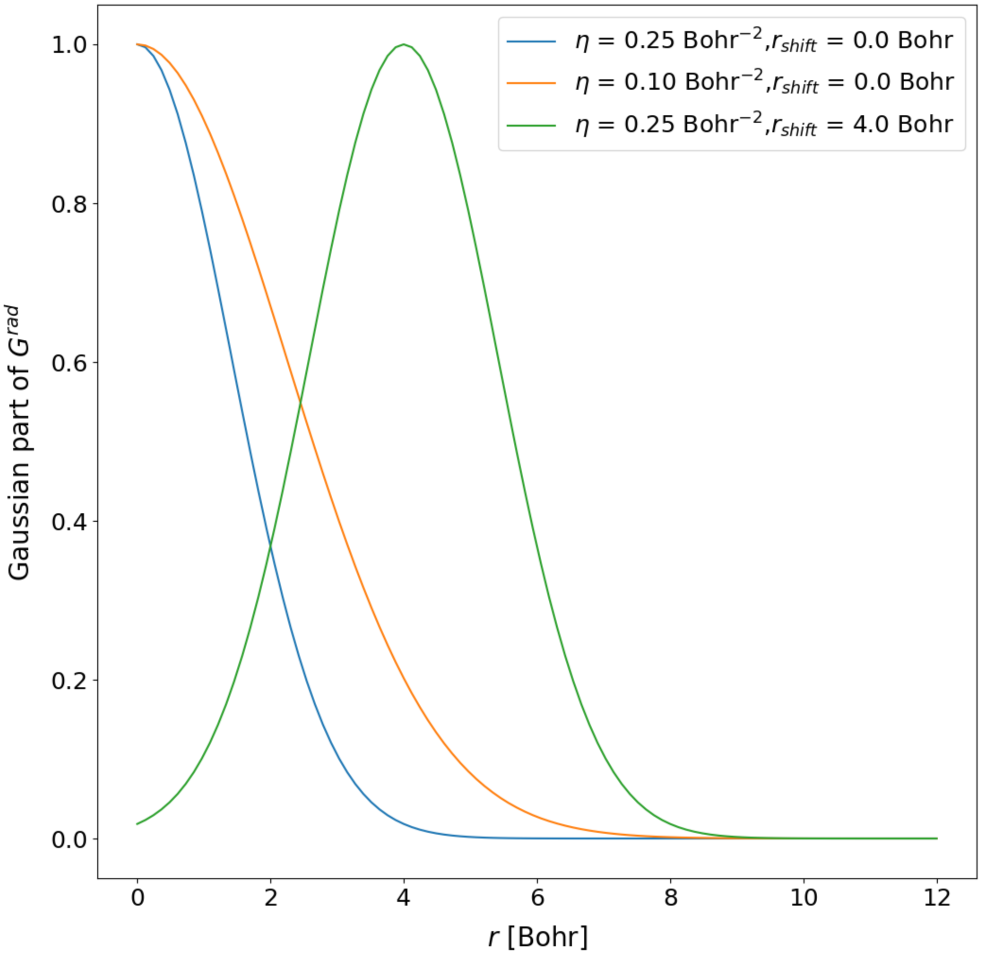


Figure 2.3: Gaussian part of radial symmetry function showing the shape of the function with varying η and r_{shift}. Here η is in Bohr^{-2} and r_{shift} is in Bohr

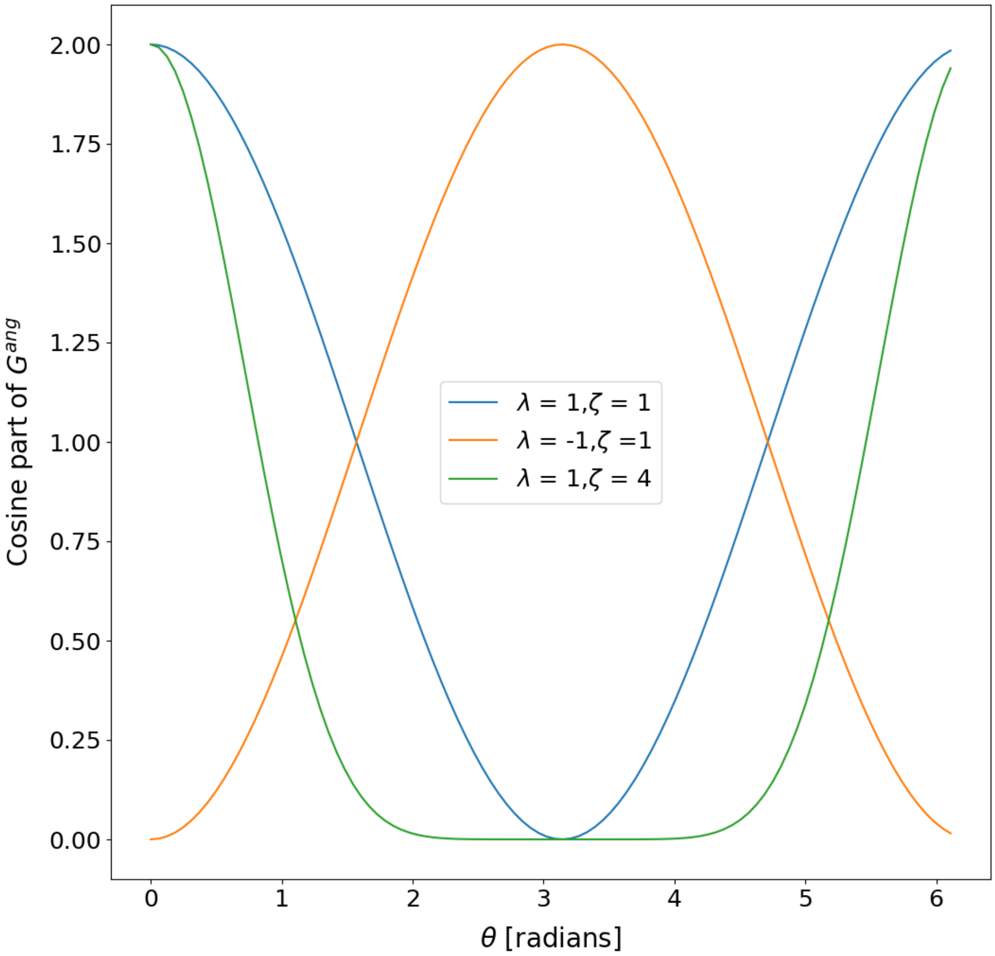


Figure 2.4: Cosine part of angular symmetry function showing the shape of the function with varying ζ and λ

atom i. The angular symmetry function is defined as

$$G_i^{ang} = 2^{1-\zeta} \sum_j \sum_k [1 + \lambda \cdot cos(\theta_{ijk})]^{\zeta} \cdot e^{-\eta(r_{ij}^2 + r_{jk}^2 + r_{ik}^2)} \cdot f_c(r_{ij}) \cdot f_c(r_{jk}) \cdot f_c(r_{ik}) \quad (2.26)$$

Here, $\zeta \in \{1, 2, 4, 16\}$. It indicates the width of the cosine part. It performs similar to η for radial symmetry functions. Parameter λ has values either -1 or $+1$.

The symmetry functions are unchanged during training and their values are determined prior to the training process. The whole training process is optimising weight parameters such that an optimal sets of weights is obtained that associates the symmetry functions and final energy output.

2.5.2 Training the High-Dimensional Neural Network

Training the High-Dimensional Neural Network Potentials indicates optimisation of weight parameters and bias parameters to get output values very close to the reference values. Here, the reference values mean the energy and force values of a given geometry from a chosen theoretical method, for example ab-initio methods. The optimisation of the parameters which could also be called fitting involves iteratively optimising the weights and biases such that the geometry and its energy are matched well enough. In the case of High-Dimensional Neural Networks, all the atomic Neural Networks are trained at the same time. The optimisation involves the minimisation of an error function:

$$f_{error} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} (E_i^{HDNNP} - E_i^{ref})^2 \quad (2.27)$$

The equation would be different if forces are also trained. It is to be noted that not all structures in the reference data go into training. A small portion of structures are chosen as testing data, usually 10%.

The minimisation is performed using the adaptive, global, extended Kalman filter [101]. The obtained HDNNP needs to be evaluated to see its quality. In case of the HDNNP not produce the desired output, the reference data can be extended which is referred to as active learning. The first validation step in assessing the quality of the HDNNP is root mean squared error (RMSE) which is computed both for energies and forces. It is defined as the average of root of sum of squared differences between HDNNP energies and reference method energies. RMSE is also defined for forces but here the the difference between each force component of a given structure is evaluated which is then squared and summed over all the force components of all the training structures and later taken mean root of.

$$RMSE(E^{train}) = \sqrt{\frac{1}{N_{train}} \sum_{i=1}^{N_{train}} (E_i^{HDNNP} - E_i^{ref})^2} \tag{2.28}$$

$$RMSE(f^{train}) = \sqrt{\frac{1}{N_{traincomp}} \sum_{i=1}^{N_{traincomp}} (f_i^{HDNNP} - f_i^{ref})^2} \tag{2.29}$$

There can be an issue of overfitting while using a Neural Network Potential. It is important to detect overfitting. The RMSE of training data is not a good indicator of the quality of the potential. As what is wanted is as low an RMSE as possible, it could lead to a situation where the training data is very well learned by the Neural Network but it is unable to predict a structure which is not in the training dataset. This is why the above mentioned testing data becomes important. The randomly selected testing data and its RMSE helps to avoid the pitfalls of overfitting. If the RMSE of the testing data (calculated in the same manner as testing), decreases and then later increases with a number of iterations (epochs), it is indicative of overfitting. Then, it is possible to choose the potential such that both the training and testing RMSEs are reasonable and can produce a potential which predicts energies or forces outside of the training data.

2.6 Molecular Dynamics

Molecular dynamics studies how a molecule or atom evolves with time. A given forcefield determines the interatomic potential. The trajectory of the dynamic evolution is governed by Newton's equations of motion.

$$F_{iq} = m_i a_{iq} \tag{2.30}$$

F_{iq} is obtained by multiplying mass of an atom m_i with the acceleration along q coordinate, a_{iq}. The force component can be obtained by negative derivative of the energy and as such depends on the interatomic potential. It is impossible to solve the problem analytically when the system size is large and Molecular Dynamics uses numerical methods to propagate the system through time.

The positions and accelerations at time t is used to calculate new positions after a time step dt. Various integrator methods are used to propagate the system through finite time steps. One such example is the Velocity Verlet algorithm.

$$q_i(t+dt) = q_i(t) + v_i(t)dt + \frac{f_i(t)}{2m_i}dt^2 \tag{2.31}$$

$$v_i(t+dt) = v_i(t) + \frac{f_i(t) + f_i(t+dt)}{2m_i}dt \tag{2.32}$$

Here, v_i is the velocity of an atom and q_i is its position. Once the new position q_{t+dt} is determined, the forces at the position are calculated using the negative gradient of the potential. And then the velocities are updated as well. Velocity Verlet has numerous advantages and is hence widely used. It is numerically stable and also time reversible without being more computationally costly compared to some other methods.

2.7 Theoretical Vibrational Spectroscopy

2.7.1 Harmonic Frequencies

Vibrational Spectra of polyatomic molecules can be described in terms of harmonic oscillator approximation. For harmonic oscillator description, normal coordinates are to be implemented.

For a molecule having N atoms, $3N$ coordinates are required to define its position. Removing the translational degrees of freedom and rotational degrees of freedom, a non linear molecule would have $3N-6$ vibrational degrees of freedom. It is $3N-5$ for a linear molecule. Each of the vibrational degrees of freedom is associated with a frequency and the vibration is called the normal mode of vibration. The potential energy of a molecule in principle is a function of the N_{vib} vibrational coordinates and can be expanded using Taylor series.

$$\begin{aligned} V(q_1, q_2..q_{N_{vib}}) &= V(0,0,...,0) + \frac{1}{2}\sum_{i=1}^{N_{vib}}\sum_{j=1}^{N_{vib}} \frac{\partial^2 V}{\partial q_i \partial q_j} q_i q_j + \cdots \\ &= \frac{1}{2}\sum_{i=1}^{N_{vib}}\sum_{j=1}^{N_{vib}} f_{ij} q_i q_j + \cdots \end{aligned} \tag{2.33}$$

Here, $V(0,0,...,0)$ is the potential energy at equilibrium. The first order derivative at equilibrium is 0 and hence that term is not included in the equation. f_{ij} is the generalised force constants. For the harmonic oscillator approximation, only up to the second derivative of the equation is considered which simplifies the equation considerably. Yet, the cross terms make finding the solution to the Schrödinger equation very difficult. Therefore, a new coordinate is introduced known as the normal coordinate Q_i which are basically mass weighted displacements. The normal coordinate is often scaled by a factor to make it dimensionless for simpler expression of energy which is known as dimensionless normal coordinate. Each normal coordinate corresponds to a normal mode of vibration. This enables to write the potential energy as :

$$V(Q_1, Q_2..Q_{N_{vib}}) = \frac{1}{2}\sum_{j=1}^{N_{vib}} F_j Q_j^2 \tag{2.34}$$

The vibrational Hamiltonian including both kinetic energy and potential energy terms turn out to be:

$$\hat{H}_{vib} = \sum_{j=1}^{N_{vib}} \left(-\frac{\hbar^2}{2\mu}\frac{\mathrm{d}^2}{\mathrm{d}Q_j^2} + \frac{1}{2}F_j Q_j^2 \right) = \sum_{j}^{N_{vib}} \hat{H}_{vib,j} \tag{2.35}$$

The vibrational wavefunction can be written as:

$$\psi_{vib}(Q_1, Q_2.., Q_{N_{vib}}) = \psi_{vib,1}(Q_1)\psi_{vib,2}(Q_2)\cdots\psi_{vib,N_{vib}}(Q_{N_{vib}}) \tag{2.36}$$

Solving the Schrödinger equation, the expressions of energy levels and wavefunction can be determined.

$$E_n = \hbar\omega(n + \frac{1}{2}) \tag{2.37}$$

Here, $n = 0, 1, \cdots$ which indicate the vibrational level with $n = 0$ being ground level. At $n = 0$ for all the modes, the system will still have some energy. This is called zero-point energy. And ω is the vibrational harmonic frequency of a particular mode. The wavefunction for a given vibrational level is:

$$\psi_n(Q) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} \frac{1}{\sqrt{2^n n!}} H_n e^{-\frac{m\omega Q^2}{2\hbar}} \left(\sqrt{\frac{m\omega}{\hbar}} Q\right) \tag{2.38}$$

Here, H_n is the hermite polynomial. Harmonic Oscillator model has a number of inadequacies. It cannot model bond dissociations. It also cannot describe overtones or hotbands and as such anharmonicity should be included to get accurate depiction of vibrational spectra.

2.7.2 Second-Order Vibrational Perturbation Theory

Second-order Vibrational Perturbation Theory (VPT2) [102, 103] is a widely used method to include anharmonicity in molecular vibrations. Anharmonicity is how molecular vibrations differ from harmonic oscillator model. In VPT2, the anharmonicity is added as a small perturbation ($\hat{H}_{anh}$) to a Hamiltonian whose eigenfunctions are known. Here, the zeroth-order Hamiltonian is the harmonic oscillator Hamiltonian (H_{HO}). Therefore,

$$\hat{H}_{VPT2} = \hat{H}_{HO} + \hat{H}_{anh} \tag{2.39}$$

The anharmonic potential is a Taylor series expansion of electronic energy in the normal coordinate. The terms with the third and fourth derivative of the electronic energy is considered. The force constants associated with these derivatives are called as cubic force constants and quartic force constants.

$$\hat{H}_{anh} = \frac{1}{3!}\sum_{ijk} F_{ijk} Q_i Q_j Q_k + \frac{1}{4!}\sum_{ijkl} F_{ijkl} Q_i Q_j Q_k Q_l \tag{2.40}$$

Solving this, the vibrational energy for a given vibrational level n is as follows:

$$E_n = E_0 + \sum_i \omega_i \left(n_i + \frac{1}{2}\right) + \sum_i \sum_{j \geq i} x_{ij} \left(n_i + \frac{1}{2}\right) \left(n_j + \frac{1}{2}\right) \tag{2.41}$$

ω_i is the harmonic frequency. x_{ij} are anharmonic constants.

2.7.3 Variational Vibrational Computation

Variational calculations are more expensive than other anharmonic treatments. It involves various steps including selection of a coordinate type like normal or curvilinear coordinates. A potential energy surface is also required. The Kinetic energy operator is defined and the wavefunction is expanded as a linear combination of basis functions. The nuclear Schrödinger equation is solved variationally. Though limited by the quality of the potential energy surface, this method can give the best possible solution to the problem at hand. It is very expensive and limited to very small molecules.

Chapter 3

Computational Details

3.1 Molpro

The reference data is constructed using Molpro 2015 for Formic Acid Monomer and Molpro 2019 for Formic Acid Dimer [104]. The various versions of Molpro were used to be consistent while extending the dataset.

The explicitly correlated CCSD(T)-F12c [105] with a cc-pVTZ-F12 basis set [106] was used for the reference calculations for Formic Acid Monomer. The Hessian was constructed numerically using default values. The default for energy convergence is ENERGY=10−6 and for the gradient, GRADIENT=3.10^{-4}. The default for the neglect of two electron integrals is TWOINT=10^{-14}. The default Hessian step size is 0.01 Bohr in Molpro.

The calculations for Formic Acid Dimer was performed using explicitly correlated CCSD(T)-F12a [107–109] using aug-cc-pVTZ [110] basis set for Carbon and Oxygen and cc-pVTZ [111] basis set for Hydrogen. This basis set will be addressed as haTZ. For the resolution of identity approximation, a VTZ/JKFIT [112] basis has been used. There are two settings used in the benchmarking for Formic Acid dimer, a default setting and a tight setting. The tight setting has the following conditions: a tighter threshold for two electron integrals (TWOINT=10^{-16}) and energy (ENERGY=10^{-10}). It also used tighter convergence criteria for geometry optimisation: GRADIENT=10^{-6} and STEP=10^{-6}. It also uses a four point numerical gradient with a step size of 0.005 Bohr for optimization and numerical Hessian.

3.2 RuNNer

The HDNNPs were constructed using the in-house software RuNNer [19, 20]. The ACSFs [17] are listed in the appendix. For the training 90% of reference data goes into the training set and the rest into the testing set. These data are selected randomly based on a random seed initially specified. The optimisation utilises a global adaptive extended Kalman filter [101]. The hidden layers utilises a hyperbolic tangent function as the activation function whereas the output layer uses a linear activation function. For both Formic Acid Monomer and Formic Acid Dimer, the energies alone are trained based on reference theoretical method.

The HDNNP for Formic Acid Monomer has a cutoff radius of 12 Bohr. Every

HDNNP here has two hidden layers with a maximum of 10 nodes per layer. The HDNNP for Formic Acid Dimer has a cutoff radius of up to 15 Bohr which includes all the interatomic distances in the dataset. The number of nodes per layer is up to 18 and there are two hidden layers for all the HDNNPs discussed here.

3.3 Construction of the HDNNP

Construction of the HDNNP is performed using the flowchart given in Fig. 3.1. This is called active learning [113–117]. The motivation behind active learning is that randomly adding data points to improve a potential leads to waste of computational resources and redundant data points in the dataset. When the computation of each data point is expensive, it is desirable to avoid the redundancy. It is important to add the most relevant data points to get a good potential.

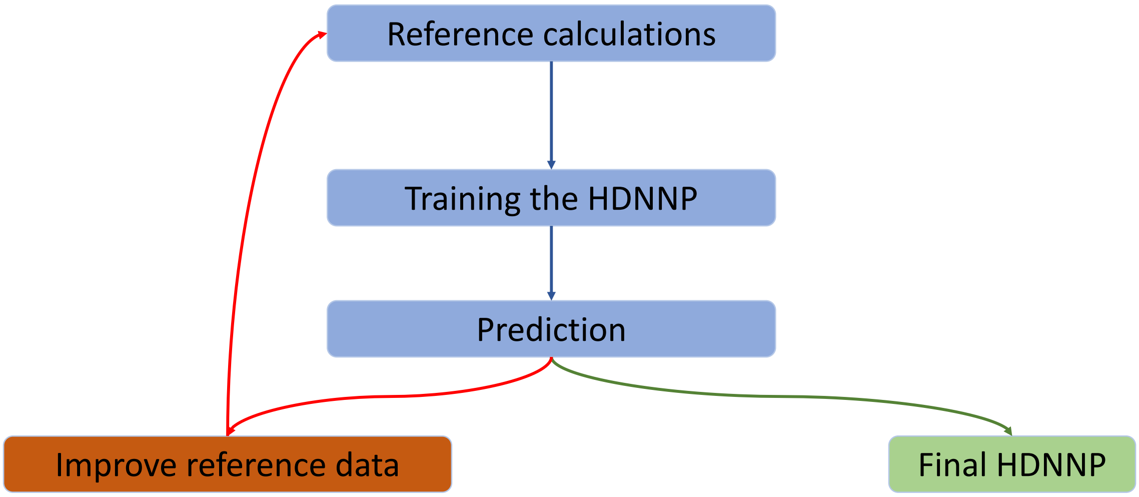


Figure 3.1: The figure explains the basic outline of the construction of the HDNNP. There is an initial reference data based on DFT or wavefunction based methods. The energies and forces in the reference data is used to train the HDNNP. The prediction or validation step showcases the quality of the potential for the given system. If the potential which is assessed by numerous methods such as RMSE is of sufficient quality, it is selected. If not, poorly described regions of the potential need to improved and geometries in these regions are probed and added to the reference data. Then, the same process repeats until a good quality HDNNP is obtained.

Here, in the case of Formic Acid Monomer and Formic Acid Dimer, the focus is on obtaining a global potential energy surface and accurate vibrational frequencies. For this active learning is employed by probing relevant areas of the potential energy surface by two best HDNNPs of a particular iteration. If the two HDNNPs differ in energy more than a predefined threshold for a given structure in the probed area, the structure is added to the reference data. The structures are probed using various sampling methods through various iterations. After this, the new HDNNP is validated using various parameters like RMSEs or harmonic frequencies. If the HDNNP does not meet the standard, the cycle is repeated.

3.4 Finite Difference

The harmonic frequencies from the HDNNPs were performed at the equilibrium structure of the corresponding potential. Finite difference was used to construct the Hessian and obtaining the harmonic frequencies. Hessian matrix is a $3NX3N$ matrix where each element is the second derivative of energy. N is the number of atoms in a system. The Hessian matrix is based on Cartesian coordinates. Each element of the Hessian matrix is as follows:

$$H_{i,j} = \frac{\partial^2 E}{\partial x_i \partial x_j} \tag{3.1}$$

In places where analytical gradients and double derivatives are not available, finite difference is used. Here, for the calculation of harmonic frequencies, the central differences method is used. Each atom is displaced by a small displacement Δx_j along an atomic coordinate x_i, then each element of the Hessian is:

$$H_{i,j} = \frac{(\frac{\partial E}{\partial x_i})_{0.5\Delta x_j} - \frac{\partial E}{\partial x_i})_{0.5} - \Delta x_j}{\Delta x_j} \tag{3.2}$$

Diagonalisation of the mass weighted Hessian provides eigen values from which the frequencies can be calculated.

3.5 Molecular Dynamics Simulations

Molecular Dynamics (MD) Simulations were performed for the validation step in active learning with intermediate HDNNPs using n2p2 [118] and LAMMPS [119]. For Formic Acid Dimer, the MD simulations were performed at 100K and 300K.

Chapter 4

Results and Discussion

4.1 Formic Acid Monomer

Formic Acid Monomer is the initial system chosen for Neural Network potential construction for the theoretical calculation of vibrational frequencies. Being a simple system with a few published potentials [57], it is an interesting benchmark as it is studied extensively in theory and experiments. Not only that, it is a good system to build up to Formic Acid Dimer. HDNNPs can be constructed for any system and hence offers and advantage over potentials constructed with a particular system in mind. Hence, it would be interesting to see the capabilities of a High-Dimensional Neural Network in describing vibrational frequencies and also benchmark them against available potentials.

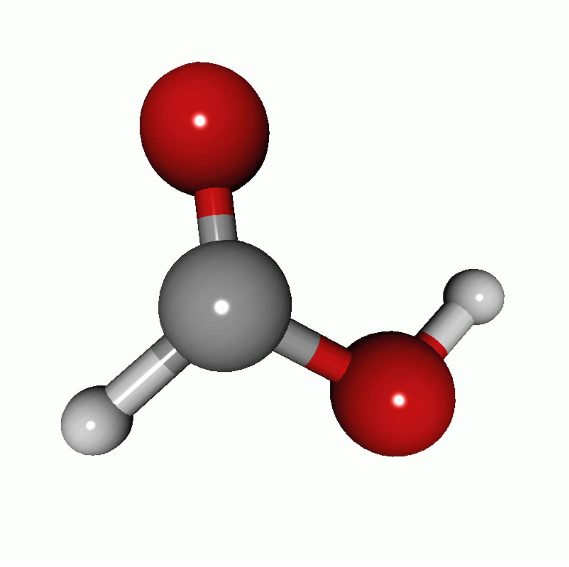


Figure 4.1: Schematics of trans Formic Acid Monomer

The initial dataset was obtained from David Tew [57]. It contains 23985 structures using which David Tew constructed an analytical potential with extremely accurate vibrational harmonic frequencies. It is interesting to see how the HDNNP would fare against such a potential. A comparison between the HDNNP and the Tew Potential will be shown in a later section. Also for this system, the goal is to construct a potential with a maximum of 10 cm^{-1} deviation for each vibrational mode for harmonic frequencies.

The reference data is constructed using CCSD(T)-F12c/cc-pVTZ-F12. For constructing the HDNNP, 10% of the structures were used for testing purposes and the rest of the structures are employed for training. Two hidden layers are used for the architecture of the HDNNP and a cutoff radius of 12.0 Bohr is used which covers all the interatomic distances. The reference data does not contain forces and as such only energy training is used to construct the potential.

4.1.1 Initial HDNNP

The initial HDNNP was constructed with the reference data used by David Tew for his potential [57]. Of course, it is not necessary that the same data would work for the Neural Network. However, it is important to get a feeling on the performance of the first HDNNP, focusing on its shortcomings and their solutions.

The first HDNNP had 7 nodes per layer. It has 21577 structures in the training set and 2408 structures in the testing set. The RMSEs are given in Table 4.1. The obtained RMSE values are around 1 meV/atom which is usually expected for a good fit. The training RMSE is 1.534 meV/atom and the testing RMSE is 1.821 meV/atom. We can also observe that both the RMSEs are of similar magnitude which is a sign that there is no overfitting. Still, it is better to have a lower RMSE value because then it is evident that the HDNNP value for energies are much closer to the reference energy values.

Table 4.1: Energy root mean squared errors (RMSE) of the training and test sets for the HDNNP trained using initial dataset for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
FAM-HDNNP1	23985	1.534	1.821	123.7	146.9

Fig. 4.2 shows the difference in energy between Coupled Cluster and FAM-HDNNP1 for each structure in the dataset. The training energies are in blue and the

Table 4.2: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for FAM-HDNNP1 with frequencies from CCSD(T)-F12c/cc-pVTZ-F12 level of theory.

Mode	FAM-HDNNP1	CCSD(T)-F12c/cc-pVTZ-F12
1 (ω_7)	641.6	632.50
2 (ω_9)	683.9	672.68
3 (ω_8)	1081.6	1056.78
4 (ω_6)	1146.4	1140.07
5 (ω_5)	1329.5	1318.39
6 (ω_4)	1422.7	1409.98
7 (ω_3)	1812.8	1816.76
8 (ω_2)	3102.2	3092.82
9 (ω_1)	3780.7	3765.31

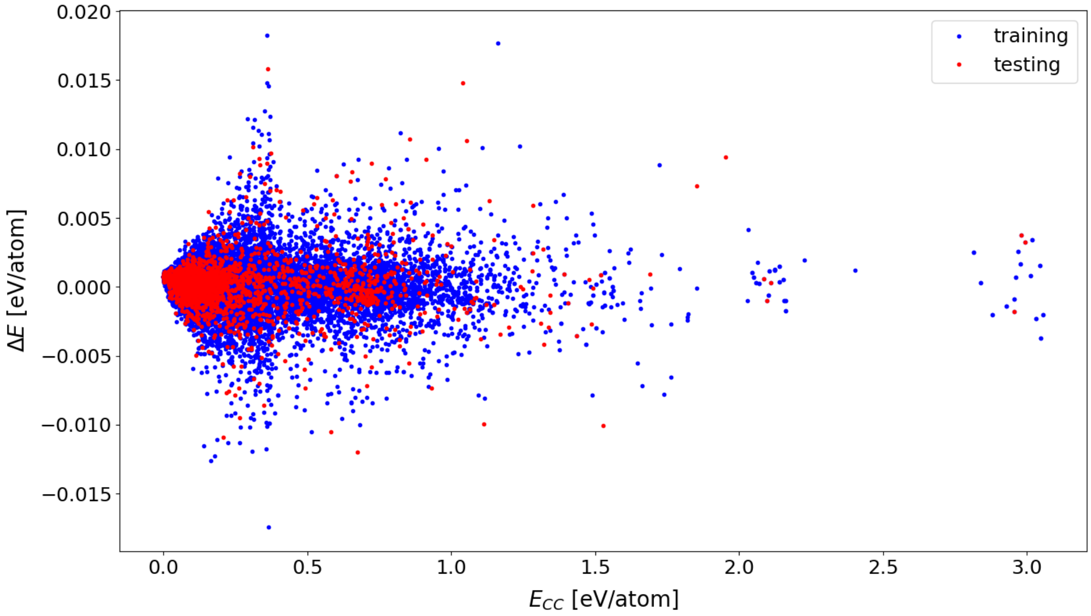


Figure 4.2: Energy difference $\Delta E = E_{CC} - E_{FAM\text{-}HDNNP1}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP are provided in Table 4.1

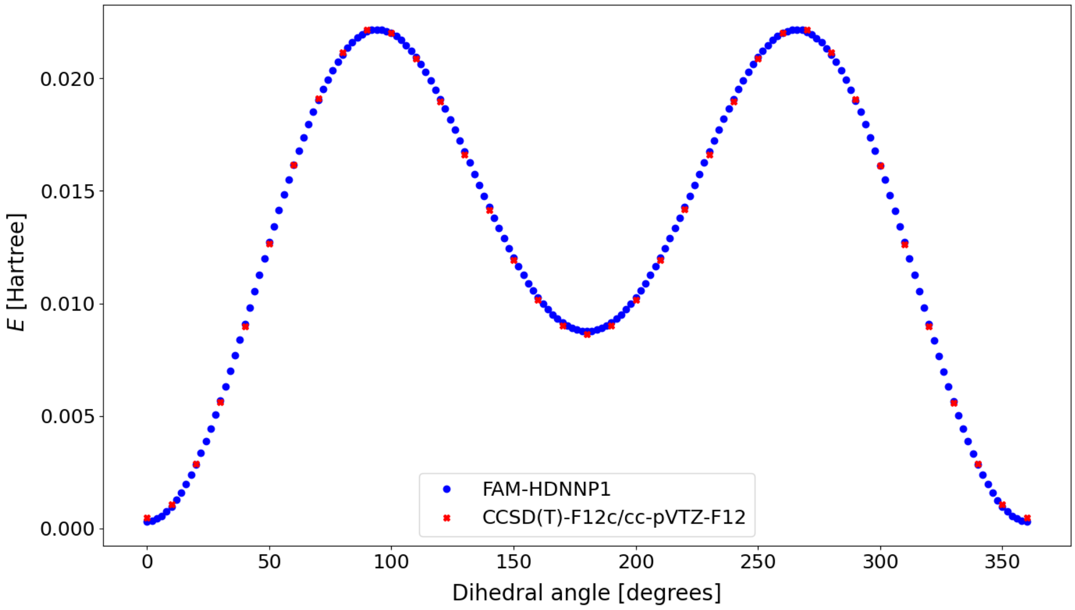


Figure 4.3: The energies according to Coupled Cluster and FAM-HDNNP1 along rotation of C-O bond of Formic Acid Monomer

testing energies are in red. We observe that most structures fall between 5 meV/atom deviation. There are a few structures that have very high deviation which are in-

Table 4.3: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for FAM-HDNNP1 with frequencies from CCSD(T)-F12c/cc-pVTZ-F12 level of theory.

Mode	FAM-HDNNP1
1 (ω_7)	-9.1
2 (ω_9)	-11.22
3 (ω_8)	-24.82
4 (ω_6)	-6.33
5 (ω_5)	-11.11
6 (ω_4)	-12.72
7 (ω_3)	3.96
8 (ω_2)	-9.38
9 (ω_1)	-15.39

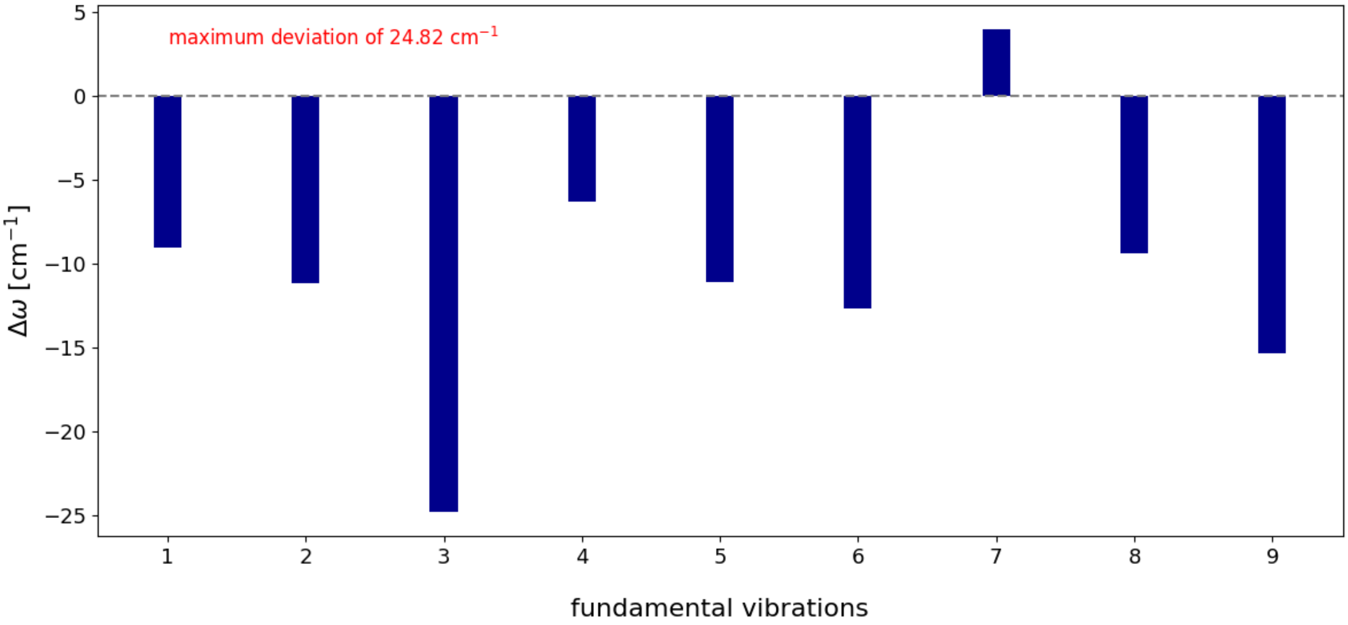


Figure 4.4: Deviations of the harmonic vibrational frequencies ω_i with respect to the reference CCSD(T)-F12c/cc-pVTZ-F12 frequencies. $\Delta\omega = \omega_{\text{CC}} - \omega_{\text{FAM-HDNNP1}}$

discriminate of training or testing. These deviations go up to 0.02 eV.atom. But, overall the spread of deviation is broad which is more indicative of the fit not being accurate yet.

Fig. 4.3 shows how the HCOOH rotation along the internal C-O bond varies the energies. The energies are shown for both Coupled Cluster and FAM-HDNNP1. We can see here that the HDNNP describes this well despite not yet at the best quality possible. Nevertheless a look at harmonic frequencies is incumbent here.

The harmonic frequencies according to FAM-HDNNP1 and CCSD(T)-F12c/cc-pVTZ-F12 are given in Table 4.2. There are 9 fundamental harmonic frequencies for Formic Acid Dimer. At the first glance, it can be observed that the frequencies from FAM-

HDNNP1 seem reasonable as the deviations are in the same order of magnitude as the threshold which is predefined. But, it can be seen that there are differences which are notable for the lowest vibrational mode for example. Mode 7 seem to be well described with FAM-HDNNP1.

Table 4.3 shows the deviation of harmonic frequencies from CCSD(T)-F12c/cc-pVTZ-F12 level of theory. We see that the highest deviation is for mode 3 with nearly 25 cm^{-1}. As this is also a low lying mode, the relative difference in frequencies is high. We also see that four modes of vibrations have a deviation below 10 cm^{-1}. It is also to be noted that the Root Mean Square Deviation (RMSD) for the frequencies is 12.9 cm^{-1}. Fig. 4.4 shows the deviations in the above table. This is pictorially represented in Fig. 4.4. As can be seen, there is much improvement needed in the Neural Network Potential as the maximum deviation and also the deviations of other modes are too high and it does not reproduce the frequencies to an acceptable degree.

The energy vs coordinate plot for nine normal modes of vibrations of Formic Acid Monomer is given in Fig. 4.5. The vibrations are from low frequency modes to high frequency modes. Even though the calculation of Hessian only uses small displacements and as such a small energy range, in order to obtain a global PES which can be used for high level vibrational computations and molecular dynamics, it is necessary to have a decent description of a large range in energetics. Though the energy range runs high, we can still observe certain deviations from the reference energies. This is especially visible in mode 6. In order to take a closer look, the difference in energy between HDNNP and Coupled Cluster is shown in Fig. 4.6. Here, we can observe that the energy difference is in hundreds of meV/atom. Nevertheless, the difference is much smaller in the energy range close to the equilibrium. There, it is below 50 meV/atom in most cases. The last two high frequency modes show an especially high deviation. But, as the energy range runs really high in the regions of high deviation, this is to be expected. Most of the structures in the reference data are in the range of 0.1 Hartree. Therefore, it is no surprise to have such high deviations. Yet, the fact that the normal modes show much deviation from the reference indicate that adding structures alongside them might help in improving the potential.

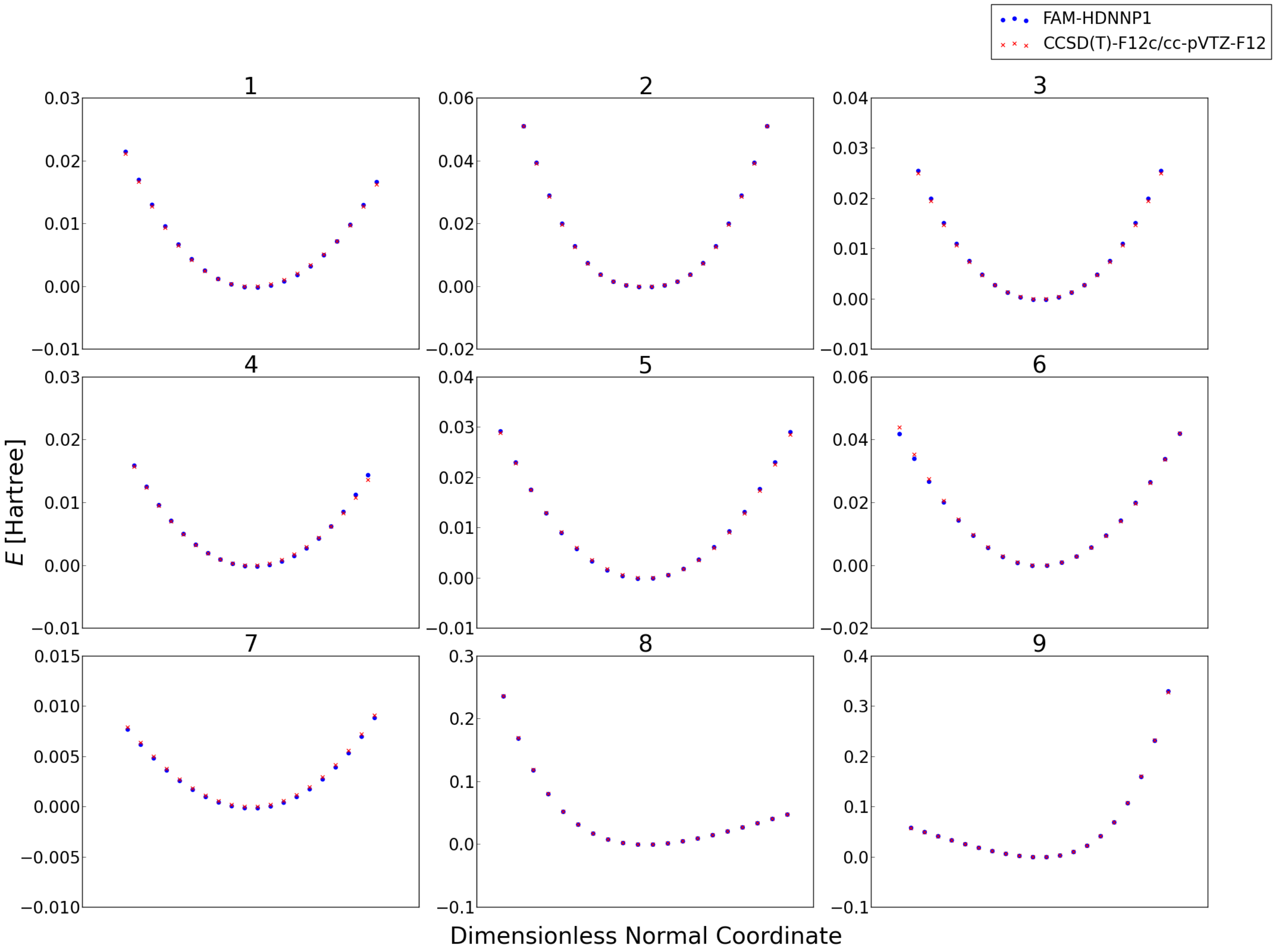


Figure 4.5: Nine normal modes of Formic Acid Monomer from low frequency vibrations to high frequency vibrations

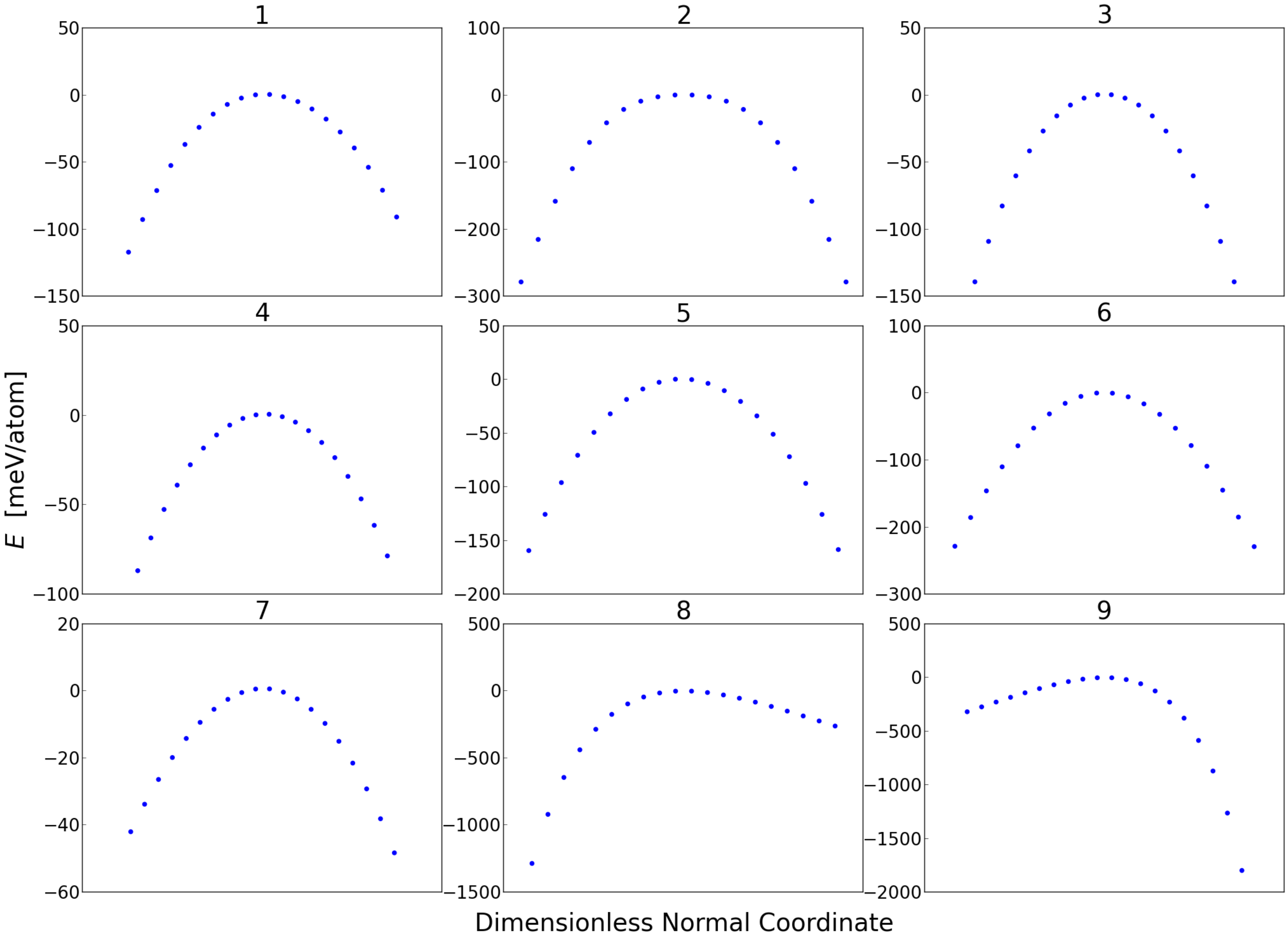


Figure 4.6: Energy difference $\Delta E - E_{\text{HDNNP}} = E_{\text{CC}}$ using FAM-HDNNP1 a for the nine normal modes of Formic Acid Monomer.

4.1.2 Final HDNNP

Table 4.4: Energy root mean squared errors (RMSE) of the training and test sets for the HDNNP trained using final dataset for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
FAM-HDNNP	25983	0.943	1.033	76.0	80.9

The final HDNNP was constructed by including additional 1998 structures along the normal modes of the Formic Acid Monomer. The Formic Acid Monomer was displaced along the nine normal modes one at a time using eigenvectors obtained from Coupled Cluster frequency calculation at the equilibrium structure. 23381 structures were used for training and 2602 structures were used for the testing set. The HDNNP was constructed with two hidden layers of 10 neurons each and the cutoff radius remains the same 12 Bohr. The quality of the fit in terms of RMSE is given in Table 4.4. We can see that the training RMSE is 0.943 meV/atom and the testing RMSE is 1.033 meV/atom. It can be observed that adding structures along the nine normal modes lead to an improvement from the previous iteration of the HDNNP for Formic Acid Monomer. Previously, the training and testing RMSEs were 1.534 and 1.821 meV/atom respectively. It is again a good sign that both the RMSEs are of the same order as it indicates lack of overfitting.

An important parameter to observe is how the energy of each structure looks like according to the HDNNP. Fig. 4.7 shows the difference in energy between HDNNP and Coupled Cluster. The spread of deviation is slightly less than the previous iteration as there the ΔE varied from -0.015 eV/atom to 0.02 eV/atom. Here, the spread is from -0.01 eV/atom to 0.02 eV/atom. Overall, we see that more structures fall in a narrower region of spread compared to the previous HDNNP thus indicating the improvement of energetics compared to previous HDNNP.

Since the potential is being developed for spectroscopic benchmarking, the look at frequencies is necessary. Table 4.5 shows the harmonic frequencies of HDNNP and Coupled Cluster in increasing order of wavenumbers. In the first glance, it is evident there is an improvement from the previous HDNNP. A closer look can be observed in Table 4.6. Here, we can observe that highest deviation is 7.71 cm^{-1} and this is also for the highest value frequency mode. As all the deviations which are also shown in Fig. 4.8 are below 10 cm^{-1}, the potential offers promise for spectroscopic applications. A discussion and comparison with Tew Potential will be shown in the next section. Six vibrational modes have deviation below 5 cm^{-1} with first, fifth, sixth and eighth mode being below 3 cm^{-1}. Also the RMSD of the frequencies from HDNNP while comparing to Coupled Cluster is 4.69 cm^{-1}. This is a good improvement from the previous HDNNP's RMSD which was 12.87 cm^{-1}.

Table 4.5: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for FAM-HDNNP with frequencies from CCSD(T)-F12c/cc-pVTZ-F12 level of theory.

Mode	FAM-HDNNP	CCSD(T)-F12c/cc-pVTZ-F12
1 (ω_7)	635.2	632.50
2 (ω_9)	676.9	672.68
3 (ω_8)	1051.9	1056.78
4 (ω_6)	1133.5	1140.07
5 (ω_5)	1316.1	1318.39
6 (ω_4)	1407.2	1409.98
7 (ω_3)	1811.5	1816.76
8 (ω_2)	3090.3	3092.82
9 (ω_1)	3757.6	3765.31

Table 4.6: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for FAM-HDNNP with frequencies from CCSD(T)-F12c/cc-pVTZ-F12 level of theory.

Mode	FAM-HDNNP
1 (ω_7)	-2.7
2 (ω_9)	-4.22
3 (ω_8)	4.88
4 (ω_6)	6.57
5 (ω_5)	2.29
6 (ω_4)	2.78
7 (ω_3)	5.26
8 (ω_2)	2.52
9 (ω_1)	7.71

1

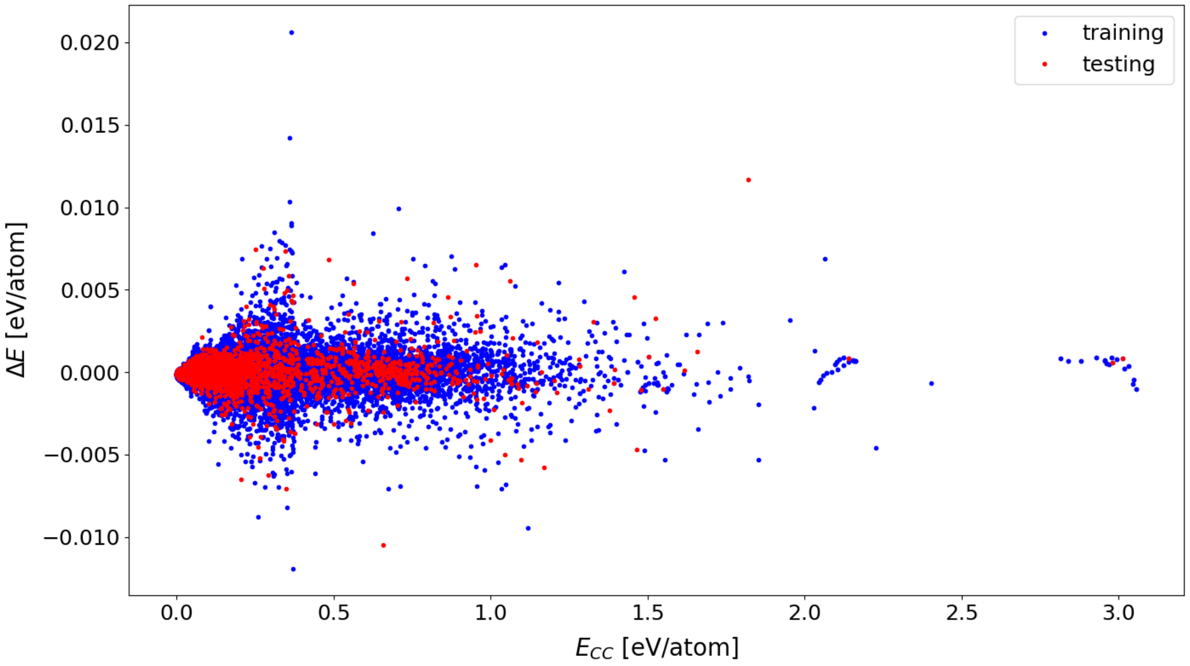


Figure 4.7: Energy difference $\Delta E = E_{\text{CC}} - E_{\text{FAM-HDNNP}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP are provided in Table 4.4

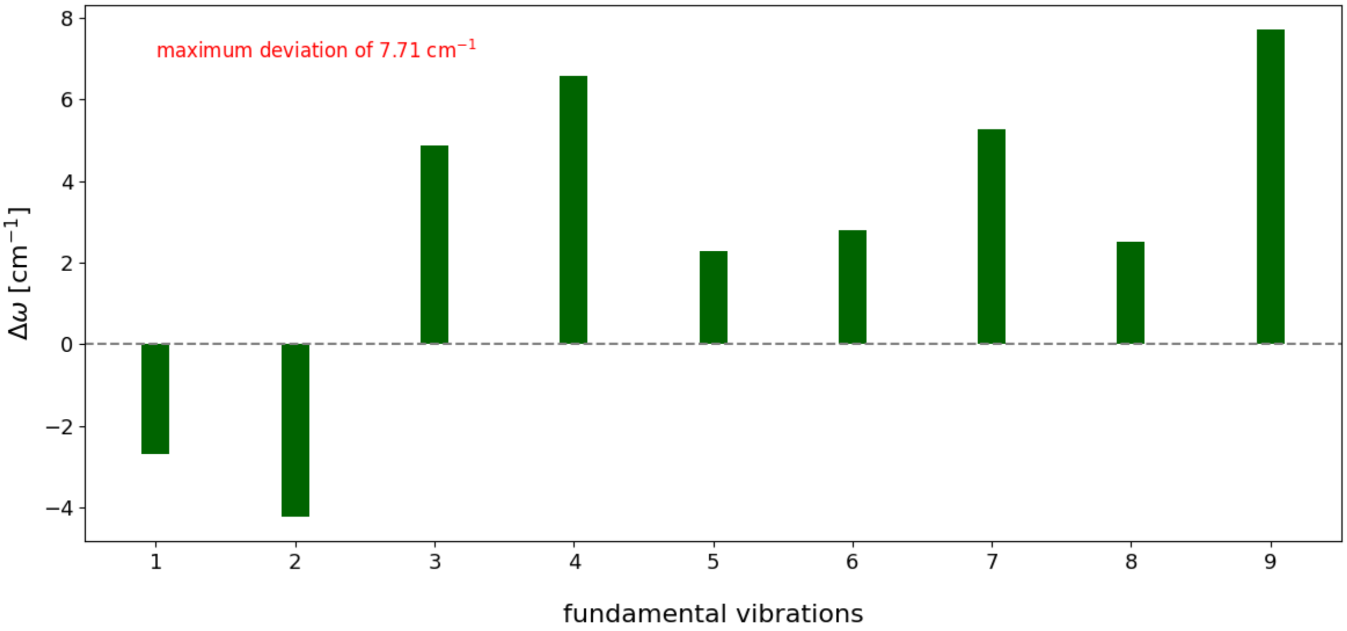


Figure 4.8: Deviations of the harmonic vibrational frequencies ω_i with respect to the reference CCSD(T) -F12c/cc-pVTZ-F12 frequencies. $\Delta\omega = \omega_{\text{CC}} - \omega_{\text{FAM-HDNNP}}$

4.1.3 Comparison with Tew Potential

The potential used by David Tew is an analytical potential. It is a LASSO-based regression model and has remarkable accuracy with an RMSE of 9 cm^{-1} [57] up to 15,000 cm^{-1} above equilibrium. The harmonic frequencies from all the potentials are given in Table 4.7. The comparison of Harmonic frequencies with respect to reference method is shown in Fig. 4.9. The tabulated deviations are given in Table 4.8 showing the two Neural Network potentials discussed above and the Tew potential are shown here. As we can observe the Tew Potential gives better frequencies with a maximum deviation of 4.61 cm^{-1}. The initial HDNNP had a maximum deviation of nearly 25 cm^{-1} while the final HDNNP improved to 7.71 cm^{-1}. This is higher than that of the Tew Potential which is shown in the figure. The Tew potential gives less than 1 cm^{-1} deviation for five vibrational modes. While comparing the RMSD of the harmonic fundamentals with respect to the CCSD(T)-F12c/cc-pVTZ-F12 fundamental harmonic frequencies, the initial HDNNP had an RMSD of 13 cm^{-1}, 5 cm^{-1} for the final HDNNP and 2 cm^{-1} for Tew Potential. Clearly the Tew Potential has a better performance. On the other hand, the HDNNP can still be improved if need be with more sampling or other criteria for fitting. Nevertheless, the current HDNNP shows good performance with deviations in the range of one digit wavenumbers for the fundamental harmonic frequencies.

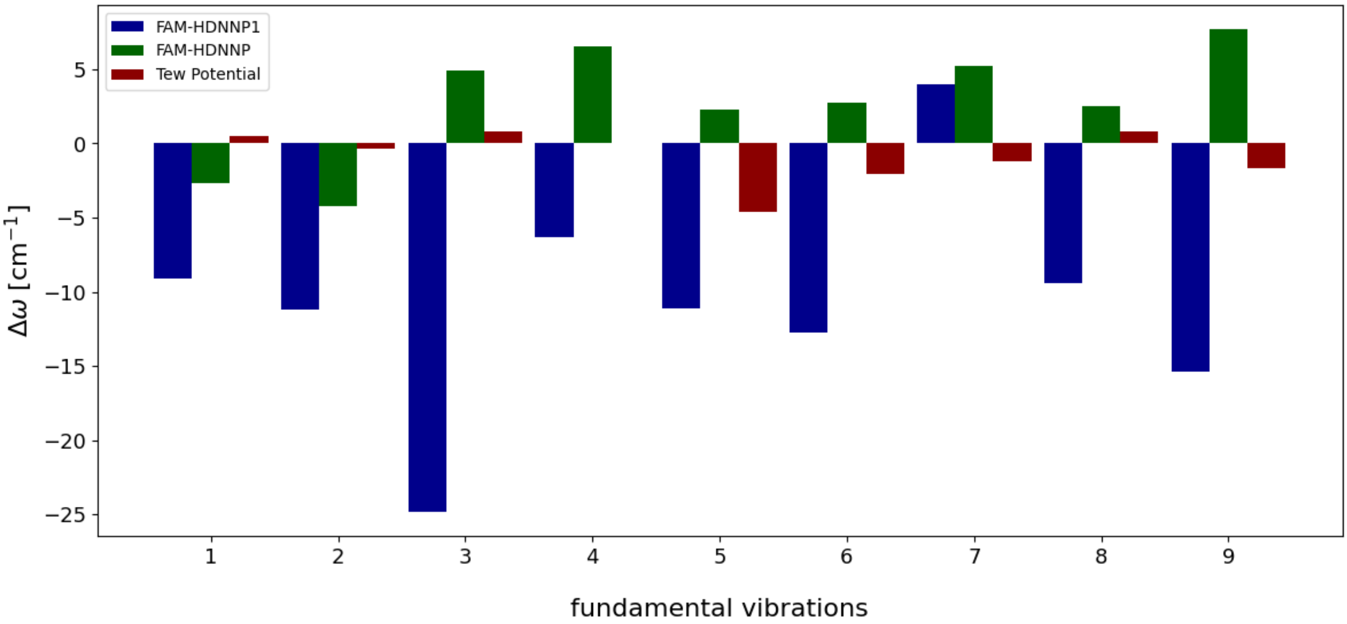


Figure 4.9: Deviations of the harmonic vibrational frequencies ω_i with respect to the reference CCSD(T) -F12c/cc-pVTZ-F12 frequencies. $\Delta\omega = \omega_{\text{CC}} - \omega_{\text{PES}}$

Table 4.7: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for different potentials with frequencies from CCSD(T)-F12c/cc-pVTZ-F12 level of theory.

Mode	Tew Potential	FAM-HDNNP1	FAM-HDNNP	CCSD(T)-F12c/cc-pVTZ-F12
1 (ω_7)	632	641.6	635.2	632.50
2 (ω_9)	673	683.9	676.9	672.68
3 (ω_8)	1056	1081.6	1051.9	1056.78
4 (ω_6)	1140	1146.4	1133.5	1140.07
5 (ω_5)	1323	1329.5	1316.1	1318.39
6 (ω_4)	1412	1422.7	1407.2	1409.98
7 (ω_3)	1818	1812.8	1811.5	1816.76
8 (ω_2)	3092	3102.2	3090.3	3092.82
9 (ω_1)	3767	3780.7	3757.6	3765.31

Table 4.8: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for various potentials from frequencies from CCSD(T)-F12c/cc-pVTZ-F12 level of theory.

Mode	Tew Potential	FAM-HDNNP1	FAM-HDNNP
1 (ω_7)	0.5	-9.1	-2.7
2 (ω_9)	-0.32	-11.22	-4.22
3 (ω_8)	0.78	-24.82	4.88
4 (ω_6)	0.07	-6.33	6.57
5 (ω_5)	-4.61	-11.11	2.29
6 (ω_4)	-2.02	-12.72	2.78
7 (ω_3)	-1.24	3.96	5.26
8 (ω_2)	0.82	-9.38	2.52
9 (ω_1)	-1.69	-15.39	7.71

4.2 Formic Acid Dimer

Formic Acid Dimer is an interesting system for construction of Potential Energy Surface with the intention of usage for spectroscopic applications. With the wealth of experimental data available and the system being at the forefront of development of Machine Learning Potentials for calculating vibrational frequencies, Formic Acid Dimer becomes a very convenient choice to construct High-Dimensional Neural Networks to benchmark vibrational frequencies.

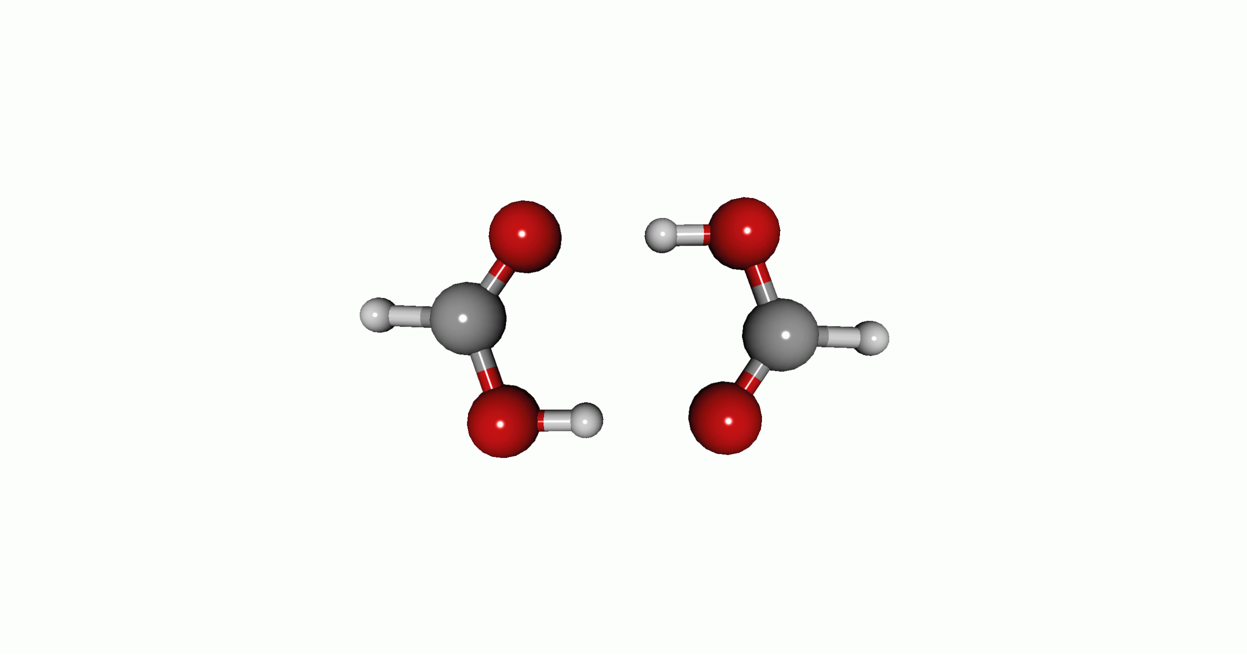


Figure 4.10: Schematics of Formic Acid Dimer

4.2.1 Construction of HDNNP

The efficient construction of an HDNNP requires a systematic approach. It begins with an initial dataset, the choice of settings that guides the construction of the Neural Network Potential and then assessment and analysis of the quality of the Neural Network Potential. In the case of a spectroscopic quality potential, the energetics alone cannot determine the choice of a final potential or even the method of refinement of the potential. The harmonic frequencies are a very important parameter that need to meet a standard. In the case of Formic Acid Dimer, along with the usual standard of what is expected as the standard of RMSE (1-2 meV/atom), we determine a maximum deviation of 10 cm^{-1} as what is required of the potential.

The construction of a potential that meets the requirements went through several iterations of assessment and improvement. Here, it starts with the analysis of the initial dataset of Formic Acid Dimer. The initial dataset was obtained from Joel Bowman [60] and contains 13475 structures of Formic Acid Dimer computed at CCSD(T)-F12a/haTZ level of theory.The energy distribution of the structures is such that the higher energy region is sparsely populated. Qu and Bowman have constructed a Permutationally Invariant Potential [60] using this dataset for Formic Acid Dimer which will be used for comparison with the HDNNP. This potential will be referred to as Bowman Potential or QB16.

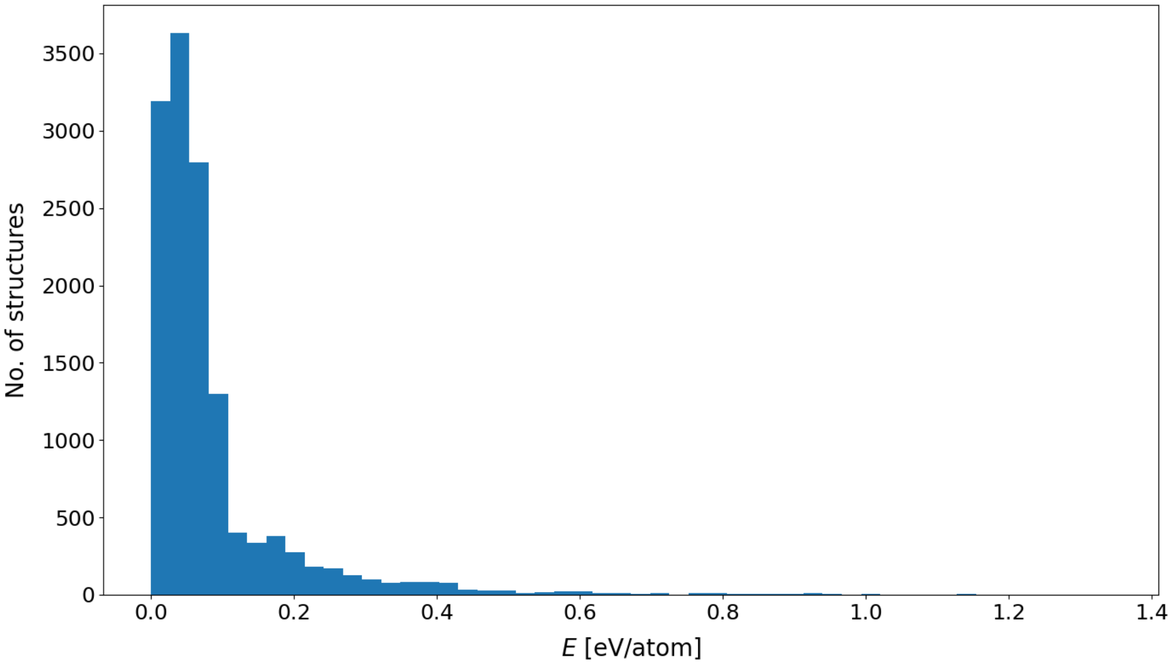


Figure 4.11: Energy distribution of Formic Acid Dimer structures in the initial dataset obtained from Prof. Joel Bowman

As seen in Fig. 4.11, most of the structures are below 0.2 eV/atom which is 16,000 cm^{-1}. As the Neural Network is heavily dependent on the underlying data, this can lead to poor description in the higher energy regions which might manifest as certain artefacts while assessing the quality of the constructed Neural Network Potential. Nevertheless, as we are aiming for spectroscopic accuracy, the lower energy regions or regions as high as 30,000 cm^{-1} are sufficient to be well described.

As discussed in the method section, the number of hidden layers is two for all the HDNNPs discussed here. The main differentiating factor for the construction of HDNNPs to arrive at a quality potential is the underlying dataset. Another differentiating factor is the architecture of the HDNNP. As the dataset gets bigger, in order to account for the necessary bigger flexibility, the Neural Network also often needs to be of bigger size.

4.2.2 HDNNP1

Here, the first reasonable HDNNP constructed using the raw data as shown in Fig. 4.11 is discussed. The HDNNP at this iteration will be referred to as HDNNP1 and the two HDNNPs discussed here will be HDNNP1a and HDNNP1b. Tables 4.9 and 4.10 give values for the parameters that define the Atom Centered Symmetry Functions (ACSFs) for HDNNP1.

Table 4.9: Radial ACSF parameters η for HDNNP1. The various parameters are described in the theory section.

element pair	η[Bohr^{-2}]
H-H	0, 0.003320, 0.007822, 0.014296, 0.024263, 0.040982, 0.072561, 0.144102
O-O	0, 0.002331, 0.005208, 0.008869, 0.013680, 0.020235, 0.029556, 0.043520
C-C	0, 0.000964, 0.002013, 0.003161, 0.004425, 0.005824
H-C	0, 0.003763, 0.009087, 0.017202, 0.030743, 0.056242, 0.113944, 0.295433
O-C	0, 0.003648, 0.008752, 0.016415, 0.028926, 0.051756, 0.100815, 0.240433
H-O	0, 0.003910, 0.009520, 0.018245, 0.033218, 0.062638, 0.134133, 0.395239

Table 4.10: Angular ACSF parameters η for all element combinations and HDNNPs.The various parameters are described in the theory section.

No.	η[Bohr^{-2}]	ζ	λ
1	0.0	1.0	1.0
2	0.0	2.0	1.0
3	0.0	4.0	1.0
4	0.0	16.0	1.0
5	0.0	1.0	-1.0
6	0.0	2.0	-1.0
7	0.0	4.0	-1.0
8	0.0	16.0	-1.0

The initial dataset was used to construct Permutationally Invariant Polynomials (PIPs) for Formic Acid Dimer [60]. As the same dataset is used to construct HDNNPs, it may not work as intended. Further refinement of the dataset is required. Even though both are Machine Learning algorithms, both are different approaches for learning the potential energy curve of a system and as such require a different method for selecting the data.

The two HDNNPs at this iteration are HDNNP1a and HDNNP1b. For both HDNNP1a and HDNNP1b, a cutoff radius of 14.901 Bohr is used. This is so that the longest atom-atom distance also falls under the definition of atomic environments. And as a result, we can be assured that all interatomic distances and interactions are learnt by the Neural Network.

Though various HDNNPs are constructed at this stage, HDNNP1a and HDNNP1b are selected because they describe the energetics of each structure in a dataset of 13,475 structures better than the other HDNNPs in this iteration. For HDNNP1a, 12138 structures are used for training and 1337 for testing. For HDNNP1b, 12062 structures went into training and 1413 structures were used for testing. Here, the structures are selected randomly for training or testing. The RMSE of HDNNP1a is 2.43 meV/atom for training and 13.99 meV/atom for testing whereas HDNNP1b has 2.75 meV/atom for training and 10.22 meV/atom for testing. For selecting an HDNNP at this stage, the energetics provide a reasonable indicator of the quality. But, here for a selection between HDNNP1a and HDNNP1b, we cannot rely on RMSE alone. As can be seen, HDNNP1a has a better training RMSE whereas HDNNP1b has a better testing RMSE. We need to take a look at how energies are described for each structure for narrowing down our selection further. Secondly, we need to look at harmonic frequencies which will be discussed later here. Because of the uneven distribution of structures over a wide energy range, RMSE though a very reasonable indicator, may not provide a full picture. The RMSEs of the HDNNPs are shown in Table 4.11.

HDNNP1a and HDNNP1b are constructed with the same dataset and definitions of Atom Centered Symmetry Functions. They differ in the initial weights used for optimisation as these are selected randomly. Another difference is the structures used for training and testing. But, the major differentiating factor between them is the architecture of the Neural Network. Both HDNNP1a and HDNNP1b have two hidden layers. HDNNP1a has 11 nodes per layer as opposed to 9 neurons for HDNNP1b. As can be immediately observed, HDNNP1a has greater flexibility than HDNNP1b.

Table 4.11: Energy root mean squared errors (RMSE) of the training and test sets for the HDNNPs trained using initial dataset for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
HDNNP1a	13475	2.43	13.99	196	1129
HDNNP1b	13475	2.75	10.22	222	824

Usually the expected RMSE value is 1 meV/atom. The training RMSEs are around 2-5 meV/atom here and the testing RMSE is a very high number, above 10 meV/atom. As the high energy regions are sparsely represented, a high testing RMSE is not unusual. Still, here we must take a look at the harmonic frequencies and see which of these HDNNPs describe it better and how good the values of frequencies are.

In Fig.4.12 and 4.13, the energetics of HDNNP1s are showing the training energies in blue and the testing energies in red. The energy units are in eV/atom. The training energies are shown in blue and the testing energies are shown in red. As can be also seen in the RMSEs, the training energies are very close to the CCSD(T)-F12a/haTZ energies. The testing energies show very high deviation above 0.2−0.3 eV/atom. This is well above 20,000 cm^{-1}. But we can also see that for certain structures the difference between Coupled Cluster and HDNNP values are above 0.1 eV/atom. Overall, this is not the quality we would want for the energetics. Even though, it is expected that the high energy region will not be well described in the PES, we still would want to have a narrower range of error. In order to do other applications like dynamics or even high level anharmonic calculations, a better accuracy is required.

Also in the figures, we can see how the energetics differ for HDNNP1a and HDNNP1b. We can see that the testing errors are very high for HDNNP1a compared to HDNNP1b. High energy structures in the testing set have double the deviation in some cases for HDNNP1a compared to HDNNP1b. This agrees with what we see in the testing RMSEs as HDNNP1a has a higher testing RMSE. But, in the case of both the HDNNPs, the training set shows relatively low error and also the lower energy regions are well described. Therefore, we can conclude that both the HDNNPs perform relatively similar if we are only looking at the energetics. And the selection at this stage would solely depend on the frequencies.

Table 4.12 shows the harmonic frequencies obtained from HDNNP1a and HDNNP1b along with harmonic frequencies from reference Coupled Cluster method. As we can see, HDNNP1a is closer to reference Coupled Cluster frequencies. But, in order to see this more closely we need to take a look at Table 4.13. The largest deviation for HDNNP1a harmonic frequencies is 48.88 cm^{-1} for ω_6. Also, ω_4, ω_{18}, ω_{20} and ω_{21} show deviations above 40 cm^{-1}. A few vibrational modes do behave well with this potential. ω_{12}, ω_{24}, ω_2, ω_7 and ω_8 are below 5 cm^{-1} deviation. But, still this is very far from the accuracy desired.

Now, we can take a look at how HDNNP1b behaves and see how we can select a potential at this stage. HDNNP1b has similar RMSE to HDNNP1a with a better error for the testing set. And it would be interesting to see how the quality of the frequencies compare. When we look at Table 4.13, we see that the maximum deviation for HDNNP1b while comparing to Coupled Cluster frequencies is 75.86 cm^{-1} for ω_{21}. This is quite a high deviation compared to HDNNP1a. We can also see that ω_4, ω_{10}, ω_{19} and ω_{20} have deviations above 50 cm^{-1}. Even though, the testing region is energetically better described and the HDNNP1b energies are also quite close to the Coupled Cluster energies, HDNNP1b does perform inferior to HDNNP1a when we look at the frequencies. Of course, having good energetics is the first step in selecting a potential, but for spectroscopic purposes, that alone is not a sufficient measure as can be seen here. Also, Fig. 4.14 gives a visual representation of the harmonic frequencies. It shows the difference between HDNNPs and Coupled Cluster harmonic frequencies. And there, the clear difference between the two HDNNPs can be seen. We can conclude that HDNNP1a is a better potential at this stage and also that further refinement of HDNNPs is required to obtain a good quality potential. Nevertheless, a first potential with a raw dataset does give reasonable numbers and provide motivation that a better potential can be obtained.

Table 4.12: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP1a and HDNNP1b with frequencies from CCSD(T)-F12a/haTZ level of theory.

Mode	Sym.	HDNNP1a	HDNNP1b	CCSD(T)-F12a/haTZ
ω_1	A_g	3217.5	3194.4	3203.36
ω_2	A_g	3107.9	3091.3	3104.59
ω_3	A_g	1708.4	1712.5	1717.13
ω_4	A_g	1443.2	1550.3	1483.92
ω_5	A_g	1376.3	1379.9	1413.1
ω_6	A_g	1207.9	1265.2	1256.78
ω_7	A_g	689.2	664.8	687.78
ω_8	A_g	216.1	221.1	211.28
ω_9	A_g	163.6	157.6	170.96
ω_{10}	B_g	1073.2	1030.6	1085.04
ω_{11}	B_g	936.9	965.3	959.6
ω_{12}	B_g	256.9	261.9	257.76
ω_{13}	A_u	1128.4	1069	1102.03
ω_{14}	A_u	979.5	1015.9	986.46
ω_{15}	A_u	213.4	210.9	185.95
ω_{16}	A_u	68.5	73.2	76.36
ω_{17}	B_u	3323.8	3285.3	3305.25
ω_{18}	B_u	3060.4	3087.4	3100.56
ω_{19}	B_u	1772.6	1728.9	1781.57
ω_{20}	B_u	1411.4	1517.2	1455.96
ω_{21}	B_u	1357.5	1329.2	1405.06
ω_{22}	B_u	1253.9	1295.4	1260.06
ω_{23}	B_u	703.9	733.5	715.81
ω_{24}	B_u	280.9	267.9	278.07

Table 4.13: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP1a and HDNNP1b from CCSD(T)-F12a/haTZ frequencies. $\Delta\omega = \omega_{\mathrm{CC}} - \omega_{\mathrm{HDNNP}}$

Mode	Sym.	HDNNP1a	HDNNP1b
ω_1	A_g	-14.14	8.96
ω_2	A_g	-3.31	13.29
ω_3	A_g	8.73	4.63
ω_4	A_g	40.72	-66.38
ω_5	A_g	36.8	33.2
ω_6	A_g	48.88	-8.42
ω_7	A_g	-1.42	22.98
ω_8	A_g	-4.82	-9.82
ω_9	A_g	7.36	13.36
ω_{10}	B_g	11.84	54.44
ω_{11}	B_g	22.7	-5.7
ω_{12}	B_g	0.86	-4.14
ω_{13}	A_u	-26.37	33.03
ω_{14}	A_u	6.96	-29.44
ω_{15}	A_u	-27.45	-24.95
ω_{16}	A_u	7.86	3.16
ω_{17}	B_u	-18.55	19.95
ω_{18}	B_u	40.16	13.16
ω_{19}	B_u	8.97	52.67
ω_{20}	B_u	44.56	-61.24
ω_{21}	B_u	47.56	75.86
ω_{22}	B_u	6.16	-35.34
ω_{23}	B_u	11.91	-17.69
ω_{24}	B_u	-2.83	10.17

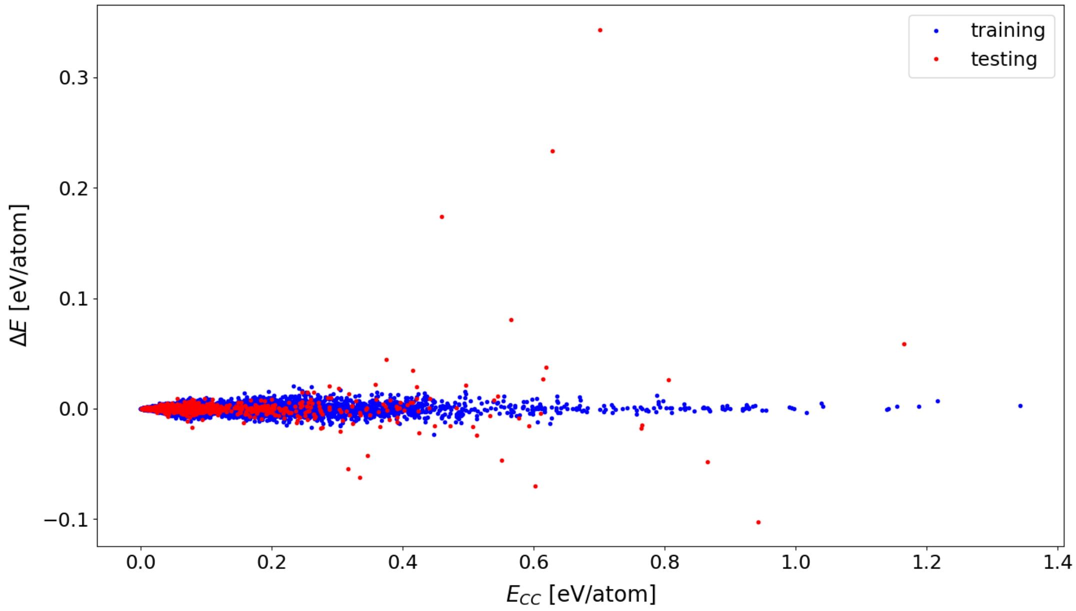


Figure 4.12: Energy difference $\Delta E = E_{\mathrm{CC}} - E_{\mathrm{HDNNP1a}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP1a are provided in Tab. 4.11.

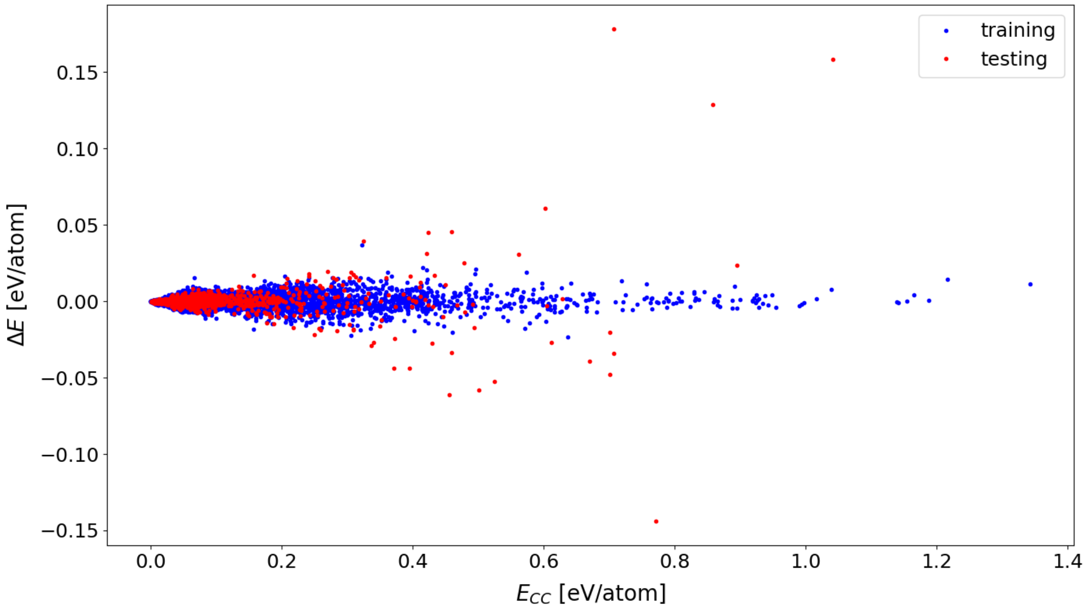


Figure 4.13: Energy difference $\Delta E = E_{\mathrm{CC}} - E_{\mathrm{HDNNP1b}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP1b are provided in Tab. 4.11.

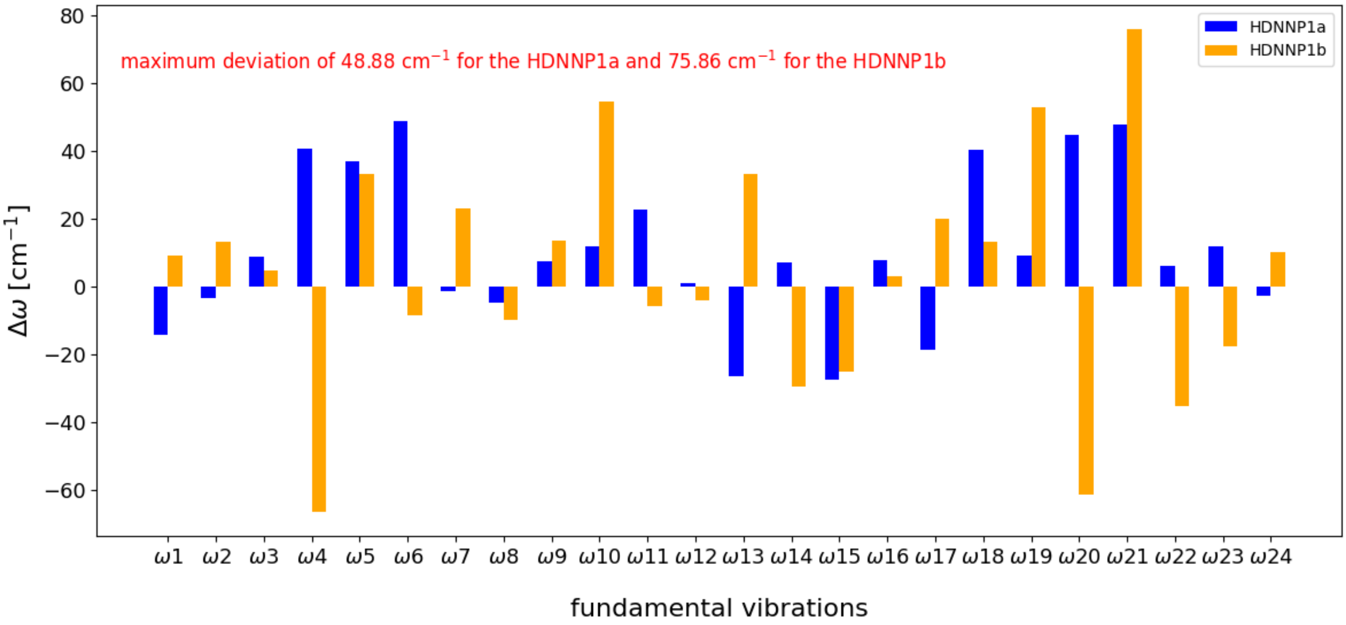


Figure 4.14: Deviations of the harmonic vibrational frequencies ω_i with respect to the reference CCSD(T)-F12a/haTZ frequencies. $\Delta\omega = \omega_{\text{CC}} - \omega_{\text{HDNNP}}$ for different HDNNPs

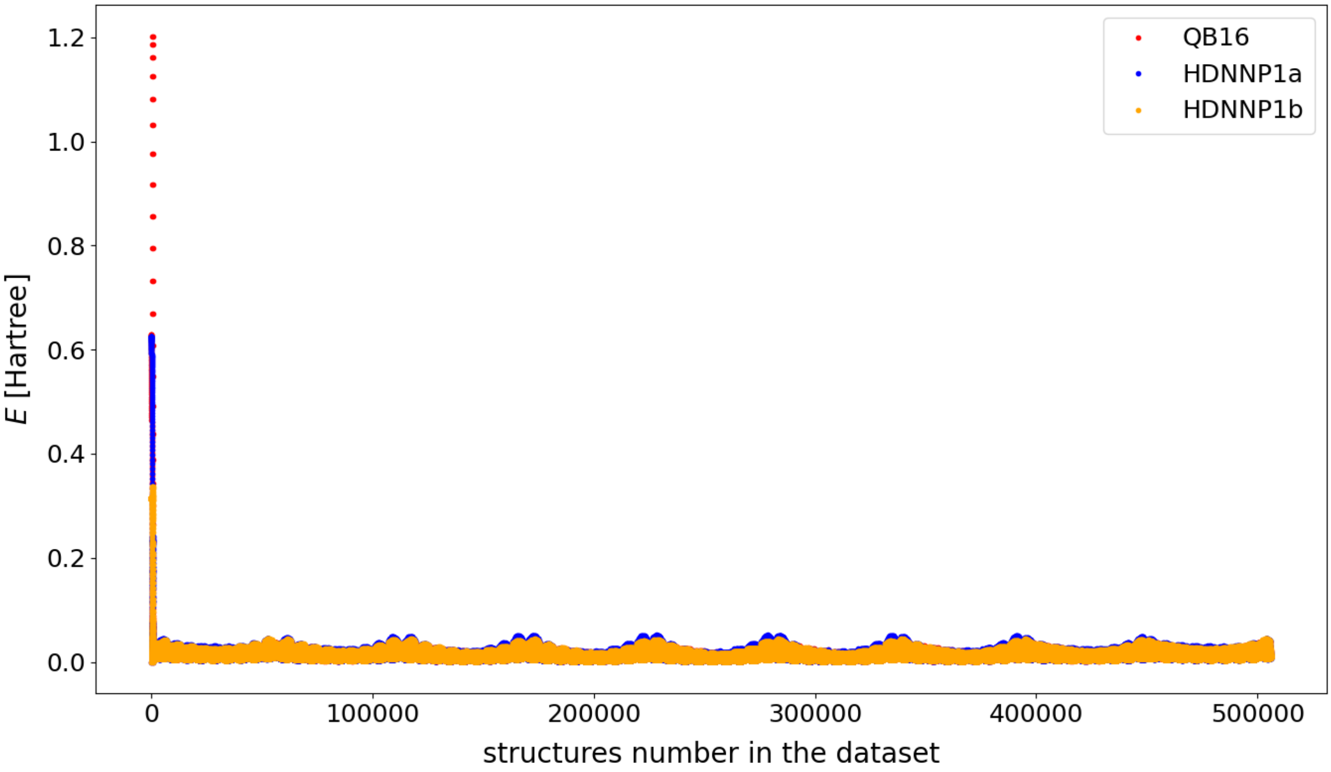


Figure 4.15: Bowman Potential and HDDNP1 energies for all the structures from the dataset obtained from Edit Matyus [79]

4.2.3 Development of HDNNP2

In order to develop the next iteration of HDNNP which will be henceforth addressed as HDNNP2, several sampling methods and some modification to the description of Neural Network were used. They will be addressed one by one.

As a result of collaboration with Prof. Matyus, an expansive dataset of structures was obtained which contains 500,000 structures from direct product grid needed for variational vibrational calculations [79]. In the dataset, the energies were computed using QB16. These structures were used to improve the underlying dataset for developing a Neural Network Potential. Here, we use the technique of active learning to select the structures which will later be recomputed using ab-initio methods to add to the dataset.

The procedure is using HDNNP1a and HDNNP1b to predict the energies of these 500,000 structures. Then we take the difference between energies of HDNNP1a and HDNNP1b and choose structures where both the HDNNPs differ greatly. The philosophy behind this is that if both the HDNNPs trained with the same data identify the energies with a big deviation, this means the region is not reasonably sampled.

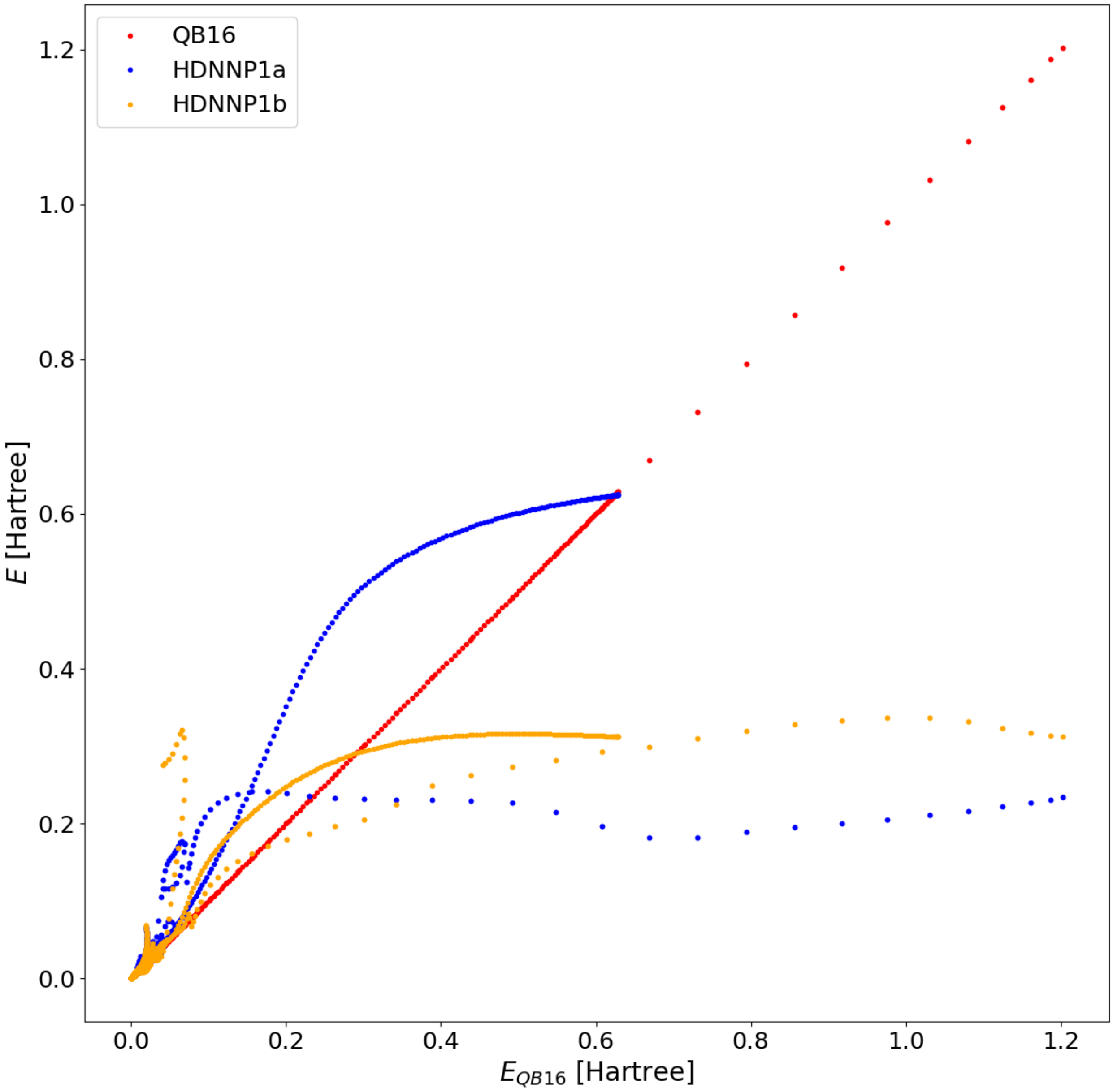


Figure 4.16: Bowman Potential and HDNNP1 energies for all the structures from the dataset obtained from Edit Matyus [79] using Bowman Potential energies as the x axis.

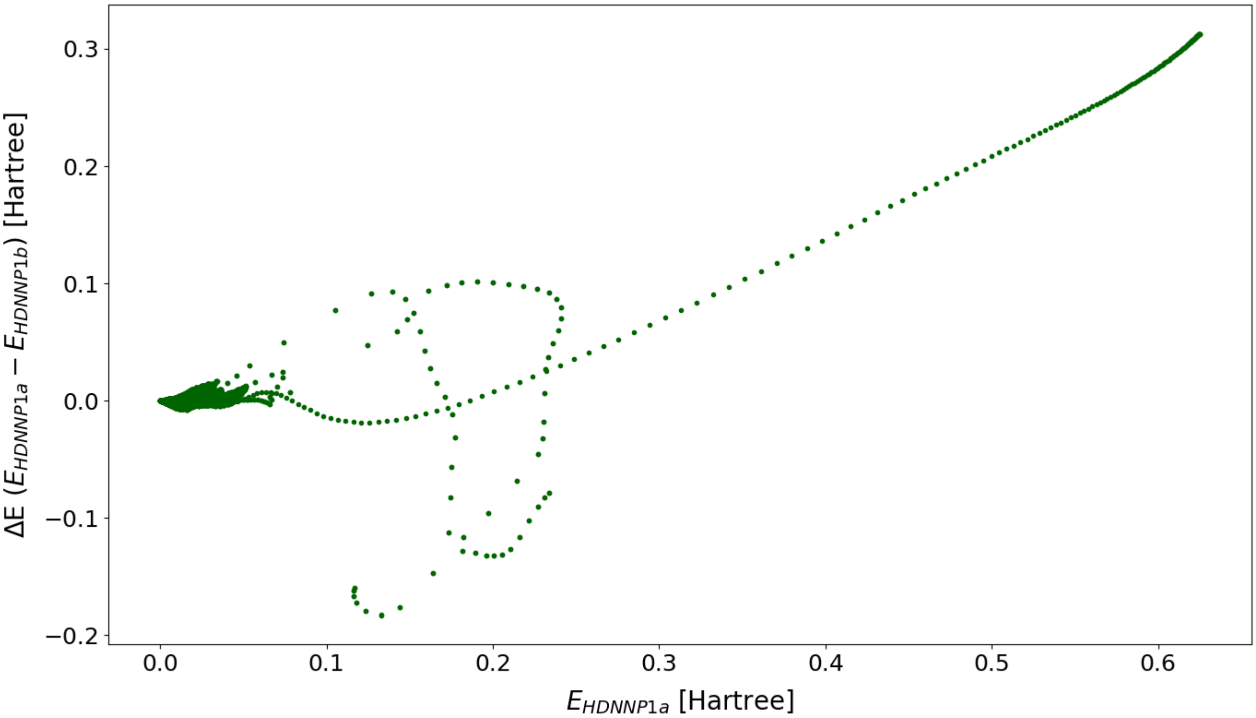


Figure 4.17: $\Delta E = E_{\mathrm{HDNNP1a}} - E_{\mathrm{HDNNP1b}}$ for all the structures from the dataset obtained from Edit Matyus [79]

Fig.4.15 shows the full dataset obtained from Edit. Here, the energies of the structures using Bowman's Potential, HDNNP1a and HDNNP1b are shown. Fig. 4.16 shows the same energies with respect to the Bowman Potential energies in the x-axis. Here, we can see that especially at high energy regions the two HDNNPs differ very much from Bowman Potential. As it cannot be deduced yet how accurate Bowman Potential is because of the sheer size of the dataset, we cannot yet infer how important the information is. But, if we look at the two HDNNPs, we see that they agree somewhat better with each other compared to Bowman Potential.

In order to see that closely, Fig. 4.17 becomes useful. Here, in Fig. 4.17, the full dataset with energy difference between the two HDNNPs is shown as a function of HDNNP1a energies. As we can see there are regions with very high deviation. Ab-initio data is not available for these structures because of the sheer size of the dataset. Also, we only need to sample structures in spectroscopically relevant energy range. Hence, we avoid the structures above 0.1 Hartree according to Bowman Potential as shown in Fig. 4.15.

In Fig. 4.18, the $\Delta E = E_{HDNNP1a} - E_{HDNNP1b}$ is shown for the relevant energy range. Here we can see that there are even structures with a deviation above 0.05 Hartree. And it becomes imperative to look at the distribution of ΔE since we cannot afford to run Coupled Cluster calculations for hundreds of thousands of structures. After looking at the energy distribution, we see that there are 2067 structures above 0.02 eV/atom, The distribution of energies with a $\Delta E > 0.02$ eV/atom is given in Fig. 4.19. This is a very high deviation but at the same time it is important not to enlarge the dataset very quickly as HDNNP should be constructed as efficiently as possible. These structures were added to the training data.

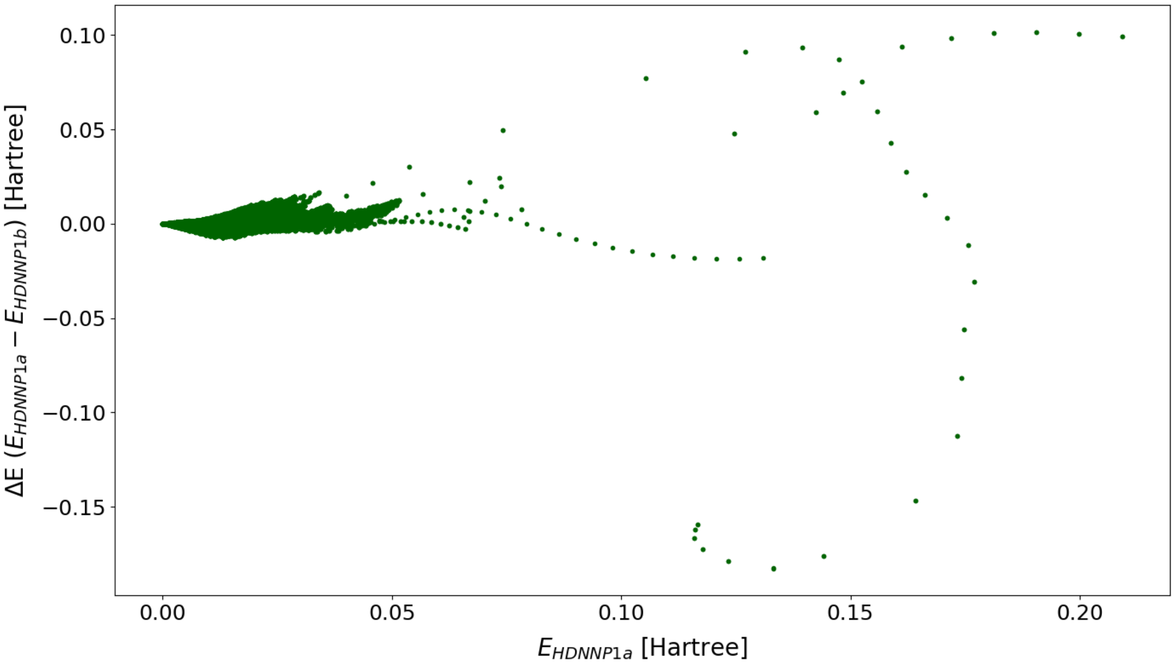


Figure 4.18: $\Delta E = E_{\text{HDNNP1a}} - E_{\text{HDNNP1b}}$ for structures below 0.1 Ha according to Bowman Potential in the dataset obtained from Edit Matyus [79]

Another change made to improve the HDNNP was to manually change the radial symmetry functions from the values obtained from the $RuNNerMakesym$ tool of RuNNer. This was done for the radial symmetry functions of Carbon-Carbon interactions. The two carbons of Formic Acid Dimer are part of the individual Formic Acid Monomers and as such the shortest distance between them in the dataset of all structures is still a high number which is 6.2330 Bohr. This results in the ACSFs covering only a narrow range of interactions. This is shown in Fig. 4.20.

Table 4.14: New C-C Radial ACSF parameters η for HDNNPs

η[Bohr^{-2}]
0, 0.003747, 0.009066, 0.017212, 0.030893, 0.056909

We can compare the new C-C parameter values given in Table 4.14 with that of the earlier ACSFs given in Table 4.9. As can be seen in Fig. 4.21, the manually changed symmetry functions are more flexible and cover a bigger region of values. This has also helped to improve the Neural Network.

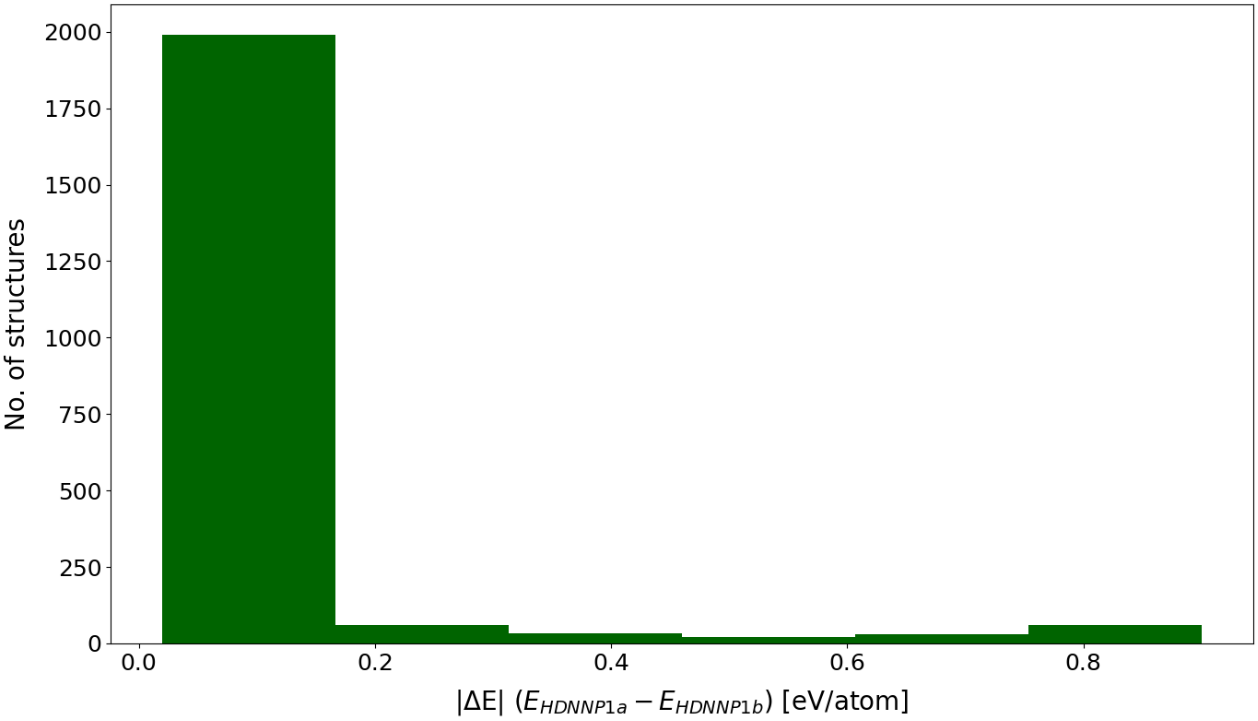


Figure 4.19: Distribution and number of structures above a difference 0f 0.02 eV/atom between HDNNP1a and HDNNP1b for structures in the dataset obtained from Edit Matyus [79]

Another sampling used to improve the Neural Network is adding structures along the 24 normal modes of Formic Acid Dimer. The Formic Acid Dimer was displaced along each of the normal mode using eigenvectors obtained from Coupled Cluster and these structures were computed using the same Coupled Cluster method. Also, the normal modes were used as a method of analysing the quality of the High-Dimensional Neural Network Potential.

The energy vs coordinate plot of the 24 normal modes of Formic Acid Dimer are shown in Fig. 4.22. The energies are shown for CCSD(T)-F12a, HDNNP1a and HDNNP1b. In this plot, we can see that both the HDNNP1s describe the normal modes fairly well even though explicit inclusion of this information is not available in the training set. But more information on the description of normal modes can be seen in 4.23. Here, the difference between HDNNP and Coupled Cluster energies are shown in meV/atom. And this gives a better picture and also tells us why we need to give a better description of the normal modes in training our dataset. We see that ω_{11}, ω_{13}, ω_4, ω_{18}, ω_2, ω_1 and ω_{17} have energy errors equal to or above 10 meV/atom. For ω_{18}, ω_2, ω_1 and ω_{17}, the energy range probed is very high. For ω_{11}, ω_{13} and ω_4, it is a bit more concerning because the probed energy range is a bit lower. HDNNP1a and HDNNP1b seem to have similar behaviour in how it differs from Coupled Cluster energies. But we can also observe that structures close to equi-

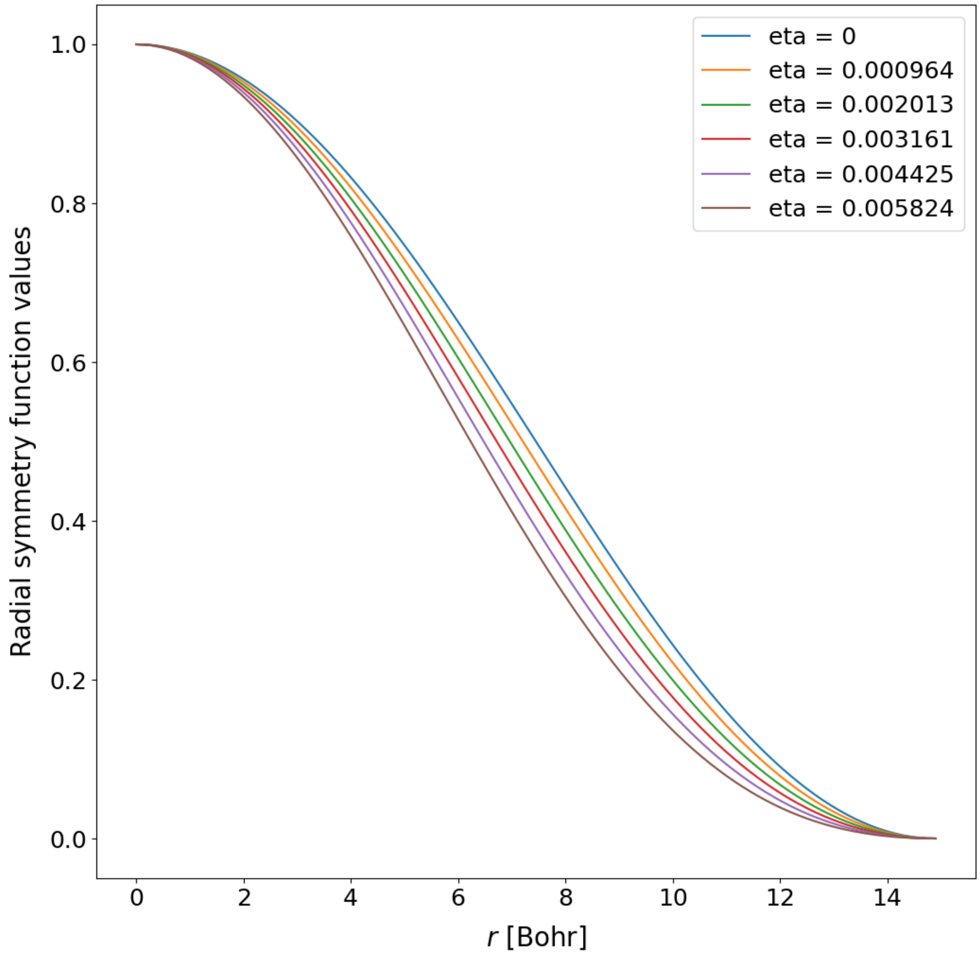


Figure 4.20: C-C radial symmetry functions for the shortest bond distance of 6.2330 Bohr with a value of 0 for r_{shift}

librium still have 1 meV/atom deviation in most cases. This reaffirms that sampling the normal mode displacements may help in improving the neural network potential.

The structures with the above changes were used to train some intermediate HDNNPs. These and additional sampling methods to improve the Neural Network to give HDNNP to is shown in next section.

4.2.3.1 Intermediate HDNNPs

MD simulations were performed using previous HDNNPs; HDNNP-ia and HDNNP-ib. One of the applications of High-Dimensional Neural Network Potentials is that they can be used to perform Molecular Dynamic simulations. Here, again active learning is employed. MD simulations were performed at 100K and 300K. The structures were the two HDNNPs differed more than 1 meV/atom were selected for

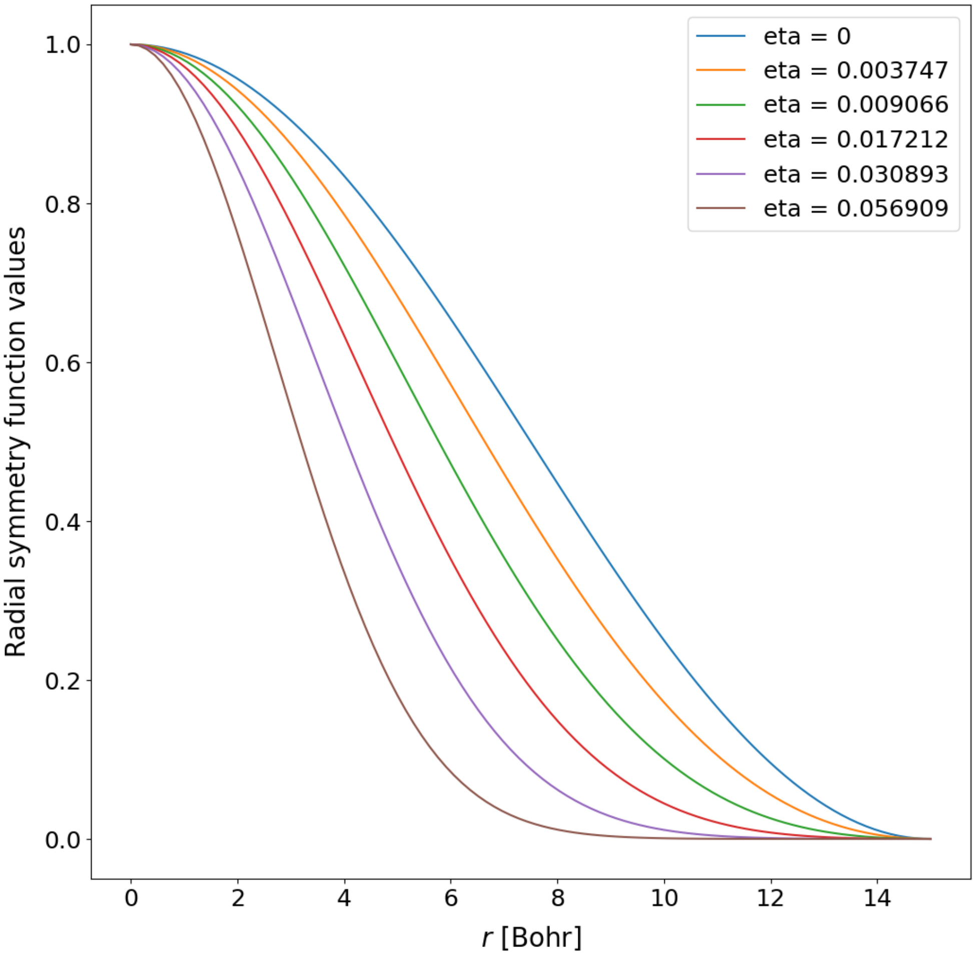


Figure 4.21: New symmetry functions for C-C distances with a cutoff of 15.0 Bohr. Here $r_{min} = 3.1907$ Bohr and $r_{shift} = 0$

adding to the training set. HDNNP-ia and HDNNP-ib have structures along normal modes, structures from Edit dataset and also the changed symmetry functions. HDNNP-ia has 14782 structures in training set and 1616 structures in testing set. HDNNP-ib has 14804 structures in training and 1594 structures in testing. More information on the two intermediate HDNNPs is given in Appendix.

The RMSEs of the two intermediate HDNNPs are given in Table 4.15. As we can see the two intermediate HDNNPs have improved RMSEs compared to the previous iteration. HDNNP-ia has 1.24 meV/atom for training set and 5.94 meV/atom for testing set whereas HDNNP-ib has 1.4 meV/atom for training and 6.16 meV/atom for testing.

In addition to the structures from MD simulations using the above HDNNPs, structures with displacement for various Hessian step sizes were also added in the next stage of training which results in HDNNP2 discussed in the next section.The structures were displaced each atom along just one Cartesian coordinate at a time with the displacement $d = 0.001, 0.0025, 0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, 0.1$ in Bohr.

Table 4.15: Energy root mean squared errors (RMSE) of the training and test sets for the intermediate HDNNPs trained using dataset for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
HDNNP-ia	16398	1.24	5.94	100	479
HDNNP-ib	16398	1.40	6.16	113	497

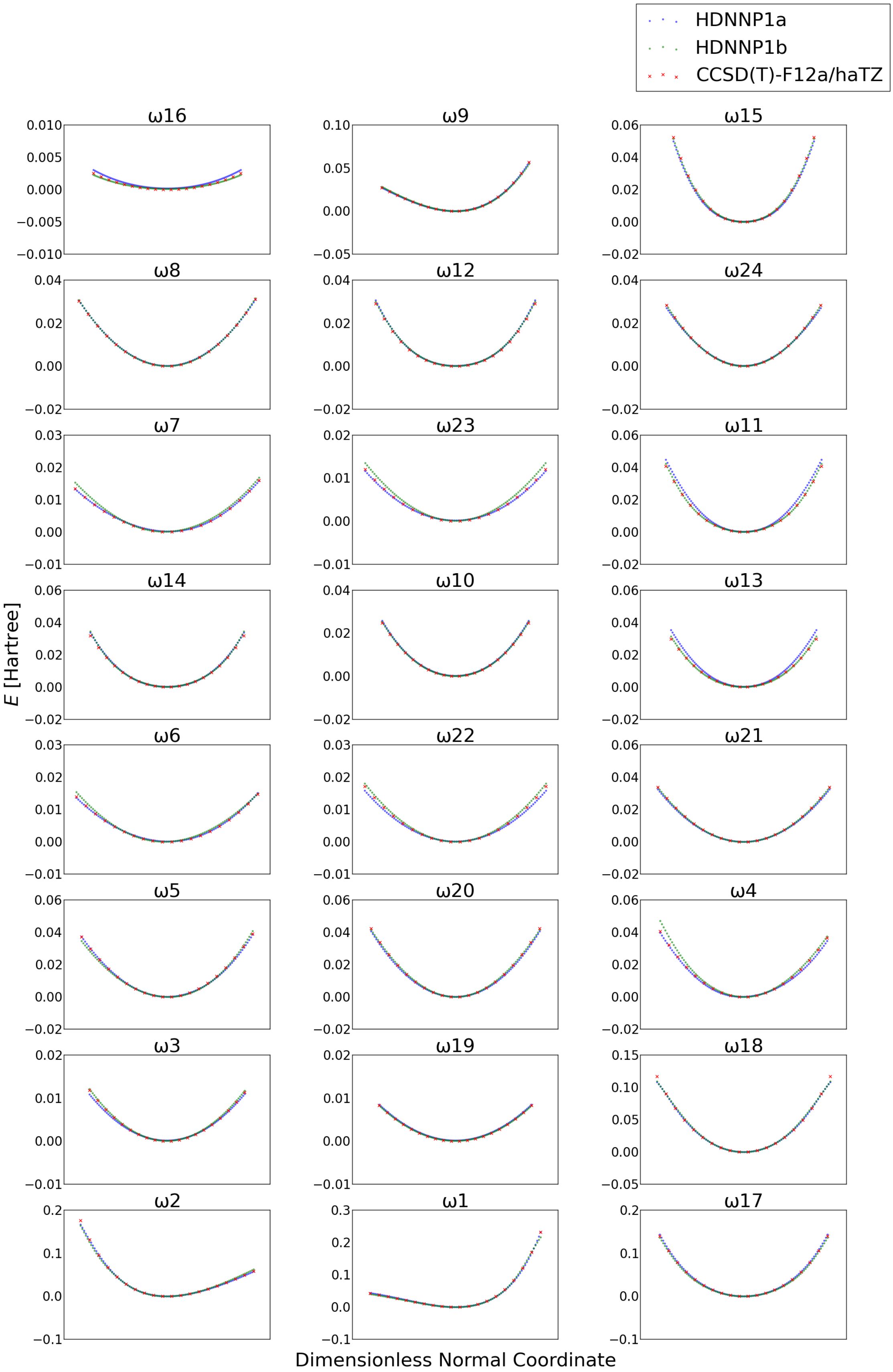


Figure 4.22: 24 normal modes of Formic Acid Dimer from low frequency vibrations to high frequency vibrations

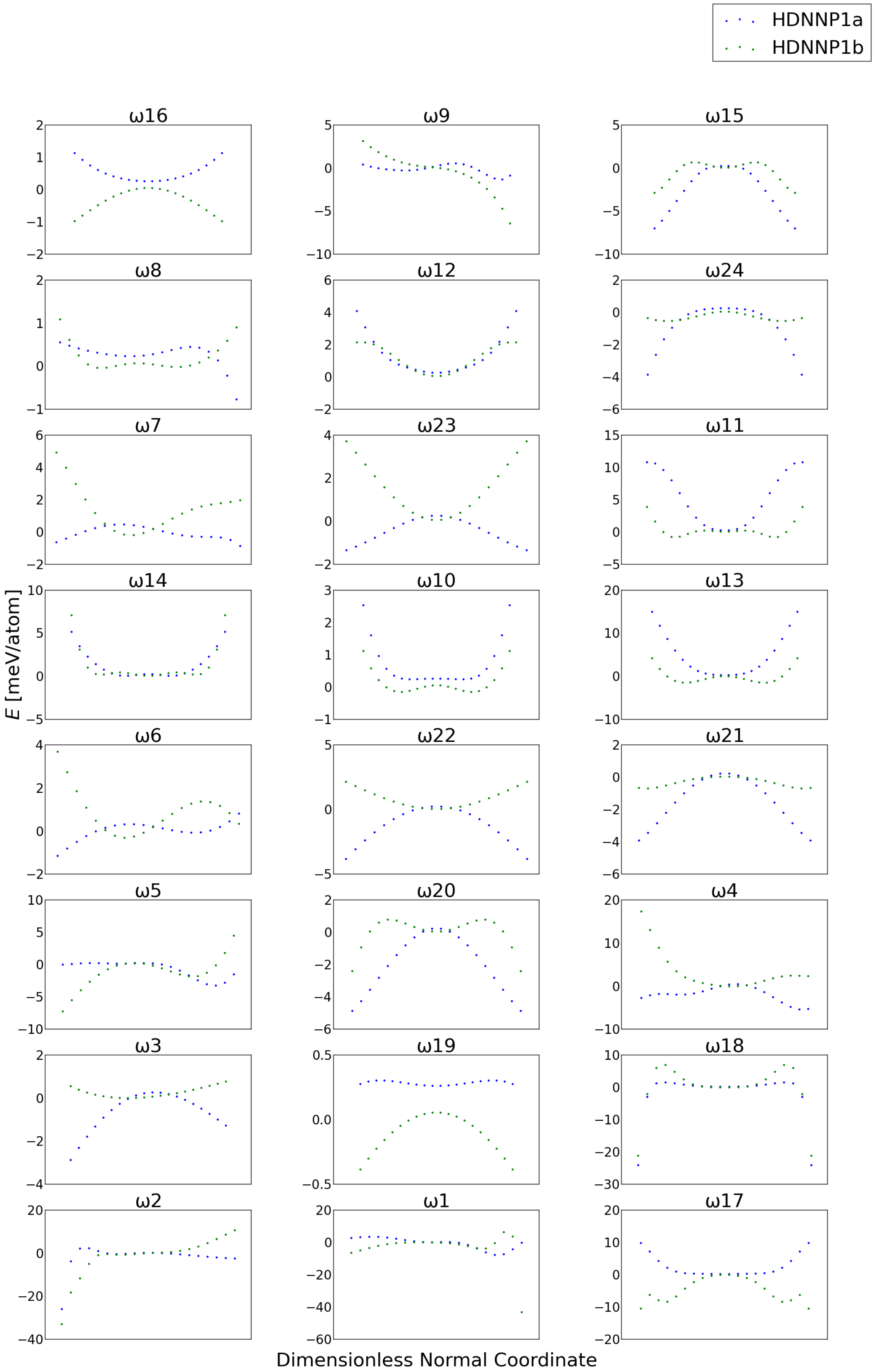


Figure 4.23: Energy difference $\Delta E - E_{\mathrm{HDNNP}=E_{\mathrm{CC}}}$ using HDNNP1a and HDNNP1b for the 24 normal modes of Formic Acid Dimer.

4.2.4 HDNNP2

Two good HDNNPs were selected at this by iteratively and systematically improving the Neural Network as described in the previous section. The two HDNNPs are HDNNP2a and HDNNP2b. These were trained with the same dataset of 27372 structures but with a different randomly selected distribution of training and testing dataset. HDNNP2 has additional 13897 structures from HDNNP1. HDNNP2a has 24629 training structures and 2743 testing structures. It is constructed with 2 hidden layers with10 nodes each. HDNNP2b has 24634 training structures and 2738 testing structures. It has 2 layers with 15 neurons each. The quality of the fit with respect to the RMSEs is shown in Table. 4.16. As we can see, HDNNP2a has 0.92 meV/atom RMSE for training and HDNNP2b has 0.71 meV/atom. In both cases, it is a marked improvement from HDNNP1a which has 2.43 meV/atom RMSE for the training set which was shown in Table 4.11. HDNNP2a has 1.88 meV/atom RMSE for the testing set whereas HDNNP2b has 1.99 meV/atom which shows considerable improvement while considering that HDNNP1a had an RMSE of 13.99 meV/atom for testing set. This also very much supports our hypothesis that Neural Networks can be improved with the right analysis and sampling even while dealing with a system requiring accuracy in the range of a few cm^{-1}. Here, at this stage both the HDNNPs have a similar quality with HDNNP2a has a better testing RMSE and HDNNP2b with a better training RMSE.

Table 4.16: Energy root mean squared errors (RMSE) of the training and test sets for HDNNP2 for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
		full energy range			
HDNNP2a	27372	0.92	1.88	74	158
HDNNP2b	27372	0.71	1.99	57	160

As before, we need to see how the energetics look for each structure for the two fits. These are shown in Figures 4.24 and 4.25. Both HDNNP2a and HDNNP2b shows a better performance for the energetics compared to the previous HDNNPs. In the case of HDNNP1a and HDNNP1b, the deviation had a spread from 0.15 eV/atom for HDNNP1b to even 0.3 eV/atom for HDNNP1a. Here, in both the cases, the deviations spread up to 0.02 eV/atom. But, it is to be noted that in both the cases the high deviations are above 0.2 eV/atom ab-initio energies. We also see that in both cases the testing structures shown in red are the ones most deviated from our reference energies. As before, the explanation remains that the sparsely populated areas are the ones showing up as such artefacts in our analysis. At a single glance, it may seem like HDNNP2a is more spread out than HDNNP2b. But as we can see that is a result of how scaled the energy range is. Yet, we can see that in the case of HDNNP2b, there are two testing structures with a deviation above 0.04 eV/atom whereas HDNNP2a has two testing structures hovering around 0.03 eV/atom. And these ones are the structures with the biggest deviation. While taking a close look at these structures, no inconsistency was observed except these are structures where there are bigger displacements of the bonds. As we can see, at this stage, the two HDNNP2s are performing at similar quality and hence there is a need to take a look at the harmonic frequencies. It is also to be noted that the two HDNNPs have same underlying dataset and same definitions of the Atom Centered Symmetry Functions. They both differ in the architecture such that HDNNP2b is more flexible. They also differ in the distribution of training and testing data.

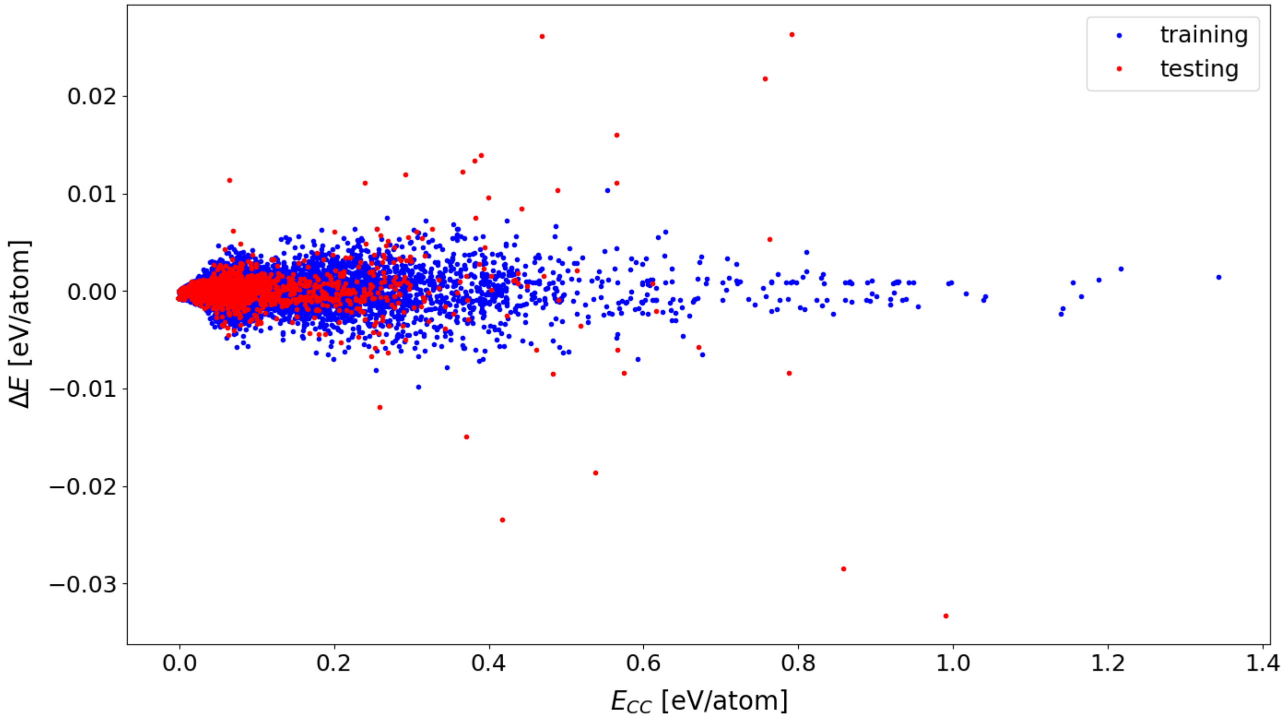


Figure 4.24: Energy difference $\Delta E = E_{\mathrm{CC}} - E_{\mathrm{HDNNP2a}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP2a are provided in Tab. 4.16

In order to choose a potential here, the harmonic frequencies of both the potentials should be looked at. This is given in Table A.12. At a first glance we can see that both the HDNNPs have better harmonic frequencies than the previous generation. To have a closer look, the deviations of the harmonic frequencies from the CCSD(T)/F12a harmonic fundamentals are given in Table A.13. We see here that HDNNP2a has a maximum deviation of 24.86 cm^{-1} for ω_{18}. Also ω_{18} has a deviation of 21.29 cm^{-1}. Also five other fundamentals have a deviation above 10 cm^{-1}. These are ω_6, ω_{12}, ω_{15}, ω_{17} and ω_{22}. We can also see that the highly deviated two frequencies are for high frequency modes which are above 3000 cm^{-1}. As such their relative deviation would be smaller. In the case of HDNNP2b, we see that the maximum deviation is 21.7 cm^{-1} for ω_{11}. This is smaller than the maximum deviation for HDNNP2a. We also see that another harmonic fundamental has a deviation above 20 cm^{-1}. This is for ω_{13}. But, seven other modes have deviation above 10 cm^{-1}. These are modes ω_3, ω_{10}, ω_{14}, ω_{17}, ω_{18}, ω_{19} and ω_{20}. Also the two modes with a deviation above 20 cm^{-1} are also relatively low frequency vibrations with a Coupled

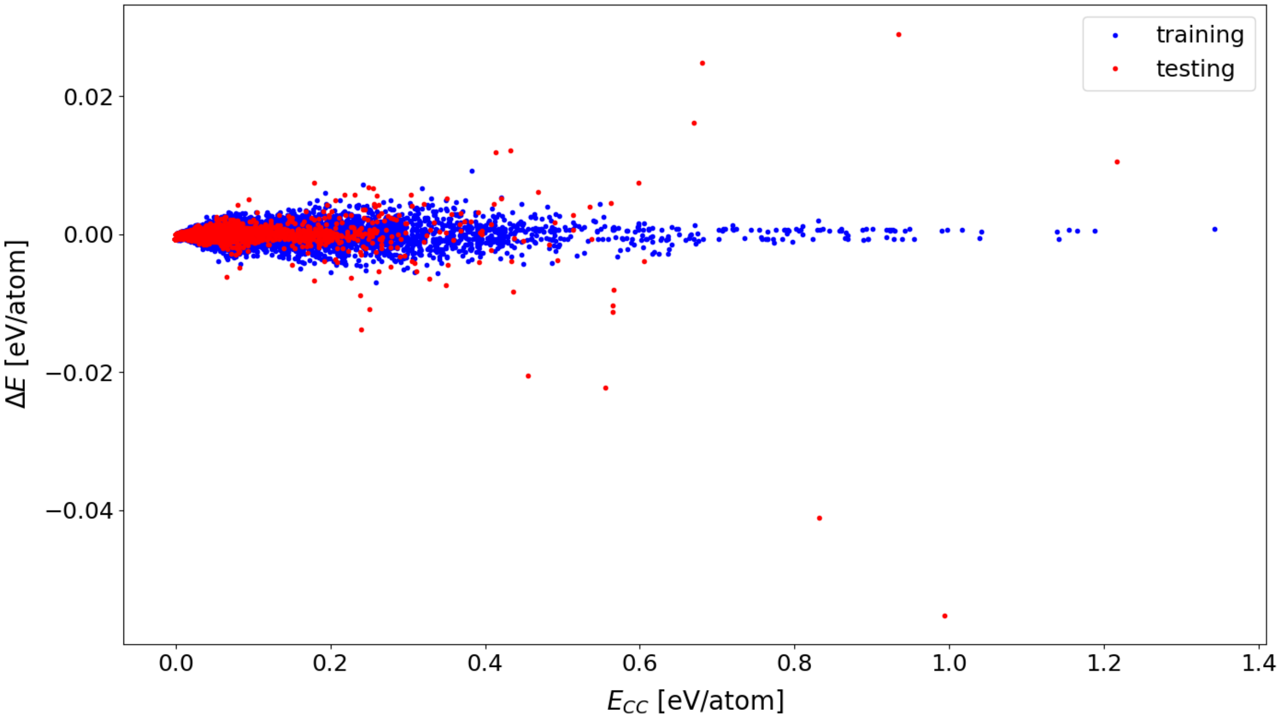


Figure 4.25: Energy difference $\Delta E = E_{\text{CC}} - E_{\text{HDNNP2b}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP2b are provided in Tab. 4.16

Cluster value of 959.6 and 1102.03 cm^{-1}. The plot of comparison of the fundamental harmonic frequencies for the two HDNNPs are given in Fig. 4.26. It is to be noted that both HDNNP2a and HDNNP2b have a comparable RMSD of 10 cm^{-1}.

After a look at the harmonic frequencies and their comparison with ab-initio values, we can now select an HDNNP at this stage. HDNNP2b has a smaller maximum deviation. But, it is also evident that the higher deviation is for a smaller vibrational frequency mode. Also, there are more modes of vibration with a higher deviation. HDNNP2a though has a maximum deviation with 3 wave numbers more, these are for higher frequency modes. Also, HDNNP2a has better description for other modes in general. Hence, HDNNP2a can be selected at this point. But, both these good HDNNPs can be used to refine our Neural Network further to obtain a fit that satisfies the stipulation set prior.

Table 4.17: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP2 with frequencies from CCSD(T)-F12a/haTZ level of theory.

Mode	Sym.	HDNNP2a	HDNNP2b	CCSD(T)-F12a/haTZ
ω_1	A_g	3213.2	3212	3203.36
ω_2	A_g	3083.3	3105.4	3104.59
ω_3	A_g	1722.5	1704.9	1717.13
ω_4	A_g	1486.8	1477.8	1483.92
ω_5	A_g	1418.5	1413.4	1413.1
ω_6	A_g	1245.8	1250.3	1256.78
ω_7	A_g	680.7	686.5	687.78
ω_8	A_g	214.7	214.4	211.28
ω_9	A_g	161.6	168.3	170.96
ω_{10}	B_g	1083.3	1097.4	1085.04
ω_{11}	B_g	958.9	937.9	959.6
ω_{12}	B_g	240.0	266.3	257.76
ω_{13}	A_u	1104.8	1122.2	1102.03
ω_{14}	A_u	981.5	968.1	986.46
ω_{15}	A_u	172.8	178	185.95
ω_{16}	A_u	71.2	69.3	76.36
ω_{17}	B_u	3323.7	3323	3305.25
ω_{18}	B_u	3075.7	3089.9	3100.56
ω_{19}	B_u	1783.0	1769.4	1781.57
ω_{20}	B_u	1461.6	1443.8	1455.96
ω_{21}	B_u	1415.0	1400	1405.06
ω_{22}	B_u	1248.0	1251	1260.06
ω_{23}	B_u	718.0	715.8	715.81
ω_{24}	B_u	275.5	278.9	278.07

Table 4.18: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP2 from CCSD(T)-F12a/haTZ frequencies. $\Delta\omega = \omega_{\mathrm{CC}} - \omega_{\mathrm{HDNNP}}$

Mode	Sym.	HDNNP2a	HDNNP2b
ω_1	A_g	-9.84	-8.64
ω_2	A_g	21.29	-0.81
ω_3	A_g	-5.37	12.23
ω_4	A_g	-2.88	6.12
ω_5	A_g	-5.4	-0.3
ω_6	A_g	10.98	6.48
ω_7	A_g	7.08	1.28
ω_8	A_g	-3.42	-3.12
ω_9	A_g	9.36	2.66
ω_{10}	B_g	1.74	-12.36
ω_{11}	B_g	0.7	21.7
ω_{12}	B_g	17.76	-8.54
ω_{13}	A_u	-2.77	-20.17
ω_{14}	A_u	4.96	18.36
ω_{15}	A_u	13.15	7.95
ω_{16}	A_u	5.16	7.06
ω_{17}	B_u	–18.45	-17.75
ω_{18}	B_u	24.86	10.66
ω_{19}	B_u	-1.43	12.17
ω_{20}	B_u	-5.64	12.16
ω_{21}	B_u	-9.94	5.06
ω_{22}	B_u	12.06	9.06
ω_{23}	B_u	-2.19	0.01
ω_{24}	B_u	2.57	-0.83

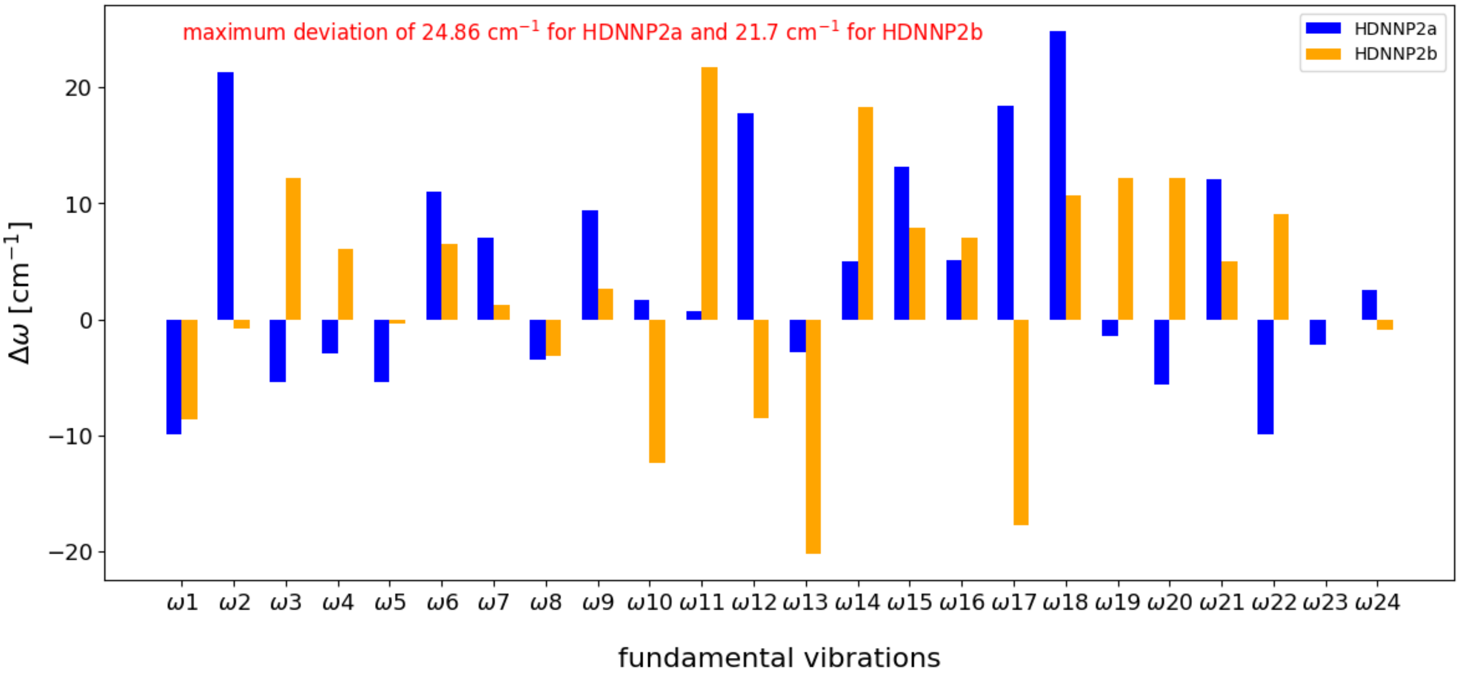


Figure 4.26: Deviations of the harmonic vibrational frequencies ω_i with respect to the reference CCSD(T)-F12a/haTZ frequencies for HDNNP2. $\Delta\omega = \omega_{\mathrm{CC}} - \omega_{\mathrm{HDNNP}}$ for different HDNNPs

4.2.5 HDNNP3

HDNNP3 was developed after using HDNNP2a and HDNNP2b, described in the previous section, for sampling a different set of structures. HDNNP3 was developed by coupling 24 normal modes of Formic Acid Dimer. Formic Acid Dimer was displaced along 2 normal modes at a time for all the different combinations, i.e. 276 combinations. Along each of the coupled mode, 100 structures were sampled. Then, HDNNP2a and HDNNP2b were used to find energies of all the structures. Here, again, active learning is employed. A threshold of 2 meV/atom was used to select structures for training the dataset.

The difference in energies between HDNNP2a and HDNNP2b is shown in Fig. 4.27 and 4.28. The blue and red show energy regions where HDNNP2a and HDNNP2b diverge greatly. White regions show where both the HDNNP2s agree well in predicting the energies. It is to be noted that there are regions in the plots where the structures are sampled at very high energy ranges. So, first selection criteria to improve the fit was to find the structures below 0.1 Ha according to HDNNP2a . Among those structures, the ones which has an energy difference greater than or equal to 2 meV/atom were selected.

The following HDNNP was selected after training with the above described dataset which now comes out to have 29162 structures which includes additional 1790 structures when compared to HDNNP2. HDNNP3 has 14 nodes per layer. It has 26221 structures going into training and 5882 structures going into the testing set. The quality of the fit in terms of RMSE is given in Table 4.19. The training RMSE is 0.37 meV/atom and the testing RMSE is 2.04 meV/atom. As observed in other iterations the testing RMSE is still higher than that of the training because of the energy distribution of the dataset. When we take a look at the RMSEs for structures below 0.1 Hartree above equilibrium, we see that the training RMSE is 0.35 meV/atom and the testing RMSE is 0.34 meV/atom. This affirms that for all purposes, the energy range we are interested in, which in itself is not a low range, is very well defined by the HDNNP3. It is also seen that less than 1000 structures in our dataset is above 0.1 Hartree.

Table 4.19: Energy root mean squared errors (RMSE) of the training and test sets for HDNNP3 for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
full energy range					
HDNNP3	29162	0.37	2.04	30	165
energy range below 0.1 Ha					
HDNNP3	28286	0.35	0.34	28	27

Now, it is of interest to observe how each structure in the dataset behaves energetically. This is shown in Fig. 4.29. We can see that almost all the structures have HDNNP3 energies very close to the Coupled Cluster energies. The highest deviation is in the range of 0.04 eV/atom and those are higher energy regions with structures

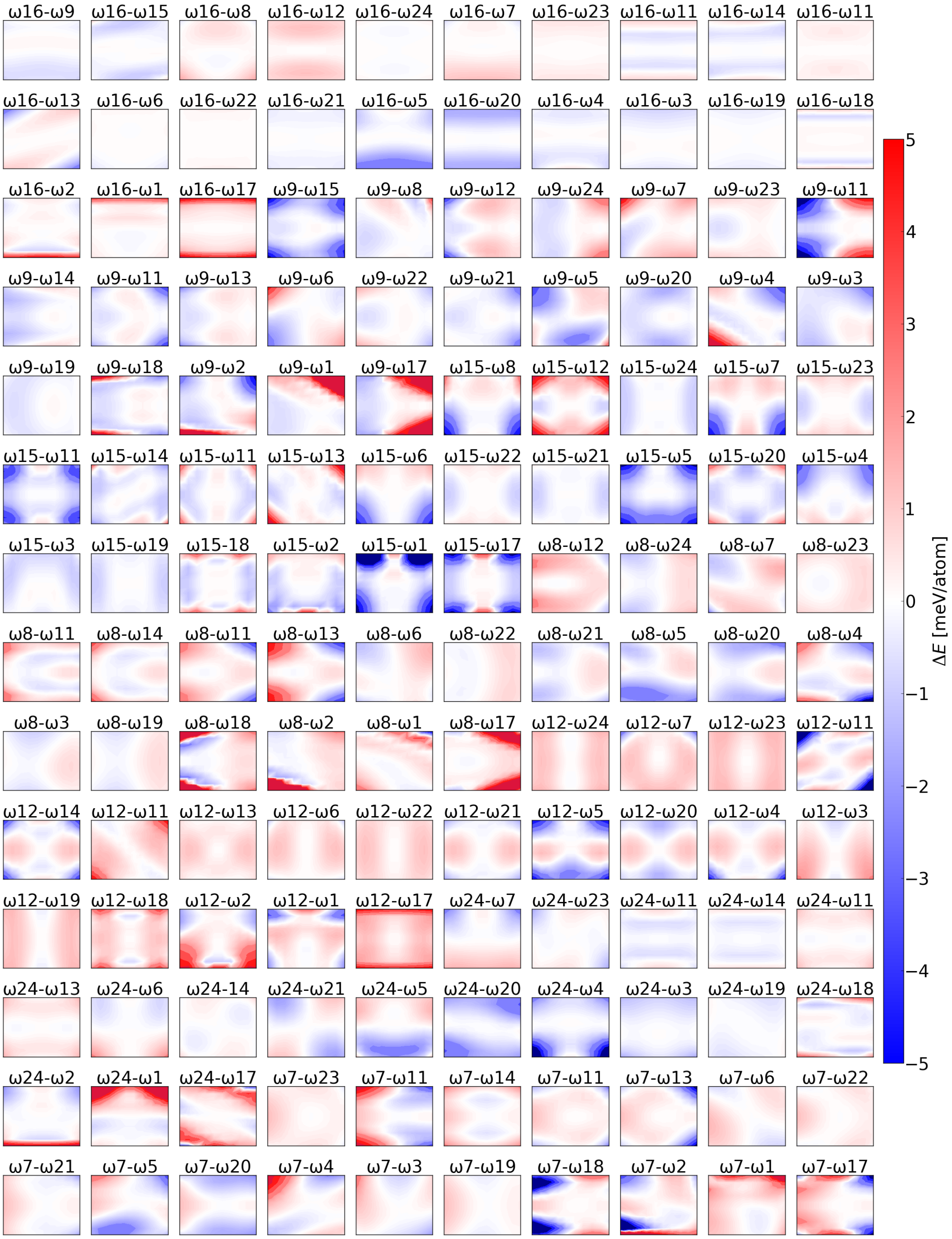


Figure 4.27: $\Delta E = E_{\text{HDNNP2b}} - E_{\text{HDNNP2a}}$ for displacements along two normal modes at a time. The displacements are made using Coupled Cluster eigenvectors. Continued on the next page

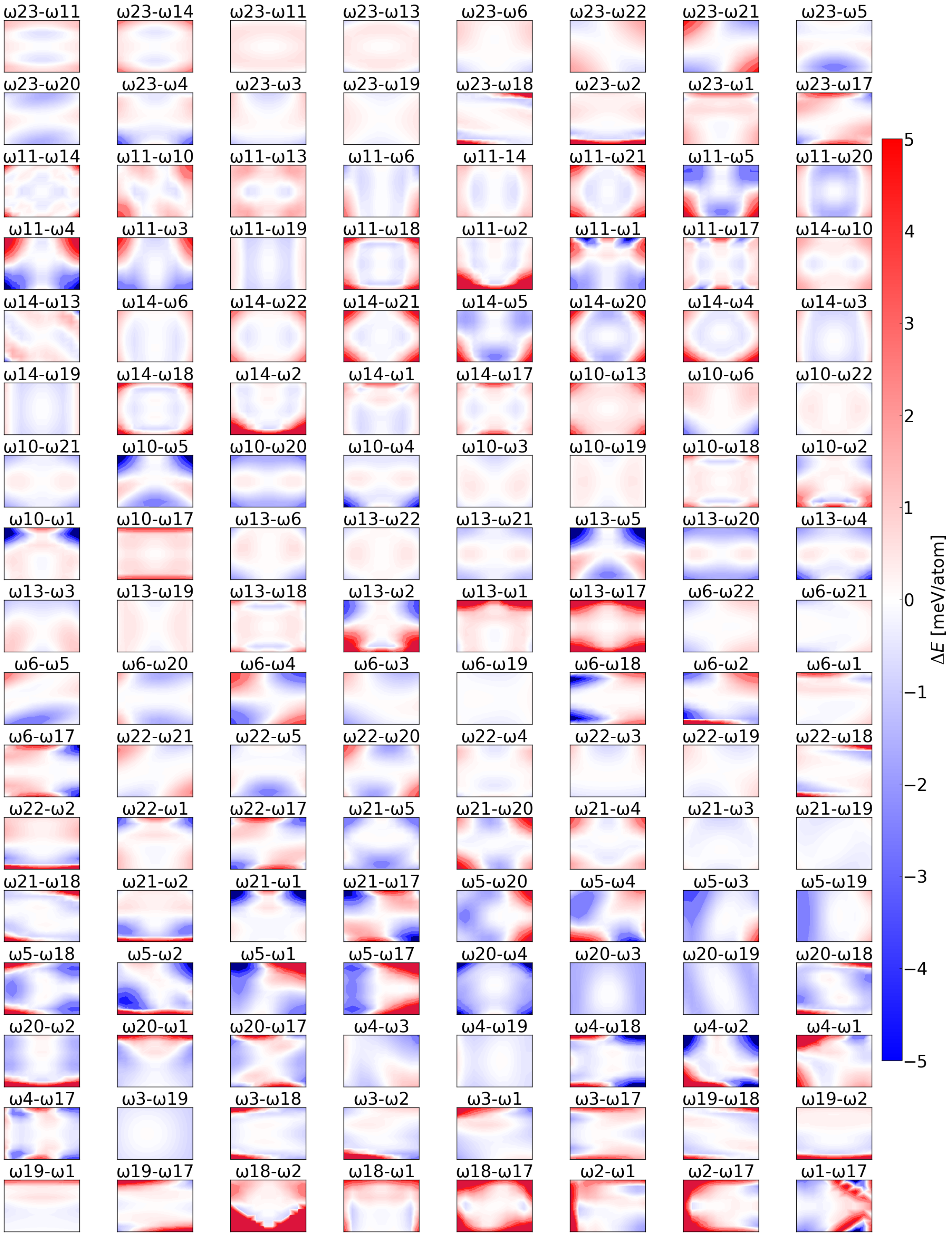


Figure 4.28: $\Delta E = E_{\mathrm{HDNNP2b}} - E_{\mathrm{HDNNP2a}}$ for displacements along two normal modes at a time. The displacements are made using Coupled Cluster eigenvectors continued from previous figure

in the testing set. In fact, all the structures which show a higher difference between HDNNP3 and Coupled Cluster are in the testing set. Also when we take a look at structures below 0.2 eV/atom in reference Coupled Cluster method, the deviation between ab-initio and HDNNP3 is really low and both the methods converge very well. So far, we can see that the energetics are well described in the third iteration of the construction of the HDNNP for Formic Acid Dimer.

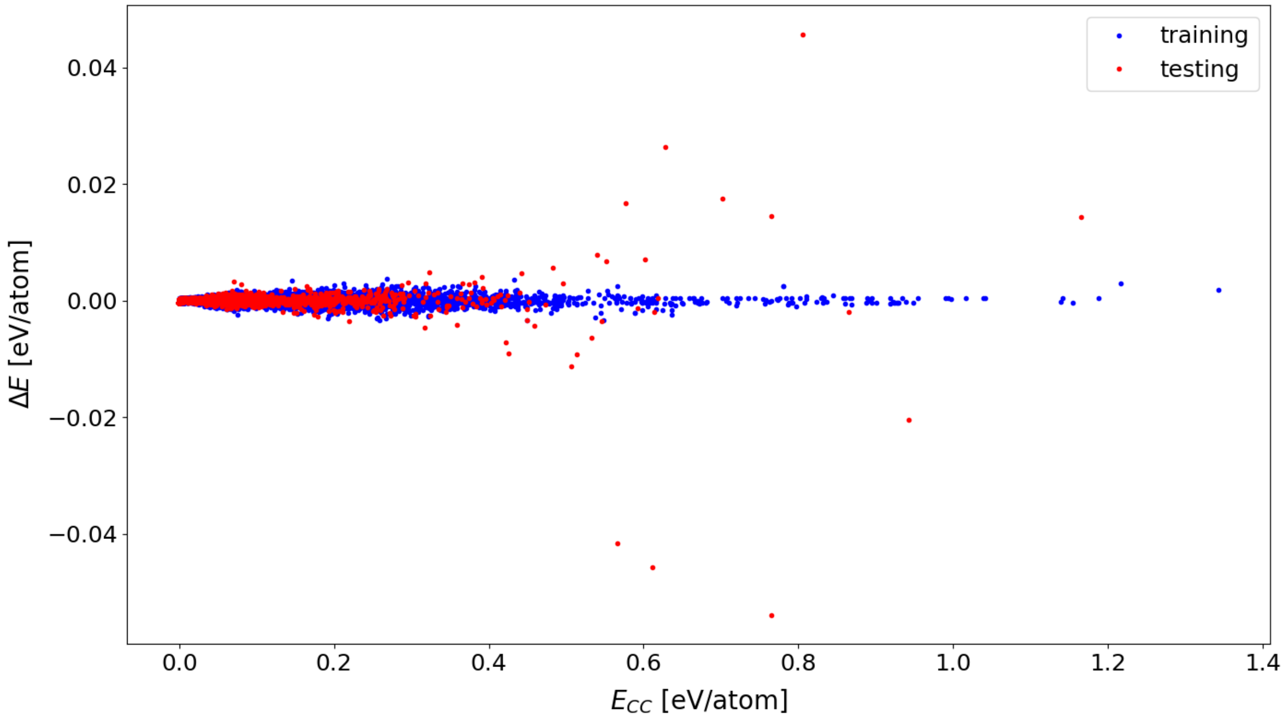


Figure 4.29: Energy difference $\Delta E = E_{\mathrm{CC}} - E_{\mathrm{HDNNP3}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP3 are provided in Tab. 4.19

Now, the final check for the quality of the potential as always is a look at the harmonic fundamentals. The harmonic frequencies from HDNNP3 are given in Table 4.20. The frequencies at first sight looks very promising. As we can see the HDNNP3 harmonic wavenumbers are very close to the CCSD(T)-F12a/haTZ numbers with a few wavenumbers uncertainty. In order to have a better understanding, Table 4.21 gives the deviation between HDNNP3 frequencies and CCSD(T)-F12a frequencies. The highest deviation between the two is 7.36 cm^{-1} for ω_{14}. This is well within the stipulated threshold of 10 cm^{-1}. There are other vibrational modes with deviation above 4 cm $^{-1}$. ω_1, ω_3, ω_9, ω_{13}, ω_{15}, ω_{16}, ω_{17}, ω_{21} and ω_{23} are the modes with deviations above 4 cm^{-1}. The plot of the deviations of frequencies is given in Fig. 4.30. The HDNNP3 has an RMSD of 4 cm^{-1} while comparing to the reference harmonic frequencies. Since, the harmonic frequencies showed a very good performance, the above PES was used for further anharmonic studies in collaborations. The PES was also named FAD-HDNNP as the potential to be published in peer-reviewed work. The anharmonic frequencies and a discussion on the 3 selected iterations of HDNNPs and comparison with Bowman Potential will be discussed in the next section. Before that, a few more results of analysis of the quality of the potential will be discussed below.

Table 4.20: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP3 with frequencies from CCSD(T)-F12a/haTZ level of theory.

Mode	Sym.	HDNNP3	CCSD(T)-F12a/haTZ
ω_1	A_g	3209.0	3203.36
ω_2	A_g	3102.4	3104.59
ω_3	A_g	1721.5	1717.13
ω_4	A_g	1486.0	1483.92
ω_5	A_g	1410.2	1413.1
ω_6	A_g	1257.0	1256.78
ω_7	A_g	687.4	687.78
ω_8	A_g	213.9	211.28
ω_9	A_g	166.8	170.96
ω_{10}	B_g	1083.3	1085.04
ω_{11}	B_g	957.1	959.6
ω_{12}	B_g	257.3	257.76
ω_{13}	A_u	1108.7	1102.03
ω_{14}	A_u	979.1	986.46
ω_{15}	A_u	180.2	185.95
ω_{16}	A_u	70.9	76.36
ω_{17}	B_u	3311.7	3305.25
ω_{18}	B_u	3099.4	3100.56
ω_{19}	B_u	1784.0	1781.57
ω_{20}	B_u	1459.0	1455.96
ω_{21}	B_u	1409.7	1405.06
ω_{22}	B_u	1259.5	1260.06
ω_{23}	B_u	711.6	715.81
ω_{24}	B_u	275.2	278.07

Though we obtained good harmonic frequencies, we also need to see how smooth the potential is. Especially, at lower energy regions and at a very minute level, it is possible that the potential is not smooth. This can arise out of overfitting. This is especially interesting to look at for the potential discussed here. As we have seen the testing RMSE for HDNNP3 is higher than that of training RMSE. We have theorised that this is because of uneven energy distribution in the testing set. But, usually a higher testing RMSE is indicative of overfitting in machine learning potentials. Here, we are looking at how the frequencies behave at various step sizes for finite difference Hessian. If our frequencies are consistent at various step sizes, we can assume that the High-Dimensional Neural Network Potential is smooth. Then, we can also affirm that there is no overfitting and hence the potential having the testing RMSE is because of the energy distribution. The harmonic frequencies for various Hessian step sizes is given in Table A.14 in the Appendix. The step sizes

Table 4.21: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP3 from CCSD(T)-F12a/haTZ frequencies. $\Delta\omega = \omega_{\mathrm{CC}} - \omega_{\mathrm{HDNNP}}$

Mode	Sym.	HDNNP3
ω_1	A$_g$	-5.64
ω_2	A$_g$	2.19
ω_3	A$_g$	-4.37
ω_4	A$_g$	-2.08
ω_5	A$_g$	2.9
ω_6	A$_g$	-0.22
ω_7	A$_g$	0.38
ω_8	A$_g$	-2.62
ω_9	A$_g$	4.16
ω_{10}	B$_g$	1.74
ω_{11}	B$_g$	2.5
ω_{12}	B$_g$	0.46
ω_{13}	A$_u$	-6.67
ω_{14}	A$_u$	7.36
ω_{15}	A$_u$	5.75
ω_{16}	A$_u$	5.46
ω_{17}	B$_u$	-6.45
ω_{18}	B$_u$	1.16
ω_{19}	B$_u$	-2.43
ω_{20}	B$_u$	-3.04
ω_{21}	B$_u$	-4.64
ω_{22}	B$_u$	0.56
ω_{23}	B$_u$	4.21
ω_{24}	B$_u$	2.87

shown are 0.001, 0.005, 0.01, 0.025 and 0.05 Bohr. Among these, the default and usually used step size is 0.01 Bohr. This is also the step size used in calculating harmonic frequencies using CCSD(T)-F12a/haTZ level of theory. A first look at the table indicates that the frequencies are more or less the same across varying Hessian step sizes. In order to have a closer look, the deviations of harmonic frequencies from that of default Hessian step size is given in Table 4.22. We can see that at 0.001 Bohr and 0.005 Bohr, the frequencies are very close to the default step size values. The difference is in decimal point values. This indicates the potential is very smooth. As we go to 0.025 Bohr the highest deviation is 1.5 cm^{-1} for ω_1. And for a higher step size of 0.05 Bohr, the deviation goes up to 7 cm^{-1} for ω_1 and ω_{17}. ω_{18} and ω_2 show 4.5 cm^{-1} deviation and the rest fall below 2.1 cm^{-1} deviation. This is to be expected for higher step sizes as the finite difference relies on the assumption that the step size is quite small. Analytical frequencies are also available, which are given in the appendix.

The energy vs coordinate plot of normal modes according to FAD-HDNNP is given in Fig. 4.31. It can be observed that the displacement of Formic Acid Dimer along

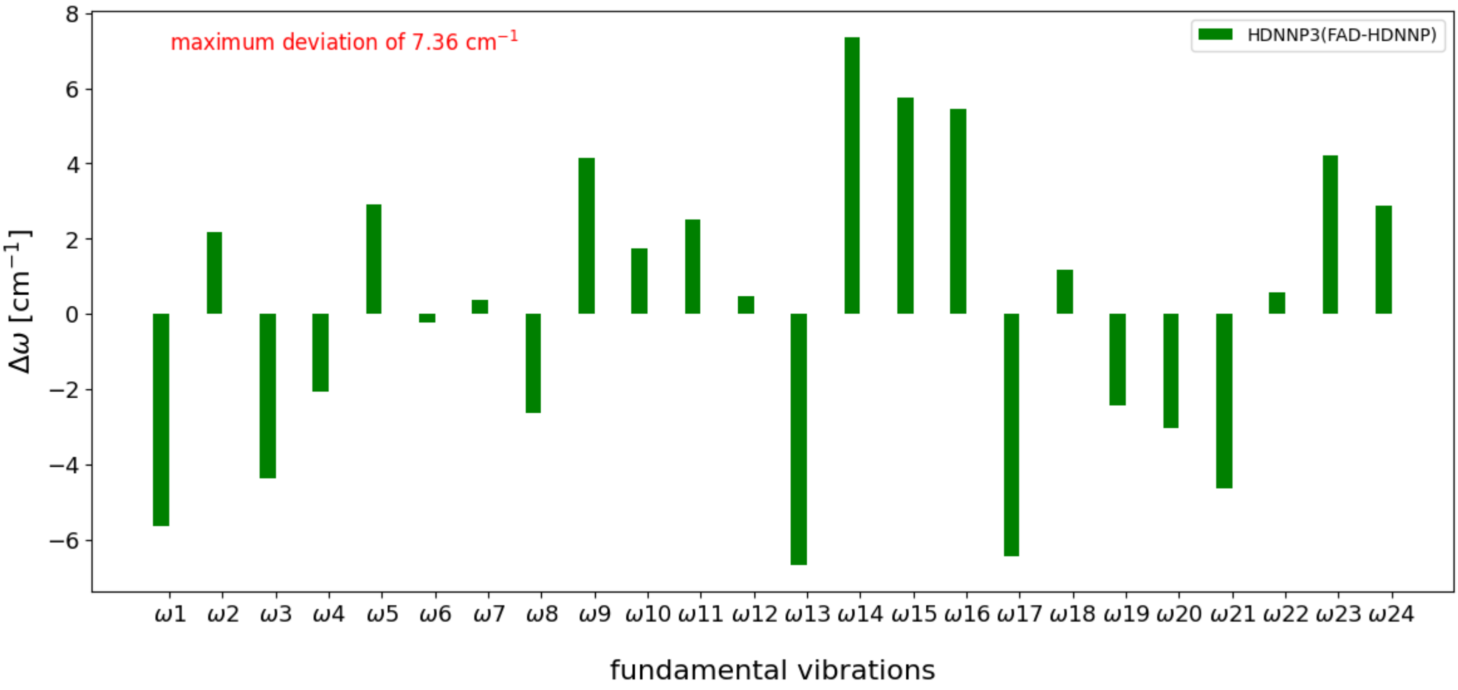


Figure 4.30: Deviations of the harmonic vibrational frequencies ω_i with respect to the reference CCSD(T)-F12a/haTZ frequencies for HDNNP3. $\Delta\omega = \omega_{\text{CC}} - \omega_{\text{HDNNP}}$ for different HDNNPs

its 24 normal modes is very well described by FAD-HDNNP. In the Figure, the FAD-HDNNP and Coupled Cluster energies are shown. They align very well. It should be noted that this is a improvement from the description of normal modes while comparing to HDNNP1a and HDNNP1b where just a first look itself shows that in some modes, the curvature wasn't fully aligned. Along with this plot, Figure 4.32 gives a better picture of the normal modes. Here, the difference between Coupled Cluster and FAD-HDNNP energies are shown in meV/atom. We see that for most of the normal modes, the deviation is in the range of 0.1 to 0.4 meV/atom. ω_{18} though shows a higher deviation up to 1.2 meV/atom. Also, ω_2 shows a deviation of up to 2 meV/atom and ω_1 shows the highest with 6 meV/atom but while comparing to the energy range in Figure 4.31, we can see it is a structure with high energy in range of 0.3 Ha. In the case of ω_2 and ω_{18} as well, the structures with high deviations are the ones above 0.1 Ha energy. This is to be expected since most of the sampled structures in the training is below that energy range. Even then, we can see that the normal modes behave quite reasonably. For HDNNP1a and HDNNP1b, the deviations were even up to 10 meV/atom. ω_1 had one structure with deviation up to 50 meV/atom. Therefore, we can see that this is a improvement in the description of the normal modes. A comparison of the three iterations would be discussed in the next section. Also, further sampling of structures to include 3 mode couplings were done in improving the HDNNP. This will also be discussed.

Table 4.22: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP3 using various Hessian step sizes in Bohr from the default step size of 0.01 Bohr. Here $\Delta\omega = \omega_d - \omega_{default}$

Mode	Sym.	0.001	0.005	0.025	0.05
ω_1	A_g	-0.3	-0.2	1.5	6.8
ω_2	A_g	-0.2	-0.2	0.9	4.4
ω_3	A_g	0	0	0.3	1.2
ω_4	A_g	0	-0.1	0.3	1.3
ω_5	A_g	-0.1	-0.1	0.1	0.6
ω_6	A_g	-0.1	-0.1	0.3	1.5
ω_7	A_g	0	0	0	-0.1
ω_8	A_g	0.1	0	-0.1	-0.4
ω_9	A_g	0.1	0.1	-0.1	-0.5
ω_{10}	B_g	0	-0.1	0.2	0.8
ω_{11}	B_g	-0.1	-0.1	0.4	1.8
ω_{12}	B_g	-0.1	-0.1	0.3	1.3
ω_{13}	A_u	-0.1	-0.1	0.1	0.8
ω_{14}	A_u	0	-0.1	0.3	1.6
ω_{15}	A_u	-0.3	0	0.5	2.1
ω_{16}	A_u	-0.1	0	0.4	2
ω_{17}	B_u	-0.3	-0.2	1.4	6.7
ω_{18}	B_u	-0.2	-0.2	1	4.5
ω_{19}	B_u	-0.1	-0.1	0.3	1.3
ω_{20}	B_u	0	0	0.3	1.2
ω_{21}	B_u	0	0	0.1	0.5
ω_{22}	B_u	-0.1	-0.1	0.3	1.5
ω_{23}	B_u	0	0	0	-0.1
ω_{24}	B_u	0	0	-0.2	-0.7

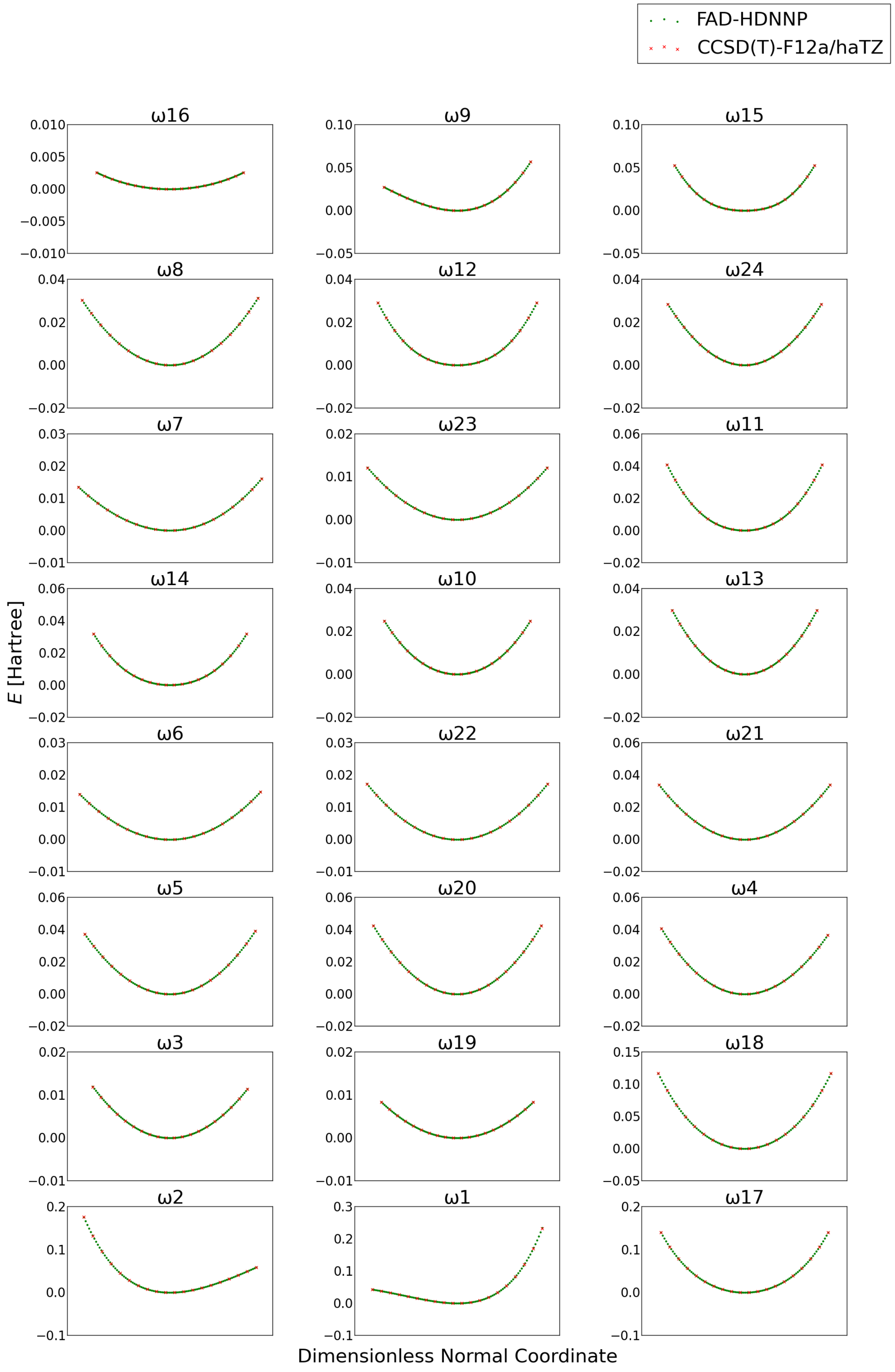


Figure 4.31: 24 normal modes of Formic Acid Dimer from low frequency vibrations to high frequency vibrations

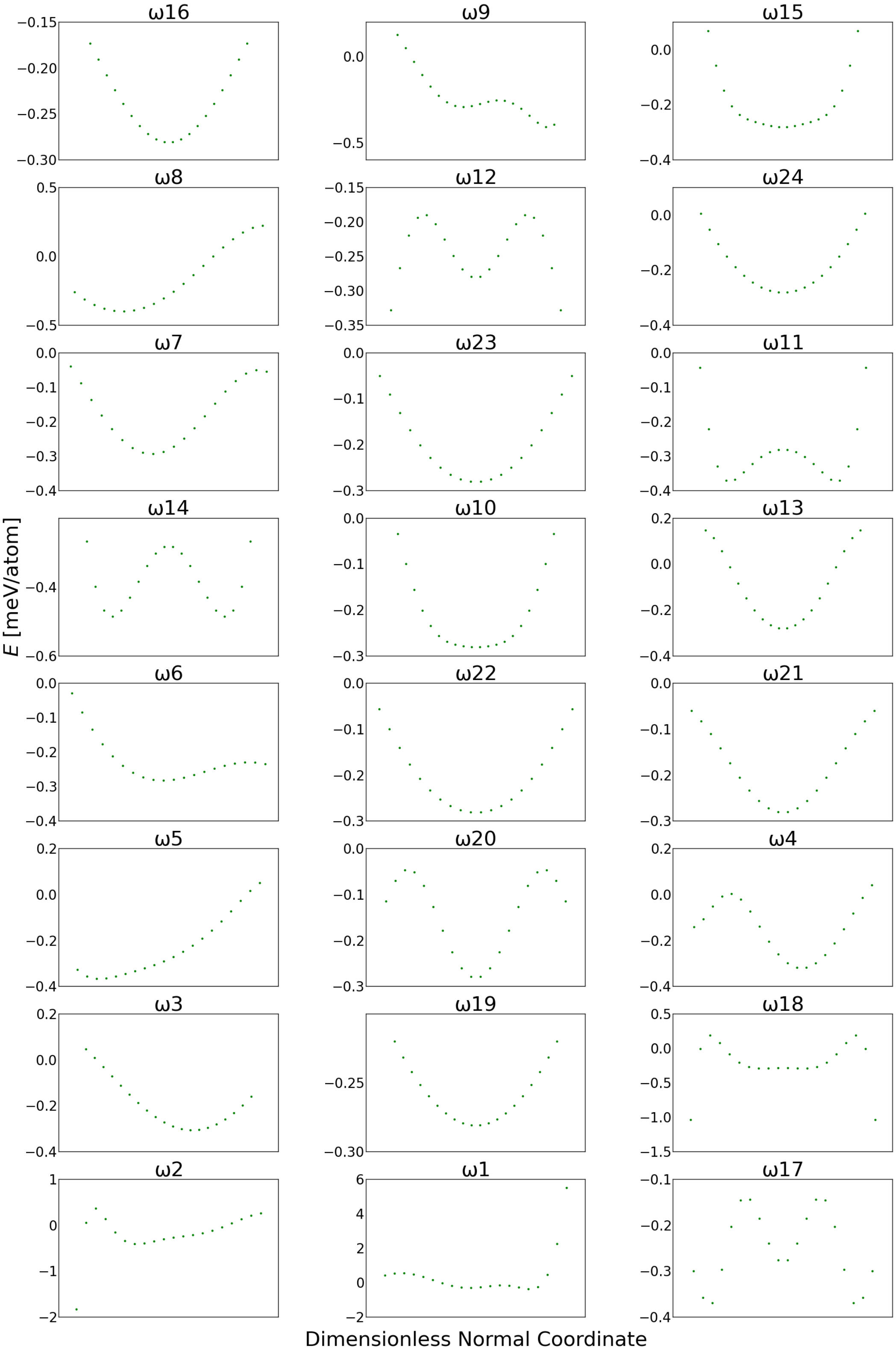


Figure 4.32: Energy difference $\Delta E - E_{\mathrm{HDNNP}} = E_{\mathrm{CC}}$ using FAD-HDNNP for the 24 normal modes of Formic Acid Dimer.

4.2.6 Discussions on the Various HDNNP Iterations and their Frequencies

After the development of the potential, it is interesting to see how it evolved overtime and also how it compares to the other available potentials. The potential for Formic Acid Dimer which has been most widely used in literature is the Poly Invariant Potential constructed by Bowman (QB16) [77]. The potential gives very accurate energies and reasonable frequencies. This potential is also interesting to compare because the reference method used in its construction is the same as the one used here. Fig. 4.33 show the difference between PES and Coupled Cluster energies for structures below 25,000 cm^{-1}. We can see how the HDNNPs improve over time. HDNNP3/FAD-HDNNP shows extreme accuracy for both testing and training. We can see how HDNNP1 initially had around 2000 cm^{-1} deviation which improved to a range of 1000 cm^{-1} for HDNNP2. When we look at QB16, we see that it also has remarkable accuracy comparable to that of FAD-HDNNP.

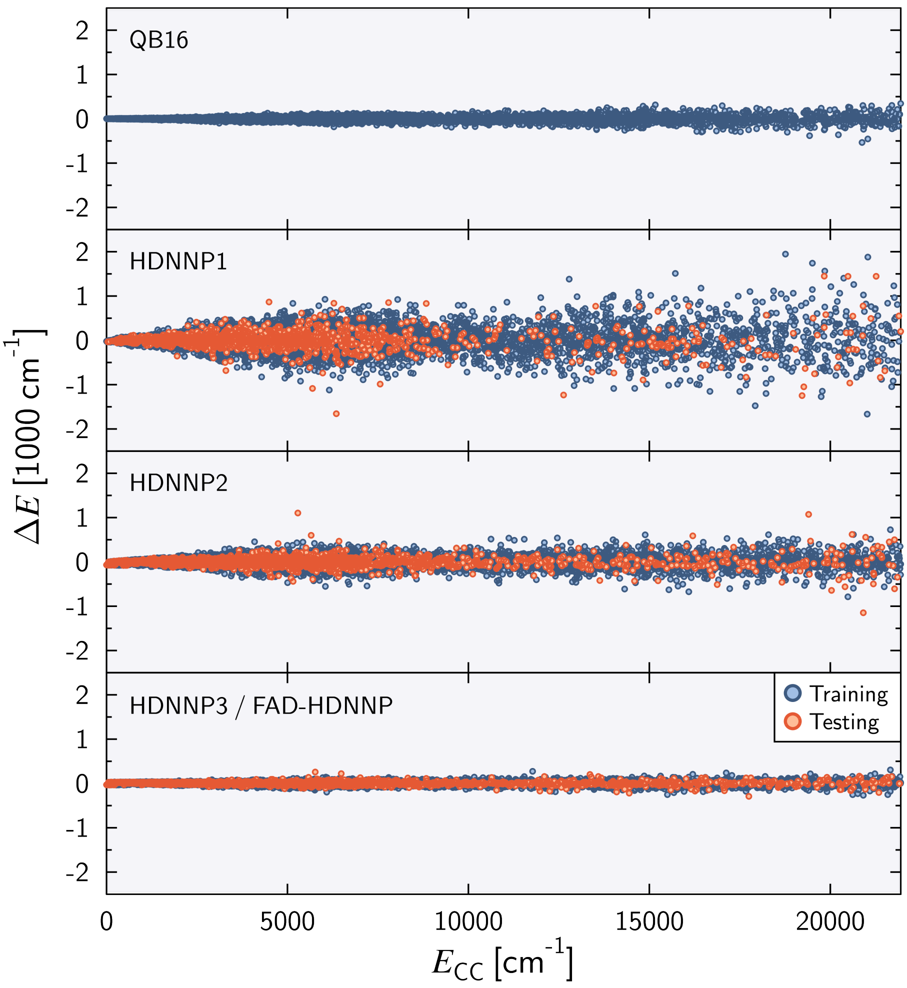


Figure 4.33: The three selected HDNNPs of various iteration up to 21950 cm^{-1} energies in the dataset. The first panel also shows Bowman Potential energies. Here, $\Delta E = E_{\mathrm{CC}} - E_{\mathrm{potential}}$. Reproduced from Ref. [80] with permission from the PCCP Owner Societies.

We can also compare RMSEs for the three potentials which is given in Table 4.23. The RMSE for HDNNP1 was 2.43 meV/atom for training set which reduced to 0.92 meV/atom for HDNNP2 and later 0.37 meV/atom for FAD-HDNNP. Similarly, the very high testing RMSE of 13.99 meV/atom for HDNNP1 reduced to 1.88 meV/atom for HDNNP2 and 2.04 meV/atom for FAD-HDNNP. While observing the more relevant energy range as shown in the table, the testing RMSEs are more reasonable. Below 0.1 Ha, HDNNP1 has 2.15 meV/atom for training and 2.42 meV/atom for testing. HDNNP2 has 0.85 meV/atom training RMSE and 1.04 meV/atom testing RMSE. FAD-HDNNP has very similar accuracy for training and testing in this energy range, 0.35 meV/atom and 0.34 meV/atom. It is to be noted that QB16 has an RMSE 0f 0.91 meV/atom. Overall, the improvement in RMSE has been systematic and we can also see that below 0.1 Ha, even HDNNP1 has a reasonable RMSE value for the testing set.

Table 4.23: Energy root mean squared errors (RMSE) of the training and test sets for the three iterations of HDNNPs. The reference dataset is increased for each iteration using various sampling methods. The RMSEs for both full dataset and the data below 0.1 Ha above equilibrium structure are given.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
full energy range					
HDNNP1	13475	2.43	13.99	196	1129
HDNNP2	27372	0.92	1.88	74	158
HDNNP3	29162	0.37	2.04	30	165
energy range below 0.1 Ha					
HDNNP1	12725	2.15	2.42	174	195
HDNNP2	26531	0.85	1.04	68	83
HDNNP3	28286	0.35	0.34	28	27

Table 4.24 shows the geometrical parameters for the HDNNPs in comparison with default and tight CCSD(T)-F12a/haTZ. We can see that FAD-HDNNP is closer to the ab-initio values as opposed to HDNNP1 and HDNNP2. The only exception being $r(\mathrm{O}\cdots\mathrm{O})$. But, overall all the three HDNNPs give reasonable geometries for the equilibrium Formic Acid Dimer structure. If we compare FAD-HDNNP with QB16, we can see that in most cases except $r(\mathrm{C\text{-}H})$, $r(\mathrm{O}\cdots\mathrm{O})$ and $\angle\mathrm{O\text{-}H}\cdots\mathrm{O}$, FAD-HDNNP gives a closer structure to default Coupled Cluster geometry. Also, $\angle\mathrm{O{=}C\text{-}O}$ is equally well described by both FAD-HDNNP and QB16.

The comparison of the harmonic frequencies using the the HDNNPs and default and tight ab-initio settings is given in Table 4.25. Here, we can see how the frequencies have evolved with various iterations of the HDNNP to give very accurate values. But, in order to see the comparison with the tight settings and other PESs available for the system at hand, we can look at Fig 4.34. As we have already done extensive comparison with the default settings of Coupled Cluster, it is more interesting to look at how the HDNNPs compare with tight settings. The first panel of

Table 4.24: Geometrical parameters of the formic acid dimer minimum structure optimized at the reference CCSD(T)-F12a/haTZ level of theory and various potentials described in the section. Bond lengths are provided in Ångström and angles in degrees.

	Ab initio					
Parameter	default	tight	QB16	HDNNP1	HDNNP2	HDNNP3
r(O–H)	0.9934	0.9932	0.9927	0.9945	0.9925	0.9936
r(C–H)	1.0929	1.0930	1.0929	1.0927	1.0937	1.0927
r(C–O)	1.3113	1.3114	1.3116	1.3104	1.3121	1.3112
r(O$\cdots$O)	2.6748	2.6758	2.6778	2.6729	2.6791	2.6709
r(C=O)	1.2177	1.2176	1.2174	1.2192	1.2172	1.2178
∠O=C–O	126.14	126.14	126.15	126.26	126.13	126.13
∠O=C–H	122.02	122.03	122.05	122.04	122.15	121.99
∠C–O–H	109.77	109.76	109.73	109.42	109.93	109.79
∠O–H$\cdots$O	178.93	178.93	178.95	179.51	179.01	178.86

the figure shows the comparison of various HDNNPs' harmonic fundamentals with tight ab-initio settings. The RMSE as given improves from 27 cm^{-1} to 9 cm^{-1} to 4 cm^{-1} for HDNNP1, HDNNP2 and HDNNP3(FAD-HDNNP) respectively. We can see how in the final FAD-HDNNP, all the frequency deviations lie below 10 cm^{-1}.

The lower panel of Fig. 4.34 shows the available PESs for Formic Acid Dimer. The figure shows how the harmonic fundamentals deviate from the respective reference methods. As can be easily observed, only FAD-HDNNP gives a deviation below 10 cm^{-1} for all the modes of vibration for Formic Acid Dimer. The QB16 [60] with which FAD-HDNNP shares level of theory for reference data has a maximum deviation of 27 cm^{-1}. It has two other fundamentals above the threshold of 10 cm^{-1} deviation. Nevertheless QB16 shows very good accuracy for all the other fundamentals. PES$_{TL}$ [59] has various fundamentals above the threshold of 10 cm^{-1} and has a RMSD of 14 cm^{-1}. FAD-HDNNP has the lowest RMSD followed by QB16 with 8 cm^{-1}. Overall, we see that when it comes to harmonic frequencies for Formic Acid Dimer with respect to the reference method, FAD-HDNNP shows the best performances so far. Of course when it comes to spectroscopy, anharmonic frequencies that compare well with the available experimental data is of the utmost importance. Still, while preparing a PES for applications in spectroscopy, the first milestone to overcome is energetics and harmonic frequencies of a certain quality.

The VPT2 frequencies using FAD-HDNNP and the analysis of it was done by Benjamin Schröder in a collaboration which is published in a recent article [80]. The results from the collaboration are discussed here as an exercise in benchmarking and to further expound upon the features and merits of FAD-HDNNP. Benjamin was able to get VPT2 frequencies from CCSD(T)-F12a and FAD-HDNNP. This is shown in comparison to experimental values in Table 4.26. The FAD-HDNNP VPT2 frequencies show good agreement with the experimental frequencies with a RMSD of 14 cm^{-1}. For v_1, v_2, v_{17} and v_{18}, there are not yet highly reliable experimental values and also VPT2 cannot give reliable values because of anharmonic couplings and resonances. If these modes are excluded, FAD-HDNNP has an RMSD of 9 cm^{-1}.

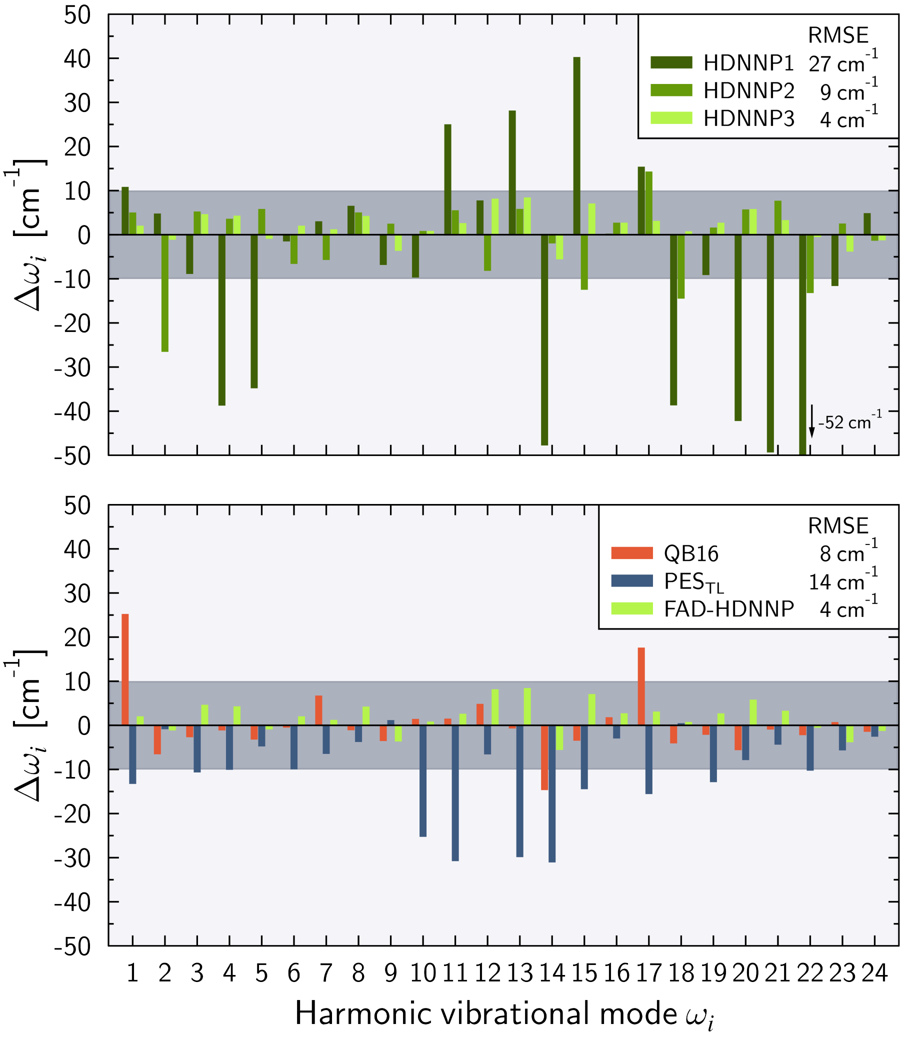


Figure 4.34: Deviation of each of the harmonic frequencies of potentials ω_i from reference tight Coupled cluster frequency. Top panel shows the deviations of various HDNNPs from CCSD(T)-F12a/haTZ harmonic frequencies. $\Delta\omega = \omega_i - \omega_{CC}$. The bottom panel compares the differences in harmonic frequencies from the QB16 PES by Qu and Bowman [77], transfer-learned potential by Käser and Meuwly [59] and the final FAD-HDNNP results. The reference ab initio frequencies are based on the level of theory used to construct the potential. CCSD(T)/aug-cc-pVTZ for PES$_{TL}$ and CCSD(T)-F12a/haTZ for QB16 and FAD-HDNNP. RMSEs are also given in the plot for easy comparison. Reproduced from Ref. [80] with permission from the PCCP Owner Societies.

In collaboration with Edit Matyus and Alberto Martín Santa Daría, 8D intermolecular-plus-torsion vibrational computations were performed on FAD-HDNNP. These are also shown in the paper that is published [80]. Contrary to QB16, the HDNNP1, HDNNP2 and FAD-HDNNP were able to produce potential energy curves along the intermolecular coordinates without holes in the potential. This is shown in Fig.

Table 4.25: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for the three HDNNPs against reference CCSD(T)-F12a/haTZ results. The CCSD(T)-F12a/haTZ frequencies are calculated using default settings and tight convergence criteria.

				FAD-HDNNP	CCSD(T)-F12a	
Mode	Sym.	HDNNP1	HDNNP2	HDNNP3	default	tight
ω_1	A_g	3218	3212	3209	3203	3207
ω_2	A_g	3060	3077	3103	3105	3103
ω_3	A_g	1708	1723	1722	1717	1718
ω_4	A_g	1443	1485	1486	1484	1482
ω_5	A_g	1376	1417	1410	1413	1411
ω_6	A_g	1254	1249	1257	1257	1255
ω_7	A_g	689	681	688	688	686
ω_8	A_g	216	215	214	211	210
ω_9	A_g	164	173	167	171	170
ω_{10}	B_g	1073	1083	1083	1085	1083
ω_{11}	B_g	980	960	957	960	955
ω_{12}	B_g	257	241	257	258	249
ω_{13}	A_u	1128	1106	1109	1102	1101
ω_{14}	A_u	937	983	979	986	985
ω_{15}	A_u	213	161	180	186	173
ω_{16}	A_u	68	71	71	76	68
ω_{17}	B_u	3324	3323	3312	3305	3309
ω_{18}	B_u	3108	3085	3100	3101	3099
ω_{19}	B_u	1773	1783	1784	1782	1782
ω_{20}	B_u	1411	1459	1459	1456	1453
ω_{21}	B_u	1358	1414	1410	1405	1407
ω_{22}	B_u	1208	1247	1260	1260	1260
ω_{23}	B_u	704	718	712	716	715
ω_{24}	B_u	281	275	275	278	276

Table 4.26: Comparison of the VPT2 fundamental frequencies (in cm^{-1}) obtained from the FAD-HDNNP with experimental data

Mode	Sym.	*ab initio*	FAD-HDNNP	Exp.
ν_1	A$_g$	2909	2920	-
ν_2	A$_g$	2942	2948	-
ν_3	A$_g$	1672	1677	1664
ν_4	A$_g$	1431	1433	1430
ν_5	A$_g$	1375	1375	1375
ν_6	A$_g$	1225	1229	1224
ν_7	A$_g$	679	682	681
ν_8	A$_g$	194	197	194
ν_9	A$_g$	157	164	161
ν_{10}	B$_g$	1061	1058	1058
ν_{11}	B$_g$	923	934	911
ν_{12}	B$_g$	241	247	242
ν_{13}	A$_u$	1072	1074	1069
ν_{14}	A$_u$	959	964	939
ν_{15}	A$_u$	162	166	168
ν_{16}	A$_u$	67	68	69
ν_{17}	B$_u$	3044	3041	-
ν_{18}	B$_u$	2935	2941	-
ν_{19}	B$_u$	1741	1745	1741
ν_{20}	B$_u$	1406	1416	1407
ν_{21}	B$_u$	1372	1375	1372
ν_{22}	B$_u$	1234	1233	1234
ν_{23}	B$_u$	704	706	708
ν_{24}	B$_u$	262	264	264

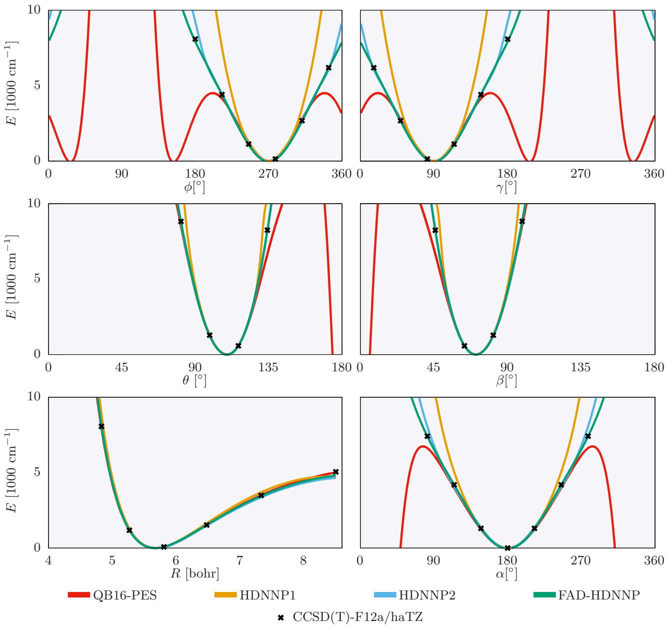


Figure 4.35: The three selected iterations of HDNNPs and Bowman Potential along the 1-D cuts of inter-molecular coordinates of the formic acid dimer used in the reduced-dimensionality variational computations. Reproduced from Ref. [80] with permission from the PCCP Owner Societies.

4.35 The obtained vibrational frequencies are given in Table 4.27. Both QB16 and FAD-HDNNP show very good performance, with only few cm^{-1} deviation from experiments. But, both the potentials still show a blue shift for v_8 and v_9 fundamental frequencies. Edit Matyus and Alberto Martín Santa Daría hypothesise two possibilities for the blue shift. Since the FAD-HDNNP has been well tested and also gave good results with VPT2, one possible reason could be the constrained coordinates used in the variational calculations. If that is the case, the relaxation of the constrained coordinates could fix the problem. The other solution could be increasing the degrees of vibrational freedom used in GENIUSH [120], the software used for computing variational frequencies. But, this could be computationally very costly.

Table 4.27: Vibrational energies with respect to the zero-point vibrational energy (in cm^{-1}) obtained with the 8D($\mathcal{I}$t) intermolecular-torsional model in the GENIUSH program using the Bowman Potential and FAD-HDNNP. These are compared to the experimental values.

Assignment	$\tilde{\nu}_{\mathrm{QB16}}$	$\tilde{\nu}_{\mathrm{FAD\text{-}HDNNP}}$	$\tilde{\nu}_{\mathrm{expt}}$
ν_{16}	70	70	69.2
$2\nu_{16}$	141	140	139
ν_{15}	162	171	168.5
ν_9/ν_8	191	190	161
ν_8/ν_9	208	210	194
$3\nu_{16}$	211	210	
$\nu_{15}+\nu_{16}$	232	240	
ν_{12}	239	243	242
ν_{24}	253	253	264
$\nu_9+\nu_{16}$	262	260	
$\nu_8+\nu_{16}$	277	279	
$4\nu_{16}$	281	280	
$\nu_{15}+2\nu_{16}$	303	309	
$\nu_{12}+\nu_{16}$	310	311	311
$\nu_{24}+\nu_{16}$	323	322	
$2\nu_{15}$	324	330	336
$\nu_9+2\nu_{16}$	332	340	
$\nu_8+2\nu_{16}$	347	348	

4.2.7 Construction of HDNNP with 3D Coupling Terms

After obtaining FAD-HDNNP, there was still an attempt to see whether further improvement of the potential was possible. Since, the previous FAD-HDNNP had two dimensional coupling, 3 dimensional coupling were included to enhance the data-set. Structures were displaced along 3 modes at a time. 125 structures were created per mode coupling. After this, the energy of the structures were predicted using two HDNNPs of the previous iteration; FAD-HDNNP and HDNNP3-i which is described in the Appendix. Structures with a ΔE of more than 3 meV/atom between FAD-HDNNP and HDNNP3-i were selected for adding to the existing reference data. 2522 additional structures are used to train HDNNP4 while comparing to FAD-HDNNP.

Table 4.28: Energy root mean squared errors (RMSE) of the training and test sets for HDNNP4 for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
HDNNP4a	31684	0.324	1.675	26	135
HDNNP4b	31684	0.341	1.328	27	107

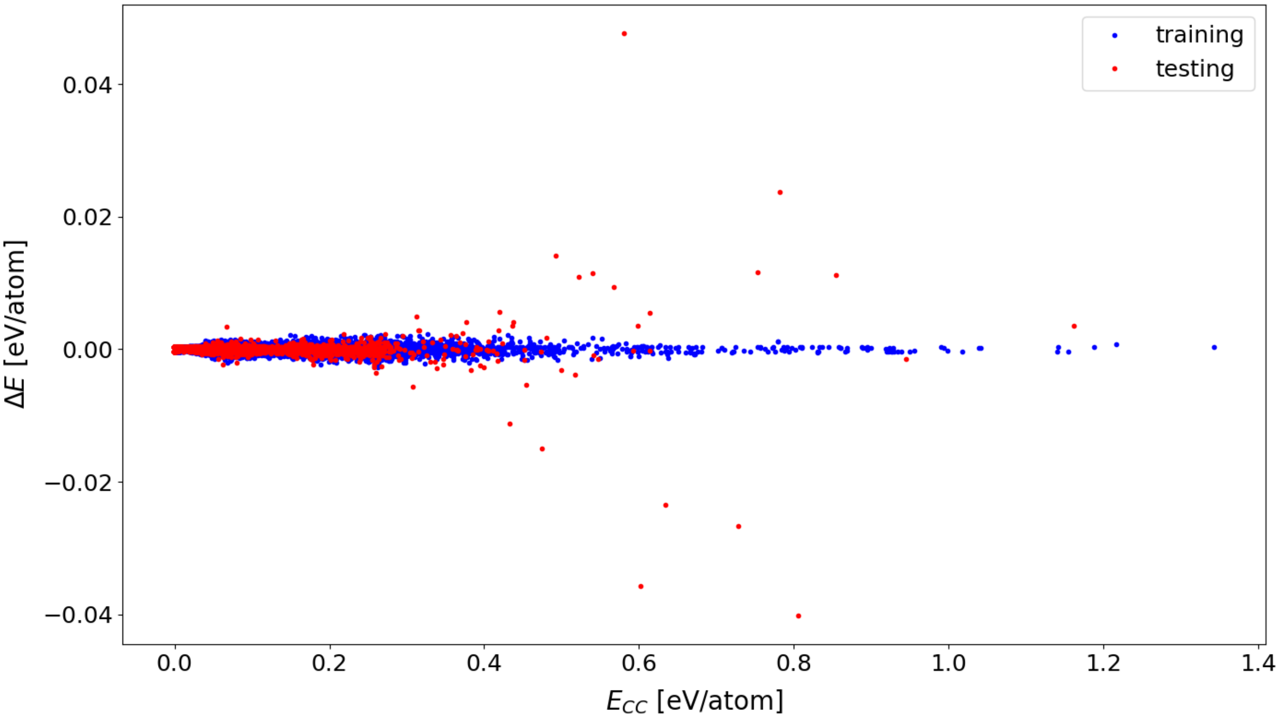


Figure 4.36: Energy difference $\Delta E = E_{\mathrm{CC}} - E_{\mathrm{HDNNP4a}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP4a are provided in Tab. 4.28

The two HDNNPs; HDNNP4a and HDNNP4b of the current iteration are to be assessed in a similar manner as described in the previous sections. Table 4.28

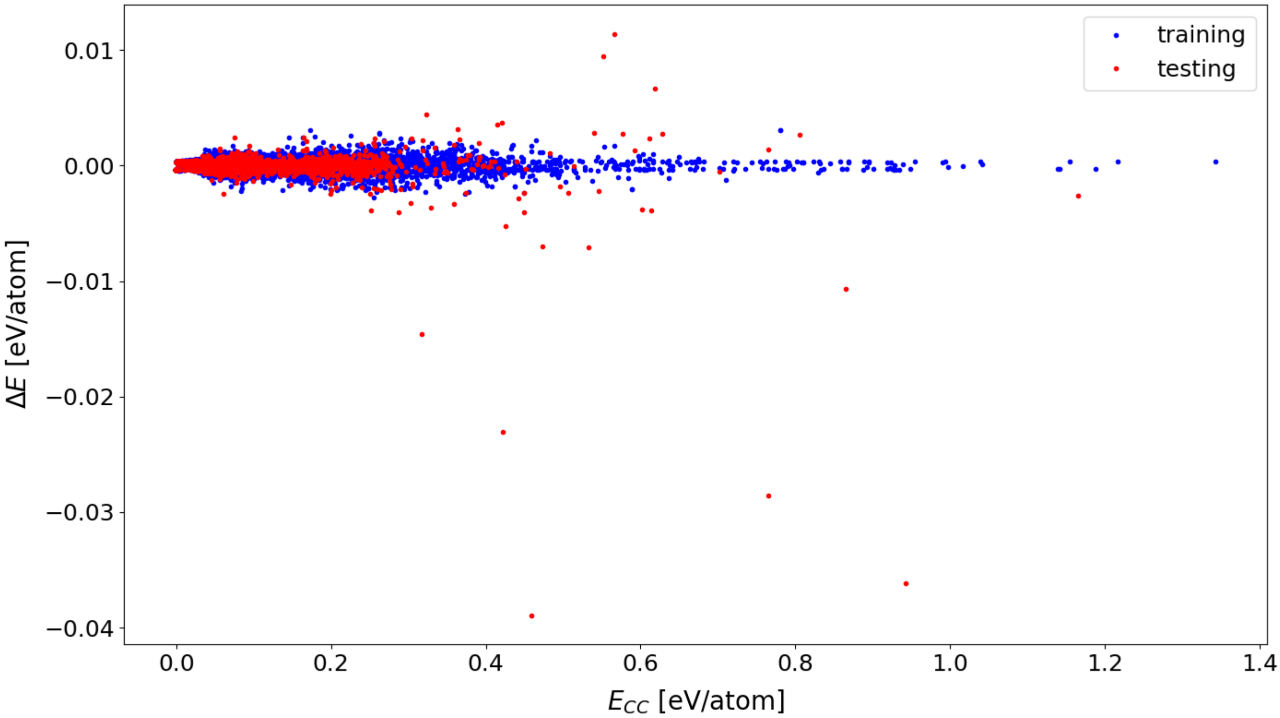


Figure 4.37: Energy difference $\Delta E = E_{\mathrm{CC}} - E_{\mathrm{HDNNP4b}}$ as a function of the reference energy E_{CC}. The root-mean-squared errors (RMSE) for the HDNNP4b are provided in Tab. 4.28

gives the RMSEs for the two HDNNPs. We can immediately observe that the two HDNNPs are of similar quality. The training RMSEs are 0.324 meV/atom and 0.341 meV/atom for HDNNP4a and HDNNP4b respectively. This is an improvement, albeit slight, from FAD-HDNNP which had an RMSE of 0.37 meV/atom. Here, HDNNP4a has a better training RMSE. When we look at the testing RMSE, HDNNP4a has 1.675 meV/atom and HDNNP4b has 1.328 meV/atom. This is an improvement from that of FAD-HDNNP which had 2.04 meV/atom as testing RMSE. Now, HDNNP4b has a better testing RMSE. This again brings the same question on which do we choose if we had to at this point. Both the HDNNPs have similar architecture with 2 hidden layers and 18 nodes per layer. Their only difference in being the randomly selected distribution of training and testing data and the initial weight parameters. HDNNP4a has 28500 structures in training and 3184 structures in testing. HDNNP4b has similar training data and testing data. But, the structures that are selected for training and testing are different.

Figures 4.36 and 4.37 show the difference in energy between Coupled Cluster and HDNNP4s for each structure in the dataset. The structures go up to 1.4 eV/atom reference energy. When we compare HDNNP4a and HDNNP4b, we see that HDNNP4b has a narrower range of deviation going from -0.04 to 0.01 eV/atom as compared to -0.04 to 0.04 eV/atom for HDNNP4a. The structures showing bigger deviations are the ones in the testing set and that is to be expected. FAD-HDNNP has a similar range of deviation going from -0.04 eV/atom to 0.04 eV/atom. This is shown in previous section in Fig. 4.29. Here, even though it seems like HDNNP4b has better description of energies, it is to be noted that the structures that differ greatly for HDNNP4a are in the higher energy region. Therefore, other than the fact that both the HDNNPs are quite close to reference energy, we cannot infer much about the a difference in quality of both the potentials at this point.

Table 4.29 tabulates the harmonic frequencies from HDNNP4a, HDNNP4b and CCSD(T)-F12a/haTZ. It can be observed that the numbers are reasonable. In order to see the deviations from reference harmonic frequencies, Table 4.30 is useful. HDNNP4a has a maximum deviation of 7.74 cm^{-1} whereas HDNNP4b has -8.87 cm^{-1}. Also, RMSD for HDNNP4a is 3.89 cm^{-1} and for HDNNPb it is 4.75 cm^{-1}. FAD-HDNNP had a maximum deviation of 7.36 cm^{-1} and an RMSD of 4 cm^{-1}. We can very well see that HDNNP4a has a very similar quality of frequency with similar maximum deviation and RMSD. Even though the testing energies are the best for HDNNP4b, the frequencies determine that it is slightly worse than HDNNP4a. Nevertheless the difference is in less than 2 cm^{-1} and it is difficult to choose a potential or determine its quality at this stage.

For HDNNP4a, only six other modes show a deviation above 5 cm^{-1}. These are ω_1, ω_2, ω_4, ω_{13}, ω_{16} and ω_{21}. Four modes show a deviation below 1 cm^{-1}. These are ω_3, ω_5, ω_8 and ω_9. When it comes to HDNNP4b, seven other modes show a deviation above 5 cm^{-1}. These are ω_1, ω_3, ω_{10}, ω_{11}, ω_{14}, ω_{15} and ω_{17}. Three modes show a deviation below 1 cm^{-1}. These are ω_{12}, ω_{22} and ω_{23}. Overall, FAD-HDNNP, HDNNP4a and HDNNP4b are of similar quality in terms of energetics and frequencies. The frequency comparison between these three potentials is shown in Fig. 4.38. However, a question can be raised whether it was necessary to invest time in developing these two potentials. First of all, if possible it is always interesting to see how much the quest for accuracy can be pushed. Secondly, to see whether the potential really has improved by including further coupling terms, anharmonic calculations would need to be performed which often requires higher dimensional grids.

Table 4.29: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP4 with frequencies from CCSD(T)-F12a/haTZ level of theory.

Mode	Sym.	HDNNP4a	HDNNP4b	CCSD(T)-F12a/haTZ
ω_1	A_g	3209.3	3210.8	3203.36
ω_2	A_g	3097.1	3106.6	3104.59
ω_3	A_g	1717.6	1723.6	1717.13
ω_4	A_g	1489.1	1485.3	1483.92
ω_5	A_g	1412.6	1411.1	1413.1
ω_6	A_g	1257.9	1253.3	1256.78
ω_7	A_g	684.6	685.7	687.78
ω_8	A_g	211.8	214.2	211.28
ω_9	A_g	171.3	168.3	170.96
ω_{10}	B_g	1087.2	1090.4	1085.04
ω_{11}	B_g	962.5	952.4	959.6
ω_{12}	B_g	259.0	257.8	257.76
ω_{13}	A_u	1107.1	1110.9	1102.03
ω_{14}	A_u	994.2	978.9	986.46
ω_{15}	A_u	182.9	177.3	185.95
ω_{16}	A_u	70.5	74.0	76.36
ω_{17}	B_u	3307.4	3312.5	3305.25
ω_{18}	B_u	3096.6	3096.1	3100.56
ω_{19}	B_u	1778.2	1785.6	1781.57
ω_{20}	B_u	1457.3	1453.2	1455.96
ω_{21}	B_u	1410.5	1407.7	1405.06
ω_{22}	B_u	1263.7	1260.2	1260.06
ω_{23}	B_u	714.0	715.4	715.81
ω_{24}	B_u	281.0	276.9	278.07

Table 4.30: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP4 from CCSD(T)-F12a/haTZ frequencies. $\Delta\omega = \omega_{\mathrm{CC}} - \omega_{\mathrm{HDNNP}}$

Mode	Sym.	HDNNP4a	HDNNP4b
ω_1	A_g	-5.94	-7.44
ω_2	A_g	7.49	-2.01
ω_3	A_g	-0.47	-6.47
ω_4	A_g	-5.18	-1.38
ω_5	A_g	0.5	2
ω_6	A_g	-1.12	3.48
ω_7	A_g	3.18	2.08
ω_8	A_g	-0.52	-2.92
ω_9	A_g	-0.34	2.66
ω_{10}	B_g	-2.16	-5.36
ω_{11}	B_g	-2.9	7.2
ω_{12}	B_g	-1.24	-0.04
ω_{13}	A_u	-5.07	-8.87
ω_{14}	A_u	-7.74	7.56
ω_{15}	A_u	3.05	8.65
ω_{16}	A_u	5.86	2.36
ω_{17}	B_u	-2.15	-7.25
ω_{18}	B_u	3.96	4.46
ω_{19}	B_u	3.37	-4.03
ω_{20}	B_u	-1.34	2.76
ω_{21}	B_u	-5.44	-2.64
ω_{22}	B_u	-3.64	-0.14
ω_{23}	B_u	1.81	0.41
ω_{24}	B_u	-2.93	1.17

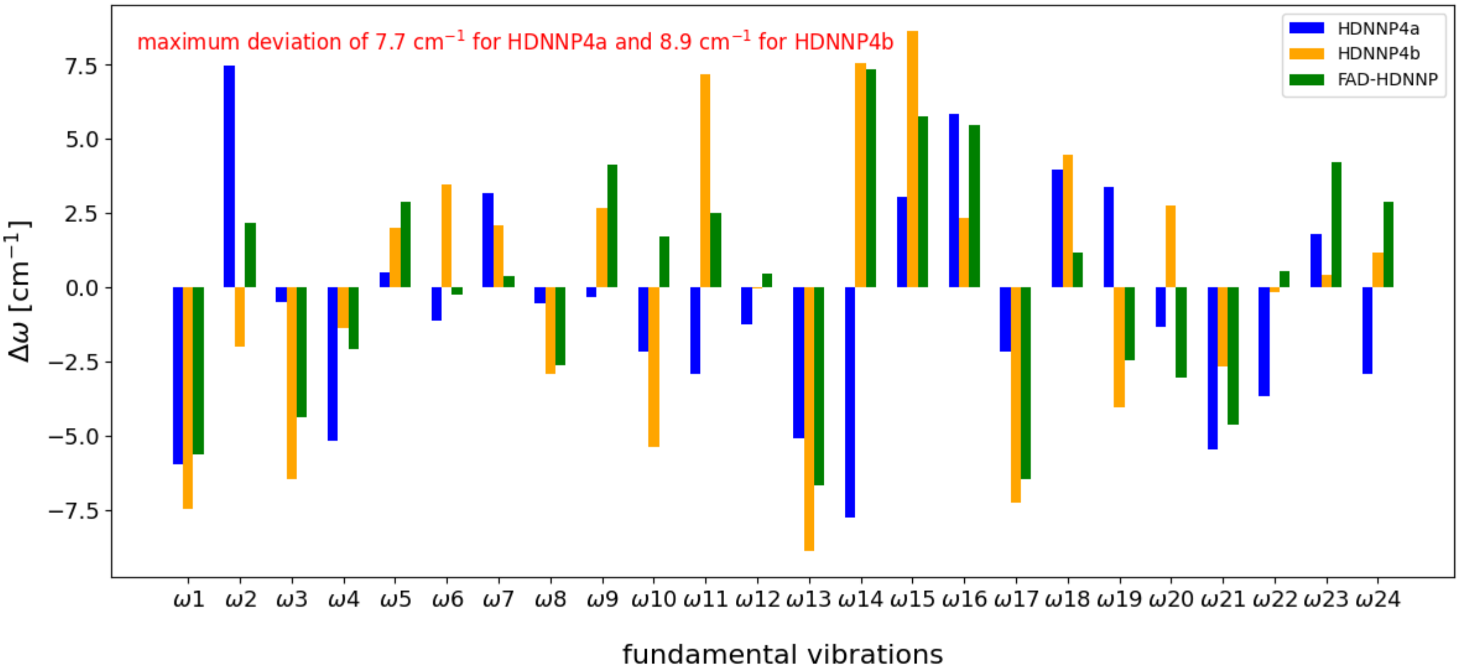


Figure 4.38: Deviations of the harmonic vibrational frequencies ω_i with respect to the reference CCSD(T)-F12a/haTZ frequencies for HDNNP4. $\Delta\omega = \omega_{\mathrm{CC}} - \omega_{\mathrm{HDNNP}}$ for different HDNNPs

Chapter 5

Conclusion and Outlook

The goal of the project was to present a methodology to construct High-Dimensional Neural Network Potentials solely for use in Vibrational Spectroscopy. This includes the procedure to build a new potential, analysing it at each step along the way and to devise approaches for sampling structures and improving the potential. The two systems studied in the project are Formic Acid Monomer and Formic Acid Dimer. Both the systems are heavily studied experimentally and theoretically. Moreover, there are widely used potentials designed for the systems thus providing extensive benchmarking opportunities. Another goal of the project is to test the capabilities and limitations of High-Dimensional Neural Networks especially in a question requiring highly accurate fine tuning like that of vibrational frequencies.

For Formic Acid Monomer, the HDNNPs are compared against the potential from David Tew. The initial dataset obtained from David Tew uses CCSD(T)-F12c/cc-pVTZ-F12 level of theory. The initial HDNNP had a maximum deviation of 24.82 cm^{-1}. The goal is to obtain an HDNNP where all the fundamental Harmonic frequencies are below 10 cm^{-1} deviation from that of the Coupled Cluster fundamentals. The Tew Potential constructed with an analytical potential meant for quantum dynamics has 4.6 cm^{-1} maximum deviation and an RMSD of 2 cm^{-1}. The initial HDNNP has an RMSD of 13 cm^{-1}. This was improved by sampling structures along normal modes using Coupled Cluster eigenvectors to obtain a final HDNNP of expected quality. This HDNNP has a maximum deviation of 7.71 cm^{-1} and an RMSD of 5 cm^{-1}. Though anharmonic frequencies are not computed for the HDNNP, it is possible and would give opportunities for benchmarking with experimental frequencies.

Formic Acid Dimer is a well studied system of particular interest because of the system is used as a benchmark for constructing Potential Energy Surfaces for spectroscopic purposes. Even though it is a ten atom system of not particularly large size and the PES is 24 dimensional, the anharmonic frequencies using high level electronic structure methods are challenging and costly. Therefore, developing a potential of Coupled Cluster quality is essential in providing avenues to perform high level vibrational computations. One of the most widely used Potential for Formic Acid Dimer is the one developed by Joel Bowman. It has been used in computing various anharmonic frequencies like VCI, variational and so on. The goal in constructing the HDNNP was to provide an alternate potential for Formic Acid Dimer which reproduces Harmonic Frequencies more accurately. Thus, in this project the potential by Bowman has been used as a comparison against the HDNNP. At the same time, it was desirable to develop a potential which again produces Harmonic Frequencies with a maximum deviation of 10 cm^{-1} compared to the reference electronic structure method.

The initial dataset obtained from Joel Bowman had 13475 structures with a sparsely populated high energy region. The reference data was calculated using CCSD(T)-F12a/haTZ level of theory. The sparsely populated regions in the dataset can give certain artefacts while evaluating the trained potential. The initial HDNNP has testing structures with higher deviations than the rest which arose as a consequence of the distribution of structures according to energy. The Bowman Potential has a maximum deviation of 27 cm^{-1} with respect to the reference Coupled Cluster fundamentals. The initial HDNNP had a maximum deviation of 49 cm^{-1}. This is to be expected with a sparse dataset.

At this initial stage of developing a potential, all the factors that go into constructing a potential need to be analysed. At the same time, a step by step analysis procedure must be followed to make sure the potential fulfills all the criteria necessary. For example, it is not enough to have a potential which gives accurate frequencies if it fails to describe a large part of the energy surface accurately. Therefore, an analysis of energy and harmonic frequencies need to be done.

The second iteration of the HDNNP was constructed by addressing the weak-points of the previous potential. The symmetry functions were made more flexible as a first step. The collaboration with Edit Matyus gave access to a large dataset from which poorly described structures in the reference data were identified. Sampling structures along normal modes was also done. Along with this, adding structures from MD simulations and structures used in Hessian construction helped in significantly improving the reference data. The thus constructed iteration of HDNNP had a much better description of the harmonic frequencies and energies. The maximum deviation for harmonic fundamentals was 24.86 cm^{-1}.

The third iteration of the HDNNP involved adding two dimensional couplings to the dataset. This has a significant consequence. A good description of couplings is necessary in a potential for getting good quality VPT2 frequencies. The final HDNNP has a maximum deviation of 7.36 cm^{-1} for the harmonic frequencies. Along with this, the good description of energies ensures that this a potential which can be utilised to compute anharmonic frequencies in collaboration with Benjamin Schröder and Edit Matyus. The potential gave good results with VPT2 and Variational methods.

Further development of the HDNNP with three dimensional couplings added give very good results comparable to that of the previous iteration but the further validation of their quality and their would depend on computing anharmonic frequencies.

Overall, in this project a three step procedure is used to construct a good quality potential for spectroscopy use. Initially, the quality of the energies is assessed using RMSE and careful probing of the global potential. This ensures that the potential does not have holes or other unnatural artefacts. The potential is then further developed by adding carefully selected structures employing an active learning method to get all the harmonic fundamental frequencies within a deviation of 10 cm^{-1}. Next step is to ensure that the couplings are well represented by performing VPT2 calculations on the potential and benchmarking with ab-initio VPT2 frequencies.

This potential has applications in computing high level anharmonicities. At the current stage the FAD-HDNNP is the potential available for Formic Acid Dimer with the best description of harmonic frequencies. It also represents a global potential with good quality VPT2 and Variational frequencies [80]. The procedure outlined here is a recipe for constructing good quality potentials for molecular systems which can be used in quantum dynamics and spectroscopic applications. The construction of an accurate and robust potential is time taking but provides opportunities to do high level vibrational calculations with Coupled Cluster qualities. Moreover, if at all the potential has weaknesses, it arises from the quality of the reference method. Hence, a careful selection of reference level of theory and development of an HDNNP based on it can provide dynamical studies of systems for which the system size is a hindrance to perform high quality vibrational calculations.

Bibliography

[1] Mater, A. C. & Coote, M. L. Deep learning in chemistry. *J. Chem. Inf. Model.* **59**, 2545–2559 (2019).

[2] Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).

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

[4] Gasteiger, J. & Zupan, J. Neural networks in chemistry. *Angew. Chem. Int. Ed.* **32**, 503–527 (1993).

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

[6] Blank, T. B., Brown, S. D., Calhoun, A. W. & Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **103**, 4129–4137 (1995).

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

[8] Deringer, V. L., Caro, M. A. & Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **31**, 1902765 (2019).

[9] Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine learning for molecular simulation. *Ann. Rev. Phys. Chem.* **71**, 361–390 (2020).

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

[11] Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater* **20**, 750–761 (2021).

[12] Behler, J. & Csányi, G. Machine learning potentials for extended systems - a perspective. *Eur. Phys. J. B* **94**, 142 (2021).

[13] Handley, C. M. & Popelier, P. L. A. Potential energy surfaces fitted by artificial neural networks. *J. Phys. Chem. A* **114**, 3371–3383 (2010).

[14] Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).

[15] Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).

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

[17] Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).

[18] Artrith, N. & Behler, J. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Phys. Rev. B* **85**, 045439 (2012).

[19] Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **115**, 1032–1050 (2015).

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

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

[22] Bartók, A. P. & Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry* **115**, 1051–1057 (2015).

[23] Thompson, A., Swiler, L., Trott, C., Foiles, S. & Tucker, G. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics* **285**, 316–330 (2015).

[24] Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).

[25] Artrith, N., Morawietz, T. & Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **83**, 153101 (2011).

[26] Morawietz, T., Sharma, V. & Behler, J. A neural network potential-energy surface for the water dimer based on environment-dependent atomic energies and charges. *J. Chem. Phys.* **136**, 064103 (2012).

[27] Ghasemi, S. A., Hofstetter, A., Saha, S. & Goedecker, S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Physical Review B* **92**, 045131 (2015).

[28] Xie, X., Persson, K. A. & Small, D. W. Incorporating electronic information into machine learning potential energy surfaces via approaching the ground-state electronic energy as a function of atom-based electronic populations. *Journal of Chemical Theory and Computation* **16**, 4256–4270 (2020).

[29] Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 398 (2021).

[30] Prudente, F. V., Acioli, P. H. & Soares Neto, J. J. The fitting of potential energy surfaces using neural networks: Application to the study of vibrational levels of h^{3+}. *J. Chem. Phys.* **109**, 8801–8808 (1998).

[31] Bittencourt, A. C. P., Prudente, F. V. & Vianna, J. D. M. The fitting of potential energy and transition moment functions using neural networks: transition probabilities in oh ($a^2\sigma^+$ to $x^2\pi$). *Chem. Phys.* **297**, 153–161 (2004).

[32] Carrington Jr., T. Perspective: Computing (ro-)vibrational spectra of molecules with more than four atoms. *J. Chem. Phys.* **146**, 120902 (2017).

[33] Malshe, M., Narulkar, R., Raff, L. M., Hagan, M., Bukkapatnam, S., Agrawal, P. M. & Komanduri, R. Development of generalized potential-energy surfaces using many-body expansions, neural networks, and moiety energy approximations. *J. Chem. Phys.* **130**, 184102 (2009).

[34] Manzhos, S. & Carrington Jr., T. Using redundant coordinates to represent potential energy surfaces with lower-dimensional functions. *J. Chem. Phys.* **127**, 014103 (2007).

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

[36] Quaranta, V., Hellström, M., Behler, J., Kullgren, J., Mitev, P. & Hermansson, K. Maximally resolved anharmonic oh vibrational spectrum of the water/zno($10\bar{1}0$) interface from a high-dimensional neural network potential. *J. Chem. Phys.* **148**, 241720 (2018).

[37] Morawietz, T., Marsalek, O., Pattenaude, S. R., Streacker, L. M., Ben-Amotz, D. & Markland, T. E. The interplay of structure and dynamics in the raman spectrum of liquid water over the full frequency and temperature range. *J. Phys. Chem. Lett.* **9**, 851–857 (2018).

[38] Shepherd, S., Lan, J., Wilkins, D. M. & Kapil, V. Efficient quantum vibrational spectroscopy of water with high-order path integrals: From bulk to interfaces. *J. Phys. Chem. Lett.* **12**, 9108–9114 (2021).

[39] Sommers, G. M., Calegari Andrade, M. F., Zhang, L., Wang, H. & Car, R. Raman spectrum and polarizability of liquid water from deep neural networks. *Phys. Chem. Chem. Phys.* **22**, 10592–10602 (2020).

[40] Khare, P., Kumar, N., Kumari, K. M. & Srivastava, S. S. Atmospheric formic and acetic acids: An overview. *Reviews of Geophysics* **37**, 227–248 (1999).

[41] Zuckerman, B., Ball, J. A. & Gottlieb, C. A. Microwave Detection of Interstellar Formic Acid. *The Astrophysical Journal* **163**, L41 (1971).

[42] Lattanzi, V., Walters, A., Drouin, B. J. & Pearson, J. C. Submillimeter spectrum of formic acid. *The Astrophysical Journal Supplement Series* **176**, 536 (2008).

[43] Snyder, L. E. Interferometric observations of large biologically interesting interstellar and cometary molecules. *Proceedings of the National Academy of Sciences* **103**, 12243–12248 (2006).

[44] Bertie, J. E. & Michaelian, K. H. The raman spectra of gaseous formic acid -h2 and -d2. *The Journal of Chemical Physics* **76**, 886–894 (1982).

[45] Olbert-Majkut, A., Ahokas, J., Lundell, J. & Pettersson, M. Raman spectroscopy of formic acid and its dimers isolated in low temperature argon matrices. *Chemical Physics Letters* **468**, 176–183 (2009).

[46] Millikan, R. C. & Pitzer, K. S. Infrared spectra and vibrational assignment of monomeric formic acid. *The Journal of Chemical Physics* **27**, 1305–1308 (1957).

[47] Luiz, G., Scalabrin, A. & Pereira, D. Gas phase infrared fourier transform spectra of h12 cooh and h13cooh. *Infrared Physics & Technology* **38**, 45–49 (1997).

[48] Tan, T., Goh, K., Ong, P. & Teo, H. Rovibrational constants for the $\nu 6$ and $2\nu 9$ bands of hcood by fourier transform infrared spectroscopy. *Journal of Molecular Spectroscopy* **198**, 110–114 (1999).

[49] Maçôas, E. M., Lundell, J., Pettersson, M., Khriachtchev, L., Fausto, R. & Räsänen, M. Vibrational spectroscopy of cis- and trans-formic acid in solid argon. *Journal of Molecular Spectroscopy* **219**, 70–80 (2003).

[50] Redington, R. L. Vibrational spectra and normal coordinate analysis of isotopically labeled formic acid monomers. *Journal of Molecular Spectroscopy* **65**, 171–189 (1977).

[51] Madeja, F., Markwick, P., Havenith, M., Nauta, K. & Miller, R. E. Rotationally resolved infrared spectroscopy of h2- and d1-formic acid monomer in liquid he droplets. *The Journal of Chemical Physics* **116**, 2870–2878 (2002).

[52] Baskakov, O. I., Markov, I. A., Alekseev, E. A., Motiyenko, R. A., Lohilahti, J., Horneman, V.-M., Winnewisser, B. P., Medvedev, I. R. & Lucia, F. C. D. Simultaneous analysis of rovibrational and rotational data for the 41, 51, 61, 72, 81, 7191 and 92 states of hcooh. *Journal of Molecular Structure* **795**, 54–77 (2006).

[53] Baskakov, O., Horneman, V.-M., Alanko, S. & Lohilahti, J. Ftir spectra of the $\nu 6$ and $\nu 8$ bands of 13c formic acid molecule—assignment of fir-laser lines. *Journal of Molecular Spectroscopy* **249**, 60–64 (2008).

[54] Cazzoli, G., Puzzarini, C., Stopkowicz, S. & Gauss, J. Hyperfine structure in the rotational spectra of trans-formic acid: Lamb-dip measurements and quantum-chemical calculations*. *A&A* **520**, A64 (2010).

[55] Nejad, A., Suhm, M. A. & Meyer, K. A. E. Increasing the weights in the molecular work-out of cis- and trans-formic acid: extension of the vibrational database via deuteration. *Phys. Chem. Chem. Phys.* **22**, 25492–25501 (2020).

[56] Käser, S., Boittier, E. D., Upadhyay, M. & Meuwly, M. Transfer learning to ccsd(t): Accurate anharmonic frequencies from machine learning models. *Journal of Chemical Theory and Computation* **17**, 3687–3699 (2021).

[57] Tew, D. P. & Mizukami, W. Ab initio vibrational spectroscopy of cis- and trans-formic acid from a global potential energy surface. *The Journal of Physical Chemistry A* **120**, 9815–9828 (2016).

[58] Richter, F. & Carbonnière, P. Vibrational treatment of the formic acid double minimum case in valence coordinates. *The Journal of Chemical Physics* **148**, 064303 (2018).

[59] Käser, S. & Meuwly, M. Transfer learned potential energy surfaces: accurate anharmonic vibrational dynamics and dissociation energies for the formic acid monomer and dimer. *Phys. Chem. Chem. Phys.* **24**, 5269–5281 (2022).

[60] Qu, C. & Bowman, J. M. An ab initio potential energy surface for the formic acid dimer: zero-point energy, selected anharmonic fundamental energies, and ground-state tunneling splitting calculated in relaxed 1–4-mode subspaces. *Phys. Chem. Chem. Phys.* **18**, 24835–24840 (2016).

[61] Richardson, J. O. Full- and reduced-dimensionality instanton calculations of the tunnelling splitting in the formic acid dimer. *Phys. Chem. Chem. Phys.* **19**, 966–970 (2017).

[62] Millikan, R. C. & Pitzer, K. S. The infrared spectra of dimeric and crystalline formic acid. *J. Am. Chem. Soc.* **80**, 3515–3521 (1958).

[63] Bertie, J. E. & Michaelian, K. H. The Raman spectra of gaseous formic acid-h_2 and -d_2. *J. Chem. Phys.* **76**, 886 (1982).

[64] Nejad, A., Meyer, K. A. E., Kollipost, F., Xue, Z. & Suhm, M. A. Slow monomer vibrations in formic acid dimer: Stepping up the ladder with FTIR and Raman jet spectroscopy. *J. Chem. Phys.* **155**, 224301 (2021).

[65] Georges, R., Freytes, M., Hurtmans, D., Kleiner, I., Vander Auwera, J. & Herman, M. Jet-cooled and room temperature ftir spectra of the dimer of formic acid in the gas phase. *Chem. Phys.* **305**, 187–196 (2004).

[66] Matylitsky, V. V., Riehn, C., Gelin, M. F. & Brutschy, B. The formic acid dimer (hcooh)2 probed by time-resolved structure selective spectroscopy. *J. Chem. Phys.* **119**, 10553–10562 (2003).

[67] Ito, F. & Nakanaga, T. Jet-cooled infrared spectra of the formic acid dimer by cavity ring-down spectroscopy: observation of the O-H stretching region. *Chem. Phys.* **277**, 163–169 (2002).

[68] Bertie, J. E., Michaelian, K. H., Eysel, H. H. & Hager, D. The raman-active O–H and O–D stretching vibrations and raman spectra of gaseous formic acid-d1 and -od. *J. Chem. Phys.* **85**, 4779–4789 (1986).

[69] Birer, O. & Havenith, M. High-resolution infrared spectroscopy of the formic acid dimer. *Annu. Rev. Phys. Chem.* **60**, 263 (2009).

[70] Kollipost, F., Larsen, R. W., Domanskaya, A. V., Nörenberg, M. & Suhm, M. A. Communication: The highest frequency hydrogen bond vibration and an experimental value for the dissociation energy of formic acid dimer. *J. Chem. Phys.* **136**, 151101 (2012).

[71] Ortlieb, M. & Havenith, M. Proton transfer in (hcooh)2: An ir high-resolution spectroscopic study of the antisymmetric c–o stretch. *J. Phys. Chem. A* **111**, 7355 (2007).

[72] Zielke, P. & Suhm, M. A. Raman jet spectroscopy of formic acid dimers: Low frequency vibrational dynamics and beyond. *Phys. Chem. Chem. Phys.* **9**, 4528 (2007).

[73] Xue, Z. & Suhm, M. A. Probing the stiffness of the simplest double hydrogen bond: The symmetric hydrogen bond modes of jet-cooled formic acid dimer. *J. Chem. Phys.* **131**, 054301 (2009).

[74] Herman, M., Georges, R., Hepp, M. & Hurtmans, D. High resolution fourier transform spectroscopy of jet-cooled molecules. *Int. Rev. Phys. Chem.* **19**, 277–325 (2000).

[75] Nejad, A. & Suhm, M. A. Concerted pair motion due to double hydrogen bonding: The formic acid dimer case. *J. Ind. Inst. Sci.* **100**, 1–15 (2020).

[76] Qu, C. & Bowman, J. M. Ir spectra of (HCOOH)2 and (DCOOH)2: Experiment, VSCF/VCI, and ab initio molecular dynamics calculations using full-dimensional potential and dipole moment surfaces. *J. Phys. Chem. Lett.* **9**, 2604–2610 (2018).

[77] Qu, C. & Bowman, J. M. High-dimensional fitting of sparse datasets of CCSD(T) electronic energies and MP2 dipole moments, illustrated for the formic acid dimer and its complex IR spectrum. *J. Chem. Phys.* **148**, 241713 (2018).

[78] Qu, C. & Bowman, J. M. Quantum and classical ir spectra of (HCOOH)2, (DCOOH)2 and (DCOOD)2 using ab initio potential energy and dipole moment surfaces. *Faraday Discuss.* **212**, 33–49 (2018).

[79] Martín Santa Daría, A., Avila, G. & Mátyus, E. Fingerprint region of the formic acid dimer: variational vibrational computations in curvilinear coordinates. *Phys. Chem. Chem. Phys.* **23**, 6526–6535 (2021).

[80] Shanavas Rasheeda, D., Martín Santa Daría, A., Schröder, B., Mátyus, E. & Behler, J. High-dimensional neural network potentials for accurate vibrational

frequencies: the formic acid dimer benchmark. *Phys. Chem. Chem. Phys.* – (2022).

[81] Born, M. & Heisenberg, W. Zur quantentheorie der molekeln. *Annalen der Physik* **379**, 1–31 (1924).

[82] Born, M. & Oppenheimer, R. Zur quantentheorie der molekeln. *Annalen der Physik* **389**, 457–484 (1927).

[83] Ritz, W. Über eine neue methode zur lösung gewisser variationsprobleme der mathematischen physik. *Journal für die reine und angewandte Mathematik* **135**, 1–61 (1909).

[84] Hartree, D. R. The wave mechanics of an atom with a non-coulomb central field. part i. theory and methods. *Mathematical Proceedings of the Cambridge Philosophical Society* **24**, 89–110 (1928).

[85] Fock, V. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Zeitschrift fur Physik* **61**, 126–148 (1930).

[86] Roothaan, C. C. J. New developments in molecular orbital theory. *Rev. Mod. Phys.* **23**, 69–89 (1951).

[87] Roothaan, C. C. J. Self-consistent field theory for open shells of electronic systems. *Rev. Mod. Phys.* **32**, 179–185 (1960).

[88] Hall, G. G. & Lennard-Jones, J. E. The molecular orbital theory of chemical valency viii. a method of calculating ionization potentials. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **205**, 541–552 (1951).

[89] Coester, F. & Kümmel, H. Short-range correlations in nuclear wave functions. *Nuclear Physics* **17**, 477–485 (1960).

[90] Čížek, J. On the correlation problem in atomic and molecular systems. calculation of wavefunction components in ursell-type expansion using quantum-field theoretical methods. *The Journal of Chemical Physics* **45**, 4256–4266 (1966).

[91] Fliegl, H., Klopper, W. & Hättig, C. Coupled-cluster theory with simplified linear-r12 corrections: The ccsd(r12) model. *The Journal of Chemical Physics* **122**, 084107 (2005).

[92] Fliegl, H., Hättig, C. & Klopper, W. Inclusion of the (t) triples correction into the linear-r12 corrected coupled-cluster model ccsd(r12). *International Journal of Quantum Chemistry* **106**, 2306–2317 (2006).

[93] Kong, L., Bischoff, F. A. & Valeev, E. F. Explicitly correlated r12/f12 methods for electronic structure. *Chemical Reviews* **112**, 75–107 (2012).

[94] Tew, D. P., Klopper, W., Neiss, C. & Hättig, C. Quintuple-ζ quality coupled-cluster correlation energies with triple-ζ basis sets. *Phys. Chem. Chem. Phys.* **9**, 1921–1930 (2007).

[95] Persson, B. J. & Taylor, P. R. Accurate quantum-chemical calculations: The use of gaussian-type geminal functions in the treatment of electron correlation. *The Journal of Chemical Physics* **105**, 5915–5926 (1996).

[96] Manby, F. R., Werner, H.-J., Adler, T. B. & May, A. J. Explicitly correlated local second-order perturbation theory with a frozen geminal correlation factor. *The Journal of Chemical Physics* **124**, 094103 (2006).

[97] Ditchfield, R., Hehre, W. J. & Pople, J. A. Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules. *The Journal of Chemical Physics* **54**, 724–728 (1971).

[98] Jensen, F. Polarization consistent basis sets: Principles. *The Journal of Chemical Physics* **115**, 9113–9125 (2001).

[99] Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen. *The Journal of Chemical Physics* **90**, 1007–1023 (1989).

[100] Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *Journal of Physics: Condensed Matter* **26**, 183001 (2014).

[101] Blank, T. B. & Brown, S. D. Adaptive, global, extended kalman filters for training feed-forward neural networks. *J. Chemometrics* **8**, 391–407 (1994).

[102] Mills, I. M. 3.2 - vibration–rotation structure in asymmetric- and symmetric-top molecules. In RAO, K. N. & MATHEWS, C. W. (eds.) *Molecular Spectroscopy*, 115–140 (Academic Press).

[103] NIELSEN, H. The vibration-rotation energies of molecules. *REVIEWS OF MODERN PHYSICS* **23**, 90–136 (1951).

[104] Werner, H.-J. *et al.* Molpro, version , a package of ab initio programs. See https://www.molpro.net.

[105] Hättig, C., Tew, D. P. & Köhn, A. Communications: Accurate and efficient approximations to explicitly correlated coupled-cluster singles and doubles, ccsd-f12. *The Journal of Chemical Physics* **132**, 231102 (2010).

[106] Peterson, K. A., Adler, T. B. & Werner, H.-J. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms h, he, b–ne, and al–ar. *The Journal of Chemical Physics* **128**, 084102 (2008).

[107] Adler, T. B., Knizia, G. & Werner, H.-J. A simple and efficient CCSD(T)-F12 approximation. *J. Chem. Phys.* **127**, 221106 (2007).

[108] Knizia, G., Adler, T. B. & Werner, H.-J. Simplified CCSD(T)-F12 methods: Theory and benchmarks. *J. Chem. Phys.* **130**, 054104 (2009).

[109] Werner, H.-J., Knizia, G., Adler, T. B. & Marchetti, O. Benchmark Studies for Explicitly Correlated Perturbation- and Coupled Cluster Theories. *Z. Phys. Chem.* **224**, 493–511 (2010).

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

[111] Dunning, T. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **90**, 1007–1023 (1989).

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

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

[114] Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science* **140**, 171–180 (2017).

[115] Browning, N. J., Ramakrishnan, R., von Lilienfeld, O. A. & Roethlisberger, U. Genetic optimization of training sets for improved machine learning models of molecular properties. *The Journal of Physical Chemistry Letters* **8**, 1351–1359 (2017).

[116] Dral, P. O., Owens, A., Yurchenko, S. N. & Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *The Journal of Chemical Physics* **146**, 244108 (2017).

[117] Peterson, A. A., Christensen, R. & Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Phys. Chem. Chem. Phys.* **19**, 10978–10985 (2017).

[118] Singraber, A., Behler, J. & Dellago, C. Library-based lammps implementation of high-dimensional neural network potentials. *J. Chem. Theory Comput.* **15**, 1827–1840 (2019).

[119] Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Phys.* **117**, 1 (1995).

[120] Mátyus, E., Czakó, G. & Császár, A. G. Toward black-box-type full- and reduced-dimensional variational (ro)vibrational computations. *J. Chem. Phys.* **130**, 134112 (2009).

Appendix A

Appendix

A.1 HDNNP3-i

Number of layers = 2
Number of nodes per layer = 17

Table A.1: Energy root mean squared errors (RMSE) of the training and test sets for HDNNP3-i for the full energy range used in training.

		RMSE [meV/atom]		RMSE [cm^{-1}]	
PES	structures	training	testing	training	testing
HDNNP3-i	29162	0.573	2.254	46	182

Table A.2: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP3-i with frequencies from CCSD(T)-F12a/haTZ level of theory.

Mode	Sym.	HDNNP3-i	CCSD(T)-F12a/haTZ
ω_1	A_g	3201.8	3203.36
ω_2	A_g	3091.8	3104.59
ω_3	A_g	1719.2	1717.13
ω_4	A_g	1491.9	1483.92
ω_5	A_g	1416.6	1413.1
ω_6	A_g	1257.2	1256.78
ω_7	A_g	692.5	687.78
ω_8	A_g	208.4	211.28
ω_9	A_g	169.2	170.96
ω_{10}	B_g	1079.2	1085.04
ω_{11}	B_g	957.6	959.6
ω_{12}	B_g	259.0	257.76
ω_{13}	A_u	1107.2	1102.03
ω_{14}	A_u	988.0	986.46
ω_{15}	A_u	181.7	185.95
ω_{16}	A_u	76.4	76.36
ω_{17}	B_u	3311.2	3305.25
ω_{18}	B_u	3090.3	3100.56
ω_{19}	B_u	1778.5	1781.57
ω_{20}	B_u	1453.9	1455.96
ω_{21}	B_u	1415.2	1405.06
ω_{22}	B_u	1259.0	1260.06
ω_{23}	B_u	722.8	715.81
ω_{24}	B_u	268.4	278.07

Table A.3: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP3-i from CCSD(T)-F12a/haTZ frequencies. $\Delta\omega = \omega_{\text{CC}} - \omega_{\text{HDNNP}}$

Mode	Sym.	HDNNP3-i
ω_1	A_g	1.56
ω_2	A_g	12.79
ω_3	A_g	-2.07
ω_4	A_g	-7.98
ω_5	A_g	-3.5
ω_6	A_g	-0.42
ω_7	A_g	-4.72
ω_8	A_g	2.88
ω_9	A_g	1.76
ω_{10}	B_g	5.84
ω_{11}	B_g	2
ω_{12}	B_g	-1.24
ω_{13}	A_u	-5.17
ω_{14}	A_u	-1.54
ω_{15}	A_u	4.25
ω_{16}	A_u	-0.04
ω_{17}	B_u	-5.95
ω_{18}	B_u	10.26
ω_{19}	B_u	3.07
ω_{20}	B_u	2.06
ω_{21}	B_u	-10.14
ω_{22}	B_u	1.06
ω_{23}	B_u	-6.99
ω_{24}	B_u	9.67

A.2 Analytical Harmonic Frequencies of FAD-HDNNP

Table A.4: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for FAD-HDNNP (analytical and the ones from finite difference) with frequencies from CCSD(T)-F12a/haTZ level of theory.

Mode	Sym.	Analytical	Finite difference	CCSD(T)-F12a/haTZ
ω_1	A_g	3208.8	3209.0	3203.36
ω_2	A_g	3102.3	3102.4	3104.59
ω_3	A_g	1721.5	1721.5	1717.13
ω_4	A_g	1485.9	1486.0	1483.92
ω_5	A_g	1410.2	1410.2	1413.1
ω_6	A_g	1256.9	1257.0	1256.78
ω_7	A_g	687.4	687.4	687.78
ω_8	A_g	213.9	213.9	211.28
ω_9	A_g	166.9	166.8	170.96
ω_{10}	B_g	1083.3	1083.3	1085.04
ω_{11}	B_g	957.0	957.1	959.6
ω_{12}	B_g	257.2	257.3	257.76
ω_{13}	A_u	1108.6	1108.7	1102.03
ω_{14}	A_u	979.02	979.1	986.46
ω_{15}	A_u	180.1	180.2	185.95
ω_{16}	A_u	70.9	70.9	76.36
ω_{17}	B_u	3311.5	3311.7	3305.25
ω_{18}	B_u	3099.3	3099.4	3100.56
ω_{19}	B_u	1783.9	1784.0	1781.57
ω_{20}	B_u	1459.0	1459.0	1455.96
ω_{21}	B_u	1409.7	1409.7	1405.06
ω_{22}	B_u	1259.4	1259.5	1260.06
ω_{23}	B_u	711.6	711.6	715.81
ω_{24}	B_u	275.2	275.2	278.07

A.3 RuNNer Settings

Table A.5: RuNNer settings for HDNNPs

Keyword	Settings
nn_type_short	1
random_number_type	5
global_activation_short	t t l
cutoff_type	1
use_short_nn	
global_hidden_layers_short	2
scale_symmetry_functions	
center_symmetry_functions	

A.3.1 ACSFs

Table A.6: Radial ACSF parameters η for FAM-HDNNP1

element pair	η[Bohr^{-2}]
H-H	0, 0.007400, 0.018800, 0.038800, 0.079700, 0.187800
O-O	0, 0.005100, 0.011800, 0.021500, 0.036200, 0.060500
H-C	0, 0.008800, 0.023800, 0.054000, 0.132700, 0.475000
O-C	0, 0.007200, 0.018200, 0.037200, 0.075200, 0.171100
H-O	0, 0.009200, 0.025200, 0.058700, 0.152500, 0.641700

Table A.7: Angular ACSF parameters for all element combinations of all HDNNPs

No.	η[Bohr^{-2}]	ζ	λ
1	0.0	1.0	1.0
2	0.0	2.0	1.0
3	0.0	4.0	1.0
4	0.0	16.0	1.0
5	0.0	1.0	-1.0
6	0.0	2.0	-1.0
7	0.0	4.0	-1.0
8	0.0	16.0	-1.0

Table A.8: Radial ACSF parameters η for FAM-HDNNP

element pair	η[Bohr^{-2}]
H-H	0, 0.007328, 0.018735, 0.038709, 0.079670, 0.187744
O-O	0, 0.005023, 0.011794, 0.021453, 0.036166, 0.060482
H-C	0, 0.008774, 0.023734, 0.053938, 0.132669, 0.474999
O-C	0, 0.007159, 0.018182, 0.037177, 0.075164, 0.171025
H-O	0, 0.009147, 0.025115, 0.058628, 0.152434, 0.641661

Table A.9: Radial ACSF parameters η for HDNNP-ia and HDNNP-ib

element pair	η[Bohr^{-2}]
H-H	0, 0.004, 0.009, 0.016, 0.028, 0.049, 0.094, 0.215
O-O	0, 0.003, 0.006, 0.010, 0.015, 0.022, 0.032, 0.048
C-C	0, 0.003747, 0.009066, 0.017212, 0.030893, 0.030893
H-C	0, 0.004, 0.009, 0.018, 0.031, 0.056, 0.114, 0.296
O-C	0, 0.004, 0.009, 0.017, 0.029, 0.052, 0.101, 0.241
H-O	0, 0.004, 0.010, 0.019, 0.033, 0.063, 0.134, 0.395

Table A.10: Radial ACSF parameters η for HDNNP3

element pair	η[Bohr^{-2}]
H-H	0, 0.004, 0.009, 0.016, 0.028, 0.049, 0.094, 0.215
O-O	0, 0.003, 0.006, 0.010, 0.015, 0.022, 0.032, 0.048
C-C	0, 0.003747, 0.009066, 0.017212, 0.030893, 0.030893
H-C	0, 0.004, 0.009, 0.018, 0.031, 0.056, 0.114, 0.296
O-C	0, 0.004, 0.009, 0.017, 0.029, 0.052, 0.101, 0.241
H-O	0, 0.004, 0.010, 0.019, 0.035, 0.067, 0.149, 0.486

Table A.11: Radial ACSF parameters η for HDNNP4a and HDNNP4b

element pair	η[Bohr^{-2}]
H-H	0, 0.004, 0.009, 0.016, 0.028, 0.049, 0.094, 0.215
O-O	0, 0.003, 0.006, 0.010, 0.015, 0.022, 0.032, 0.048
C-C	0, 0.003747, 0.009066, 0.017212, 0.030893, 0.030893
H-C	0, 0.004, 0.009, 0.018, 0.031, 0.056, 0.114, 0.296
O-C	0, 0.004, 0.009, 0.017, 0.029, 0.052, 0.101, 0.241
H-O	0, 0.004, 0.010, 0.019, 0.035, 0.067, 0.149, 0.486

A.4 Harmonic Frequencies of Intermediate HDNNPs

Table A.12: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP-i with frequencies from CCSD(T)-F12a/haTZ level of theory.

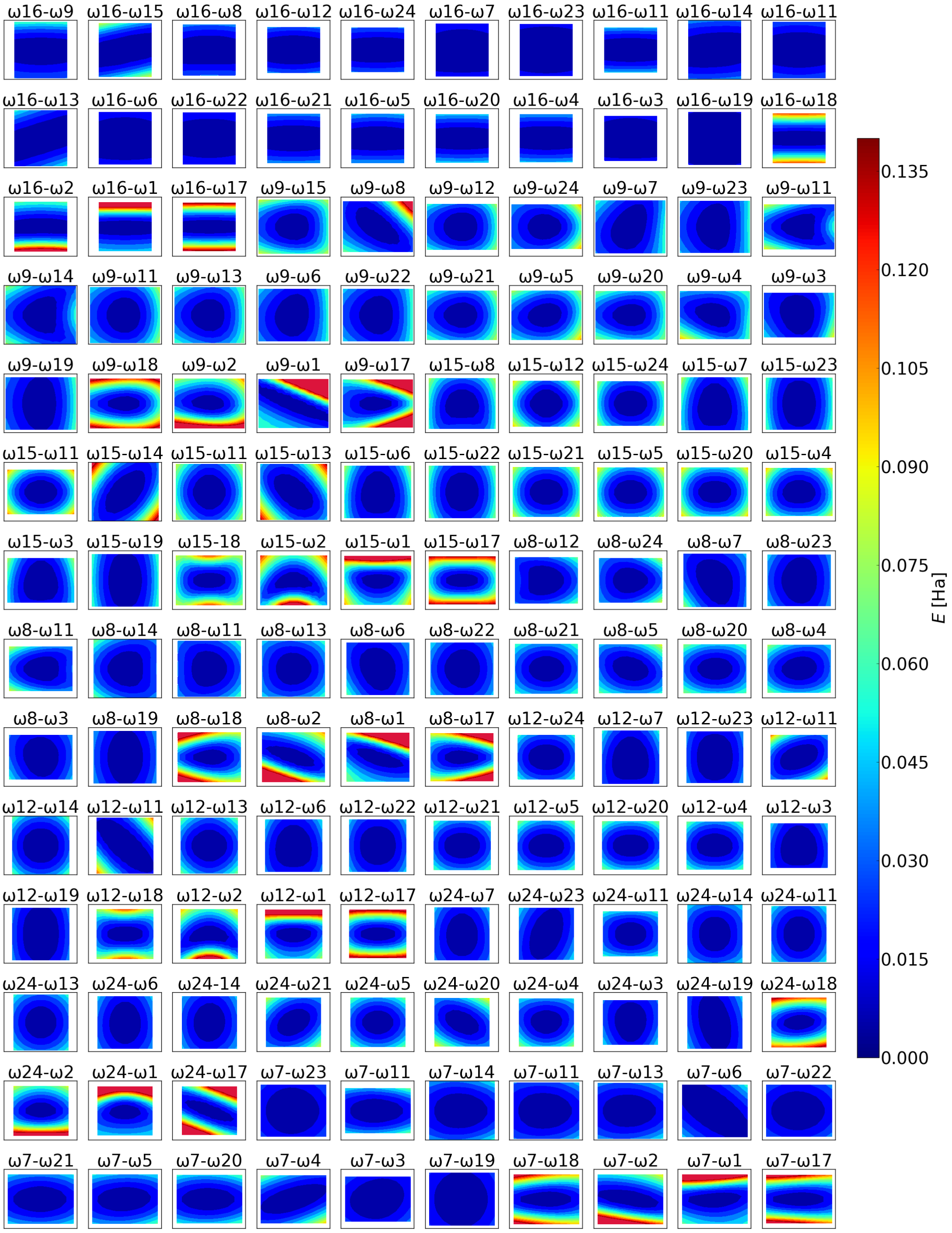
Mode	Sym.	HDNNP-ia	HDNNP-ib	CCSD(T)-F12a/haTZ
ω_1	A_g	3216.1	3203.7	3203.36
ω_2	A_g	3136.5	3139.3	3104.59
ω_3	A_g	1719.4	1708.0	1717.13
ω_4	A_g	1500.5	1480.8	1483.92
ω_5	A_g	1401.6	1407.3	1413.1
ω_6	A_g	1240.0	1248.8	1256.78
ω_7	A_g	684.8	674.5	687.78
ω_8	A_g	216.1	212.1	211.28
ω_9	A_g	132.5	146.3	170.96
ω_{10}	B_g	1072.0	1101.1	1085.04
ω_{11}	B_g	953.2	957.6	959.6
ω_{12}	B_g	257.0	252.1	257.76
ω_{13}	A_u	1134.2	1132.5	1102.03
ω_{14}	A_u	977.2	984.9	986.46
ω_{15}	A_u	165.0	165.8	185.95
ω_{16}	A_u	78.2	64.9	76.36
ω_{17}	B_u	3285.4	3280.9	3305.25
ω_{18}	B_u	3064.5	3061.7	3100.56
ω_{19}	B_u	1751.4	1774.8	1781.57
ω_{20}	B_u	1445.3	1438.3	1455.96
ω_{21}	B_u	1388.7	1361.6	1405.06
ω_{22}	B_u	1273.6	1266.0	1260.06
ω_{23}	B_u	727.3	713.2	715.81
ω_{24}	B_u	275.0	257.2	278.07

Table A.13: Deviation of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP-i from CCSD(T)-F12a/haTZ frequencies. $\Delta\omega = \omega_{\mathrm{CC}} - \omega_{\mathrm{HDNNP}}$

Mode	Sym.	HDNNP-ia	HDNNP-ib
ω_1	A_g	-12.74	-0.34
ω_2	A_g	-31.91	34.71
ω_3	A_g	-2.27	9.13
ω_4	A_g	-16.58	3.12
ω_5	A_g	11.5	5.8
ω_6	A_g	16.78	7.98
ω_7	A_g	2.98	13.28
ω_8	A_g	-4.82	-0.82
ω_9	A_g	38.46	24.66
ω_{10}	B_g	13.04	16.06
ω_{11}	B_g	6.4	2
ω_{12}	B_g	0.76	5.66
ω_{13}	A_u	-32.17	30.47
ω_{14}	A_u	9.26	1.56
ω_{15}	A_u	20.95	20.15
ω_{16}	A_u	-1.84	11.46
ω_{17}	B_u	- 19.85	24.35
ω_{18}	B_u	36.06	38.86
ω_{19}	B_u	30.17	6.77
ω_{20}	B_u	10.66	17.66
ω_{21}	B_u	6.36	43.46
ω_{22}	B_u	-13.54	-5.94
ω_{23}	B_u	-11.49	2.61
ω_{24}	B_u	3.07	20.87

Table A.14: Comparison of the harmonic frequencies ω_i (in cm^{-1}) for HDNNP3 with various Hessian step sizes in Bohr.

Mode	Sym.	0.001	0.005	0.01	0.025	0.05	CCSD(T)-F12a/haTZ
ω_1	A$_g$	3208.7	3208.8	3209.0	3210.5	3215.8	3203.36
ω_2	A$_g$	3102.2	3102.2	3102.4	3103.3	3106.8	3104.59
ω_3	A$_g$	1721.5	1721.5	1721.5	1721.8	1722.7	1717.13
ω_4	A$_g$	1486.0	1485.9	1486.0	1486.3	1487.3	1483.92
ω_5	A$_g$	1410.1	1410.1	1410.2	1410.3	1410.8	1413.1
ω_6	A$_g$	1256.9	1256.9	1257.0	1257.3	1258.5	1256.78
ω_7	A$_g$	687.4	687.4	687.4	687.4	687.3	687.78
ω_8	A$_g$	214.0	213.9	213.9	213.8	213.5	211.28
ω_9	A$_g$	166.9	166.9	166.8	166.7	166.3	170.96
ω_{10}	B$_g$	1083.3	1083.2	1083.3	1083.5	1084.1	1085.04
ω_{11}	B$_g$	957.0	957.0	957.1	957.5	958.9	959.6
ω_{12}	B$_g$	257.2	257.2	257.3	257.6	258.6	257.76
ω_{13}	A$_u$	1108.6	1108.6	1108.7	1108.8	1109.5	1102.03
ω_{14}	A$_u$	979.1	979.0	979.1	979.4	980.7	986.46
ω_{15}	A$_u$	179.9	180.2	180.2	180.7	182.3	185.95
ω_{16}	A$_u$	70.8	70.9	70.9	71.3	72.9	76.36
ω_{17}	B$_u$	3311.4	3311.5	3311.7	3313.1	3318.4	3305.25
ω_{18}	B$_u$	3099.2	3099.2	3099.4	3100.4	3103.9	3100.56
ω_{19}	B$_u$	1783.9	1783.9	1784.0	1784.3	1785.3	1781.57
ω_{20}	B$_u$	1459.0	1459.0	1459.0	1459.3	1460.2	1455.96
ω_{21}	B$_u$	1409.7	1409.7	1409.7	1409.8	1410.2	1405.06
ω_{22}	B$_u$	1259.4	1259.4	1259.5	1259.8	1261.0	1260.06
ω_{23}	B$_u$	711.6	711.6	711.6	711.6	711.5	715.81
ω_{24}	B$_u$	275.2	275.2	275.2	275.0	274.5	278.07


ω16-ω9 ω16-ω15 ω16-ω8 ω16-ω12 ω16-ω24 ω16-ω7 ω16-ω23 ω16-ω11 ω16-ω14 ω16-ω11
ω16-ω13 ω16-ω6 ω16-ω22 ω16-ω21 ω16-ω5 ω16-ω20 ω16-ω4 ω16-ω3 ω16-ω19 ω16-ω18
ω16-ω2 ω16-ω1 ω16-ω17 ω9-ω15 ω9-ω8 ω9-ω12 ω9-ω24 ω9-ω7 ω9-ω23 ω9-ω11
ω9-ω14 ω9-ω11 ω9-ω13 ω9-ω6 ω9-ω22 ω9-ω21 ω9-ω5 ω9-ω20 ω9-ω4 ω9-ω3
ω9-ω19 ω9-ω18 ω9-ω2 ω9-ω1 ω9-ω17 ω15-ω8 ω15-ω12 ω15-ω24 ω15-ω7 ω15-ω23
ω15-ω11 ω15-ω14 ω15-ω11 ω15-ω13 ω15-ω6 ω15-ω22 ω15-ω21 ω15-ω5 ω15-ω20 ω15-ω4
ω15-ω3 ω15-ω19 ω15-18 ω15-ω2 ω15-ω1 ω15-ω17 ω8-ω12 ω8-ω24 ω8-ω7 ω8-ω23
ω8-ω11 ω8-ω14 ω8-ω11 ω8-ω13 ω8-ω6 ω8-ω22 ω8-ω21 ω8-ω5 ω8-ω20 ω8-ω4
ω8-ω3 ω8-ω19 ω8-ω18 ω8-ω2 ω8-ω1 ω8-ω17 ω12-ω24 ω12-ω7 ω12-ω23 ω12-ω11
ω12-ω14 ω12-ω11 ω12-ω13 ω12-ω6 ω12-ω22 ω12-ω21 ω12-ω5 ω12-ω20 ω12-ω4 ω12-ω3
ω12-ω19 ω12-ω18 ω12-ω2 ω12-ω1 ω12-ω17 ω24-ω7 ω24-ω23 ω24-ω11 ω24-ω14 ω24-ω11
ω24-ω13 ω24-ω6 ω24-14 ω24-ω21 ω24-ω5 ω24-ω20 ω24-ω4 ω24-ω3 ω24-ω19 ω24-ω18
ω24-ω2 ω24-ω1 ω24-ω17 ω7-ω23 ω7-ω11 ω7-ω14 ω7-ω11 ω7-ω13 ω7-ω6 ω7-ω22
ω7-ω21 ω7-ω5 ω7-ω20 ω7-ω4 ω7-ω3 ω7-ω19 ω7-ω18 ω7-ω2 ω7-ω1 ω7-ω17
0.135
0.120
0.105
0.090
0.075
0.060
0.045
0.030
0.015
0.000
E [Ha]

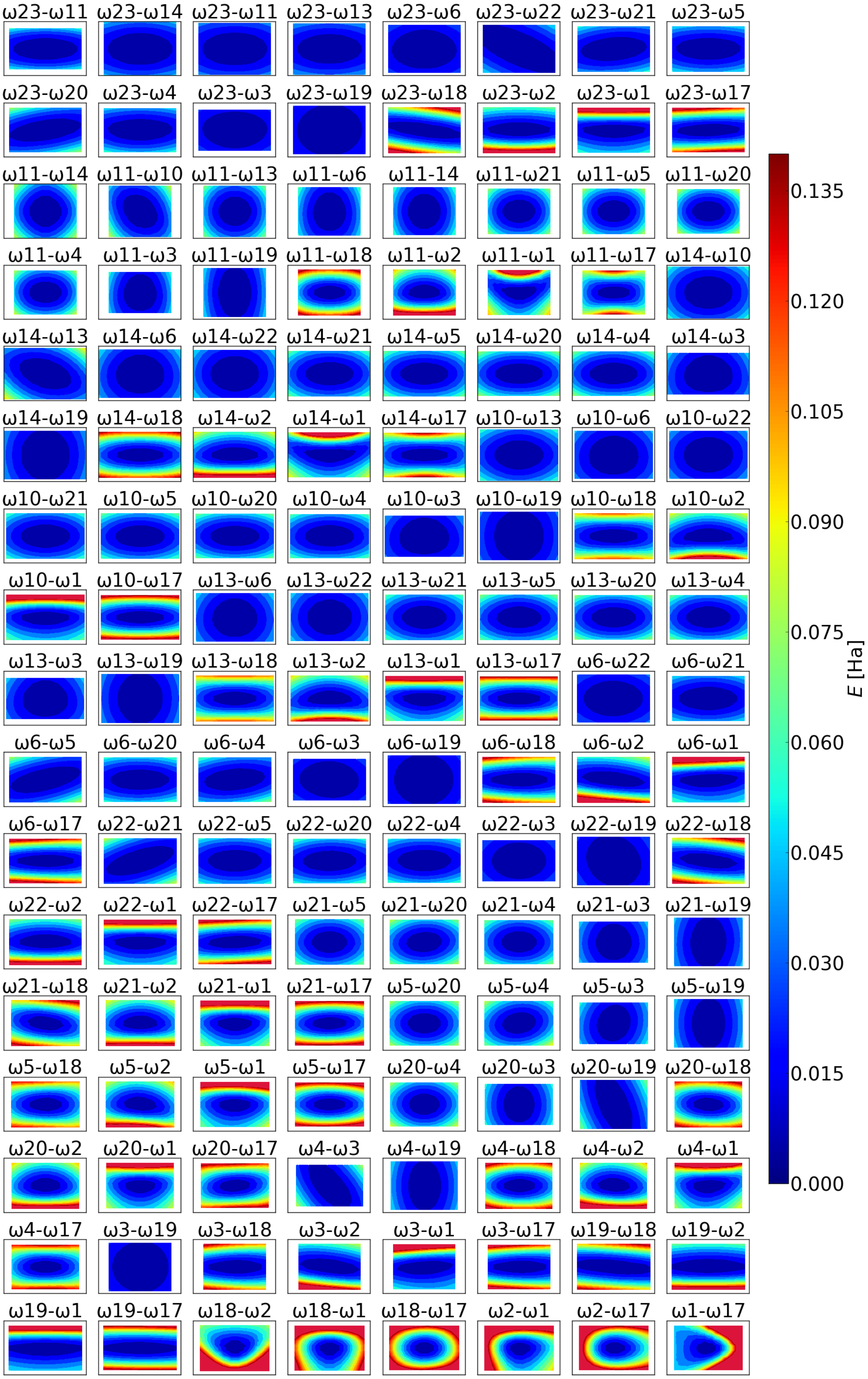
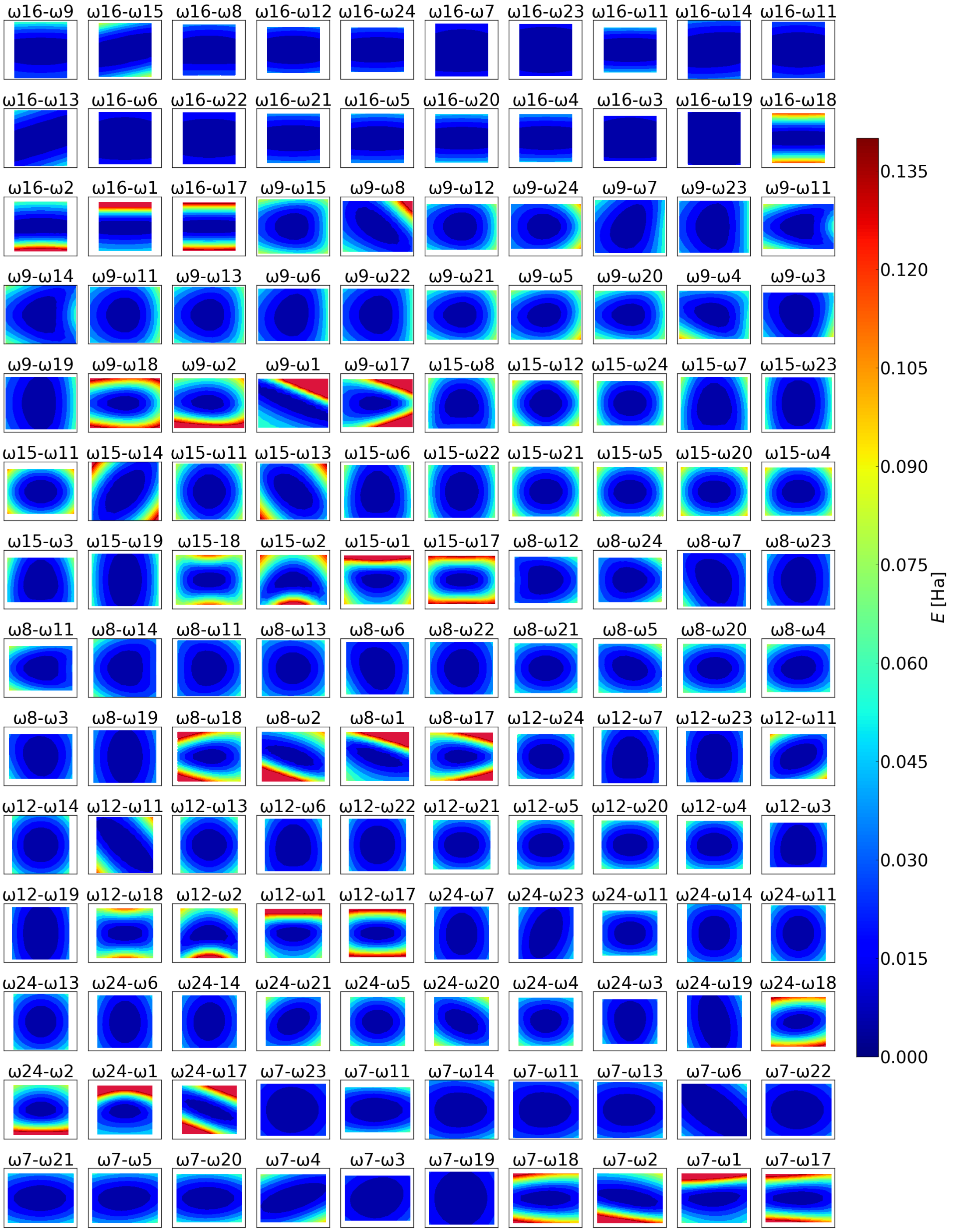


Figure A.1: $E_{HDNNP2a}$ for displacements along two normal modes at a time. The displacements are made using Coupled Cluster eigenvectors.


ω16-ω9
ω16-ω15
ω16-ω8
ω16-ω12
ω16-ω24
ω16-ω7
ω16-ω23
ω16-ω11
ω16-ω14
ω16-ω11
ω16-ω13
ω16-ω6
ω16-ω22
ω16-ω21
ω16-ω5
ω16-ω20
ω16-ω4
ω16-ω3
ω16-ω19
ω16-ω18
ω16-ω2
ω16-ω1
ω16-ω17
ω9-ω15
ω9-ω8
ω9-ω12
ω9-ω24
ω9-ω7
ω9-ω23
ω9-ω11
ω9-ω14
ω9-ω11
ω9-ω13
ω9-ω6
ω9-ω22
ω9-ω21
ω9-ω5
ω9-ω20
ω9-ω4
ω9-ω3
ω9-ω19
ω9-ω18
ω9-ω2
ω9-ω1
ω9-ω17
ω15-ω8
ω15-ω12
ω15-ω24
ω15-ω7
ω15-ω23
ω15-ω11
ω15-ω14
ω15-ω11
ω15-ω13
ω15-ω6
ω15-ω22
ω15-ω21
ω15-ω5
ω15-ω20
ω15-ω4
ω15-ω3
ω15-ω19
ω15-18
ω15-ω2
ω15-ω1
ω15-ω17
ω8-ω12
ω8-ω24
ω8-ω7
ω8-ω23
ω8-ω11
ω8-ω14
ω8-ω11
ω8-ω13
ω8-ω6
ω8-ω22
ω8-ω21
ω8-ω5
ω8-ω20
ω8-ω4
ω8-ω3
ω8-ω19
ω8-ω18
ω8-ω2
ω8-ω1
ω8-ω17
ω12-ω24
ω12-ω7
ω12-ω23
ω12-ω11
ω12-ω14
ω12-ω11
ω12-ω13
ω12-ω6
ω12-ω22
ω12-ω21
ω12-ω5
ω12-ω20
ω12-ω4
ω12-ω3
ω12-ω19
ω12-ω18
ω12-ω2
ω12-ω1
ω12-ω17
ω24-ω7
ω24-ω23
ω24-ω11
ω24-ω14
ω24-ω11
ω24-ω13
ω24-ω6
ω24-14
ω24-ω21
ω24-ω5
ω24-ω20
ω24-ω4
ω24-ω3
ω24-ω19
ω24-ω18
ω24-ω2
ω24-ω1
ω24-ω17
ω7-ω23
ω7-ω11
ω7-ω14
ω7-ω11
ω7-ω13
ω7-ω6
ω7-ω22
ω7-ω21
ω7-ω5
ω7-ω20
ω7-ω4
ω7-ω3
ω7-ω19
ω7-ω18
ω7-ω2
ω7-ω1
ω7-ω17
0.135
0.120
0.105
0.090
0.075
0.060
0.045
0.030
0.015
0.000
E [Ha]

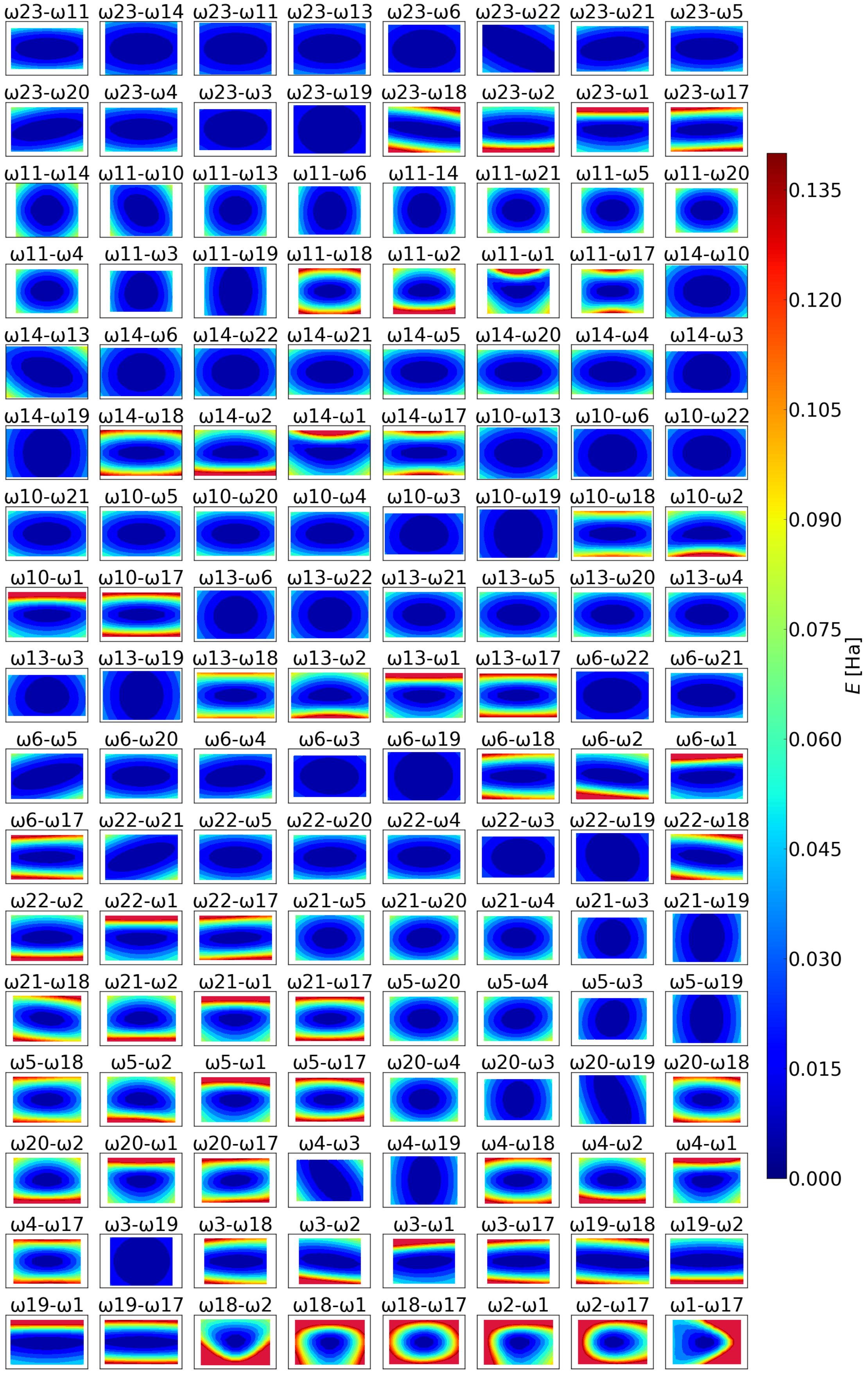


Figure A.2: $E_{HDNNP2b}$ for displacements along two normal modes at a time. The displacements are made using Coupled Cluster eigenvectors.

Curriculum Vitae

Personal Information

Full Name	**Dilshana Shanavas Rasheeda**
Date of Birth	**18.10.1994**
Place of Birth	**Kerala, India**
Citizenship	**India**
Email	dilshanasr@gmail.com

2019-2023	**Promotion** in the group of Prof. Dr.Jörg Behler
2016-2018	**Integrated Master of Science** in the group of Dr. U. Lourderaj

Göttingen, **9.12.2022**

(**Dilshana Shanavas Rasheeda**)