Dynamic structure investigation and spectra prediction of biomolecules using machine learning techniques

Vom Promotionsausschuss der
Technischen Universität Hamburg

zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation (Monografie)

von
Amir Kotobi

aus
Teheran

2024

1. Gutachter: Prof. Dr.-Ing. Robert Meißner

2. Gutachterin: Prof. Dr. Sadia Bari

Prüfungsvorsitzender: Prof. Dr. Patrick Huber

Tag der mündlichen Prüfung: 06.06.2024

Author : https://orcid.org/0000-0002-1488-2847

DOI: https://doi.org/10.15480/882.9689

Handle: https://hdl.handle.net/11420/47867

Dedicated to my family and my partner in life...

Acknowledgements

I would like to extend my deepest gratitude to those who have supported me throughout this journey, each contributing in their unique and invaluable ways.

To my beloved girlfriend, Giulia, your unwavering support, patience, and encouragement have been my anchor. Your belief in me and your continuous motivation have been instrumental in helping me stay focused and determined. I am profoundly grateful for your love and companionship, which have made this journey more bearable and rewarding. Grazie alla tua famiglia per il costante sostegno e per avermi reso più felice in questo percorso.

To my family, thank you for your constant support and encouragement. Mom, Dad, and sister, your endless love, guidance, and sacrifices have provided me with the foundation to pursue my dreams. Your confidence in my abilities has been a driving force in my life, and I owe my success to your nurturing and support.

To my brother, Amjad, your support and understanding have meant the world to me. Your words of wisdom and occasional comic relief have kept me grounded and reminded me of the importance of balance in life. Thank you for always being there for me.

To my close friends, Kaan, Mahdi, Yahya, Thea, Sebastian, Marvin, Maryam, and Dawit, your encouragement have been a source of great strength. The countless discussions, shared experiences, and the laughter we've shared have been a much-needed reprieve from the stresses of this journey. Your friendship has been a pillar of support, and I am incredibly thankful for each one of you.

This thesis would not have been possible without the collective support and love from all of you. Thank you from the bottom of my heart.

Zusammenfassung

Die Infrarot- und Röntgenabsorptionsspektroskopie haben sich als leistungsfähige experimentelle Instrumente zur die elektronischen und strukturellen Feinheiten von Biomolekülen, insbesondere Peptiden und Proteinen, aufzuklären. Parallel dazu haben die bemerkenswerten Fortschritte bei den Rechenkapazitäten die Fähigkeit beschleunigt die Fähigkeit, Chemie, Physik und maschinelles Lernen in einer echten Symbiose zu kombinieren, wodurch die präzise Modellierung und Verständnis komplexer biomolekularer Prozesse auf atomarer Ebene und die Validierung von experimentell beobachteten Spektralmerkmalen. Doch die inhärente Komplexität von Peptiden und Proteinen, gekoppelt mit den Rechenanforderungen quantenmechanischer Methoden für große Systeme stellen jedoch eine große Herausforderung dar, wenn es darum geht, die inhärenten Eigenschaften dieser Biomolekülen. Um diese Herausforderungen zu bewältigen, ist die Einbeziehung von überwachten und unüberwachten Techniken des maschinellen Lernens in die Molekulardynamik-Simulations-Toolbox erleichtert die das komplexe Zusammenspiel interatomarer und intermolekularer Wechselwirkungen zu entschlüsseln und den Weg für die den Weg für die Vorhersage verschiedener Eigenschaften dieser Systeme. Diese Dissertation befasst sich mit Feature und Techniken des unüberwachten maschinellen Lernens (z. B. Clustering und Dimensionality-Reduction), die auf atomistische Datensätze angewandt werden, um zu untersuchen, wie diese Techniken die komplexe Strukturlandschaft eines Modellpeptids beleuchten können. Darüber hinaus werden in dieser Arbeit Graph neuronale Netze als leistungsstarker und effizienter Ansatz zur Entschlüsselung der komplizierten

Abstract

The investigation of biomolecular structures and the prediction of their spectra using experimental and theoretical studies in the gas phase represent fundamental steps in comprehending their intrinsic properties and biological functions. Nonetheless, the complexity of the potential energy surface of biomolecules, combined with limitations in computational resources, limits the interpretation of experimental observations. Integrating supervised and unsupervised machine learning (ML) techniques into theoretical calculations is considered as an effective way to address these challenges.

Infrared (IR) and X-ray absorption spectroscopy (XAS) has proven to be powerful experimental techniques to study the electronic and spatial structure of biomolecules such as peptides and proteins. Reproducing and validating the features observed in spectra resulting from these experiments often requires the use of sophisticated *ab initio* calculations and comprehensive understanding of biomolecules' configurational space.

In this thesis, I introduced a novel approach in interpretation of IR experimental spectrum of a peptide which aims enhancing the exploratory power of searching configurational space by combining REMD simulations, unsupervised machine learning, and *ab initio* calculations. This scheme relies on a set of structural descriptors and data-driven clustering technique which accounts for canonical ensemble of real experimental condition to obtain an accurate computed spectrum. We show that by partitioning the configurational space into subensembles of similar conformations i.e. clusters, an accurate IR spectrum can be calculated by averaging the IR contribution of each representative conformer in each cluster, weighted according to the population of each cluster. While this approach unravels important fingerprints of experimental spectroscopic data, the calculation of IR and particularly XAS spectra, due to its inherently expensive theoretical computation, is often computationally prohibitive task for even medium-sized molecules.

To remedy the computational obstacles associated with spectra prediction, we develope a data-driven supervised ML frameworks, i.e. graph neural networks which are trained on a custom-generated XAS dataset to find a mapping between structures and spectroscopic signals, thus bypassing the need for expensive *ab initio* quantum chemistry calculations. To insure the interpretability of GNN models' predictions, we employ feature attribution to determine the respective contributions of various atoms in the molecules to the peaks observed in the XAS spectrum. Within this approach, we show that it is possible to link the peaks observed in the spectra to certain core and virtual orbitals from the quantum chemical calculations and obtain an in-depth understanding of the ML predicted XAS spectrum.

The results presented in this thesis show that the integration of supervised and unsupervised ML techniques can effectively enhance the interpretation of spectroscopic data and make efficient use of the expensive *ab initio* calculations.

Table of Contents

1

Introduction to Investigation of Biomolecules

1. Introduction to Investigation of Biomolecules

Biomolecules such as peptides and proteins are building blocks of living organisms and perform the most important functions in biological processes. The study of biomolecules is driven by the multitude of biological processes they participate in and their applications across various fields. Peptides and proteins play crucial role in numerous areas including therapeutics [1], drug delivery systems [2], biotechnology, industrial enzymes [3], agricultural applications [4], structural biology, and drug discovery [5]. The functionality of peptides and proteins relies on the chemical properties of their amino acids and their conformation. Therefore, understanding their structure is pivotal for gaining insights into their molecular-level functionality across various fields. A significant portion of biomolecule research focus on understanding the 3-dimensional structure and their impact on their function, using different experimental techniques such as X-ray crystallography [6], Nuclear Magnetic Resonance (NMR) spectroscopy [7] and gas-phase X-ray absorption spectroscopy (XAS) [8, 9]. Notably, one can observe a trend where these experimental techniques are used together with computational modeling and simulations of biomolecules to provide insight into their dynamics, folding mechanisms, and interactions, while keeping the computational expense of this investigation at affordable scale [10, 11, 12]. However, understanding the biological processes, driven by complex biological systems in both spatial and temporal scales, still remains a challenging task despite the emergence and evolution of new experimental and theoretical techniques. These challenges often arise from time complexity of simulating complex biological systems [13], difficulty in experimental validation [14], and lack of sufficient data for modelling and analysis of many biomolecules [15]. Furthermore, conformational space of biomolecules is highly complex which makes these molecules undergo transitions involving large numbers of atoms between an enormous number of different conformations [16, 17]. This poses a challenge for an accurate and quantitative description of the experimental data by theoretical approaches.

The conformational space in high-dimenesional potential energy surface (PES) of biomolecules is determined by many factors which influence the number of possible conformations and the difficulty of exploring the entire space. Number of degrees of freedom, size of the biomolecule, conformational flexibility, complex non-covalent interactions within the biomolecule, and solvent effects are just a few to mention but are often the main contributors to the complexity of PES. To explore the conformational space of peptides and proteins, sophisticated computational methods including molecular dynamics simulations, Monte carlo simulations and *ab initio* quantum mechanics are the common approaches [18, 19, 20, 21]. While exploring the PES of biomolecules is conceptually a well defined problem in obtaining a fundamental understanding of structure-property relationships, thoroughly exploring complex PES of biomolecules proves to be prohibitive in terms of computation for already many medium-sized biomolecules. To remedy this problem, enhanced sampling techniques such as metadynamics [22], umbrella sampling [23], and replica exchange molecular dynamics [24] are often employed in coarse-grained models to overcome energy barriers and sample rare events in reduced computational complexity. Enhanced sampling techniques partly elucidate the intricate relationships between biomolecule properties and complex PES, which

account for the conformational diversity observed in measured spectroscopic spectra. However, directly validating theoretical spectra and establishing connections with experimental measurements remains a formidable challenge because of the above mentioned many influencing factors which possibly need to be included [12].

A relevant portion of research on biomolecules continues to strive for advanced and efficient investigation of their intrinsic physical properties to answer fundamental questions about electron dynamics and the interplay between the electronic and spatial structures of peptides and proteins [25, 26, 9]. Infrared (IR) and soft X-ray spectroscopy which are often combined with mass spectrometry, are proposed as analytical tools to reveal the secondary structure and study the electron dynamics of peptides and proteins [27, 28, 29, 30]. A useful approach to facilitate direct comparison of experimental spectra to theoretical predictions is to neglect the solvation effects and study the biomolecule in gas phase. Investigating different secondary structure motifs in the gas phase is thus a valuable approach that eases the investigation of structure-property relationships and helps disentangling the delicate balance between enthalpic and entropic contributions in an unperturbed environment [31, 32] Additionally, these studies can provide insights into the molecule's inherent behavior in a solvent environment [33]. Among gas-phase experimental techniques, IR spectroscopy is widely used as it provides detailed insights into the three-dimensional molecular structure, as well as into intra- and intermolecular interactions of gas-phase protein ions [30, 34, 35, 36] and polymers [37]. In gas-phase experiments, electrospray ionization (ESI) [38] gives the opportunity to bring molecular ions into solvent-free gas phase in which the native-like conformations in the absence of solvent remains intact [39]. These molecular ions can then be guided in vacuum using ion manipulation tools such as ion mass filters and ion traps. Removing solvation effects in both theory and experiments simplifies the understanding of the intrinsic behavior of biomolecules. However, the low free-energy barriers between conformations and the vast conformational space of even relatively small peptides, especially for intrinsically disordered proteins and peptides, result in overlapping signals from the many nonspecific, i.e. nonhelical or sheet-like, conformations commonly encountered in experimental measurements [40, 41]. Discrepancies between theoretical predictions and experimental observations, particularly for intrinsically disordered peptides and proteins with highly complex configurational spaces, presents a significant challenge [27, 42, 43, 44, 45]. Addressing discrepancies between theory and experiments has been carried out in several studies using different organic materials [45, 44, 43, 46]. As it will be demonstrated in this thesis, considering only a few individual conformations in theoretical predictions without taking into account their correct statistical ensemble weight (which is common practice in computational spectroscopy) can hamper the interpretation of experimental results and lead to discrepancies between theory and experiment. The computational expenses of *ab initio* calculations for large biomolecules is yet another common problem faced in calculating their spectra and investigating their secondary structure motifs in the gas phase [47].

The development of Machine Learning (ML) techniques has promoted the role of data as a promising and powerful catalyst in discovery of previously unknown insights into biomolecules. ML techniques are being increasingly applied to various areas of theoretical and computational chemistry given their ability to infer structure-property relationships on the basis of large amounts of data [48, 49, 50]. Both supervised and unsupervised ML approaches have been extensively incorporated into theoretical calculations to analyze atomistic datasets and predict crucial chemical properties of biomolecules [51, 48, 52, 53]. The combined growth of computational resources and these ML techniques has enabled several investigations to avoid many costly experiments by leveraging a substantial amount of data available for training and inference [54]. For instance, the application of supervised and unsupervised ML to data coming from atomistic simulations has provided a broad spectrum of theoretical breakthroughs ranging from understanding the complex configurational space of biomolecules to predicting *ab initio* level properties such as IR or XAS spectra [55, 56, 57].

With the introduction of dimensionality reduction and clustering techniques compatible to datasets of atomic trajectories obtained from atomistic simulations, theoretical prediction of experimental observations is often initialised with obtaining the structural motifs, i.e. recurring low energy conformations of PES, from raw atomisic datasets [51, 58]. This analysis can often be hampered due to high dimensionality and redundancy of atomistic datasets. In this case, the combination of unsupervised ML and feature engineering e.g. use of atomic descriptors, on the atomistic dataset creates an efficient pipeline which helps to condense the vast amount of data in the raw atomistic dataset into meaningful information of the conformational space. This narrowed-down information can then be utilized for further theoretical calculations of important biomolecular properties. In particular, a careful choice of unsupervised ML techniques in combination with enhanced sampling simulation pave the way of incorporating a realistic representation of the canonical ensemble into spectra prediction. While understanding the conformational space of biomolecules is crucial in prediction of experimental spectra, accurately calculating XAS and IR spectra yet requires sophisticated and expensive *ab initio* quantum chemistry.

Integration of supervised ML frameworks to investigating biomolecules has opened the path of bypassing expensive electronic structure calculations and predicting accurate spectroscopic properties by training a predictive model which infers the spectra of a biomolecule based on its structure [59, 60, 61, 62, 63, 64, 65, 66]. Among those ML techniques, graph neural networks (GNN) and deep neural networks (DNN) are promising candidates to predict the properties of matter such as the electronic structure [67] at a higher computational speed already made them favorable for high-throughput calculations in materials design and drug discovery [68, 69]. Thus, the ability to perform efficient computations with high accuracy has demonstrated that ML techniques are advantageous in domains such as various types of spectroscopy including vibrational and optical [67, 59, 70, 60, 61, 62, 71, 63, 12, 64, 65, 66]. While these methods are very effective in spectra prediction by capturing both local and global chemical environment of atoms,

understanding the rationale behind the predictions made by these black-box models, remains a challenge which requires integrating explainability techniques into these ML frameworks.

The overall goal of this thesis is to implement supervised and unsupervised ML on atomistic datasets to not only understand the importance of conformational ensemble of biomolecules for an accurate and robust spectra prediction but also leverage them to extend high-throughput calculations for predicting the spectra of biomolecules beyond traditional approaches. This investigation has led to the introduction of data-driven approaches to explore and include the conformational space of a biomolecule in reproducing and interpreting experimental IR spectra. Additionally, by implementing supervised ML for spectra prediction of biomolecules, we introduce a framework to predict XAS spectra by GNN models, while integrating further the explainability of the predicted XAS peaks through these ML models.

The thesis is structured as follows: In chapter 2; an introduction of experiments as well as machine learning techniques in atmostic datasets is given; in chapter 3; after a brief introduction on general overview of using unsupervised machine learning in exploring the conformational space of a model peptide, we show the reconstruction of IR spectrum by considering the canonical ensemble of the peptide and, ultimately, in chapter 4, we introduce a recipe of integrating explainability techniques into GNN models trained on custom-generated dataset for fast and efficient XAS prediction.

2

Machine Learning and Spectroscopy Methods for Biomolecules

2.1 Experiments

Infrared multiphoton dissociation spectroscopy (IRMPD) at room temperature is a technique used in mass spectroscopy to fragment molecules in gas phase for investigating the vibrational properties, gaining insight into the molecular structure and dynamics [72]. IRMPD utilizes the principles of multiphoton absorption, where multiple photons of infrared light are absorbed by a molecule simultaneously, leading to the excitation of vibrational modes [73, 74]. Through the careful selection of laser wavelengths and intensities, it becomes possible to selectively excite specific vibrational modes within a molecule. In IRMPD experiment, the laser light which has inherently high intensity and many coherent photons, causes the intense vibration of the bonds by multi photon absorption. The excitation energy finally surpasses the dissociation energy of a particular bond, the molecule undergoes fragmentation, leading to the formation of fragment ions. To identify the vibrational modes, the yield of the fragments dependent on the photon wavelength is plotted. The technique has been previously used to obtain insights into the three-dimensional molecular structure, as well as into intra- and intermolecular interactions of gas-phase biomolecules [75, 76, 77].

For the data represented in chapter 3, room temperature IRMPD experiments in the fingerprint region, $600 - 1800\,\mathrm{cm}^{-1}$, were performed at the Free Electron Laser for Infrared eXperiments (FELIX) laboratory in Nijmegen (The Netherlands). The Free-Electron Laser (FEL) was coupled to a Bruker AmaZon ETD quadrupole ion trap mass spectrometer which was modified to have optical access to the ion trap [78]. The IR frequency was calibrated using a grating spectrometer. Protonated biomolecular ions were generated from solution by electrospray ionization (ESI) source, mass-to-charge isolated in the trapping region, and irradiated with a single infrared laser pulse from the FEL (5-100 mJ per pulse, bandwidth 0.4% of the IR frequency) to induce wavelength-dependent IRMPD. Precursor and fragment ion intensities were determined from six averaged mass spectra at each IR frequency. IRMPD spectra in the $2700 - 3700\,\mathrm{cm}^{-1}$ range were recorded at the Centre Laser IR d'Orsay (CLIO) FEL facility in Orsay (France) [79] using a modified 7 T hybrid FT-ICR mass spectrometer (APEX-Qe Bruker) coupled to the table-top optical parametric oscillator/amplifier (OPO/OPA, Laser Vision). This experimental setup for the first time here implemented with a 10 Hz Nd:YAG pumped OPO/ OPA system (Surelite II,10-Continuum, Laser Vision). The protonated biomolecular ions were delivered from the solution to the gas phase by ESI with a typical extraction voltage of 4 kV and desolvation temperature of $150 - 200\,^{\circ}\mathrm{C}$. Ions were accumulated, pre-mass-selected and thermalized in a quadrupole-hexapole interface and then pulse extracted towards the FT-ICR cell maintained under high vacuum ($< 10^{-9}$ mbar) and at room temperature. Mass-selected ions were stored and irradiated for 2 s with the OPO/OPA light in the frequency range of interest. In order to increase fragmentation, ions were also exposed to an

auxiliary broad-band CO_2 laser synchronized with the OPO/OPA (Universal Laser System, 10 W, continuous wave operation centered at $\lambda = 10.6\,\mu m$), at the beginning of the OPO/OPA irradiation period and for times varying between 7 to 12 ms. The CO_2 laser pulse length was adjusted to avoid photo-dissociation of the molecule by the CO_2 laser alone while promoting the fragmentation in the presence of the OPO/OPA. The IR spectra in both ranges are obtained by plotting the IRMPD yield as a function of IR frequency, where the yield is defined as $-\ln\left(I_p/\sum\left(I_f + I_p\right)\right)$ with the precursor and fragment ion intensities (I_p and I_f). In the 600 to 1800 cm^{-1} fingerprint region, the IR yield was further linearly corrected for the frequency dependent variation of the FEL pulse energy [80].

2.2 Atomistic simulations

This section focuses on describing the atomistic simulations which is necessary in describing the behavior and properties of biomolecules [81]. Employing atomic-level simulations on biomolecules, such as peptides and proteins, in short timescales of milliseconds or nanoseconds enables to uncover key biochemical processes including protein folding [82], drug binding [83] and conformational changes critical to the biomolecules' functions [84]. Such simulations, beside experimental techniques, serve as an important supplement to reveal biomolecular processes in spatial and temporal scales which are otherwise difficult to investigate in experiments [81].

The potential energy surface (PES) of a biomolecule that is a multi-dimensional surface describing how the energy of a biomolecular system changes based on the geometry [85] (shown schematically in Figure 2.1), is known to be highly complex due to the number of atoms and their intricate interactions in the molecule. PES can be defined classically as an energy function $E(\mathbf{R}_1, \ldots, \mathbf{R}_N)$ in which the atomic positions (x,y,z coordinates) of N atoms are $\mathbf{R}_1, \ldots, \mathbf{R}_N$. In intrinsically disordered biomolecules, often a number of local minima along with global minimum exists on the PES which is closely related to thermodynamic variables such as pressure, temperature, entropy, and volume. Elucidating the PES and free energy surface of a biomolecule, which govern its folding and structural formation, provides insights into the molecule's behavior in experiments [85]. Due to computational expensive *ab initio* approaches (which solve the many body electronic problem), force fields molecular dynamics (MD) is often used to thoroughly investigate the PES of biomolecules.

2.2.1 Force fields Molecular Dynamics in gas phase

Investigation of structure-property relationships and other important properties of the biomolecule such as protonation or interactions with ions, can be studied with gas-phase simulations, allowing

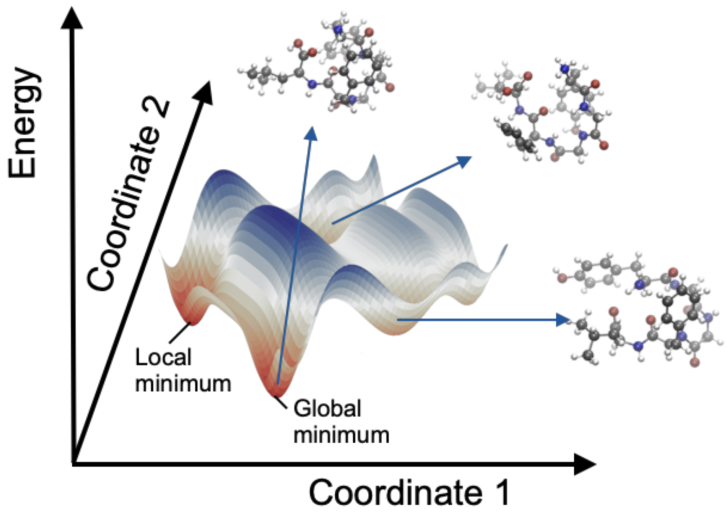

Figure 2.1: Schematic representation of a model PES with the free energy as a function of two coordinates. Examples of corresponding conformations of some local and global minima is shown, indicating the relationship between the geometry and the potential energy of the biomolecule.

a direct validation of gas-phase experimental results. Performing gas-phase molecular dynamics not only provides the opportunity to investigate various secondary structure motifs of a peptide or protein but also helps to disentangle the delicate balance between enthalpic and entropic contributions in an unperturbed environment.

In summary, MD simulations solve Newton's equations of motions for N interacting atoms in which the forces between the atoms are the negative gradients of a potential function. The equations are solved in small time steps (typically in femtosecond range) while keeping the temperature and pressure in required values. To run the MD simulations in canonical ensemble NVT (constant-number (N), constant-volume (V), and constant-energy (E)), velocity-rescaling thermostat [86] was used. This themostat is essentially a Berendsen thermostat [87] with an additional stochastic term that ensures a correct kinetic energy distribution. Temperature fluctuations can still occur with NVT at MD simulations. These fluctuations which can lead to inadequate sampling of PES [88], often occur due to force fields errors, i.e. truncation of forces and energies. Velocity-rescaling thermostat produces a correct canonical ensemble and still has the advantage of the Berendsen thermostat: first order decay of temperature deviations and no oscillations [89]. The output of the MD simulation is the coordinates of the atoms as a function of time, representing a trajectory in the high-dimensional PES of the biomolecule. To allow a computationally easier approach to explore the biomolecule's PES and estimate the proper free energies, empirical approximation,

namely force fields commonly used in MD simulations. Empirical potentials in force fields divide in different terms which correspond to different interactions exists in the system [85]. The terms in the force fields are based on several parameters that can either be obtained from experiments or from reference quantum mechanical (e.g. Density Functional Theory). The commonly employed functional form found in many biomolecular force fields can be written as,

$$E(\mathbf{R}_1, \ldots, \mathbf{R}_N) = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{electrostatic}} + E_{\text{van der waals}} \tag{2.1}$$

in which $\mathbf{R}_1, \ldots, \mathbf{R}_N$ are the atomic positions of N atoms. The first three terms correspond to bonded interactions between the atoms. Bonded interactions are based on fixed list of atoms which are not exclusively pair interactions but include 3- and 4-body interactions as well. Bond stretching (2-body), bond angle (3-body), and dihedral angle (4-body) interactions are some of these bonded interactions. The remaining terms in the above equation correspond to non-bonded interactions between the atoms. The non-bonded interactions contain a repulsion term, a dispersion term, and a Coulomb term. The repulsion and dispersion term (long range Van der Waals interactions) are combined in the Lennard-Jones. Additionally, (partially) charged atoms act through the Coulomb term.

Unlike *ab initio* methods, the electrons of the system are not explicitly considered in force fields. However, due to large computational costs of *ab initio* calculations for even small to medium size biomolecules, force fields MD simulations are often used to get statistical information and sample the large conformational space of biomolecules.

Although performing molecular dynamics (MD) simulations offers numerous advantages in understanding the intrinsic properties of peptides and proteins, there are several challenges associated with force fields based MD simulations in the gas phase. The primary problem arise from the lack of comprehensive force fields which are specifically designed for simulation of gas-phase biomolecules [90, 91]. Solution-based force fields are often used for simulating gas-phase biomolecules even if certain parameters such as charge distributions around surface-exposed atoms may differ in gas phase [92]. To remedy this problem, polarizable force fields can be used to consider the charge distributions in MD simulations [93, 94]. However, the higher computational cost of these force fields compared to simple point charge simulations, limit the use of them and make fixed charge solution force fields more favorable. As a solution force fields are parameterized from *ab initio* vacuum data of small molecules, such force fields perform quite well in the gas phase [95, 91]. Moreover, the atoms in large biomolecules such as proteins, mostly not “feel” the gas-phase environment since the atoms are buried in the interior [91, 90].

2.2.2 Temperature Replica Exchange Molecular Dynamics

Complex free-energy landscape of biomolecules with large number of local minima (or metastable states) has urged the need to enhance comformational sampling in MD simulations and to obtain reliable free-energy landscapes of biomolecules [96, 97, 98]. Replica-exchange molecular dynamics (REMD) is a method used to more extensively explore the PES of biomolecules [99, 100]. In temperature REMD simulations, copies of the biomolecule, denoted as replicas, are simulated at different temperatures in parallel. Pair of replicas exchange temperatures using the Metropolis algorithm in MD simulations [100]. Metropolis algorithm ensures that the equilibrium distribution of configurations is preserved. It calculates the acceptance probability based on the energy difference between the proposed exchange and the current configuration (atomic positions) when attempting to exchange configurations between replicas. In temperature REMD, low and high temperatures are experienced by each replica, passing the large energy barriers and thoroughly searching the conformational space of the system. Each replica searches large conformational changes at high temperature as well as stable conformations with low potential energies at low temperatures [100]. Therefore, REMD provide the opportunity to improve conformational sampling and provide accurate thermal statistical information of the system in various conditions. Figure 2.2

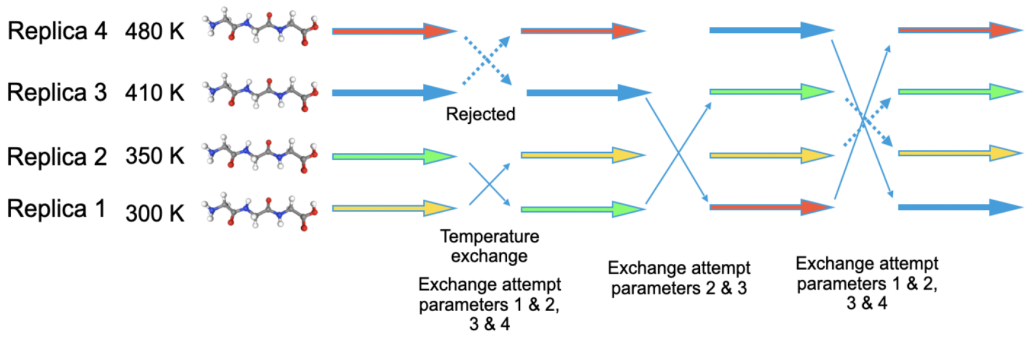

Figure 2.2: Schematic representation of temperature replica-exchange molecular dynamics. The idea of this figure is adopted from [100].

demonstrates how different replicas at various temperatures are simulated in temperature REMD. During temperature REMD, the pair of replicas exchange their temperatures when the Metropolis criteria is satisfied for the transition probability, $w(\mathbf{X} \to \mathbf{X}')$

$$w(\mathbf{X} \to \mathbf{X}') = \begin{cases} 1, & \text{for } \Delta \leq 0 \\ \exp(-\Delta), & \text{for } \Delta > 0 \end{cases} \tag{2.2}$$

where $\mathbf{X} = \{\mathbf{R}_1, \ldots, \mathbf{R}_N\}$ and $\mathbf{X}' = \{\mathbf{R}'_1, \ldots, \mathbf{R}'_N\}$ are the replica states (i.e. atomic positions) before and after the replica exchange. Assuming an exchange between replicas i and j, simulated

at temperatures T_i and T_j, the Δ can be written as,

$$\Delta = \left(\frac{1}{k_\mathrm{B} T_j} - \frac{1}{k_\mathrm{B} T_i} \right) (E_j - E_i) \tag{2.3}$$

in which k_B is the Boltzmann constant and E_i, E_j are the potential energy of the target system as a function of the replica states $\mathbf{X}_i$ and $\mathbf{X}_j$, respectively [100].

In this thesis, simulations of biomolecules in the canonical ensemble (NVT) were run for 200 ns. The leap-frog MD algorithm [101, 102] was used for the simulations in the gas phase with a time step of 0.5 fs with no cutoffs for non-bonded interactions and no periodic boundary conditions applied. In addition, temperature REMD simulations were performed on the basis of the Amber ff14SB force field [103] in GROMACS version 2018.8 to efficiently sample the PES. Exponentially distributed [104] REMD simulations were carried out at temperatures of 300, 352, 413, 481, 559, and 648 K. Temperature coupling during the simulation was realized with the velocity rescaling temperature control of Bussi, Donadio, and Parrinello [86] and in absence of any bonded constraints, i.e. no SHAKE or RATTLE [105, 106].

2.2.3 Ab initio calculations

2.2.3.1 Infrared spectra calculation

In this thesis, harmonic vibrational frequency calculations were performed to interpret IRMPD experimental results. The structural identification and comparison to IRMPD spectra were mainly performed in two steps, namely conformational search using classical force fields augmented by temperature REMD simulations and electronic structure calculations.

As the first step, the conformational space of the biomolecule was thoroughly searched by REMD simulation at room temperature, similar to conditions in IRMPD experiment. In brief, by employing unsupervised machine learning on the raw REMD output, recurring molecular motifs were extracted as introduced in chapter 3. These motifs were subsequently sorted based on their REMD energies, allowing for later structural optimizations and electronic structure calculations of infrared(IR) absorption spectra through *ab initio* calculations. The Born-Oppenheimer approximation simplifies the molecular Schrödinger equation by decoupling the electronic motion and the nuclear motion in molecules,

$$\psi(\mathbf{R}, \mathbf{r}) = \psi_e(\mathbf{R}, \mathbf{r}) \psi_n(\mathbf{R}) \tag{2.4}$$

$$H(\mathbf{R}, \mathbf{r}) = H_e(\mathbf{R}, \mathbf{r}) + H_n(\mathbf{R}) \tag{2.5}$$

in which $\mathbf{R}$ and $\mathbf{r}$ are the coordinates of the nuclei and the electrons. $H_e(\mathbf{R}, \mathbf{r})$ and $H_n(\mathbf{R})$ represent the Hamiltonian for electrons and nuclei, respectively. For a given solution of the Schrödinger

equation as the result of Born-Oppenheimer approximation,

$$H_e(\mathbf{R}, \mathbf{r})\psi_e(\mathbf{R}, \mathbf{r}) = E_e\psi_e(\mathbf{R}, \mathbf{r}) \tag{2.6}$$

vibrational energy levels can be obtained by solving the electronic Schrödinger equation. Several theoretical methods exist for calculating the electronic Schrödinger equation for a set of molecular conformations obtained from the REMD simulation. In this thesis, Density Functional Theory (DFT) was used to calculate the ground state [107]. The guiding principle of DFT theory is to substitute $3N$ dimensional wave function by the electronic density. Two fundamental theorems of DFT provides rigorous proof that all the observables of the system can be written as a function of the electronic density. First theorem states that there is a one to one correspondence between the external potential and the density of the ground-state. This means that the external potential is fully determined by the density and the other way around. Second theorem states that it is possible to show that the ground-state energy is minimum for the exact (Born-Oppenheimer) ground-state density. Kohn-Sham scheme [108] provide a practical way to solve the equations and obtain the densities and build the external potential from them. The idea of Kohn-Sham scheme is to map the system of interacting electrons into fictitious system of non-interacting electrons. The kinetic energy is addressed by splitting it into two parts, one corresponding to a system of non-interacting electrons and another corresponding to the part that accounts for the correlations. Therefore, the total energy of the system is given as the sum of kinetic energy of non-interacting electrons, external potential energy due to the nuclei and other external fields, and the exchange-correlation energy. The functional form for the exchange-correlation potential is unknown and an approximation is necessary for this term. For the description of the electron exchange energy, BP86 [109, 110] was used. Minimizing the total energy of the system with respect to the electron density using the iterative self-consistent field method [111], gives the ground-state energy.

The theory of vibrational structure is based on time-independent vibrational (nuclear) Schrödinger equation in which Born-Oppenheimer approximation still applies [112]. In this case, the electronic structure is reduced to the role of the source of an external potential upon which the motion of nuclei depends [113]. To obtain the vibrational frequencies for the equilibrium structure of the molecule, the geometry of the molecule is first optimized to its minimum energy using DFT, as explained previously. The geometry optimisation is considered as minimizing the molecule's potential energy with respect to its atomic coordinates. In case of harmonic approximation to IR calculations, the vibrational Hamiltonian is constructed by neglecting the third and higher terms in the expansion. To obtain the mass-weighted second-derivative of the potential (i.e. Hessian matrix), the potential energy in the vicinity of the equilibrium is then

approximated as a Taylor series. Diagonalization of the mass-weighted Hessian matrix yields a matrix which describes the vibrational motion of the system within harmonic approximation, the so-called mass-weighted normal modes. The intensity of IR absorption band for those normal modes is directly proportional to the square of the change in the dipole moment μ of the molecule during vibrational motion. Therefore, the IR intensities can be calculated from the dipole moment derivatives of the molecule. During vibrational motion, the dipole moment derivative quantifies the extent to which the molecule's dipole moment changes as its normal vibrational modes are excited. For detailed mathematical explanation of how to obtain the vibrational frequencies, see Appendix A.

2.2.3.2 Linear-response time-dependent density functional theory

Time-Dependent Density-Functional Theory (TDDFT) [114], as an extension to ordinary ground-state DFT, is often used to consider the influence of an external electromagnetic field to a biomolecule, i.e. excitation absorption spectra. TDDFT can be viewed as an alternative formulation of time-dependent quantum mechanics. However, the basic variable of TDDFT is the one-body electron density $n(\mathbf{r}, t)$, in contrast to the normal approach that relies on wave-functions and on the many-body Schrödinger equation. The density is a simple function which depends only on 3 variables x, y, and z, in contrast to many-body wave-function which is a very complex mathematical object since it is a function in a $3N$-dimensional space (N is the number of electrons in the system). $n(\mathbf{r}, t)$ can be obtained with the help of a fictitious system of non-interacting electrons, the Kohn-Sham system. These electrons feel an effective potential, i.e. time-dependent Kohn-Sham potential which has to be approximated. As the simplest approximation to the Kohn-Sham potential, linear-response theory can be used to study the system and calculate the spectra. In linear-response TDDFT, X-ray absorption spectra is calculated by perturbing the system with an external electromagnetic field and then calculating the linear response of the system to the field. Linear-response TDDFT is relatively efficient and can be used to calculate X-ray absorption spectra spectra for a wide range of biomolecules [115, 116]. As a general overview to linear-response TDDFT calculations, the ground-state density of the system is first calculated using DFT. The linear response function is then constructed from the ground-state density and the exchange-correlation functional used in TDDFT. Finally, the linear response equations are solved to find the excitation energies and oscillator strengths of the system. For full mathematical explanation, see Appendix A.

2.3 Unsupervised machine learning

The difficulties faced by systems relying on hard-coded knowledge suggest that artificial intelligence (AI) systems need the ability to acquire their own knowledge, by extracting patterns from raw data. This capability is known as machine learning. Machine learning is a data-driven approach where the algorithms are trained based on a large amount of data allowing them to infer correlations from the training data and identify patterns, relationships, and hidden insights.

Machine learning is a fundamentally interdisciplinary field which finds application in broad range of domains, including image and speech recognition [117], natural language processing [118], recommendation systems [119] and in particular quantum chemistry [120, 121]. The inherent power of machine learning approaches lies in their capability to generate models that capture complex phenomena through training with examples. This bypasses the requirement of finding analytical models or conducting expensive underlying calculations, whenever they are available.

In a machine learning algorithm, the process of learning involves the increase in the model's performance by gaining experience through the training data. The scope of a successful machine learning algorithm is to model the correlation between the inputs and the outputs (labels) encoded in the data, in the best possible way to ensure trust in new outputs given an unseen set of inputs.

Machine learning algorithms are broadly categorized into two main types, supervised learning and unsupervised learning. In unsupervised learning, the algorithms are trained based on unlabeled data which aims to discover hidden patterns or structures in the data without any specific guidance. The objectives of unsupervised learning can be divided into two main aspects. The first is to explore the dataset and uncover hidden correlations between input points through dimensionality reduction techniques. The second aspect involves identifying patterns and rules to group the data using clustering analysis.

2.3.1 Dimensionality reduction

Dimensionality reduction involves reducing the number of variables or features in the data, while still preserving the structure of the data and important information contained within the high-dimensional dataset. A collection of N points in a D-dimensional space is defined as, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,\ldots,N}$ with $\mathbf{x}_i \in \mathbb{R}^D$.

In datasets obtained through sampling from an underlying distribution, the data points often lie in a subspace of its full feature space. The effective dimensionality of this low-dimensional data is often called intrinsic dimensionality. The accurate estimation of intrinsic dimensionality finds application in image recognition [122], text classification [123] and protein structure prediction [124]. Therefore, the primary goal of dimensionality reduction techniques is to simplify the data

representation by removing irrelevant or redundant features, noise and correlations and find the optimal $d < D$ dimensional space which preserve the topology of full-dimensional space. Dimensionality reduction techniques are often used to transform high-dimensional data into two or three dimensions, making it easier to visualize and interpret patterns such as outliers or clusters.

2.3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a widely used statistical technique for dimensionality reduction and data exploration [125]. PCA reduces the number of correlated dimensions in the data into smaller number of uncorrelated variables, called principal components (PC), by applying the orthogonal transformation to the feature space.

The fundamental concept behind PCA is to identify uncorrelated variables that are linear combinations of the original variables. These principal components are ordered in a manner that the first component captures the highest variance present in the data. To extract the principal components from a set of N data points with dimension of D, the eigenvalue-eigenvector problem of the symmetric matrix is solved. The eigenvalue-eigenvector problem (or shortly eigenvalue problem) is defined as $\mathbf{Ax} = \lambda\mathbf{A}$ in which an eigenvalue of a square matrix $\mathbf{A}$ is scalar λ such that there exists a non-zero vector $\mathbf{x}$. The vector $\mathbf{x}$ is called an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ.

As the first step, the distribution of $\mathbf{X}$ is centered around the mean of the data by,

$$\tilde{\mathbf{X}} = \mathbf{X} - \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{2.7}$$

we can then calculate the covariance matrix,

$$\mathbf{C} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}} \tag{2.8}$$

we then perform eigendecomposition on the covariance matrix to obtain the eigenvalues and eigenvectors $\mathbf{e}_k$ of the covariance matrix. By sorting the eigenvectors in descending order of their corresponding eigenvalues, we can cut the expansion of data points until d ($d < D$) large enough eigenvalues. We can then project our data points into a new coordinates by rewriting the coordinate of every point as,

$$\mathbf{y}_i = \sum_{k=1}^{d} p_{i,k}\mathbf{e}_k \tag{2.9}$$

where $\mathbf{y}_i$ are the transformed samples and the projection of point $\mathbf{x}_i$ is written as $p_{i,k}$ in dimension k.

2.3.1.2 Kernel trick

Kernel trick is the fundamental concept of machine learning that enables algorithms to implicitly operate in high-dimensional space without explicitly computing the transformations. Using kernel trick provide the opportunity to handle non-linearly separable data and improve the performance of learning algorithms. This means that kernel trick avoids the explicit mapping needed to get linear learning algorithms to learn a nonlinear function or decision boundary. The kernel function operates in the original input space X and maps the data into higher dimensional space (feature space) V to capture the similarities and relationships between data points. The function k is referred as kernel function. A kernel function is defined as any symmetric, real-valued function which is a positive definite function. Positive definite condition amounts to finding a symmetric function k which satisfies $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i)$ and,

$$\sum_{i,j}^{N} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{2.10}$$

for every $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbf{X}$ and real-valued coefficients $c_1, \ldots, c_N \in \mathbb{R}$. The kernel can be written in the form of a feature map $\phi : X \to V$ which satisfies $k(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_V$. This means that as long as V is an inner product space, explicit representation of function ϕ is not required. Positive definite kernel ensures that it induces a valid inner product between two feature vectors in a Hilbert space. This results in simple linear kernels by considering the Gram matrix which is a symmetric matrix obtained on the feature space V. The Gram matrix which sometimes also called kernel matrix $\mathbf{K}$ with respect to $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ can be written as,

$$\mathbf{K} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x_1})^T \boldsymbol{\phi}(\mathbf{x_1}) & \boldsymbol{\phi}(\mathbf{x_1})^T \boldsymbol{\phi}(\mathbf{x_2}) & \cdots \\ \boldsymbol{\phi}(\mathbf{x_2})^T \boldsymbol{\phi}(\mathbf{x_1}) & \ddots & \\ \vdots & & \end{bmatrix}. \tag{2.11}$$

As shown in Figure 2.3, the key idea of using such mathematical construction is to define a kernel function that compute the inner product between the transformed representations of two data points without explicitly referencing to their underlying Hilbert space. Higher order correlation between the data is considered by adopting kernel functions such as Gaussian, radial basis and polynomial kernels which enables the mapping of highly non-linear input data. Alternatively, the distribution of points in the feature space can be represented through the measurement of pairwise distances between data points. The distance function is considered between the kernels,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{x}_i, \mathbf{x}_j) - 2k(\mathbf{x}_i, \mathbf{x}_j)} \tag{2.12}$$

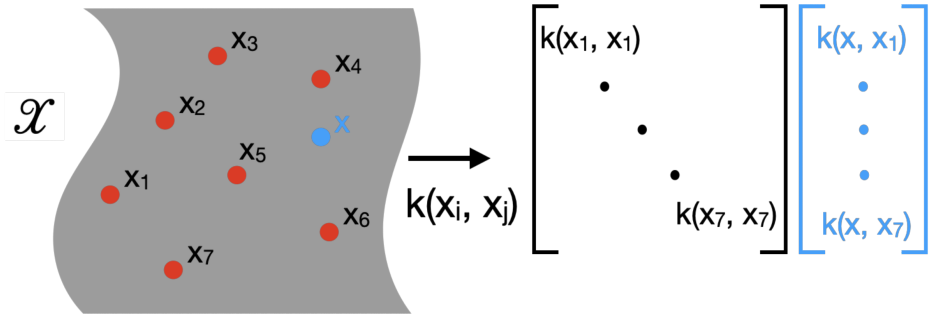

Figure 2.3: The set of points $\{\mathbf{x}_i\}_{i=1,\ldots,7}$ in the space $\mathcal{X}$ can be characterised by either their coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_7 \in \mathbf{X}$ or by a kernel representation of their pairwise similarity. The same spatial correlations can be encoded by symmetric matrix $\mathbf{K}$ where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ which has 7×7 dimensions. In order to predict a new point $\mathbf{x}$, the similarity of it is compared with respect to the initial embedding set and a new similarity matrix between this new point and the training set is built.

By using kernel representations, one can use this equation in many distance based techniques.

2.3.1.3 Kernel Principal Component Analysis

Kernel Principal Component Analysis (Kernel PCA) is the non-linear extension of PCA by using the kernel trick. Kernel PCA projects the data into higher-dimensional feature space which allows to discover the non-linear patterns in data and easier visualisation of high-dimensional manifolds. Considering that our data is mapped into feature space V, $\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)$ where $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbf{X}$ and the function is nonlinear mapping $\phi : X^D \to V$ and is centered, i.e. $\sum_{k=1}^{N} \boldsymbol{\phi}(\mathbf{x}_k) = 0$. Considering the kernel matrix in equation 2.11 in subsection 2.3.1.2, an inner product of two arbitrary elements $\mathbf{x}_i$ and $\mathbf{x}_j$ in its reproducing kernel Hilbert space can be written as,

$$K_{ij} = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle_V = \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j) \tag{2.13}$$

The covariance matrix associated with the mapped distribution is calculated by,

$$\mathbf{C}_{\boldsymbol{\phi}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_i) \tag{2.14}$$

By solving the eigenvalue problem of the covariance matrix, the eigenvectors $\mathbf{V}$ and their corresponding eigenvalues λ will be given, satisfying $\lambda \mathbf{V} = \mathbf{C}_{\boldsymbol{\phi}} \mathbf{V}$. By spanning all the solutions in $\mathbf{V}$ from the collection of $\{\boldsymbol{\phi}(\mathbf{x}_i)\}_N$ with coefficients α_i, we can rephrase the solution to eigenvalue problem,

$$\lambda(\boldsymbol{\phi}(\mathbf{x}_k) \cdot \mathbf{V}) = (\boldsymbol{\phi}(\mathbf{x}_k) \cdot \mathbf{C}_{\boldsymbol{\phi}} \mathbf{V}) \text{ and } \mathbf{V} = \sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}(\mathbf{x}_i). \tag{2.15}$$

We can combine equations 2.14 and 2.15 and write,

$$\lambda \sum_{i}^{N} \alpha_i \left(\boldsymbol{\phi}\left(\mathbf{x}_k\right) \boldsymbol{\phi}\left(\mathbf{x}_i\right)\right) = \frac{1}{N} \sum_{i}^{N} \alpha_i (\boldsymbol{\phi}\left(\mathbf{x}_k\right) \cdot \sum_{j}^{N} \boldsymbol{\phi}\left(\mathbf{x}_j\right)) \left(\boldsymbol{\phi}\left(\mathbf{x}_j\right) \boldsymbol{\phi}\left(\mathbf{x}_i\right)\right) \quad \text{for } k = 1, \ldots, N \tag{2.16}$$

where k means the principal component k. We can define the eigenvalue problem as $N\lambda\mathbf{K}\boldsymbol{\alpha} = \mathbf{K}^2\boldsymbol{\alpha}$ (where $\boldsymbol{\alpha}$ denotes the column vector with entries $\alpha_1, \ldots, \alpha_N$) by recalling the matrix element of kernel matrix is $K_{ij} = \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j)$. To find the solution of such a system, we solve the eigenvalue problem $N\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}$ for non-zero eigenvalues. Considering all the eigenvalues λ_i of $\mathbf{K}$ from above and their corresponding eigenvectors $\boldsymbol{\alpha}$, we can impose that their non-zero eigenvalues corresponding vectors in $\mathbf{V}$ are normalised, i.e. $(\mathbf{V}^k \cdot \mathbf{V}^k) = 1$ (where superscript k means the principal component k),

$$1 = \sum_{i,j=1}^{N} \alpha_i^k \alpha_j^k K_{ij} = (\boldsymbol{\alpha}^k \cdot \mathbf{K}\boldsymbol{\alpha}^k) = \lambda(\boldsymbol{\alpha}^k \cdot \boldsymbol{\alpha}^k) \tag{2.17}$$

To extract the principal component, we compute the projections of the image of a test point $\boldsymbol{\phi}_{(\mathbf{x}^{test})}$ onto to the first d eigenvectors $\mathbf{V}^k$ according to,

$$(\mathbf{V}^k \cdot \boldsymbol{\phi}(\mathbf{x}^{\text{test}})) = \sum_{i=1}^{N} \alpha_i^k (\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}^{\text{test}})) \tag{2.18}$$

The above results can be used to extend the use of PCA to non-linear mapping of the input data.

2.3.1.4 Multidimensional Scaling and sketchmap

Multidimensional Scaling (MDS) is a dimensionality reduction technique which aims to find a low-dimensional Cartesian projection while preserving their pairwise distances of points in high-dimensional space. The optimization technique used in metric Multidimensional Scaling is stress loss function which accumulates the residuals of the lower dimensional embedding compared to the high dimensional space. The stress loss function (ρ) is a residual loss function defined as,

$$\rho(x_1, x_2, \ldots, x_N) = \frac{1}{N} \sqrt{\sum_{i,j} (d_{ij}^p - ||x_i - x_j||)^2} \tag{2.19}$$

The p term is a metric scaling to weight the distances in the high-dimensional space with different intensities. d_{ij} is the dissimilarity between points, i.e. the Euclidean distance between embedded points i and j. The metric scaling assumes that the dissimilarities between data points represent actual distances in the underlying space. The effect of metric scaling is that a more accurate representation of the underlying data is produced since the magnitude of the dissimilarities (i.e.

the relationship between data points) is considered. Metric Multidimensional Scaling can be reformulated and extended to introduce elements of nonlinearity by distorting the distances in the stress function. The distortion of the distances is a way to characterise the topology of the high-dimensional space.

Sketch-map algorithm [17] is another very effective way to introduce non-linearity into dimensionality reduction of the data, in particular for datasets generated from atomistic simulations, such as MD simulations. Sketch-map is essentially multidimensional scaling in which the distances in both high- and low-dimensional spaces are transformed by sigmoid function. One advantage of using this approach in datasets of MD simulations is that one can obtain a picture of the complex free-energy landscape of the molecule by reproducing the relations and connections between nearby energy basins [126]. Sketch-map method demonstrates that there is a characteristic distance at which the most valuable topological information of free energy landscape is encoded. To maintain the relationships between points, nearby points are mapped close together while points that are farther apart are mapped farther apart. The stress function ($\mathscr{X}$) in obtaining the low-dimensional embedding is defined as,

$$\mathscr{X} = \sum_{i \neq j} \frac{1}{w_i w_j} \sum_{i \neq j} w_i w_j [F(R_{ij}) - f(r_{ij})]^2 \tag{2.20}$$

where w_i is the weight of point i, $R_{ij} = ||\mathbf{X}_i - \mathbf{X}_j||$ with $\mathbf{X} \in \mathbb{R}^D$ and $r_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$ with $\mathbf{x} \in \mathbb{R}^d$ are the distances between points i and j in the high and low-dimensional spaces, respectively. F and f are both sigmoid functions of the form,

$$s_{\sigma,a,b}(r) = (1 - (1 + (2^{a/b} - 1)(r/\sigma)^a))^{-b/a} \tag{2.21}$$

By adjusting the parameter σ, the user can specify the threshold for the characteristic distance. By setting σ, we can essentially decide what features will be displayed in the projection. For length scales less than σ, sketch-map makes little to no effort to accurately reproduce these distances [127]. The value of σ should be chosen by examining the data, i.e. deciding which short-range features should be ignored. In case of molecular dynamics trajectory data, the internal structure of energy basins, i.e. the thermal fluctuations can be ignored. The exponents a and b determine the rate at which the function approaches 0 and 1, respectively, i.e. tuning the steepness with which the points falling before or after σ. Therefore, carefully fine tuning these parameters give an opportunity to focus on more important distances between points which finally give rise to project relations and connections between energy basins and discard all the high-dimensionality, unfittable data on the internal structure of energy basins and the relative positions of distant basins [17].

Computational complexity of using MDS algorithms scales with the number of points N to embed, i.e $O(N) \sim N^2$. Therefore, a two-step process of finding the low-dimensional embedding can be applied to reduce the scaling to $O(N) \sim NM$ (M are the number of landmarks). To build such a process, an explicit embedding for a reduced set of representative configurations or landmarks M is first produced and then the stress function for distances of every other point $\mathbf{X}$ from the M landmarks is minimised:

$$\mathcal{X}^2(\mathbf{x}) = \left(\sum_{i}^{M} w_i\right)^{-1} \sum_{i=1}^{M} w_i[F(||\mathbf{X} - \mathbf{X}_i||) - f(||\mathbf{x} - \mathbf{x}_i||)]^2 \tag{2.22}$$

where w_i is the weight associated with the i-th landmark, e.g. corresponding to the number of points within Voronoi cell of the landmark.

This formulation is applicable to embedding molecular dynamics simulation, which often involve millions of points, where thermal vibrations introduce degrees of freedom that may not be crucial for the analysis. An example of application for this approach is describing the potential energy surface of a biomolecule, allows the recognition of recurring structural motifs [128].

2.3.1.5 Other dimensionality reduction techniques

t-distributed stochastic neighbor embedding (t-SNE) is another dimensionality reduction technique for visualising high-dimensional data [129]. t-SNE is a non-linear technique which is based on stochastic neighbor embedding and consists of two main steps. First, a probability distribution is constructed over the pairs of high-dimensional points, considering higher probability for similar (or closer) points while lower probability for dissimilar points. Then, a similar probability distribution over the points in low-dimensional map is defined. The two distributions with respect to the location of the points in the map is minimized by using Kullback–Leibler (KL) divergence. t-SNE first computes probabilities p_{ij}, proportional to the similarity of objects $\mathbf{x}_i$ and $\mathbf{x}_j$, for a given set of high-dimensional objects $\mathbf{x}_1, \ldots, \mathbf{x}_N$ ($\mathbf{x}_i \in \mathbb{R}^D$). The similarity of data point $\mathbf{x}_j$ to data point $\mathbf{x}_i$ is the conditional probability $p_{j|i}$ that can be written as,

$$p_{j|i} = \frac{\exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_k\right\|^2 / 2\sigma_i^2\right)} \tag{2.23}$$

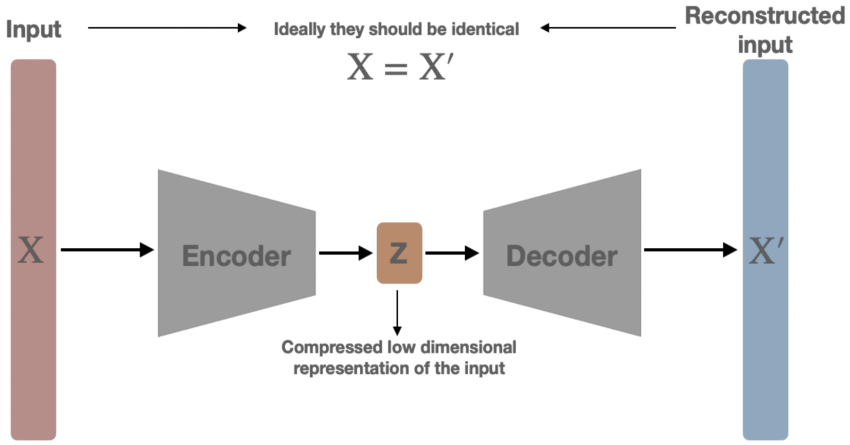

Figure 2.4: Basic architecture of an autoencoder. In autoencoder, a compressed low dimensionality representation of the input is obtained by reconstructing the input by encoder and decoder.

then to reflect the p_{ij} distribution as much as possible, t-SNE learns the d-dimensional map $\mathbf{y}_1, \ldots, \mathbf{y}_N$, with distribution q_{ij} where $\mathbf{y}_i \in \mathbb{R}^d$ and $d << D$. The q_{ij} distribution is defined as,

$$q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_k \sum_{l \neq k} \left(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2\right)^{-1}} \tag{2.24}$$

the location of the points y_i in low-dimensional map is obtained by minimising the KL divergence of the distributions P from the distribution Q,

$$\mathrm{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.25}$$

The minimisation of KL divergence results in a low-dimensional map with similar distribution of high-dimensional points.

The autoencoder is a novel approach in dimensionality reduction designed for feature learning in unsupervised ML. It is a type of artificial neural network specifically designed for this purpose. Autoencoders reconstruct the input data at the output layer, having an intermediate or hidden layer that represents the compressed or encoded representation of the input. The encoder architecture consists of encoder, decoder and the bottleneck (feature representation), as shown in Figure 2.4. This bottleneck encourages the network to capture the patterns in the data and map the input vectors into the latent space. During the training process, the autoencoder learns to minimize the difference between the input data and the reconstructed output.

In general, reducing the dimensionality is useful in visualisation and interpretation of complex data as well as removing noisy and redundant features to focus on more important aspects of the

data. In case of atomistic simulations, reducing the dimensionality of the data is an opportunity to unravel the complexity behind the data and use the obtained features for further analysis such as clustering of similar molecular structures.

2.3.2 Clustering

Clustering is another type of unsupervised machine learning which aims to automatically partition points into groups (or clusters). In clustering, a suitable metric is used to reflect the similarity between points belonging to the same group and draw differences between points of different groups. The choice of distance metric depends on the nature of the data and specific clustering algorithm being used. For example, regarding the nature of the data, if the data is categorical, the distance metric is different from when the data is continuous. The k-means clustering algorithm assumes that the data follows a spherical distribution, which is why it utilizes the Euclidean distance metric However, Euclidean distance may not be a good choice for density-based spatial clustering since it assumes that clusters are dense regions of points [51]. The goal of using clustering techniques is to discover the patterns and underlying structures of the data without prior knowledge of the labels of data. Applying clustering techniques often reflects into obtaining finite and clear number of clusters. Some common application of clustering consists of document clustering in text mining [130], image segmentation in computer vision [131], anomaly detection [132] and pattern recognition in atomistic simulations data [51]. In atomistic simulations, clustering is an effective way of representing the multidimensional probability distribution of data. At the core of each clustering technique, a cost function is used to measure the quality of grouping achieved in each iteration as a function of every point's assignment. The assignment of points to each cluster is either with boolean value (hard clustering) or with a probablity (soft clustering).

Clustering techniques in atomistic simulations data can be categorised into two different classes, namely partitioning schemes and density-based schemes [51]. In partitioning schemes, similar structural configurations are grouped into the same cluster and are different from configurations belonging to other clusters. These clusters define a partition or tessellation of the space in which the configurations are defined and every point has to be assigned to a group. k-means clustering is the simplest example of clustering technique in partitioning scheme. On the other hand, in density-based clustering, the clusters correspond to the peaks of the probability distribution from which the data are harvested (or, equivalently, to the free energy minima in case of molecular dynamics simulations) [51]. Therefore, data points of different clusters are not necessarily far as the clusters are separated with low probability density.

2.3.2.1 k-means

k-means is a popular clustering used for partitioning the data into $k_{\max}$ distinct clusters [133]. It is based on minimising the sum of squared distances between data points and their assigned cluster centroids. The k-means objective is to partition a feature data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,\ldots,N}$, with $\mathbf{x}_i \in \mathbb{R}^D$, into $k_{\max}$ ($\leq N$) sets $\mathbf{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_{k_{\max}}\}$ sets by minimising the within-cluster sum of squares,

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k_{\max}} \sum_{\mathbf{x}\in\mathbf{S}_i} ||\mathbf{x} - \mu_i||^2 \text{ with } \mu_k = \frac{1}{N_k} \sum_{\mathbf{x}\in\mathbf{S}_k} \mathbf{x} \tag{2.26}$$

where the vector μ_i represents the average value of the points in the cluster i ($x_i \in \mathbf{S}_i$). Therefore, in case of molecular dynamics data, the loss function in k-means is defined as sum of the square of the distances from each configuration in the data set to the cluster center to which the configuration is assigned. Considering this loss function, k-means is an optimisation problem of finding the best set of cluster centers. The algorithm follows an iterative procedure to find an approximate solution as follows:

1. Initialization: Randomly selecting $k_{\max}$ points from the dataset as initial cluster centroids
2. Assignment: Based on the distance metric (commonly considered as the Euclidean distance), each data point is assigned to the closest center
3. Update: The centers of the clusters are recalculated by taking the mean of the data points assigned to that cluster
4. Repeat: The assignment and update steps are repeated until the convergence.

As the outcome of the algorithm depends heavily on the initialization step, it is often necessary to start over this iterative process with different initial centers in order to obtain reasonable results [51]. Mini-batch k-means is another version of k-means clustering technique that is designed to handle large and complex datasets. In this technique, random batches of data is used to update the cluster centroids, rather than using the entire dataset. Since all the data points simultaneously used in the optimisation, faster convergence to a lower minimum of the associated cost function can be obtained and thus the problem of trapping into local minima can be alleviated [134].

2.3.2.2 Gaussian Mixture Models

In k-means, hard clustering approach is considered as a limitation since every point can be assigned to only one cluster. This limitation becomes particularly problematic when dealing with scattered distributions that exhibit varying intensities and lack clear boundaries between cluster nuclei. To

alleviate the hard clustering limitation, one can consider assignment vector as probabilities of each point being part of a cluster.

Gaussian mixture models (GMM), as probabilistic models, are used to model and represent probability distributions which have underlying normal distributions. A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. GMMs consist of a weighted sum of multiple Gaussian distributions in which each Gaussian component represent the mode of a cluster in data distribution. Normal distributions in GMMs are represented with center of the cluster and the covariance matrix which is optimised to maximize the likelihood of the spread of points around them. In GMM, the probability $P(\mathbf{x})$ of each point belonging to a certain cluster which depends on the sum of k models (Gaussians) is defined as,

$$P(\mathbf{x}) = \sum_{i=1}^{k} \phi_i \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \text{ with } \sum_{i=1}^{k} \phi_i = 1 \tag{2.27}$$

where ϕ_i is the weight of each model and $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) = \frac{1}{(2\pi)^k|\Sigma_i|} \exp\left\{\left(\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right)\right\}$ represent a multivariate normal distribution with $\boldsymbol{\mu}_i$ representing the means vector and $\boldsymbol{\Sigma}_i$ representing the covariance matrix.

Given a dataset, the parameters of GMM are estimated in an iterative algorithm called expectation-maximization (EM) algorithm. The first step is called expectation step and involves calculation of the component assignment's expectation C_k for each $\mathbf{x}_i \in \mathbf{X}$ given the set of model parameters $(\phi_k, \mu_k, \sigma_k)$ with σ_k as the spherical Gaussian. In the maximization step, the aim is to maximize the expectations coming from the previous step over the model parameters. This step results in updating the model parameters to $(\phi_{k+1}, \mu_{k+1}, \sigma_{k+1})$. This iterative process will continue until the convergence of model parameters which results in maximum likelihood estimate. In case of univariate distribution, the iterative process for finding the optimal $(\phi_k, \mu_k, \sigma_k)$ parameters is defined as follows,

1. Finding the probability of a point x_i to be in cluster k in the expectation step,

$$\gamma_{i,k} = \frac{\phi_k \mathcal{N}\left(x_i|\mu_k, \sigma_k\right)}{\Sigma_{j=1}^{k} \phi_j \mathcal{N}\left(x_i|\mu_j, \sigma_j\right)} \tag{2.28}$$

2. Updating the models' parameters in maximisation step,

$$\phi_k = \Sigma_{i=1}^{N} \frac{\gamma_{i,k}}{N} \ , \ \mu_k = \frac{\Sigma_{i=1}^{N} \gamma_{i,k} x_i}{\Sigma_{i=1}^{N} \gamma_{i,k}} \ , \ \sigma_k^2 = \frac{\Sigma_{i=1}^{N} \gamma_{i,k}(x_i - \mu_k)^2}{\Sigma_{i=1}^{N} \gamma_{i,k}} \tag{2.29}$$

This iterative process can also be extended to multivariate cases which allows to capture clusters with anisotropic distributions.

The ability to perform soft clustering and detect clusters which has non-globular (i.e. non-spherical symmetry) shapes, using anisotropy in multivariate covariance matrices, makes GMMs more versatile compared to k-means clustering.

2.3.2.3 Density-based clustering

Density-based clustering is a technique that groups data points depending solely on the density of their neighborhood. Unlike k-means clustering, this scheme does not require specifying the number of clusters in advance and can discover clusters of arbitrary shapes and sizes. In density-based clustering, the clusters are regions with high density of data points separated by regions of lower densities. To first find core points, the algorithm identifies the data points that have sufficient number of neighboring points within a specified distance. From these core points, reachable points that have a density above a certain threshold will be iteratively added to expand the clusters.

Density-based spatial clustering of applications with noise (DBSCAN) [135] is a popular algorithm to perform density-based clustering in which closely packed points are labeled into different clusters and outlying points in sparse regions are considered as noise. In DBSCAN, there are two parameters namely ϵ and $N_{\min}$, indicating the cutoff distance connecting two points and the minimum number of points defining a dense neighbourhood, respectively. By inputting a dataset in DBSCAN, points are tagged as whether they are core points, reachable points and noise, based on the previous parameters and the pairwise distances between the samples. All points below the value of ϵ are considered to be noise. The clusters are formed by all the core points and their corresponding reachable points.

DBSCAN has the ability of discovering clusters of arbitrary shape and is robust to noise and outliers, without the need to specify the number of clusters in advance. However, when it comes to distributions with regions of varying densities, DBSCAN is not so powerful. The solution to this problem has been introduced in the hierarchical DBSCAN (HDBSCAN) which does not require the definition of a ϵ distance and depends only on the defined number of $N_{\min}$.

2.3.2.4 Clustering technique for atomistic datasets

The increased complexity of data produced from atomistic simulations requires developing a data-driven and unbiased approach to fully characterize the conformational space of complex biomolecules. In typical molecular dynamics simulations, a trajectory represents the time dependent exploration of the conformational phase space. Each frame of that trajectory represents moreover a point in the high dimensional phase space. Clustering algorithms can aid to partition the

conformational ensemble into distinct subensembles according to the (usually high-dimensional) distances between points. Partitioning the free-energy landscape into groups of metastable basins, and demonstrating the connection between them, is particularly useful to reproduce qualitatively the system's dynamical behaviour and to identify the transformation pathways between different states [17, 136, 137].

Probabilistic Analysis of Molecular Motifs (PAMM) [128, 138] is a clustering technique to recognize recurring molecular conformations of different metastable states sampled during the MD simulation. PAMM is specifically designed to perform clustering on high-dimensional datasets produced from atomistic simulations and is able to partition the underlying probability distribution function into modes corresponding to different configurations. In PAMM, a kernel density estimation (KDE) is first performed on a grid representing the data which was extracted by choosing a small set of landmarks from the full data set by using a farthest point sampling (FPS) algorithm [127]. FPS attempts to span the conformation space homogeneously by selecting landmark molecular configurations $\{\mathrm{X}_i\}$, where $i = 1, \ldots, N$ and N is the total number of landmarks to select from the points in the feature dataset. An adaptive multivariate kernel bandwidth for the KDE on the landmark points results then from a localized version of Silverman's rule defined by the smoothing parameter f_{points} [128]. This parameter can be used to optimize the number of identified clusters if done with due care, as it sets the automatically selected bandwidth for kernel density estimation either by a fraction of the total number of points f_{points} or by a fraction of the variance of the entire dataset f_{spread}. Apart from this, a slight change in the quick-shift cutoff (usually by changing the so-called scaling factor α) can also help to divide the data into more clusters [138]. A careful parameter search should be first performed to find the optimal number of grid points and Gaussian kernel widths to obtain a robust and smooth estimate of the high-dimensional probability density function underlying the data [128, 139]. Following this, a localized quick-shift variant [128] is used to identify to which cluster the landmark configurations belong. PAMM produces a set of fingerprints or more specifically probability motif identifiers (PMI), which are simple Gaussian naive Bayes classifiers in the original formulation [138], trained to distinguish between different recurrent motifs in the simulation, while providing at the same time a probabilistic interpretation. Since often different metastable structures are sampled from different free-energy basins that are very close (hence very similar), it is convenient to merge multiple PMIs together in an hierarchical clustering approach. One can see this step as a coarse-graining of the structural landscape often employed to properly extract informative structure-property relationships, e.g. to capture the hierarchical nature of a biomolecule's free-energy landscape [12, 139, 140]. Briefly, identification of such macroclusters is done by bootstrapping [141] the clustering step to determine an adjacency matrix, $\mathbf{A} \equiv (a_{ij})$, capturing the overlap between clusters. The metric used to compare two distinct PMIs i and j during the hierarchical clustering is given by $d_{ij} = -\ln(a_{ij}/\sqrt{a_{ii}a_{jj}})$, where a_{ii} values

are the diagonal elements indicating how robust the determination of the i-th cluster was, while a_{ij} are the off-diagonal terms, indicating the fuzziness of the cluster borders.

2.4 Supervised machine learning

In supervised ML, the models are trained on a dataset containing features, but each sample of the dataset is also associated with a label or target. The objective of supervised learning is to observe several examples of a random vector $\mathbf{x}$ and an associated value or vector $\mathbf{y}$, and learning mapping function f to predict $\mathbf{y}$ from $\mathbf{x}$. The input for training the model can be an image (encoded with RGB arrays), graph data (set of nodes and edges) [142] or a vector, while the output can be an array of floats, a set of binaries (multi-output classification) or encoded labels. The training of the mapping function is based on learning a series of parameters, $\mathbf{W} = \{w_i\}_{i\in\text{model}}$. Considering all the input $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,\dots,N}$ and all the labels $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1,\dots,N}$, the model's loss function, $\mathcal{L}$, is minimised to find the optimal $\mathbf{W}^*$, fitting the training dataset,

$$\mathbf{W}^* = \min_w \mathcal{L}(\mathbf{W}, \mathbf{X}, \mathbf{Y}) \tag{2.30}$$

in which the minimization is done on the training dataset and the learning process of the model is analyzed and monitored from the loss values of the validation dataset that is a subset of the training dataset. The optimal choice of weights can be extracted numerically, e.g. through gradient descent optimization [143] and conjugate gradient method [144].

Classification and regression are two types of supervised ML algorithms. In classification, the model learns from the labeled data to classify new instances into discrete class labels. Decision trees, support vector machines and random forests are common classification algorithms. On the other hand, regression algorithms aim to find a function that approximate the relationship between inputs and outputs and predict continuous or numerical values. Linear regression, polynomial regression, kernel ridge regression, and neural networks are common regression algorithms although neural networks can also be used in case of classification of the inputs.

Overfitting is a common problem faced in training ML models in which the model learns the training data too well and performs poorly on new data rather than learning the underlying true patterns in the data. In this scenario, the model learns all the bias, noise and random fluctuations existing in the training data, leading to poor generalisation. Common solutions to address overfitting issues are dropout regularization [145], cross-validation [146], regularization [147] and early stopping [148]. As the most important solution to overfitting, regularization is a technique that adds a penalty term to the model's loss function. The penalty term avoids the model from assigning large weights to features, reducing the complexity of the model. L1 regularization (Lasso)

and L2 regularization (Ridge) are common regularization techniques which add the absolute or squared values of the weights as penalties, respectively. In L1 regularization (Lasso), the model is penalized more severely for having large coefficients, which can lead to the coefficients of some features being shrunk to zero. On the other hand, ridge regularization does not force any coefficients to be exactly zero, which can help to prevent the model from losing important information.

2.4.1 Feedforward Neural Networks

An artificial neural network (ANN) is a computational model, loosely inspired by the functioning and structure of biological neural networks. ANNs consists of interconnected processing units called neurons which are organised into layers and apply non-linear activation function to its inputs [149]. Figure 2.5 depicts the flow of information into an artificial neuron. A feedforward neural network is

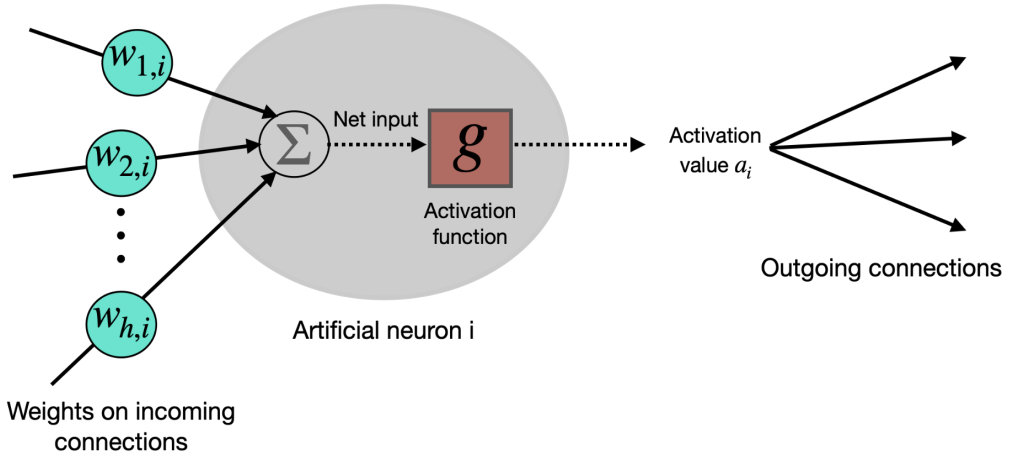

Figure 2.5: An artifical neuron in a multilayer neural network.

a type of ANN in which the flow of information move in forward direction from input layer, hidden layer(s) and to output layer. The goal of a feedforward neural network is to approximate some function f which maps input $\mathbf{X}$ to $\mathbf{Y}$. A feedforward network learns the value of the parameters $\mathbf{W}$ that result in the best function approximation. Arranging artificial neurons into groups known as layers and stacking these layers together leads to the formation of a multilayer feed-forward neural networks. The connections between the nodes (or neurons) have their associated weights which shows the strength of the connections. These models are called feedforward because information flows through the function being evaluated from $\mathbf{X}$, through the intermediate computations used to define f, and finally to the output $\mathbf{Y}$. Feedforward neural networks are typically represented by composing together many different functions. For example, we might have three functions $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$ connected in a chain, to form $f(\mathbf{x}) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(\mathbf{x})\right)\right)$. In this case, $f^{(1)}$ is called the first layer of the network, $f^{(2)}$ is called the second layer and so on. The overall depth of the chain

gives the depth of the model. The final layer of feed forward neural networks called the output layer. Feedforward networks introduce the concept of a hidden layer which requires to choose the activation functions that will be used to compute the hidden layer values. In the training process of feedforward neural networks, the weights of the connections are adjusted through backpropagation method [150] based on the loss value between the network's output and the target. This iterative process improve the ability of the model to make accurate predictions.

To explain the functioning of feedforward neural networks, we consider a very simple feedforward network with one hidden layer containing two hidden units. This feedforward network has a vector of hidden units $\mathbf{h}$ that is computed by a function $f^{(1)}(\mathbf{x};\boldsymbol{\theta},\mathbf{c})$, where $\boldsymbol{\theta}$ provides the weights of linear transformation and $\mathbf{c}$ the biases. The values of these hidden units are then used as the input for a second layer which is the output layer of the network in this case. The output layer is still just a linear regression model, but now it is applied to $\mathbf{h}$ rather than to $\mathbf{x}$. The network now contains two functions chained together: $\mathbf{h} = f^{(1)}(\mathbf{x};\boldsymbol{\theta},\mathbf{c})$ and $\mathbf{y} = f^{(2)}(\mathbf{h};\boldsymbol{\omega},\mathbf{b})$, where $\boldsymbol{\omega}$ provides the weights of the second linear transformation and $\mathbf{b}$ is the bias. The complete modeling can then be written as $f(\mathbf{x};\boldsymbol{\theta},\boldsymbol{c},\boldsymbol{\omega},\mathbf{b}) = f^{(2)}\left(f^{(1)}(\mathbf{x})\right)$. To describe the features with non-linear function, most neural networks do so using an affine transformation controlled by learned parameters, followed by a fixed, nonlinear function called an activation function. We can use this strategy by defining $\mathbf{h} = g(\mathbf{x};\boldsymbol{\theta},\mathbf{c})$, where g is the activation function. The default recommendation of activation function in modern neural networks is to use the rectified linear unit or ReLU [151, 152, 153], defined as $g(z) = \max\{0, z\}$. We can now specify the complete network as,

$$f(\mathbf{x};\boldsymbol{\theta},\boldsymbol{c},\boldsymbol{\omega},\mathbf{b}) = \boldsymbol{\theta}^{\top} \max\left\{0, \boldsymbol{\omega}^{\top}\mathbf{x} + \boldsymbol{c}\right\} + \mathbf{b} \tag{2.31}$$

Deep Neural Networks (DNN) consists of multiple hidden layers which are designed to learn complex patterns and hierarchy of features in data, employing multiple layers of non-linear transformations. The term "deep" in these ML models refer to the successive layers of representations. Deep learning models involve tens or even hundreds of successive hidden layers. Although some fundamental principles of deep learning draw inspiration from the human brain's neural structure, the learning mechanisms employed in modern deep learning models diverge from the complex processes underlying the brain's functioning. Schematic representation of DNN is shown in Figure 2.6. The flow of inputs into DNN can be seen as a process of information distillation where details of the inputs are learned through successive filters in the network.

2.4.2 Graph Neural Networks

Graph neural networks (GNN) are type of neural network models designed to learn on graph-structured data [154]. GNNs are based on representation learning in which the goal is to fully rely

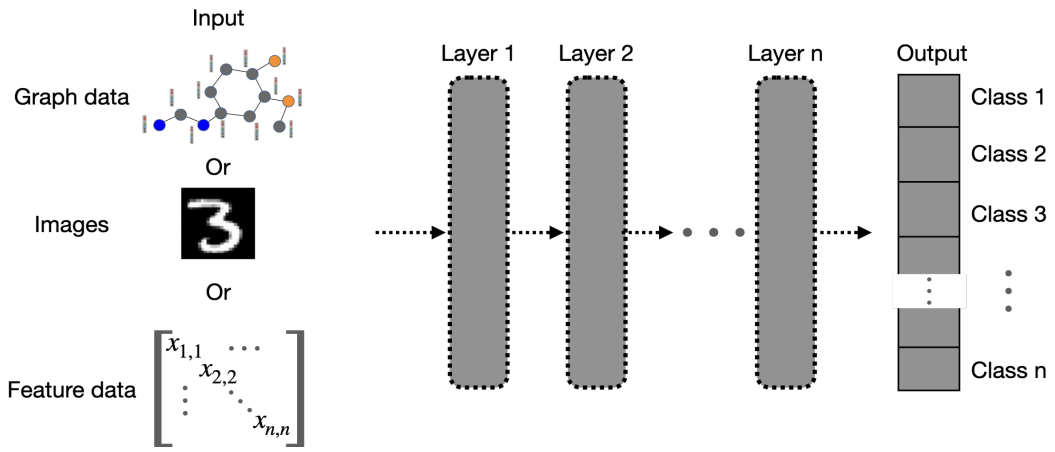

Figure 2.6: Schematic representation of deep neural network. Depending on the type of layers in DNN, the input of network can be graphs, images or feature data.

on the structure of data and extract sufficient but minimal features from data in order to avoid human efforts and biases in feature engineering based on prior knowledge and domain expertise on the data and tasks.

2.4.2.1 Graph data

Graph data structures are commonly used to model interactions, dependancies and relationships in different domains such as social networks [155], knowledge graph [156] and biological networks [157]. Graph data is represented as a collection of nodes (or vertices) connected by edges (or links). Nodes represent entities or objects and edges represent the relationship between them. Therefore, a graph formally defined as a tuple of $G = (\mathbf{V}, \mathbf{E})$ of a set of nodes $v_i \in \mathbf{V}$ and a set of edges $e_{i,j} = (v_i, v_j) \in \mathbf{E}$, which defines the connection between nodes. Further information on nodes and edges is incorporated in form of feature vectors added to tuple G. Categorical attributes of nodes and edges can be converted into numerical vectors using one-hot encoding. Biomolecules can also be represented as graphs, in which atoms and the bonds between them are represented as nodes and edges respectively. Further information about each atom and bond in a molecular graph is incorporated in the form of feature vectors added to the tuple G of each graph in the dataset. For instance, an atomic feature vector represents information such as the atom type (e.g. 'C', 'H', 'N', or 'S') or the number of hydrogen atoms attached to it. Similarly, edge feature vectors are representatives of properties such as bond lengths between two atoms or bond multiplicity. Figure 2.7 shows an example of molecular graph where atoms and their relationship with other atoms in the molecule are represented with node and edge features. Depending on the target property, certain local information (with a cutoff) surrounding each atom can be encoded into the node and edge features of the graph representation.

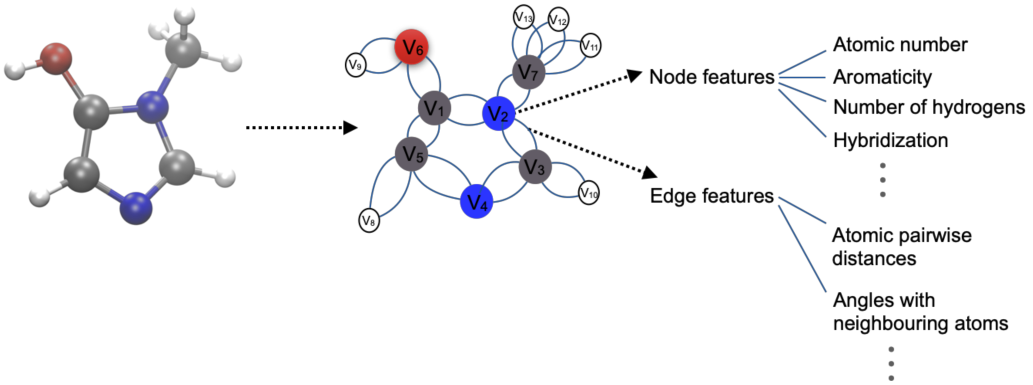

Figure 2.7: An example of a molecular graph. Depending on the prediction task, different type of information on the atoms and bonds can be added as node and edge features to the tuple G.

In GNNs, the complex relationships and dependencies among nodes in a graph are captured by aggregating and transforming information from neighboring nodes. Figure 2.8 demonstrate

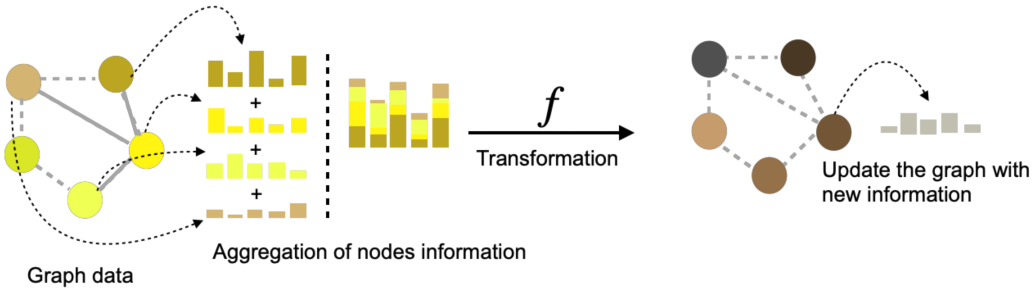

Figure 2.8: Basic representation of GNN. In GNN layer, the aggregation function aggregates information from the neighboring nodes and edges in the graph. The transformation function f (e.g. neural networks) is used to obtain the latent vector information of the node.

the process of learning the relationships between nodes and edges in a graph. A GNN layer takes as input a graph with certain node and edge features and outputs a graph with the same topology where the node, edge, and global graph information are updated. To achieve this, the node and edge information represented as feature vectors are first transformed by a transformation function, into vectors in higher dimensional space (feature space) referred to as node and edge states respectively. Depending on the GNN architecture, transformation functions can be fully connected layers, convolutional layers, or recurrent layers. As the fundamental part of GNNs, the so-called propagation (or message-passing) function is used to update these nodes' states by aggregating information from the neighboring nodes and edges in the graph. The readout function comprises the ultimate process of aggregating this updated latent information into a solitary latent vector, which can subsequently be used for predicting the target property for the graph. Depending on the GNN architecture, the propagation and readout functions vary on how the node, edge

and graph latent vectors are calculated. The next sections explain different GNN architectures according to [158].

2.4.2.2 Message Passing Neural Network

Message Passing Neural Network (MPNN) [159] is a GNN architecture which is originally proposed to learn molecular graph data, by propagating messages between nodes. Message and updating functions are two important functions in MPNNs,

$$\mathbf{H}_i^{(k)} = U^{(k)}(\mathbf{H}_i^{(k-1)}, m_i^{(k)}) \text{ with } m_i^{(k)} = \sum_{j \in \mathcal{N}_i} M^{(k)}(\mathbf{H}_i^{(k-1)}, \mathbf{H}_j^{(k-1)}, \mathbf{e}_{ij}) \tag{2.32}$$

where $\mathcal{N}_i$ denotes the connected neighboring nodes of node i and the message M_k exchanged between nodes i and j in the k-th layer depends on their respective node representations and the edge features connecting them. $\mathbf{H}_i$ is the node representation of node i. U_k is the node updating function in the k-th layer. To aggregate the messages from the neighbors and the node representation itself, U_k is used to update the node's status in the k-th layer. Aggregate function is another function in MPNNs to sum up all the messages from the neighbors.

Message passing occurs in two steps, the first step involves gathering the information of the nodes (or edges) surrounding a target node by collecting their node states. In the second step, these states, along with the state of the target node are aggregated using an aggregate function such as sum or average. If the final task is to predict the property of a graph, then these updated node states are further aggregated using a graph-level aggregation function, termed readout. A feed-forward neural network as the output layer can be used to map the updated nodes and edges of the graph to the target property.

2.4.2.3 Graph Convolutional Neural Network

Graph convolutional networks (GCN) [160] is the most popular GNN architecture due to their simplicity and effectiveness in variety of applications. In GCN, the graph input data is represented by an adjacency matrix that describes the connections between nodes in the graph, and the features associated with each node. In case of multiple layers of GCNs, each layer aggregates information from neighboring nodes in the graph, taking into account both the feature information associated with nodes and the graph structure itself. After each graph convolutional operation, a non-linear activation function such as ReLU is typically applied to introduce non-linearity into the network. The propagation rule to update the node representations in each layer is written as,

$$\mathbf{H}^{(k+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(k)}\mathbf{W}^{(k)}\right) \tag{2.33}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix of the given graph G which defines the connections between nodes as well as self-connections in the graph. The adjacency matrix allows to incorporate the node features in updating the node representations. $\mathbf{I} \in \mathbb{R}^{N \times N}$ (N is the number of nodes) is the identity matrix and $\tilde{\mathbf{D}}$ is a diagonal matrix with $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$ and $\sigma(.)$ (dot as the placeholder) is an activation function such as ReLU. $\mathbf{W}^{(k)}$ is the layer-wise linear transformation matrix in the k-th layer which will be trained during the optimization.

In the final layers of the network, fully connected layers can be added for tasks such as node classification or graph-level prediction. The final layer of the network produces the desired output, which could be node-level or graph-level predictions.

2.4.2.4 Graph attention Neural Network

Graph attention neural network (GAT) [161] implements attention mechanism instead of considering static weights between nodes. Unlike GCNs where the importance of node j is determined by the weight of their edge α_{ij}, GAT automatically learns the importance of each neighbor. Attention mechanism has been widely used in natural language processing [162] and computer vision [163].

Graph attention mechanism defines how to transfer the hidden node representations at layer $k-1$ denoted as $\mathbf{H}^{(k-1)} \in \mathbb{R}^{N \times F}$ (with F filters or number of hidden units) to the new node representation $\mathbf{H}^{(k)} \in \mathbb{R}^{N \times F'}$. Shared linear transformation, denoted as $\mathbf{W} \in \mathbb{R}^{F \times F'}$ is applied to every node to insure sufficient expressive power by transforming the node representation from lower level to higher level. Shared linear transformation involves applying a single weight matrix to the feature vectors of all nodes in the graph. This transformation projects the node features into a high-dimensional space, enabling the attention mechanism to effectively capture the relationships and dependencies between nodes. Self-attention is defined for every node in order to measure the attention coefficients for every pair of nodes through a shared attentional mechanism a [158],

$$e_{ij}^{(k)} = a(\mathbf{W}\mathbf{H}_i^{(k-1)}, \mathbf{W}\mathbf{H}_j^{(k-1)}) \tag{2.34}$$

$e_{ij}^{(k)}$ indicates the relationship strength between node i and j in the k-th layer. To take into account the graph structural information, the first-order neighbors of each node (below denoted as $\mathcal{N}_i$ for node i) are used in attention mechanism. The attention coefficients are then normalized with the softmax function to make the coefficients comparable across different nodes,

$$\alpha_{ij} = \text{softmax}_j(\{e_{ij}\}) = \frac{\exp(e_{ij})}{\sum_{l \in \mathcal{N}_i} \exp(e_{il})}. \tag{2.35}$$

α_{ij} can also be interpreted as the transition probability from node i to each of its neighbors j [158].

As it was discussed by Veličković, Cucurull, Casanova, Romero, Liò, and Bengio [161], the attention mechanism a is a single-layer feedforward neural network, parametrized by a weight vector $\mathbf{a}$, and applying a LeakyReLU nonlinear activation function. LeakyReLU is a variant of the ReLU activation function, which allows for a small, non-zero gradient when the unit is saturated and not active [164]. This can help to prevent the "dying ReLU" problem, which occurs when ReLU units consistently output zero, effectively killing off that neuron. Therefore, the attention coefficient can be calculated as,

$$\alpha_{ij}^{(k)} = \frac{\exp\Big(\text{LeakyReLU}\Big(\mathbf{a}^T\Big[\mathbf{W}\mathbf{H}_i^{(k-1)} \| \mathbf{W}\mathbf{H}_j^{(k-1)}\Big]\Big)\Big)}{\sum_{l\in\mathcal{N}_i}\exp\Big(\text{LeakyReLU}\Big(\mathbf{a}^T\Big[\mathbf{W}\mathbf{H}_i^{(k-1)} \| \mathbf{W}\mathbf{H}_l^{(k-1)}\Big]\Big)\Big)} \tag{2.36}$$

where the concatenation of two vectors is denoted by $[\dots||\dots]$. The new node representation is calculated by a linear combination of the neighboring nodes multiplied by their corresponding weights determined by the attention mechanism,

$$\mathbf{H}_i^{(k)} = \sigma\Big(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}\mathbf{H}_j^{(k-1)}\Big) \tag{2.37}$$

For learning stability of attention coefficients, multi-head attention mechanism is used in which different similarity functions over the nodes are determined instead of using only one single attention mechanism. This means that the importance of neighboring nodes to each node is calculated multiple times in parallel. The final node representation will be a concatenation (denoted as $||\dots$) of the node representations learned by different attention heads,

$$\mathbf{H}_i^{(k)} = ||_{t=1}^{T}\sigma\Big(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^t\mathbf{W}^t\mathbf{H}_j^{(k-1)}\Big) \tag{2.38}$$

where T is the total number of attention heads. α_{ij}^t and $\mathbf{W}^t$ are the calculated attention coefficient and the transformation matrix of the t-th attention head, respectively.

2.4.2.5 Graph Networks

Graph networks (GraphNet) are specifically designed to capture the relationships and interactions between entities represented as nodes and edges in a graph, enabling learning of complex graph structures [165]. Message passing is the key concept of GraphNets where nodes exchange information through messages and update their states respectively. In GraphNets, the messages are computed based on the attributes of the neighboring nodes, edges and global attributes which contain the overall properties and characteristics of the entire graph. The messages capture the local interactions and dependencies in the graph. Node, edge and global update functions incorporate

the received messages and current attributes to compute new representations for nodes, edges and global properties. To obtain the graph-level representations, the readout function aggregates the information of nodes or edges which summarises the overall individual nodes or edges. The flexibility of GraphNets allows the design of various architectures by specifying the message passing, update and readout functions. Incorporating relational inductive biases, such as symmetry, locality and compositionality enable the effective learning and generalisation on the data [165].

2.5 Explainability Artificial Intelligence

Explainable artificial intelligence (XAI) is a field of ML which consists of several techniques to understand the behavior of a model and provide explanations as to how a model made a prediction. Using XAI techniques in ML provides an opportunity to understand how the behavior of the model is influenced by its training data and to build confidence in their predictions by understanding when the model makes the wrong assumptions. XAI techniques foster trustworthy predictions by providing quantitative metrics and by illuminating the internal workings of the model, enhancing human comprehension [166, 49]. Therefore, XAI gives new abilities to improve the weaknesses of the model in predicting a specific target and make the model's actions comprehensible to everyone using the model's predictions.

As the complexity of DNN and GNN models increases, a large number of layers, as well as a high parameter count, implies that such models are black-box ML models [167], which means understanding the rationale behind predictions is a challenging task. Thus, the integration of XAI techniques becomes increasingly important in DNNs and GNNs. There are numerous techniques to incorporate XAI into DNN and GNN models [168, 169]. The focus of this thesis is on a straightforward explainability method known as attribution [170]. Attribution methods have found widespread use in applications where input data consists of images, text, and graphs [171, 172, 170, 173]. Attribution scores highlight specific regions of an image, characters or words in text, and particular nodes and edges in graphs, which have an impact on the decision-making process of the ML model employed for the task.

Feature attribution techniques which are common throughout XAI, are based on attributing the model behavior to features in the dataset. A feature attribution represents the influence of that feature on a specific target property. In case of GNNs, the attribution scores are assigned to nodes and edges to highlight the importance of them to the final prediction of the model. To visualise the attribution scores of nodes and edges, an heatmap is usually overlaid on top of the graph data. For instance, in case of a molecular graph, the attribution scores can be visualised by heatmaps to highlight the importance of individual atoms to the traget property of the molecule. From these heatmaps, one can deduce structural correlations between the model's rationale for

good or bad predictions, and compare them to existing knowledge (i.e. the ground truth) of why the prediction should be so.

Gradient-based approaches in explainability are common to obtain the feature attributions in GNNs. GradInput (GI) [174], class activation map (CAM) [175] and gradient class activation map (GradCAM) [176] are examples of explainability approaches which have been shown to successfully explain predictions made by GNN models for molecular structure-property prediction models, [170] i.e. they can reveal the contribution of individual atoms or atom pairs to the model's decision.

In this thesis, we use CAM method to obtain the feature attributions in the GNN model. To generate the CAM weights, the final GNN convolutional layer or feature map is first utilized. The weights of the fully connected layer at the output layer are then used to obtain the weighted combination of the feature map. The resulting weighted combination represents the importance of each spatial location i.e. the nodes of the graph in the feature map for the specific class [177, 178].

2.6 Machine learning in investigating the structure-property relationship

Experimental and theoretical data serve as the foundation for uncovering the intricate connection between material structure and properties [179]. Statistical analysis of these data, when available in sufficient quantities, complements this exploration by providing an additional lens to examine relationships between material structure and properties. With sophisticated but computationally expensive theoretical methods originating in ab-initio techniques, the integration of ML techniques into various areas of theoretical and computational chemistry becomes increasingly important given their ability to infer structure-property relationships on the basis of large amount of data produced by this ab-initio techniques. [48, 49, 50].

In case of biomolecules, supervised and unsupervised ML, as data-driven methods, have been shown promising in predicting the geometric and electronic structure determination in proteins and peptides [12, 67] as well as the prediction of protein folding, protein-protein interactions, and protein stability [180]. Unsupervised ML, such as dimensionality reduction and clustering algorithms, is usually used to extract information from atomistic datasets, such as molecular dynamics dataset, to understand the simulated systems on a physical level. The analysis of biomolecular simulation data, using unsupervised ML, can provide satisfactory description of thermodynamic and kinetic properties of the system [51]. Among supervised ML techniques, GNNs are designed to handle molecular graph data and capture the intricate relationships between atoms, bonds, and functional groups for prediction of quantum mechanical properties of biomolecules. Predicting spectra for biomolecules, such as IR [181, 63] and XAS spectra [67], are examples of using GNNs.

2.6.1 Feature engineering and unsupervised machine learning on atomistic datasets

The atomistic dataset obtained by molecular dynamics simulations of biomolecules are often extremely large in terms of number of data points and dimensionality, i.e. the number of configurations in the saved MD trajectory and the number of particles in the biomolecule, respectively. Extracting meaningful insights from the vast expanse of MD data requires a feature engineering technique that transforms the raw dataset into a lower-dimensional feature representation. This transformation, achieved through the selection of appropriate collective variables (CVs), can be a challenging task [182].

Due to the presence of significant number of similar configurations in raw MD datasets, it becomes crucial to account for symmetries like the invariance of physical properties under translation, rotation, or permutation of equivalent particles when examining the kinetics of complex conformational changes [183, 184, 185, 186]. To consider these physical invariances, the analysis of MD datasets of biomolecules is performed by first choosing a set of collective variables, which are usually complex and nonlinear functions of the atomic coordinates, to represent the raw dataset. However, determining important structural features characteristic of metastable states is far from trivial due to the highly dynamic exploration of the free-energy landscape. Moreover, identifying significant motifs is a tedious task given the often vast amount of simulation data. Thus, it is necessary to convert the atomic coordinates of the trajectory frames from the MD simulation into some set of useful CVs to calculate relevant statistical quantities [187].

In this thesis, the Smooth Overlap of Atomic Positions [188, 189, 190, 191] (SOAP) was used to obtain the feature representation of the MD datasets. SOAP kernels are generic descriptors of local structures which discretize three-body correlation functions centered around each atom, capturing its relationship with neighboring atoms, and the relationships between neighboring local environments [189, 192]. Moreover, SOAP kernels are designed to capture physical symmetries such as invariances to translations, rotations and permutations of atoms [193, 194]. Within the SOAP formalism, a local atomic environments ρ_i^α of an atom i of chemical species α is described by a set of atomic densities, represented by Gaussians centered on neighboring atoms j,

$$\rho_i^\alpha(\mathbf{r}) = \sum_j \exp(-\frac{|\mathbf{r} - \mathbf{r}_{ij}|^2}{2\sigma^2}) f_\mathrm{c}(|\mathbf{r}_{ij}|) \tag{2.39}$$

where f_c is a cutoff function that selects smoothly atoms from a local environment within a radius r_c, from the central atom. The width of the Gaussians is set here to σ, which is typically given in Å. To prevent the number of local environments becoming prohibitively large for further calculations, SOAP environments can be created only for certain atoms in the biomolecular backbone. It should

be noted, though, that these environments contain all the atomic contributions of other chemical species within the cutoff. A kernel between local environments of two biomolecule conformations in MD dataset, denoted A_i and B_j (i and j refer to local environment of each atom), is usually formed via a rotationally-averaged squared overlap kernel [195] of the smooth atomic densities,

$$k(\mathrm{A}_i, \mathrm{B}_j) = \int_{\mathrm{SO}(3)} \left| \sum_{\alpha} \int_{\mathbb{R}^3} \rho^{\alpha}_{\mathrm{A}_i}(\mathbf{r}) \rho^{\alpha}_{\mathrm{B}_j}(\mathbf{r}) \mathrm{d}\mathbf{r} \right|^2 \mathrm{d}\hat{R} \tag{2.40}$$

The advantage of SOAP environments is that this integral is analytically solvable with spherical harmonics, controlled by $l_{\max}$ the maximum angular degree of the spherical harmonics, and orthogonal radial basis functions, controlled by $n_{\max}$ the maximum number of the radial basis functions [188].

Pairwise environment similarities of two biomolecule conformations A and B are stored in a similarity matrix $C_{ij}(\mathrm{A}, \mathrm{B}) = k(\mathrm{A}_i, \mathrm{B}_j)/\sqrt{k(\mathrm{A}_i, \mathrm{A}_i) k(\mathrm{B}_j, \mathrm{B}_j)}$. To construct a kernel K between A and B from the SOAP kernel k, we averaged over all possible pairs of environments,

$$K(\mathrm{A}, \mathrm{B}) = \frac{1}{N_{\mathrm{A}} N_{\mathrm{B}}} \sum_{i \in \mathrm{A}, j \in \mathrm{B}} k(\mathrm{A}_i, \mathrm{B}_j) \tag{2.41}$$

where N_{A}, N_{B} are the number of atoms in molecules A and B. Having defined the kernel between the conformations, a kernel distance [196] for conformers A and B is obtained from

$$D(\mathrm{A}, \mathrm{B}) = \sqrt{K(\mathrm{A}, \mathrm{A}) + K(\mathrm{B}, \mathrm{B}) - 2K(\mathrm{A}, \mathrm{B})} \tag{2.42}$$

The kernel distance D can then be used as the metric for representing the MD dataset via dimensionality reduction and clustering the similar molecular conformations. The workflow of analysing the raw MD dataset in case of unsupervised ML is shown in figure 2.9.

2.6.2 Supervised machine learning on molecular graph datasets

Biomolecules, particularly peptides and proteins, can be represented as molecular graphs, enabling the utilization of GNNs to capture the intricate interactions between atoms and predict specific molecular properties [197]. A GNN layer takes as input a graph with node and edge features, i.e. atoms and their relations with neighboring atoms, and outputs a graph with the same topology where the node, edge, and global graph information are updated. To achieve this, the node and edge information represented as feature vectors are first transformed into vectors in higher dimensional space (feature space) referred to as node and edge states respectively, using a transformation function. Transformation functions can be fully connected layers, convolutional layers, or recurrent layers, depending on the GNN architecture. A fundamental part of GNNs is

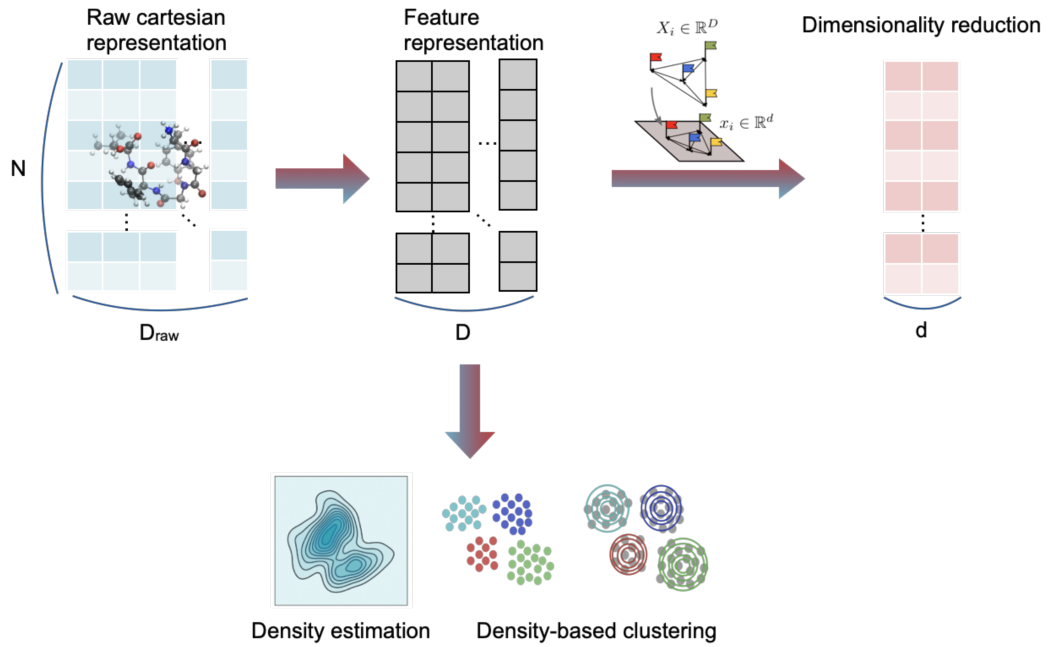

Figure 2.9: The workflow of unsupervised ML with atomistic datasets. The obtained feature dataset from feature engineering can be represented in low dimensions via dimensionality reduction techniques. Density estimation with clustering techniques can also be applied on the feature dataset to categorise the biomolecule conformations based on their similarities in the feature dataset. The idea of the figure was adapted from [17, 51].

the so-called propagation (or message-passing) process used to update these nodes' (or edge) states. Figure 2.10 shows how training a GNN model is applied to find a mapping between molecular structures and the molecular properties, in this case X-ray absorption spectrum.

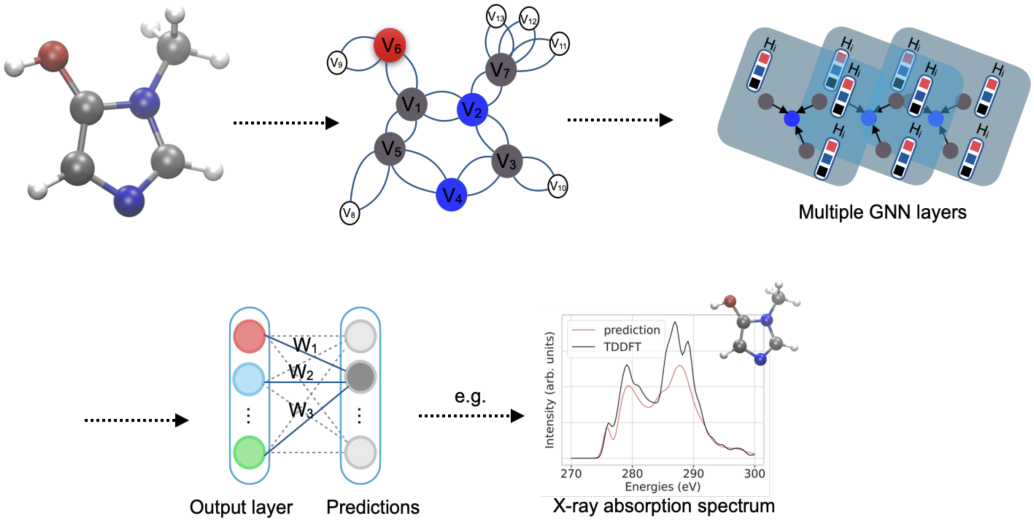

Figure 2.10: Workflow of training a GNN model based on molecular graph dataset. Several GNN layers can be used to learn the local and global structure of atoms in graph molecular datasets. Depending on the GNN architecture, each GNN layer aggregates the node features ($\mathbf{H}_i$) (for better visualisation, the node features are colored in the figure) to update each node representation. The final output layer is used to obtained the final mapping with the target property, e.g. X-ray absorption spectrum. The different colors of the neurons in the output layer denote the different weights of these neurons with respect to the activation values of last GNN layer.

3

Reconstructing the Infrared Spectrum of a Peptide using Unsupervised Machine Learning

"This chapter reproduces the contents of reference [12], with minor adjustments. The contribution of the author of this thesis is the development and design of the machine learning analysis on the atomistic dataset. Moreover, the author has performed the simulations, data processing and analysis of experimental data."

Dynamic processes in biomolecules, such as secondary structural changes in peptides and proteins, pose a challenge for the accurate and quantitative description of experimental data by theoretical approaches. Several structure-determining factors, such as the intramolecular hydrogen-bonding (H-bonding) pattern, play an important role in the rapid transformation between structural motifs. Moreover, the coupling between different vibrational modes lead to anharmonicity and complex intra- and intermolecular energy exchanges [198]. Although parts of these intricate relationships can be explained by the conformational diversity evident in all measured spectra, as shown in this chapter, it should be kept in mind that this still cannot lead to perfect agreement between experiment and theory, since anharmonicities (i.e., combination bands, overtones, and Fermi resonances) are difficult to treat even with very sophisticated theoretical approaches.

A useful approach to gain a fundamental understanding of these processes is to predict theoretical spectra and relate those to experimental measurements [199]. However, to facilitate direct validation of theoretical predictions, solvation effects are often neglected or considered by frequency maps [200, 201] and predictions are usually made on isolated systems. Frequency maps, either from *ab initio* [202], empirical [203] or using machine learning [204], correct the vibrational frequencies for solvent effects by using maps connecting the vibrational frequency shifts with the local electrostatic environment. Frequency maps have also been successfully used to correct vibrational frequencies for intermolecular H-bonding, although only for a rather simple system of liquid water [205]. Thus, although the link between gas-phase studies and biologically relevant reactions is discussed controversially, many important properties (e.g., protonation or interactions with ions and solvation effects) can be studied in gas-phase experiments, allowing a comparison with theoretical methods, such as first-principles and molecular mechanics approaches [31, 32, 206].

In this chapter, we explore the structural landscape of leucine enkephalin (LeuEnk), an experimentally well-studied biologically active endogenous opioid pentapeptide, by combining Replica-Exchange Molecular Dynamics (REMD) simulations with machine learning, as discussed in section 2.2.2 of chapter 2. Moreover, we discuss that considering only a few individual conformations in theoretical predictions without taking into account their correct statistical ensemble weight can result to discrepancies between theory and experiment. LeuEnk is a well-established standard in ESI-based biomolecular mass spectrometry [207, 208] and has been investigated extensively using IR-UV double resonance photofragment spectroscopy to obtain conformer-specific spectra at cryogenic temperatures [27, 42, 46, 209] as well as infrared multiphoton dissociation (IRMPD) at room temperature [74, 73]. LeuEnk lies here in an interesting size regime: small enough to efficiently use sophisticated computational methods while, on the other hand, large enough to allow insights into the low-lying minima of the conformational space [210, 211] and competing H-bonded networks. IRMPD experiments have been extremely useful in understanding key features of theoretical predictions [27, 44, 212].

In this chapter, comparisons with IRMPD experiments in the gas phase are performed based on representative peptide conformations identified from extensive REMD simulations and clustering. IR spectra were calculated at *ab initio* levels for each identified representative conformer. The influence of the conformational ensemble was assessed by means of a hierarchical clustering using the Probabilistic Analysis of Molecular Motifs (PAMM), as it was explained in section 2.3.2.4 of chapter 2. While spectra from LeuEnk were experimentally measured in the gas phase at room temperature using IRMPD, the above mentioned methods were used to study the conformational averaging of recurring motifs to finally predict the theoretical IR spectrum at a finite temperature with high accuracy.

3.1 Exploring the configurational space of leucine enkephalin

Principal Component Analysis (PCA) was used to project the entire structural landscape of an N-terminal protonated LeuEnk (amino acid sequence[YGGFL+H]$^+$) shown in Figure 3.1, sampled during the REMD at the experimental relevant temperature of 300 K. The first principal

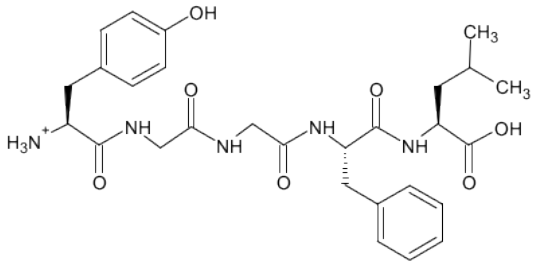

Figure 3.1: 2D structure representation of N-terminus protonated LeuEnk

component can be defined as a direction that maximizes the variance of the projected data. Hence, the cumulative explained variance ratio was used as a function of the number of components in order to reduce the dimensionality of the Smooth Overlap of Atomic Positions (SOAP) kernels. Consequently, Figure 3.2 shows the cumulative explained variance versus the first 40 principal components. On the basis of the first two principal components, the SOAP feature space results

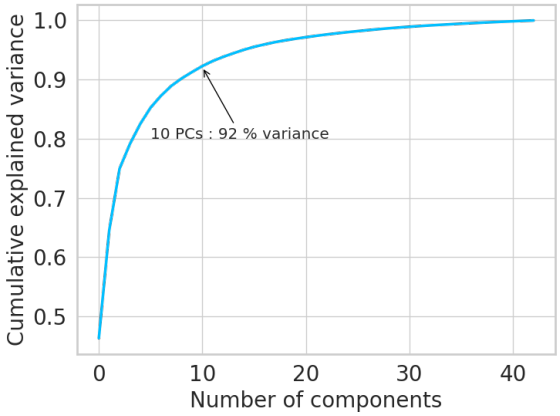

Figure 3.2: Relationship of the cumulative explained variance of the SOAP kernels to the number of principal components (eigenvectors) of the PCA. The cumulative explained variance is a statistical measure of how much information, i.e. variation, can be retained in the data set as more and more principal components are included.

in an easily interpretable two-dimensional map given in Figure 3.3a, where points are coloured additionally according to the PAMM cluster they belong to. The conformation with the lowest

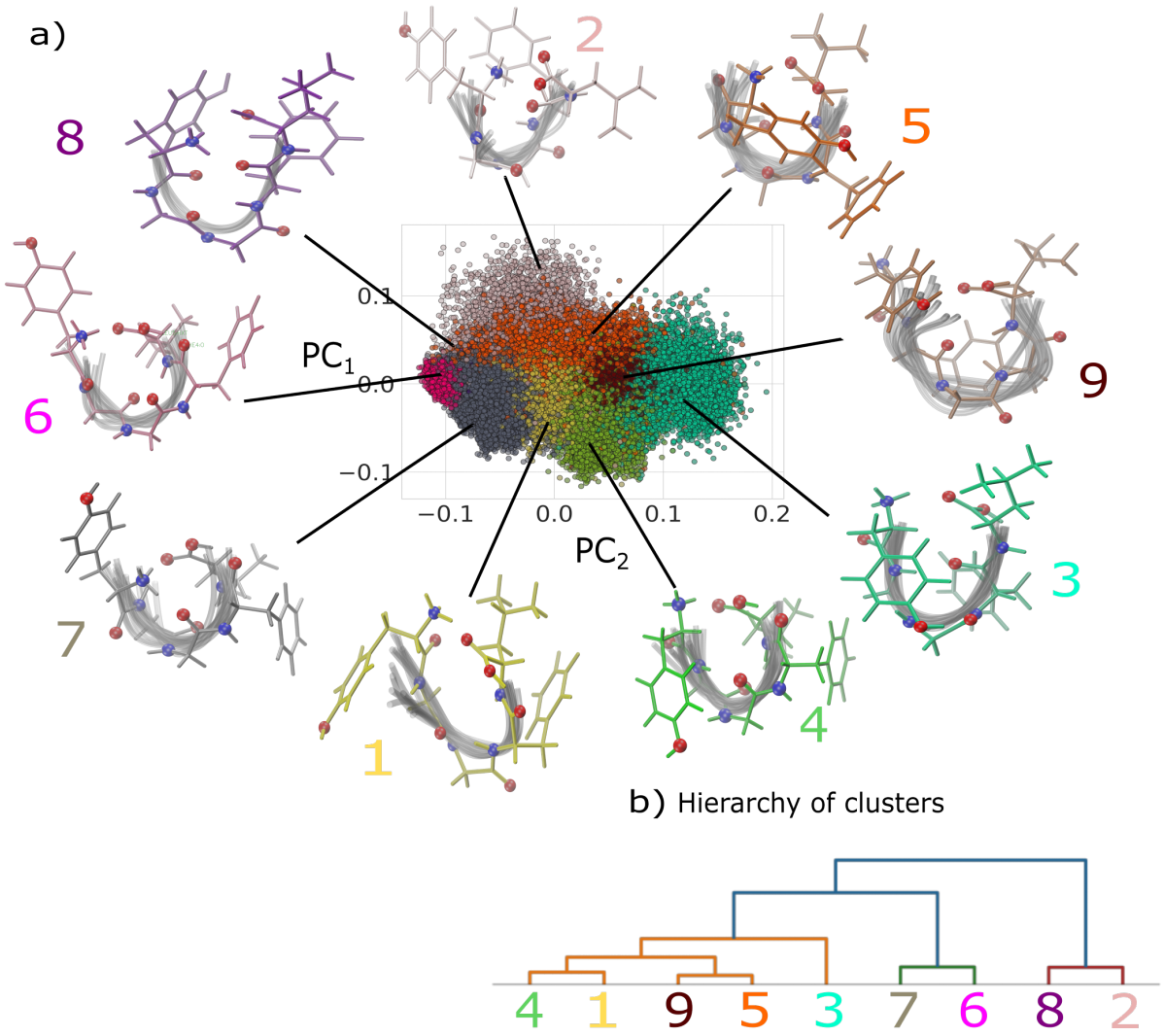

Figure 3.3: Low-dimensional conformer map of the gas phase LeuEnk SOAP feature space at 300 K. The first two Principal Components (PC) of a Principal Component Analysis (PCA) and the classification identified using PAMM are shown (a). More details on the PCA are found in the SI. Colored molecular line representations correspond to the most energetically favorable (i.e. representative) conformer of each cluster, while the gray shaded configurations were randomly selected from the same cluster and shown for comparison. Hierarchy of clusters (b) illustrating the similarities of clusters as they result from hierarchical clustering to facilitate grouping of similar conformer clusters. Color coding corresponds to molecular conformers and to the dots in the low-dimensional conformer map in (a).

energy of each identified PAMM conformer cluster was selected, and after confirming that they belonged to the same PAMM cluster after geometry optimization with DFT, its SOAP distance was recalculated. The geometry-optimized conformers were embedded in the two-dimensional map shown in Figure 3.4. These selected conformers with the lowest energy of each cluster is

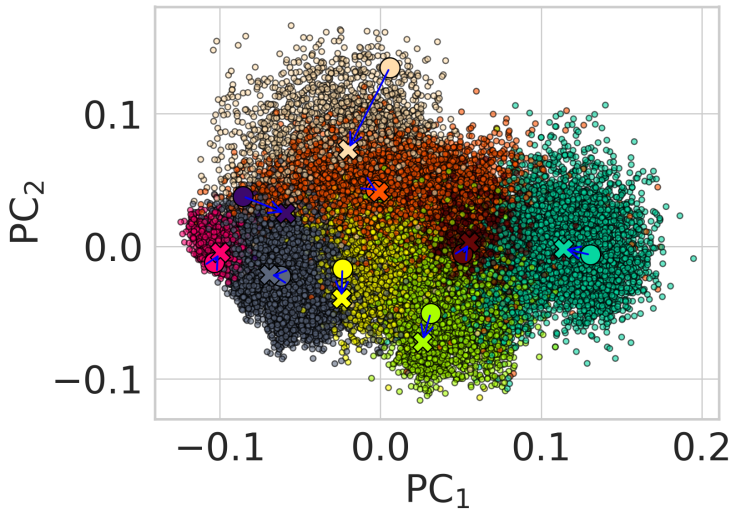

Figure 3.4: 2D-PCA representation of LeuEnk conformers from a REMD simulation along with an embedding of DFT geometry optimized conformers of each identified representative PAMM conformer. Circles and crosses indicate the position of the original PAMM conformer and the geometry optimized conformers in the 2D PCA map, respectively.

considered as the representative motif of the corresponding cluster and those are shown around the map in Figure 3.3a. The dendrogram in Figure 3.3b shows the hierarchical merging of conformer clusters and the structural metastability represented by this dendrogram reveals the connection between the free-energy basins (and associated conformers). The colors correspond to the coloring of the molecules in Figure 3.3a and are intended to visually help to group similar conformer clusters together. The hierarchical merging is done in a controlled way by checking how fuzzy the cluster boundaries are [128]. As in reality biomolecules actively explore the free-energy landscape continuously [140], a weighting factor w which estimates the statistical relevance of the various metastable states is introduced. w resembles in principle a canonical ensemble weight and is obtained from the population of structures within a cluster normalised with the overall population and is given in percentage in Table 3.1. Remarkably, cluster **7**, the conformer with the highest relevance, is also the one identified in a previous work of Burke et al.[27, 42] in terms of its secondary structure and H-bond configuration. While cluster **5**, the cluster with the second highest relevance, is quite similar to cluster **7** in its secondary structure, the side chain angles differ (a detailed conformer comparison of both cluster is shown in Figure 3.5).

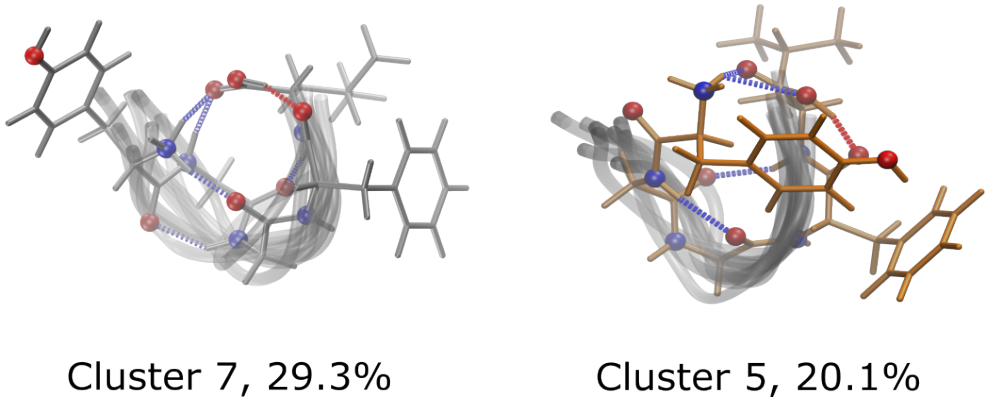

Figure 3.5: 3D-representation of the two clusters (**5** and **7**) with the highest weight, given in percent. Oxygen and nitrogen atoms are highlighted with blue and red colors, respectively. Hydrogen bonds formed between NH···O and OH···O are indicated by corresponding colored dotted lines.

3.2 Hydrogen bonding statistics

In general, the REMD simulation showed highly dynamic changes in the H-bonding patterns throughout the simulation. Figure 3.6 shows the H-bonding free energies of $N-H\cdots O$ at 300 K obtained from the full REMD trajectory at that temperature. As will be explained a little later, $N-H\cdots O$ is the predominant H-bonding motif in the gas phase of LeuEnk where no water is present. In Figure 3.6, s_A and s_D denote the number of H-bonding accepted and donated in $N-H\cdots O$ triplets, while s_H is the number of hydrogen bonds in which a hydrogen is involved. More specifically, the H-bonding definition provided by PAMM consists of a continuous function that given in input a specific triplet [138] returns a real number between 0 and 1, with 1 being a perfect match with the typical H-bonding pattern found in the training data. Thus, summing and averaging over all the possible triplets in which a specific tagged atom is involved we can define a collective variable that effectively corresponds to an H-bonding counting function. In the example reported in Figure 3.6, s_D is defined as the sum over all the possible triplets in which a generic tagged N is acting as donor in a $N-H\cdots O$ pattern. Similarly, s_A is defined as the sum of all the possible triplets in which a generic tagged O is acting as acceptor in a $N-H\cdots O$ pattern. s_A, s_D and s_H are calculated for each snapshot of the REMD trajectory using the HBPAMM [138] implementation from the PAMM package. The probabilities P for these values are obtained from the normalized and smoothed histograms of these quantities, i.e. the probability distributions, and consequently used to calculate the free energy equivalent via $F = -k_B T \ln(P)$, where P denotes an unbiased probability of the different H-bond configurations, as obtained from the histograms of s_A, s_D, and s_H calculated from the REMD trajectory at 300 K, while k_B is the Boltzmann constant and T is the temperature. Note that s_D indicates the number of hydrogens donated by nitrogen in general, independent of the specific nitrogen and the total number of nitrogens in the molecule

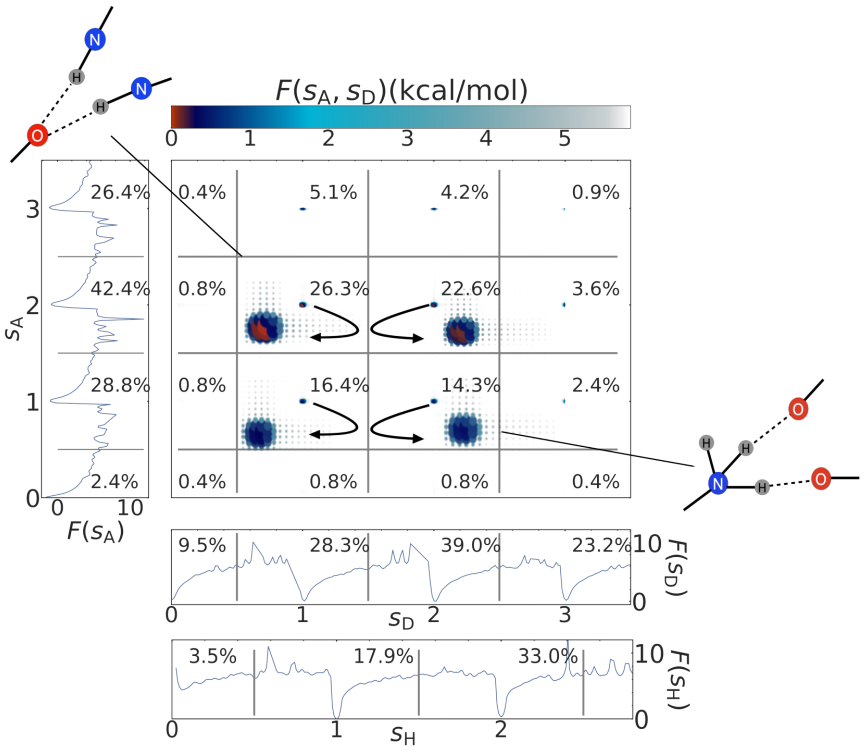

Figure 3.6: Hydrogen bonding statistics of LeuEnk from the REMD trajectory at 300 K. Probability distributions have been smoothed with triangular kernel of width 0.025 and are represented in terms of $F = -k_B T \ln(P)$, expressed in kcal/mol. We also report the integrated (joint) probabilities (in percent) for the corresponding region of different integer values of s_A, s_D and s_H. Schematic representations of two exemplary H-bonding configurations are given in the insets.

(nitrogens that do not donate hydrogens are ignored in the calculations). This is similarly true for s_A and s_H. A consequence of this definition is that whenever s_D is greater than 1, the N-terminal NH_3^+ is definitely involved in H-bonding, since otherwise there are only NH groups in the molecule that can donate only a single hydrogen. On the other hand, if s_D is very close to two or even above, it means that basically only the NH_3^+ group is involved in H-bonding (another possibility which would give $s_D = 3$ would be a rather unlikely configuration in which NH_3^+ donates three hydrogen bonds while another single NH donates one). $s_D = 0$ means, obviously, that no H-bonding occurs in the $N - H \cdots O$ triplets.

Integration over probability distributions in the different H-bonding regions in Figure 3.6 resulting in the mentioned percentages is used to explain the (often just subtle) differences between the identified PAMM clusters. Percentages reported in the free energy surface of Figure 3.6 denote the integrated joint probability of finding a configuration in the vicinity of the different integer numbers of donors and acceptors. The percentages in the free energy profiles on the sides of Figure 3.6 denote the integrated probabilities in the vicinity of the integer number of hydrogen bond acceptors (or donors), regardless of the number of donors (or acceptors, respectively). Consequently, the (averaged) H-bonding pattern of $N - H \cdots O$, denoted by $\langle s_D \rangle$ and $\langle s_A \rangle$, and the most probable H-bonding configuration (s_A, s_D) along with its joint probability are calculated for each subensemble and are given for each identified cluster in Table 3.1. While the averaged

Cluster	$\langle s_A \rangle$	$\langle s_D \rangle$	Most probable (s_A, s_D)	w %
1	1.82 ± 0.64	1.69 ± 0.78	(1,1) 49%	13.0%
2	1.71 ± 0.56	1.92 ± 0.74	(1,3) 52%	4.6%
3	1.41 ± 0.26	1.67 ± 0.58	(1,1) 37%	9.8%
4	1.54 ± 0.40	1.69 ± 0.57	(1,1) 40%	14.2%
5	1.59 ± 0.42	1.77 ± 0.48	(1,2) 30%	20.1%
6	1.53 ± 0.40	1.65 ± 0.41	(1,1) 34%	2.5%
7	1.34 ± 0.29	1.74 ± 0.31	(1,2) 42%	29.3%
8	1.95 ± 0.01	0.95 ± 0.01	(2,1) 60%	0.2%
9	1.67 ± 0.42	1.82 ± 0.58	(2,1) 30%	6.3%

Table 3.1: Summary of LeuEnk H-bonding. $\langle s_D \rangle$ and $\langle s_A \rangle$ denotes the weighted averaged number of donated and accepted $NH \cdots O$ and its standard deviation. w quantifies the weight of each cluster with respect to the relative population of each cluster to the total number of conformers found in the REMD trajectory at 300 K. Most probable (two-dimensional) H-bonding configurations are given as tuples along with their joint probability.

H-bonding values $\langle s_A \rangle$ and $\langle s_D \rangle$ show some differences between subensemble clusters, the most probable H-bonding patterns (s_A, s_D) are more distinct and are quite different for each conformer indicating a more or less pronounced involvement of the NH_3^+ group. Moreover, the observed high standard deviations of the averaged values are further evidence of the quite dynamic H-bonding observed during the REMD. Both the N- and C-termini are usually involved in H-bonding and

thus LeuEnk often resembles a β-hairpin or α-helix secondary structure motif. Clearly, the NH_3^+ group plays an important role in forming these secondary structures due to the engagement of this group with several nucleophilic regions of the molecule. Interaction between amide hydrogens and carbonyls and the strong COO−H$\cdots$O=C at the C-terminal are characteristic of γ-turns and noticeable interactions for most clusters. Although the involvement of the carboxyl or phenol in the formation of $O-H\cdots O$ hydrogen bonds is often of great importance, especially when water is nearby, we neglect this type of otherwise quite important H-bonding in our analysis for simplicity, since H-bonding of the type $N-H\cdots O$ is able to explain most of the conformer structural differences. Moreover, if there are $O-H\cdots O$ H-bonds (in cluster **2**, **3** and **9** this H-bonding is not apparent) they are always formed between the first and second glycine and the carboxyl and phenyl group, respectively. Thus, this type of H-bonding does not contribute to the highly dynamic H-bonding behavior. Motifs of the clusters **1** and **4**, which have lower probability, form usually α-helix-like motifs. However, the motif of cluster **3** contains probably a δ-turn, which is sterically quite improbable. It is worth pointing out that clusters **2**, **3** and **9** lack a strong COO−H$\cdots$O=C at the C-terminal. While the carbonyl in the third glycine of LeuEnk shows typically just one H-bond with amides for most clusters, the motif of cluster **9** has interactions of this group with phenylalanine and leucine amides and an absent H-bonding between N- and C-terminus.

Overall, H-bonding strongly correlates with the wavenumber and intensity of different vibrational bands[213, 214]. Thus, combining the information on the contribution of H-bonding in both the C=O and N−H groups to the overall IR bands in the following elucidates the impact of the H-bonding of the different conformers on the final IR signal.

3.3 Infrared spectroscopy of identified recurring structural motifs

Armed with the information on the LeuEnk conformers and their H-bonding network, we are now in the position to evaluate the influence of the subensembles from the REMD simulation on the IR signatures of LeuEnk. It should be noted, however, that the broadening and shape intensity of the peaks in the experiment are influenced only partly by effects related to conformational dynamics of the molecules [35]. One has to take into account that the formation of the lowest energy conformation is driven by a balance between kinetic, enthalpic, and entropic effects. The kinetic trapping, specific to the experimental conditions, can be entering into the final balance and cause the formation of different conformers [76]. The conformational distribution depends thus strongly on the thermodynamic conditions in the experiments as these have a strong impact on the formation pathway to the metastable conformations. Another issue that should be taken into consideration is that experimental IRMPD spectra are usually compared to calculated linear

absorption spectra. In reality, IRMPD spectra depend on very complex mechanisms involving sequential photon absorption, stimulated and spontaneous emission, energy distribution, and fragmentation, all of which have their own time scales and will affect the peak shape [14]. Despite the theoretical complexity of the interpretation, it has been shown that IRMPD spectra agree with calculated ones if the comparison is performed carefully. Hence, it is generally accepted that the peak positions obtained from theory can be trusted [41]. As it is shown in this chapter, consideration of representative conformers can improve the IRMPD spectrum prediction significantly.

In this chapter, harmonic vibrational modes of representative conformers for each identified PAMM cluster have been calculated in order to compare the vibrational modes of the obtained conformational families – represented by a single conformer – to experimentally observed spectroscopic fingerprints. It should be noted, however, that while this approach should be able to capture the main features of the vibrational spectra of the corresponding conformational family, some particular features may be missing as a result of neglecting other important local minima, anharmonicity, and other effects. Figure 3.7 shows the weighted average IR spectra of the amide I/II/III/V and amide A/B regions of the 9 PAMM conformers from Figure 3.3 using their corresponding weighting factor and compares the prediction with the experimental IRMPD spectra along with Pendry reliability factors. Table 3.2 reports the IR shift of amide I peak of PAMM

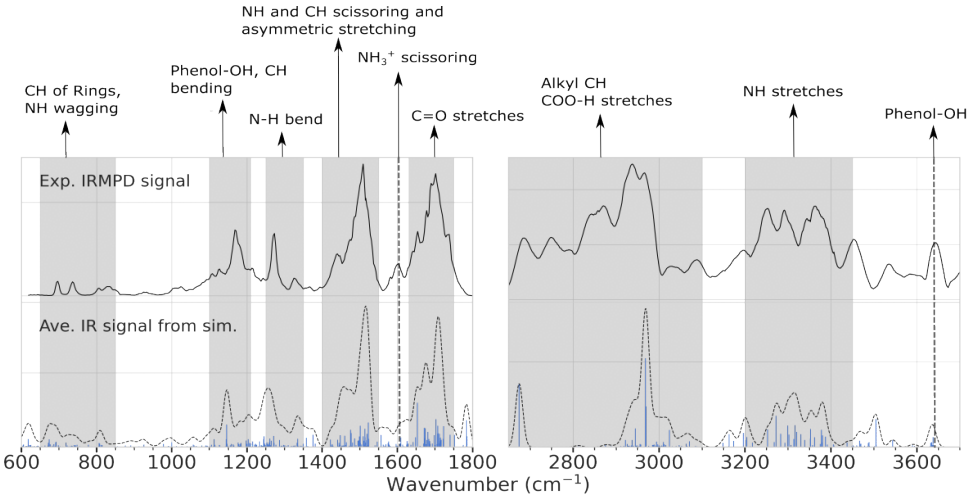

Figure 3.7: Comparison between experimental IRMPD and predicted average IR spectra using the PAMM conformers. Calculated IR intensities are convoluted by Gaussians with a 10 cm^{-1} full-width of half-maximum. Pendry reliability factors R_P are 0.56 for the left panel and 0.7 for the right panel.

representative conformations with respect to IRMPD experimental spectrum. Pendry reliability factors [215] R_P are often used for unambiguous comparisons of theoretical and experimental IR spectra [41]. In short, the R_P factors are sensitive to peak positions because they take into account information about the intensities and approximate half-widths of the peaks. A perfect match between two spectra gives $R_P = 0$, while $R_P = 1$ means no correlation. As R_P is quite sensitive

Cluster	w%	Amide I peak shift (cm^{-1})
1	13%	3.6
2	4.6%	10.1
3	9.8%	13.8
4	14.2%	10.2
5	20.1%	2.3
6	2.5%	23.2
7	29.3%	3.1
8	0.2%	18.2
9	6.3%	28.6

Table 3.2: IR shift of amide I peak of PAMM representative conformations with respect to IRMPD experimental spectrum.

to small kinks, the rather noisy high wavenumber part of the spectrum was smoothed and is shown consistently in the following, while the low wavenumber part of the experimental spectrum was smoothed only for calculating R_P factors and is otherwise shown unsmoothed. A detailed explanation of this factor and practical examples can be found elsewhere [212]. The averaged estimated IR spectrum shows an overall good agreement in the intensity pattern, particularly for the lower wavenumber range as indicated by the low R_P value, indicating that the probabilistic PAMM clustering is able to provide adequate statistical weights for the prediction of the IR spectra.

The assumption of representing the vibrational modes of obtained conformational families by a single conformer from each cluster was further examined by assessing whether spectra within a cluster are more similar to each other than to other clusters. We calculated R_P values for all conformers from cluster **7** – selected using farthest point sampling (FPS) (explained in section 2.3.2.4 of chapter 2) – and all other representative conformers, and summarized them in Table 3.4. Comparison of the values in the table shows that the similarity with the representative IR signal of FPS-selected conformers within the cluster **7** is generally better than for other clusters, as the average values are higher than the value within the cluster for both low and high wavenumber regions. The same procedure was applied to the FPS-selected conformers of all other clusters, with the similar result that the R_P were more consistent within each cluster than to others.

IR peaks in the amide I/II/III region are characteristic of the covalent bonds of the peptide backbone [216]. Normal modes in the amide III region ($1200-1400\,cm^{-1}$), which are combinations of N−H in-plane bending, C−N stretching, and C^{α}−N bending vibrations, are complexly interrelated but are considered structurally sensitive bands for polypeptides [217]. The complexity of amide III vibrational modes makes deciphering the correlation between different secondary structures and this region of the vibrational spectrum much more difficult than for the amide I/II regions [218]. Not surprisingly, most DFT studies that proposed to relate the intensity in this range to the full range of possible backbone dihedral angles did not provide satisfactory insight into the structural

sensitivity [217]. This is most likely due to the contribution of several effects, such as the dependence of the bond strengths on the backbone dihedral angles, the role of coupling between the amide III and $C^{\alpha}-H$ vibrations, and, most probably, the effect of intramolecular H-bonding. Peak positions in the amide I/II regions ($\sim 1500 - 1700\,\mathrm{cm}^{-1}$) are sensitive to various secondary structures but considered only weakly affected by side-chain conformations. [44, 219]. However, this region is particularly interesting when solvent comes into play as it is influenced strongly by it. Many studies have hence investigated this particular region and developed frequency maps that correct the peaks for solvent effects [200, 201, 202, 203, 204]. To establish a link to these frequency maps and illustrate the effects of specific H-bonding in the representative conformers on the IR band of amide I, we summarize the peak shift along with the specific H-bonding pattern of the representative conformers to the experimental spectra in Table 3.2. Interestingly, the most favorable clusters show the smallest peak shift. Moreover, a strong coupling between NH_3^+ scissoring modes and C=O stretches in this region has been reported [220]. Apart from this, it should be noted that the use of more than one conformation for the prediction of IR spectra of peptides in the amide I/II regions leads to better agreement with experiments, as shown previously[44, 221]. However, given the overlap of bands in this region due to the many different interactions between C=O and NH_3^+ of the different conformers, direct mixing of these signals in the amide I/II regions was not yet possible because the relative weights of the conformers were not previously known. In contrast, the excellent agreement of these regions with the experimental IRMPD spectrum in Figure 3.7 elucidates these complex interactions by incorporating the many different contributions of the conformations of the protonated LeuEnk through a thorough exploration of the potential energy surface.

Peaks in the higher wavenumber range between 2700 and $3600\,\mathrm{cm}^{-1}$ are often influenced by a more dynamic H-bonding and are inherently more complex to interpret due the anharmonicity. However, the agreement of the intensities in the $2900 - 3100\,\mathrm{cm}^{-1}$ range shows that an appropriate weighting of the different contribution via the PAMM conformer clusters, implicitly taking into account the H-bond statistics, is able to reproduce partly this normally highly red-shifted range after applying the standard scaling factor [222]. Particularly, it has been reported by Burke, Redwine, Dean, Mcluckey, and Zwier [27] that vibration modes from N−H of NH_3^+ interacting with phenol rings of the side chains or C=O are broadened, although the extent of their shift has been underestimated by DFT calculations [209, 42]. The presence of a broad low-intensity band at $3100\,\mathrm{cm}^{-1}$ is consistent with the dynamic H-bonds observed across the PAMM clusters, where NH_3^+ is highly likely to donate two hydrogen atoms, indicating that NH_3^+ (and the other amides) interact strongly with electron-rich regions of the peptide.

Closer examination of the contributions of individual conformations in Figure 3.8 shows that the calculated IR spectra of the **4**, **5**, and **7** clusters agree well with the IRMPD result, indicating

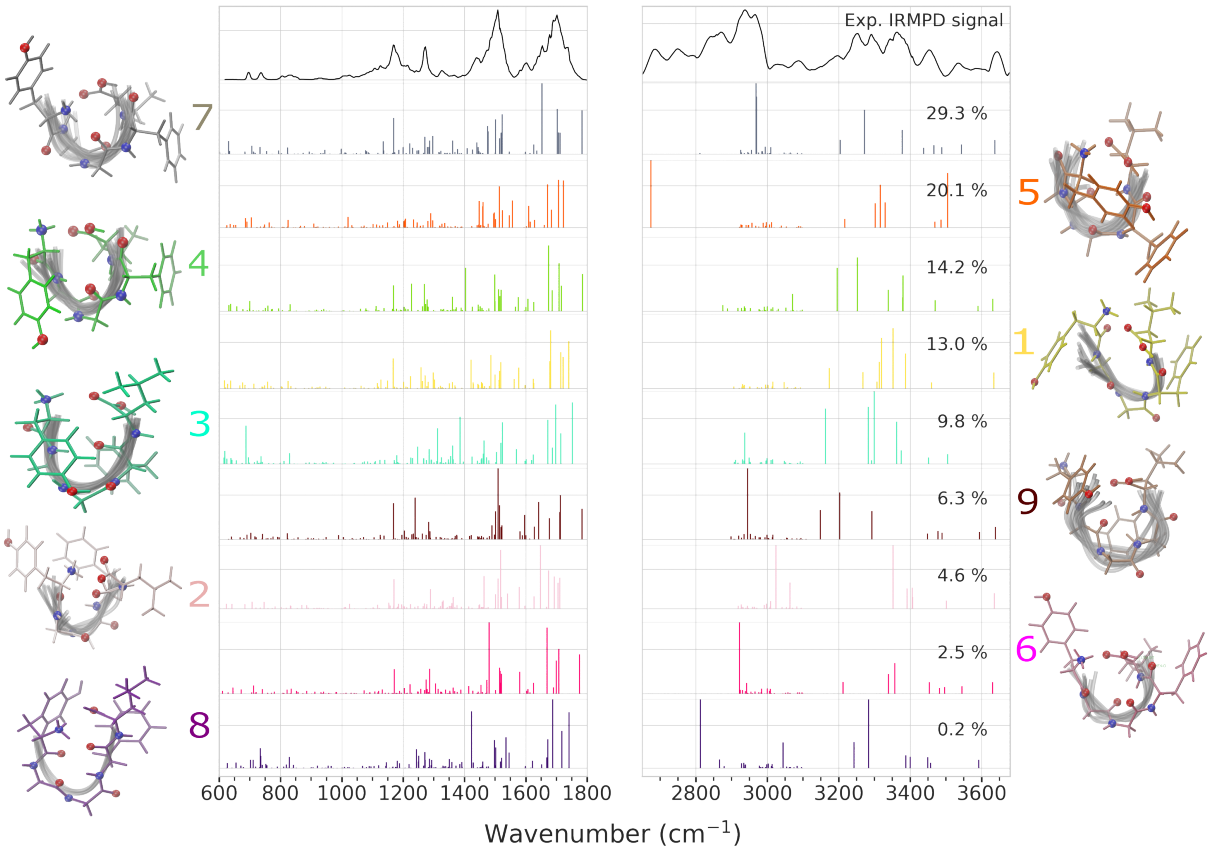

Figure 3.8: IR spectra of all PAMM conformer clusters and IRMPD spectrum. Spectra are colored according to the assigned representative conformer of each cluster which is shown next to the corresponding spectrum. Associated weights of each conformer cluster are shown in each panel.

that the higher-weight motifs are the most abundant conformers in the experiment. The peak in the amide A region at 3643 cm^{-1} is commonly assigned to phenol O−H stretches of tyrosine, which is well reflected in the averaged spectra in terms of peak position and intensity. The phenol O−H vibrational mode was previously reported to be rather insensitive to conformational variations in LeuEnk [74], which is confirmed when looking at individual contributions of conformers as shown in Figure 3.8. Surprisingly, however, there the clusters **3** and **5** actually have no signature of this peak around 3643 cm^{-1} due to the involvement of the phenol O−H in H-bonding.

The tendency of involving COO−H and the phenylalanine carbonyl in H-bonding, visible in almost all conformers of Figure 3.8 with the exception of clusters **2**, **3** and **9**, is represented by a strongly shifted peak at 3012 cm^{-1} from the "free" position at 3584 cm^{-1}[223, 27]. Near the N-H stretch region, the shoulder-like feature between 3400 and 3500 cm^{-1}, which has not been observed for single conformation spectra, could be related to the presence of weak and slightly longer NH$\cdots$C=O H-bonds or free NH stretches (excluding NH_3^+), which shift to higher wavenumbers and are less abundant in the peptide and consequently not as prominent[224, 223]. The broadening and nature of the multiple peaks between 3200 and 3400 cm^{-1} suggest that it is unlikely to assign a single conformation with a specific NH$\cdots$O=C bonding to this feature. However, the here presented conformational weighting sheds light on this particular region in terms of the intensity pattern and shape of the peaks.

The selected representative conformers via the lowest REMD energy were further investigated to verify whether they robustly identify the corresponding geometry optimized *ab initio* conformer with comparably low energy. Farthest Point Sampling (FPS) was applied to the distance kernel D (equation 2.42) to achieve a homogeneous exploration and sampling of the high-dimensional phase space. The 10 conformers indicated by large dots in Figure 3.9 were randomly selected from the **7** cluster in this way. To calculate vibration frequencies using *ab initio* methods, the conformers selected by FPS had to be geometrically optimized, which are indicated by the crosses in Figure 3.9. Table 3.3 summarizes the changes in pairwise kernel distances as well as the relative DFT energies of the FPS conformers to the representative conformer. Taking into account the initial position of the conformers, the position of the geometrically optimized conformers illustrates the highly complex transition pathways between conformations and the complex and shallow landscape of conformational free energies. In summary, random selection with FPS usually does not lead to the global minimum of a representative cluster. However, selection via the conformer with the lowest energy from the REMD energies seems to work very well here, since the resulting *ab initio* energy of the representative conformer is still lower than all geometry-optimized FPS-selected conformers.

The resulting IR spectra of the FPS-selected conformers, ordered by their pairwise distances from the representative conformer of cluster **7**, are shown in Figure 3.10. The distinct peak position and shape of the amide I band in Figure 3.10 is an indicator that all conformers possess a

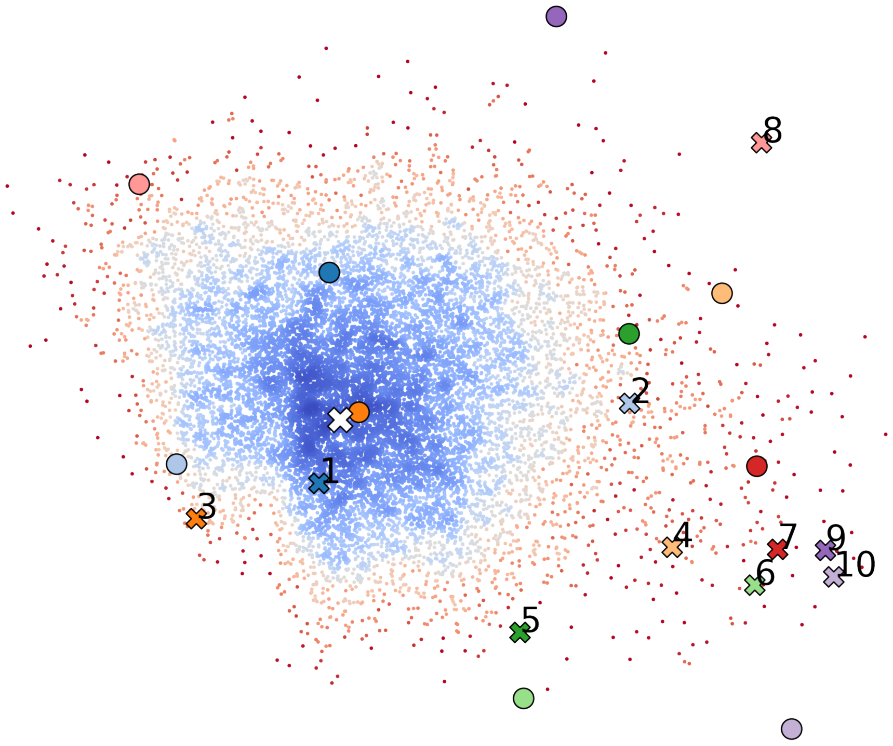

Figure 3.9: Scatter plot of cluster **7**. Points are colored according to the logarithm of the KDE determined from PAMM, which illustrates the free energy landscape (blue: low free energy; red: high free energy). FPS selected conformers, before and after *ab inito* geometry optimization, are indicated with circles and crosses, respectively. The white cross indicates the position of the representative conformer.

Conformer	D^{REMD}	D^{DFT}	ΔE^{DFT} / eV
1	0.0052	0.0018	0.46
2	0.0089	0.0043	0.38
3	0.0053	0.0067	0.68
4	0.0114	0.0077	0.24
5	0.0159	0.0078	0.57
6	0.0113	0.0109	0.16
7	0.0075	0.0114	0.16
8	0.0058	0.0123	0.40
9	0.0192	0.0155	0.10
10	0.0101	0.0162	0.13

Table 3.3: Kernel distances (equation 2.42) of FPS-selected conformers to the representative conformer of cluster **7** before, denoted by D^{REMD}, and after *ab inito* geometry optimization, denoted by D^{DFT}. The relative *ab initio* energy difference between the representative conformer and FPS selected conformers is given in the last column.

similar backbone conformation. The σ-NH-scissoring mode at 1500 cm^{-1} [213], which is sensitive to the angle between the donor and acceptor of the H-bond, appears in all FPS-selected conformer spectra, but in different shapes and intensities, indicating, however, some backbone flexibility. Controversially, the IR peaks in the amide A/B regions of FPS-selected conformers are highly variable and probably strongly influenced by dynamic H-bonding, which is observed even in the same cluster. This becomes clear when looking at the main peaks of the N−H and C−H stretching modes between 3000 and 3300 cm^{-1}, which are more or less visible in all IR spectra of the FPS-selected conformers, but vary greatly in intensity and shape.

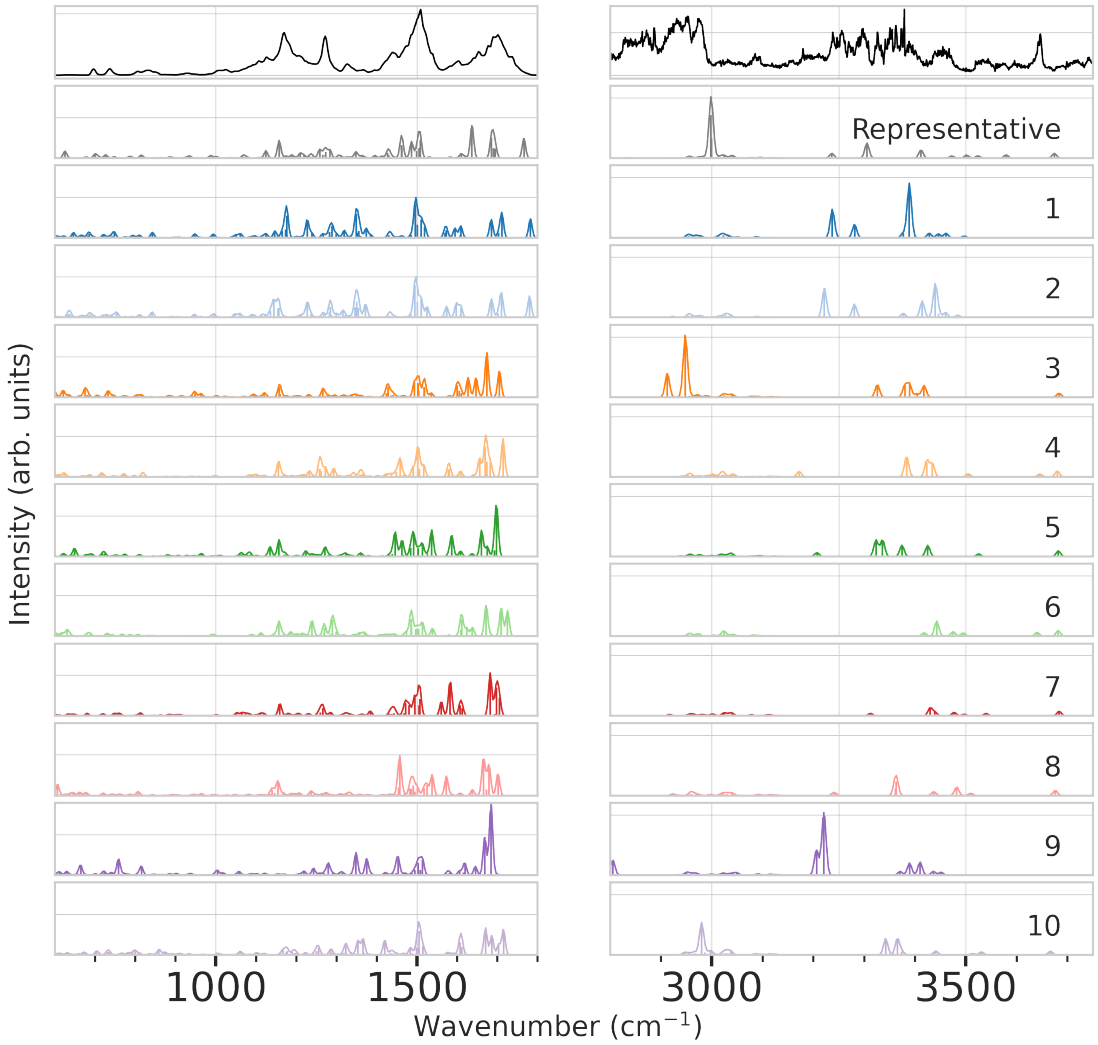

Figure 3.10: IR spectra of the 10 FPS-selected conformers from cluster **7**.

In addition, FPS was used to select 10 conformers from each of the clusters. For each conformer, the *ab initio* IR spectrum was calculated, and the resulting (unweighted) average spectrum is shown

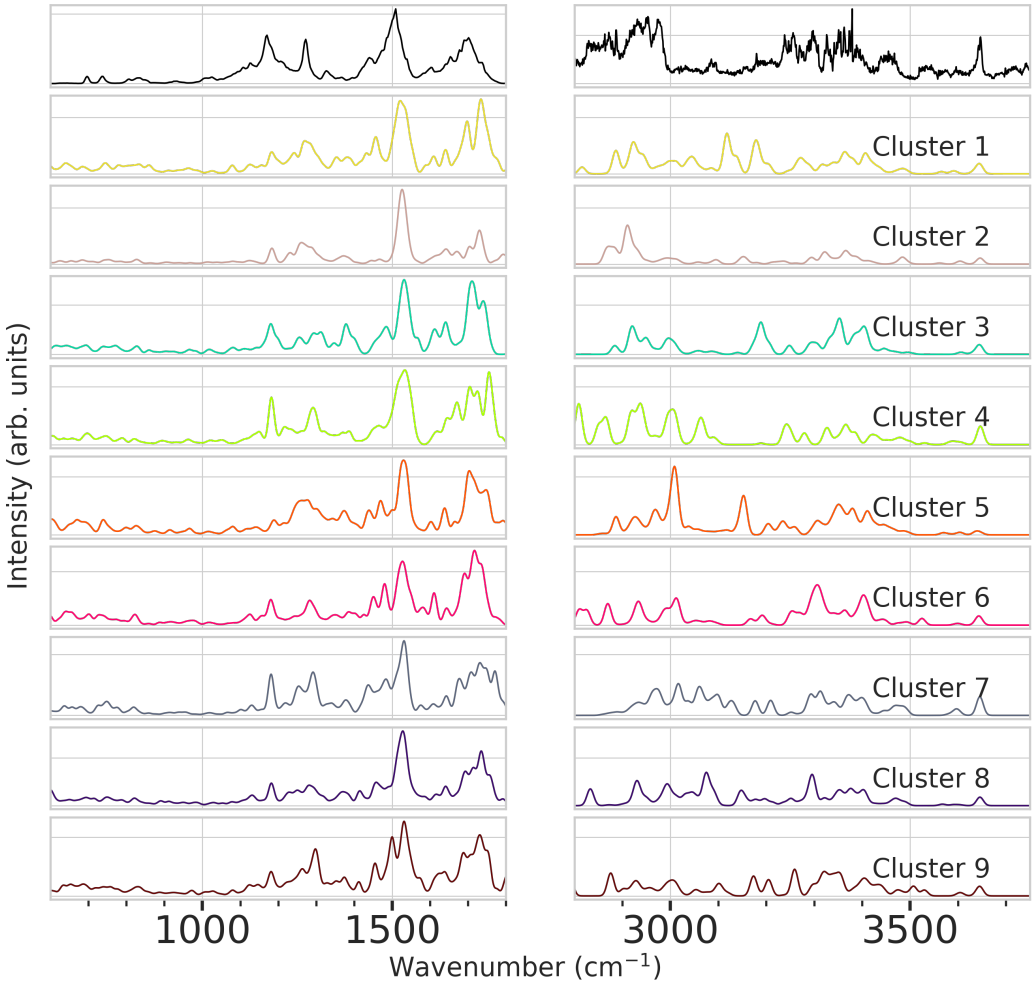

Figure 3.11: Averaged IR spectra for each cluster calculated from 10 conformers selected via FPS from all REMD conformers of each cluster.

in Figure 3.11. It should be noted that the geometry optimization of the conformers performs a kind of implicit, and physical, weighting, as the 10 conformers selected with FPS are shifted to their nearest local minimum. The bands in the low-frequency region are relatively consistent with each other, although differences are observed in the low-frequency region of the amide bands, suggesting different backbone configurations between the clusters – as a comparison with the conformers shown in Figure 3.3a would suggest. Changes in amide I/II regions (1200 - 1750 cm^{-1}) in the averaged IR spectra of Figure 3.11 could serve as a marker for conformational changes when comparing the signals to different conformers shown in Figure 3.3a. In addition, the variations in the high-frequency region are quite large, which can be attributed to the larger hydrogen bonding dynamics of the peptide motifs in this region.

Table 3.4 summarizes the R_P values of IR spectra for all FPS-selected conformers from cluster **7** and all other representative conformers.

FPS conformer from **7**	$\langle R_P \rangle$ to all other representatives (low/high)	R_P to representative of **7** (low/high)
1	0.73 ± 0.02 / 0.92 ± 0.03	0.66 / 0.85
2	0.75 ± 0.01 / 0.94 ± 0.01	0.70 / 0.87
3	0.71 ± 0.03 / 0.95 ± 0.03	0.65 / 0.81
4	0.74 ± 0.05 / 0.94 ± 0.02	0.68 / 0.89
5	0.78 ± 0.03 / 0.96 ± 0.04	0.72 / 0.90
6	0.80 ± 0.04 / 0.98 ± 0.02	0.75 / 0.95
7	0.78 ± 0.04 / 0.97 ± 0.01	0.74 / 0.95
8	0.73 ± 0.03 / 0.95 ± 0.01	0.69 / 0.94
9	0.70 ± 0.05 / 0.91 ± 0.02	0.64 / 0.89
10	0.74 ± 0.03 / 0.93 ± 0.02	0.67 / 0.90

Table 3.4: Pendry reliability factors, R_P, comparison for the low and high wavenumber (low/high) IR signal of FPS-selected conformers of cluster **7** with IR signals of representative conformers of all other clusters. $\langle R_P \rangle$ values denote average values along with their standard deviation in the absence of the representative IR signal of cluster **7**. The "R_P to **7**" column denotes the R_P of the FPS-selected conformers to the representative IR signal of the cluster **7**. Note that smaller R_P values indicate a better match.

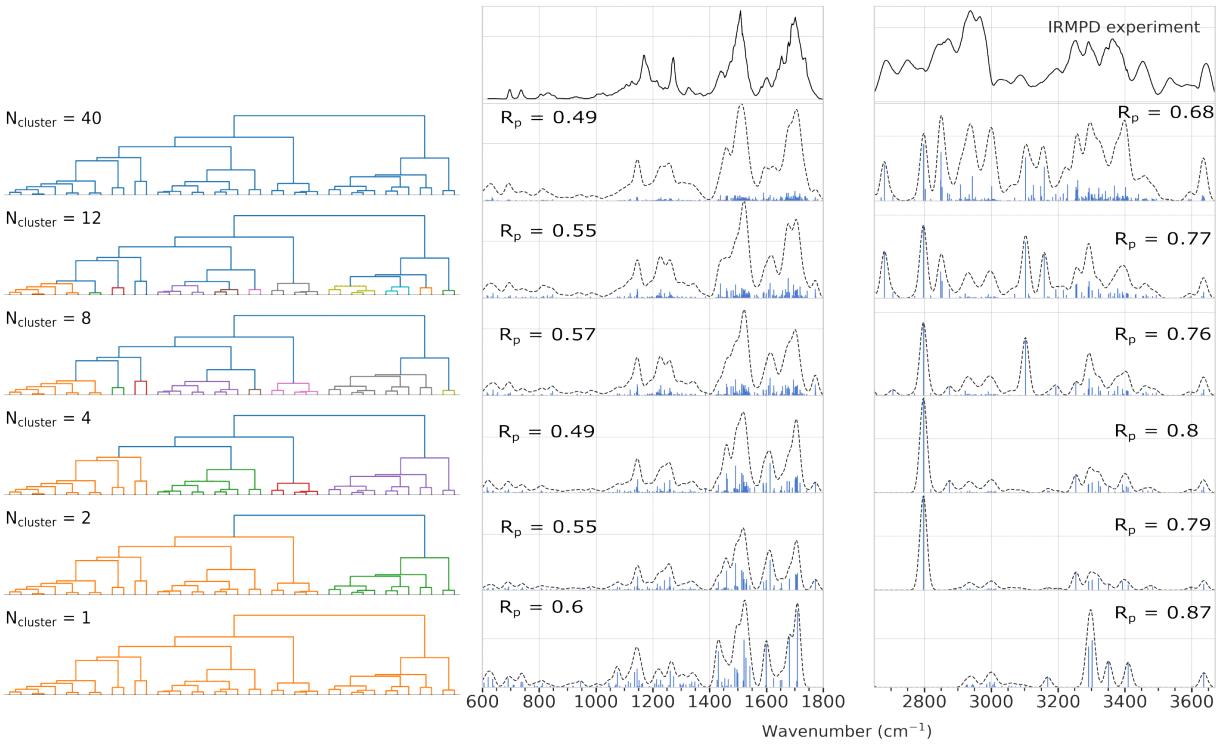

Figure 3.12: Average IR spectra for different number of merged PAMM clusters compared to the experimental IRMPD spectra of LeuEnk. Spectra are ordered by the number of PAMM clusters used and merged hierarchically as indicated by the dendrograms in the left panel. The top panel shows the spectrum considering all conformer clusters, while the bottom shows the spectrum of a single merged conformer cluster. Reliability factors R_P are reported individually for both amide A/B (middle) and amide I/II/III/V (right) regions.

3.4 Hierarchical Infrared Spectroscopy Prediction

Hierarchical clustering was used as illustrated by the dendrograms in Figure 3.12 to mix similar conformer clusters based on their adjacency and Ward's linkage criterion [51]. The conformers with the lowest energy of the resulting macroclusters were again selected as representative conformers for each cluster, and *ab initio* IR spectra were calculated for them as before. Since PAMM searches for peaks in the probability distribution using a Gaussian smearing (with a user-selected width to obtain a localized version of the Silverman rule), a narrower "probe" distribution is used to increase the number of peaks identified (i.e., $f_\mathrm{points} = 0.008$ and, additionally, a quick-shift cutoff scaling of $\alpha = 0.9$). Again, the lower wavenumber range between $600 - 1800\,\mathrm{cm}^{-1}$ shows good agreement according to the R_P factors. Upon closer inspection, the peak positions of the amide I and II regions agree better with the IRMPD spectrum when more conformer contributions are included. It should be noted, however, that the Pendry reliability factors here serve more as a quantitative comparison to show improvements and differences between the various obtained spectra, and their

absolute values should not be overrated, as they are highly dependent on smoothing. The relatively broad band at $1630\,\mathrm{cm}^{-1}$, corresponding to the umbrella vibrations of the $NH_3{}^+$ group, shows improved similarity with the experimental signal in terms of intensity pattern and broadening when more conformers are included. Inclusion of more conformers also improves the amide II region and leads to a more similar shoulder as observed in the experiment at wavenumbers below $1500\,\mathrm{cm}^{-1}$. This band could be attributed to a σ-NH scissoring mode, sensitive to the orientation of the H-bonds in which its found[219].

In the amide A region, the phenol OH stretching mode remains unchanged, further demonstrating the insensitivity of this band to conformational changes. On the other hand, the amide A/B regions, which include localized CH and NH stretching vibrations, improve dramatically as more conformations are included in the averaged IR spectrum. High-frequency CH and NH vibrational modes of molecules are often coupled with degenerate overtone and combination bands, as the vibrational configuration interaction causes intermixing of the high-frequency stretching states.

To support the argument that a single representative conformer from each cluster represents the contribution of the whole cluster to the complete spectra, a comparison of the IR signal of 10 randomly selected conformers within cluster **7** is given in Figure 3.10. After geometry optimization, these conformers are generally pushed to some local minima within the cluster, which could be identified by finer probing of the conformational phase space, i.e., by tuning f_{points}, with PAMM. Thus, differences are expected in the calculated IR spectra from the individual conformers within each cluster (cf. Figure 3.10). Although the low-frequency region does not seem to be affected as much as the high-frequency region, as shown by the averaging of the spectra in Figure 3.11, suggesting greater structural similarity associated with this wavenumber range. Most importantly is that the representative structure is the energetically preferred structure even after geometry optimization (cf. Table 3.3). Signals from other conformers in this cluster will thus most likely not contribute to the same extent as the representative conformer. Related to this, increasing the number of clusters to a still reasonable number seems to improve the IR prediction as more local minima and their associated IR signals are considered. Keeping this in mind, it is in general quite difficult to assign specific conformers to IR signals, especially when multiple conformers with different H-bonding patterns are present in the experiment but not in the conformational ensemble (cf. high frequency part of the IR spectrum in Figure 3.10). However, further breakdown of the cluster facilitates the interpretation of this region in terms of intensity patterns and peak positions. Nevertheless, one must carefully consider when to stop adding conformations, as this may result in adding conformations that are incorrectly weighted if too large a number of clusters are identified and clusters are incorrectly populated due to missing samples.

3.5 Conclusion and discussion

At finite temperatures, multiple peptide conformers separated by low-energy barriers and affected by different intramolecular H-bonding are present in a macroscopic ensemble and hence contribute to the experimental observations [225, 40, 226, 199]. In this chapter, the importance of conformational ensembles for a robust IR prediction was investigated which resulted in a good qualitative agreement with the experimental IRMPD spectrum for protonated LeuEnk in the gas phase. In particular, unsupervised machine learning with replica-exchange molecular dynamics (REMD) simulations and *ab initio* calculations was combined to incorporate a realistic representation of a canonical ensemble in the spectra prediction. An important part of the study of conformations generated by free-energy methods such as REMD is the partitioning of the conformational space using an agnostic analysis approach to avoid potential biases that could lead to contrived results and thus misleading conclusions [128, 227]. As a possible solution to this, it was shown how a PAMM clustering analysis applied to a calculated collective variable space of SOAP kernels is able to categorize the conformational ensemble into different recurrent molecular motifs based on a kernel density estimation. Furthermore, it is shown that investigating the H-bonding dynamics, as obtained by a PAMM analysis, shed further light on specific H-bonding influences on the secondary structure motifs of LeuEnk.

In gas-phase experiments with biomolecules, the intact H-bonding network and the enhanced electrostatic interactions can render native-like conformations that are stable in solution metastable, leading to rather unusual transitions pathways between conformations [228, 229]. Thus, the experimental preparation is crucial, as it can affect the final outcome of the spectroscopic experiment, e.g., through kinetic trapping [76]. To address this better, previous IRMPD spectroscopy experiments with LeuEnk were repeated, with particular attention to the preparations. In contrast to previous experimental and theoretical studies of LeuEnk [27, 73], the current preparation leads to a very good agreement with our predicted IRMPD spectrum, which is probably related to a better agreement between the two ensembles in the theoretical and experimental approach. As it was shown, considering only a few lowest energy conformers and neglecting the relative importance of the conformers, as it is often common practice in simulation approaches, leads to an inadequate description of the conformational space explored by LeuEnk in gas-phase experiments and, consequently, to a misleading interpretation of IR spectroscopy. Averaging the calculated IR spectra of representative motifs, based on the weighting factor obtained from the relative population of a PAMM cluster, can be used as a reasonable approach for representing the general structural canonical ensemble. Specifically, the comparison between the averaged IR and the IRMPD spectra illustrates that the averaged spectrum reproduces well the intensity pattern and the main peak positions, capturing the main features of the vibrational spectra although some

particular features may be missing due to the specific way a representative conformer is chosen here. Especially in the high wavenumber region that is strongly influenced by H-bonding, it was concluded that the existence of multiple low-energy conformers should not be neglected. For a further evaluation of the importance of the conformational ensemble, a hierarchical clustering was performed based on the information of a PAMM analysis to evaluate the effect of breaking down "macroclusters" on the averaged IR spectra. The improvements observed in the Pendry reliability factor R_{P} for amide A/B and amide I/II/III regions indicate that the importance of unravelling non-Gaussian features in the potential energy surface to explain the IR fingerprints are strongly correlated to changes in the backbone conformation and H-bonding patterns. However, the stabilizing role of the $NH_3{}^+$ group in H-bonding networks defines a cut-off in the number of conformations contributing to the experimental IR spectra due to the appearing discrepancies in peak positions and intensity patterns in amide A/B regions.

4

Integrating Explainability into Graph Neural Network Models for Prediction of X-ray Absorption Spectroscopy

"This chapter reproduces the contents of reference [12], with minor adjustments. The author contribution is the development and design of the machine learning model, data engineering/analysis, and training of the models. Moreover, the author has performed the simulations and data processing for creating the datasets. The code used to train the models and generate the figures in this chapter is publicly available at https://github.com/AI-4-XAS/XASNet-XAI. The QM9-XAS is available at https://dx.doi.org/10.5281/zenodo.8276902."

4. Integrating Explainability into Graph Neural Network Models for Prediction of X-ray Absorption Spectroscopy

X-ray absorption spectroscopy (XAS) is an important characterization technique in chemical analysis to unveil the atomic structure of matter, having a broad range of applications in material science [230], biomedical research [231] and identification of metals and solids [232]. XAS is particularly useful in the investigation of the electronic and geometric structure of bio-molecules, nanoparticles, and metal complexes [9, 233, 234]. However, understanding XAS spectra obtained experimentally is entangled with complexity, which arises from the complex electronic structure interplay with the adsorption of X-ray photons by atoms and is influenced by factors such as the chemical environment of the atom, the presence of solvents, and the energy of the incident X-rays [235]. Therefore, sophisticated – but computationally also expensive – theoretical methods from *ab-initio* quantum chemistry can accurately predict XAS and are often a necessary complement to interpret experimental results [236].

Several studies have focused on X-ray spectroscopy using ML methods with the additional aim to improve the understanding of the contribution of different atomic environments to the peaks occurring in the spectra [64, 65, 66]. Accurate prediction of XAS spectra has been accomplished by employing some of the more sophisticated ML models such as graph neural networks (GNN) and deep neural networks (DNN) [237, 61, 238]. However, a large number of layers in the underlying neural network, as well as a high parameter count implies such models are black-box [167], which means it is only possible to reason about the input and output of ML models, but not to understand the rationale behind predictions. While these models are extremely efficient at predicting spectra, it is not straightforward to understand which correlations in the feature data contribute to specific predictions. On the other hand, ML models designed to predict XAS spectra must provide clear peak assignments, as this option for interpretation is typically required in spectroscopy experiments and often necessitates theoretical calculations. The comprehensibility of why ML models can achieve this peak assignment capability must be transparent to users to ensure trust in the predictions, given the diverse range of applications of XAS in material and biochemical sciences [239, 235, 240]. It is therefore imperative to develop an understanding of the XAS predictions made by complex ML models and ascertain whether the predictions align with human logic and decision-making as incorporated in the quantum-mechanical equations. This can be achieved using explainable artificial intelligence (XAI) methods, which provide a window into the ML model's decision-making process and correlations uncovered by the model through data analysis [241].

To incorporate explainability in GNN model, we use attribution method [170]. We implement class activation map (CAM) as an attribution method to highlight the importance of particular regions in the graph with respect to the target property. Validating explanations for chemical property prediction is challenging since a property is often the result of a complex interplay between the geometric and electronic structure of the atoms in a molecule. This gives rise to intricate structure-property connections within molecules, especially complex properties such as X-ray

absorption spectra, which only find interpretation by the examination of each individual peak detected through a combination of experiments and simulations [12]. Therefore, the validation of explanations generated using attributions also requires the creation of a robust 'ground-truth' benchmark using such domain-specific knowledge, which is often a challenging task in molecular-property prediction.

In this chapter, a framework that uses a combination of graph attributions and ground-truth data generated from linear-response time-dependent density functional theory (TDDFT) [114] is introduced, to provide explainability on GNN models trained to predict carbon K-edge XAS spectra of organic molecules. Carbon atoms play a central role in the structure and function of numerous molecules, making them an ideal choice for training the GNN models in this case. Therefore, carbon K-edge XAS offers a unique perspective, providing valuable insights into the structure, function, and reactivity of these molecules [242, 243]. Additionally, among the XAS calculations, K-edge spectroscopy on a main group element is less complicated than, e.g., the spectroscopy on the transition metal L edge, and can be computed via TDDFT on a time scale that allows the creation of a large dataset. To train the different GNN architectures, an in-house QM9-XAS dataset, based on a subset of the QM9 dataset of small organic molecules, [244] was set up. The performance of the trained models was compared in predicting the XAS spectra on the test dataset. In order to evaluate the explainability of GNN models, the ability of these models to identify the contribution of atoms and their surrounding environment toward the distinct peaks in the XAS spectrum was analyzed. For creating the 'chemical' ground truth pertaining to XAS, a data pipeline was created which inputs the output of TDDFT calculations and renders the labels to atoms, indicating whether or not an atom contributes to a specific excited state in XAS. These ground-truth values are then finally quantitatively compared with the attribution scores obtained from GNNs. By integrating domain expertise with interpretability, the approach here establishes statistical and instance-based correlations between explainability and model performance. Applying this method to different GNN models, specific GNN architectures, which incorporate both global and local information on atoms, offered superior explanations for the peaks observed in carbon K-edge XAS spectra. Additionally, the robustness of the GNN models, by randomly perturbing molecules in the test dataset, was investigated to rationalize the difference in the explainability power of various used GNN architectures.

4.1 The QM9-XAS dataset

While X-ray absorption spectroscopy is a popular technique in chemistry, to the best of our knowledge there is no XAS data set for organic molecules that is large enough and available for training ML models. Therefore, the QM9 dataset [244] containing 132,531 organic molecules composed of the first- and second row of main group elements H, C, N, O, and F, was used. A random subset of the QM9 dataset, containing 56,000 molecules, was chosen, termed as QM9-XAS for the purpose of the dataset in this study. Carbon K-edge XAS spectra was calculated with the linear response time-dependent density functional theory (TDDFT) [245] method (explained in section 2.2.3.2 of chapter 2), which is in general a useful complement to experiments and allows for the interpretation of spectral peaks. More specifically the ORCA electronic structure package [246] was used to calculate TDDFT at the B3LYP/TZVP [109, 247] level of theory. All calculated XAS spectra were obtained in the energy range $E_{min} = 270$ eV and $E_{max} = 300$ eV and peaks broadened using Gaussians of widths 0.8 eV. The resulting curves were discretized into $N_{grid} = 100$ points. This step ensures that the length of the target output to be learned for ML applications is consistent across all spectra. Further processing is then performed to generate tuples of molecular graphs and their spectra to convert them into a format optimal for training GNN models. Molecular graphs were generated from the SMILES strings of the molecules, which were available in the original QM9 dataset using the RDKit [248] python library. Since the models developed here are implemented using the Pytorch Geometric [249] library, the graph and spectrum tuples were converted into the native dataset class of this library.

4.2 Molecular graph data

As explained in section 2.4.2.1, molecular structures can be represented as graphs by considering atoms and bonds, more abstract, the relationships between atoms as nodes and edges, respectively. In this work, atom types exists in the QM9-XAS dataset as well as the number of hydrogens attached to the atoms were presented as node feature vectors. The bond lengths between two atoms and bond multiplicity were used to represent the edge (bond) feature vectors. One-hot encoding was used to convert most of the node and edge features, including atom types and other categorical attributes, into numeric vectors. All encodings used in this chapter are summarized in Table 4.1. In this work, the GNN models were trained based on three different architectures, namely GCN, [250], GraphNet, [165], and the multi-head graph attention network (GATv2) [251]. In GAT, the attention weights for each node are fixed throughout the entire training process. On the other hand, GATv2 addresses this limitation by introducing dynamic attention weights. Using GATv2 layers makes the model more expressive and able to better capture the complex

Table 4.1: Features of node and edges (atoms and bonds) as represented in the encoded vector in conjunction with their respective type of encoding.

Node feature	**Encoding**
Atom type	One hot
Hybridisation	One hot
Aromaticity	One hot
Number of H atoms	Integer
Edge feature	**Encoding**
Bond distance	Real
Bond type	One hot

relationships between nodes. Training a multi-head GATv2 converges faster at a moderately higher computational cost, while also increasing the robustness of the final model since it is in principle trained on multiple attention instances in parallel. The details of these GNN algorithms are discussed in chapter 2.

4.3 Training

In order to assess various trained models, the QM9-XAS dataset was shuffled and divided into a training set of 50k samples and a test set of 6k samples, with the training data further partitioned into an 80:20 ratio for training and validation. The GNNs and all fully connected layers were trained for 1000 epochs with a learning rate of 1×10^{-3}, and a batch size of 100 samples. A learning rate scheduler was implemented to reduce the learning rate by a factor of 0.8 every 100 epochs. For all the models, three GNN hidden layers with sizes of 128, 256, and 512 were used for node updates, and a fully connected layer as the output layer for predictions. The AdamW optimizer [252] and the root mean squared error (RMSE) as the loss function to train the models were implemented. In order to keep track of over-fitting, the RMSE loss on the validation set after every 50 epochs was monitored. All models were trained on a single NVIDIA Tesla A100 64GB GPU. The model which has the best RMSE loss and relative spectral error (RSE) [71] was selected on the validation data set. RSE is obtained by dividing the RMSE among the target y^{tar} and the predicted y^{pred} intensities of the signal at energy E, by the total spectral energy of the target. In the discretized spectrum in steps of $\Delta E = (E_{\text{max}} - E_{\text{min}})/N_{\text{grid}}$, the RSE is approximated as

$$\text{RSE} = \frac{\sqrt{\sum_i^N (y_i^{\text{tar}} - y_i^{\text{pred}})^2 \cdot \Delta E}}{\sum_i^N y_i^{\text{tar}} \cdot \Delta E} \tag{4.1}$$

A small relative spectral error indicates that the predicted spectrum is a good prediction of the original spectrum. The quality of XAS spectra predictions made by different GNN architectures was compared by calculating the average RSE on the test dataset.

4.4 Graph attribution

Attributions or feature attributions are one of the most popular techniques used to explain a model's predictions [253]. The attribution method assigns scores to each input feature that reflects the contribution of that feature to an ML model's prediction, thereby explaining the role played by that feature in the prediction [254, 255, 173]. In the case of GNNs, attribution methods assign attribution scores to graph nodes and edges based on their contributions to the final prediction of the model. As explained in section 2.5 of chapter 2, we used CAM attributions to integrate explainability into our GNN models.

4.5 Ground truth evaluation

In addition to evaluating attributions, it is crucial to establish a ground truth logic that enables the assessment of attribution quality. Hence, the agreement between CAM weights of the model's prediction and ground truth logic should be quantified. To this end, a definition for a numerically measurable ground truth for the excitations underlying the spectra is needed. In other instances of XAI in chemistry, a suitable ground truth was developed by directly considering the molecular fragments or functional moieties that experts knew to be important for decision making [256], such as binding mechanism learned by DNNs [173]. Nevertheless, when it comes to predicting XAS, comparing attribution scores to ground truth becomes more complex since it necessitates careful examination of all atoms in the molecule and a comprehensive understanding of the quantum mechanics behind X-ray excitations. Furthermore, delocalized molecular orbitals present yet another challenge for understanding the precise contribution of atoms to virtual orbitals in excitation states of XAS [257]. Therefore, in this work, a method is developed, which assigns the ground-truth contributions of various atoms in a molecule to a peak in the TDDFT spectrum. It uses a combination of orbital populations of all the initial and final states underlying the respective X-ray excitations and their oscillator strengths to obtain the contribution of each atom to a specific peak in the XAS spectrum. To derive atomic contributions in the ground truth, the core excitations were computed within the energy range of the peak in XAS spectra and then the atoms contributing to both core and virtual orbitals of a certain excitation state were determined. The atom contributions were weighted according to the oscillator strength of the corresponding excited state as well as the atom population per molecular orbital. As explained in section 2.2.3.2

of chapter 2, in XAS, the oscillator strength of the corresponding excited state is a measure of the probability of a core electron being excited to that state by an X-ray photon. In cases where the calculated weights in the ground truth necessitate (i.e. being above the average weights as a threshold) the presence of particular atoms in a peak of the XAS spectrum, those atoms were labeled as 1 and all other atoms as 0. Figure 4.1 depicts the process of obtaining the ground-truth

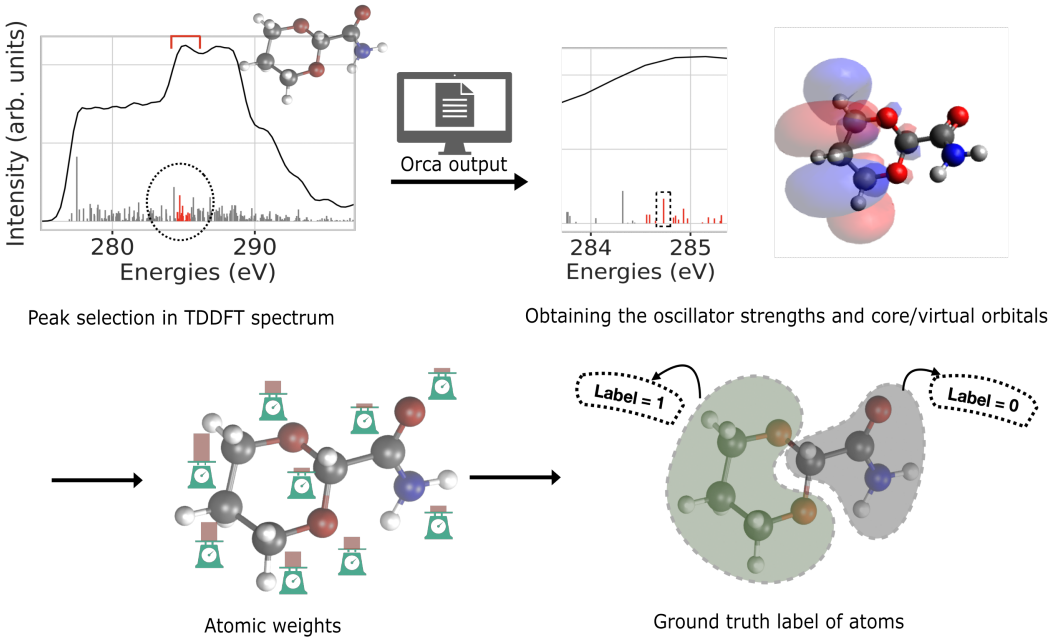

Figure 4.1: Ground truth evaluation based on TDDFT data. The process of evaluating the TDDFT data starts with selecting a specific peak in the XAS spectrum. The oscillator strength and orbital contributions for each excitation state in the peak are used to determine the final atomic contributions to the peak. Atoms in the molecule are then labeled based on the calculated weights, i.e. 1 for atoms contributing to the peak and 0 otherwise.

labels for atoms. Given the fact that the core level transitions obtained from TDDFT are discrete lines and that the ML spectra are distributed on a grid and have wide peaks, for the comparison it is necessary to unify all CAM scores of a given peak to a given line from TDDFT spectrum. Hence, the CAM scores were summed up of all atoms in the molecule for all energy points in a range equivalent to the full width of half maximum of a peak.

4.6 Model explainability

Explaining a model's predictions involves comparing the ground truth to the attributions obtained from the model using an XAI method. To measure to what extent the ML models learn the correct atomic contributions to the XAS spectra, the area under the curve (AUC) of the receiver operating characteristic (ROC) was used [258, 173]. The ROC itself is a curve formed by plotting the rate of true positive outcomes and that of false positive ones at various classification thresholds, that divide the assignments between the true and false classes. A true positive outcome occurs when a model tasked with distinguishing two or more classes, correctly predicts the class to which an instance belongs. In this case, the CAM weight assigned to an atom at a certain peak matches the ground truth of the atom belonging to an orbital. Similarly, a false positive occurs when the class under investigation is incorrectly predicted by the model, i.e. when atom contribution in ground truth and CAM disagree. The AUC thus quantifies the performance of a classification model into a single value between 0 and 1, where an area of 0.5 means a model works only as good as a random classifier. A value of 1.0 means that the model has the ability to perfectly discriminate between different classes. In this particular case, the AUC is indicative of whether the model can correctly identify if an atom contributes to a peak in the spectrum or not.

Figure 4.2 illustrates the workflow to make a GNN prediction of a spectrum, determine the CAM attribution, and compare it in the last step to the ground truth, i.e. the contribution of atoms to core and virtual orbitals obtained from TDDFT, here shown for a prediction made by the multi-head GATv2 model. More explicitly, a model with a large AUC close to 1.0 would perfectly assign the labels 0 and 1 to each atom in the spectrum for all the molecules in the test set. Moreover, we identify the baseline of AUC as 0.5 which is basically a model classifier that randomly assigns these labels to the atoms in a molecule.

The attribution AUC values were computed at each peak in a TDDFT spectrum and were averaged over all the peaks to arrive at a final score that explains the degree of agreement between ground truth logic and CAM attribution scores. The AUC is determined for the different model architectures in section 4.8. To demonstrate that the explainability method is stable, we perturb a randomly chosen set of molecules from the test dataset and evaluate the change in attribution AUC in section 4.9.

4.7 Model performance

To first visualize the predictions made by these GNN models, the best, average, and worst prediction of the XAS spectrum is demonstrated for each model based on RSE values in Figure 4.3 (b). While the best prediction across all models is a near-perfect replica of the TDDFT spectrum, the average,

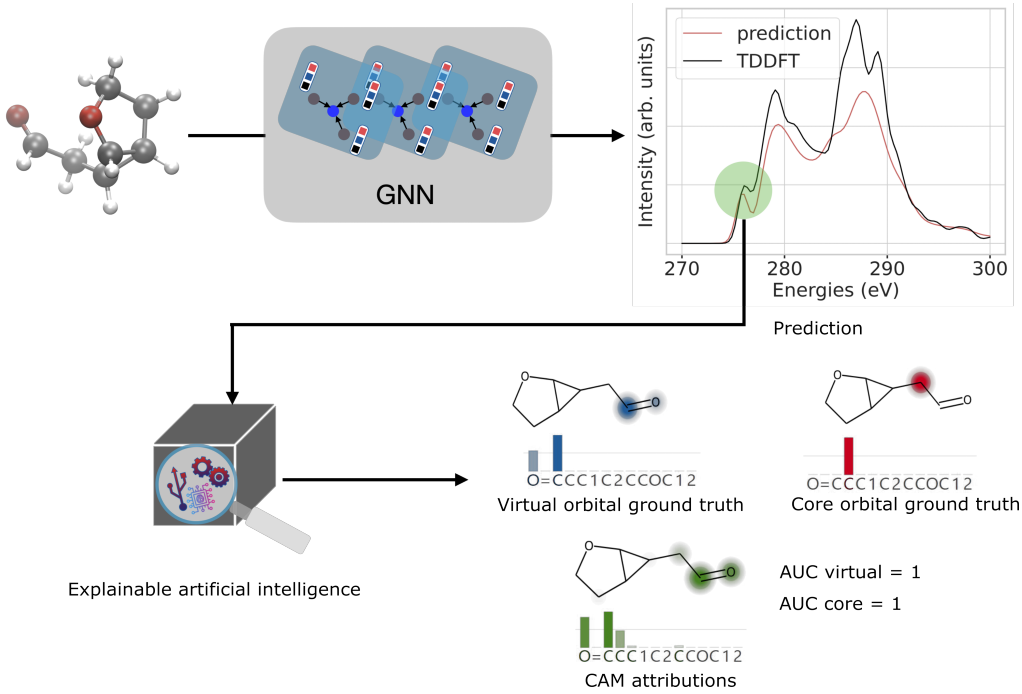

Figure 4.2: Workflow of the ML and explainability of XAS spectrum. This process consists of converting a molecule to a molecular graph, training a GNN, comparison of the ML predicted and TDDFT spectra for obtaining the RSE, and finally applying the XAI technique to obtain here the CAM weights (green). In this example, the CAM weights are compared to ground truth attributions for core (red) and virtual (blue) orbitals at the highlighted 277 eV peak of the spectrum, using a heatmap [259] on the molecular structure. These ground truth labels are then compared to CAM weights, giving the AUC values for core and virtual contributions.

and worse ones predict general features of the spectrum correctly, but miss out on the finer peak structure or incorrectly predict peak intensities. In Figure 4.3 (a) all RSE for one model are plotted in a histogram and the average RSE is determined. The GATv2 model has a slightly lower average RSE value of 0.031 compared to 0.042 for GraphNet and 0.047 for GCN. The distributions look similar. They have their onset with a small slope at RSE = 0.0 and then quickly grow to their maximum around the average RSE. The decline is slow following the shape of a skewed distribution with a long tail leading to a low number of structures with RSE values above 0.1. Such structures are fewer for models with the GATv2 and the GraphNet GNN architectures, demonstrating their superiority for XAS predictions compared to GCN.

The above results are consistent with the findings of earlier research, which suggest that integrating an attention mechanism [162] and applying combinatorial generalization [165], i.e. enabling the network to reason about the global structure of a graph, while learning the graph representation, as done in the GATv2 and GraphNet models, help enhance the learning of target properties related to both local and global structures of the graph [260, 261]. In the case of the GATv2 model, computing the importance of the neighboring atoms for a target atom in a molecule using the weighted attention mechanism assigns relevance to a local region of the molecule to a specific excitation energy in the spectrum, which differs from traditional GCN layers with fixed weights for connections between atoms. On the other hand, by incorporating relationships and interactions among nodes, edges, and global graph attributes, GraphNet significantly improves the acquisition of structure-properties relationships in XAS spectra [165].

4.8 Explainability of XAS predictions

While comparing the prediction performance of different ML models is crucial, the similarity observed in the RSE distributions in the previous section motivates the exploration of the interpretability of these models. Figure 4.4 illustrates the peak assignment via core and virtual orbitals from the TDDFT calculation as red and blue spheres on participating atoms and via the CAM scores given as green spheres. The AUC values for the respective orbitals quantify this assignment. We compare an accurate GATv2 prediction at about 288 eV, in which the intensities of both curves lie on top of each other, with one with a larger deviation from the TD-DFT data at about 292 eV. In both cases, the core orbitals are accurately matched by the CAM score giving AUC values of above 0.9, significant quality differences occur for the virtual orbitals. Those contribute the most to an XAS spectrum in general. Hence, a good prediction comes with a good assignment of the peak with a large AUC of 0.88. By contrast, the poorer XAS prediction with about 10 % peak intensity differences also leads to a much reduced AUC of 0.52 only. In this case, one can already visually see that the CAM is much more significantly spread over the entire

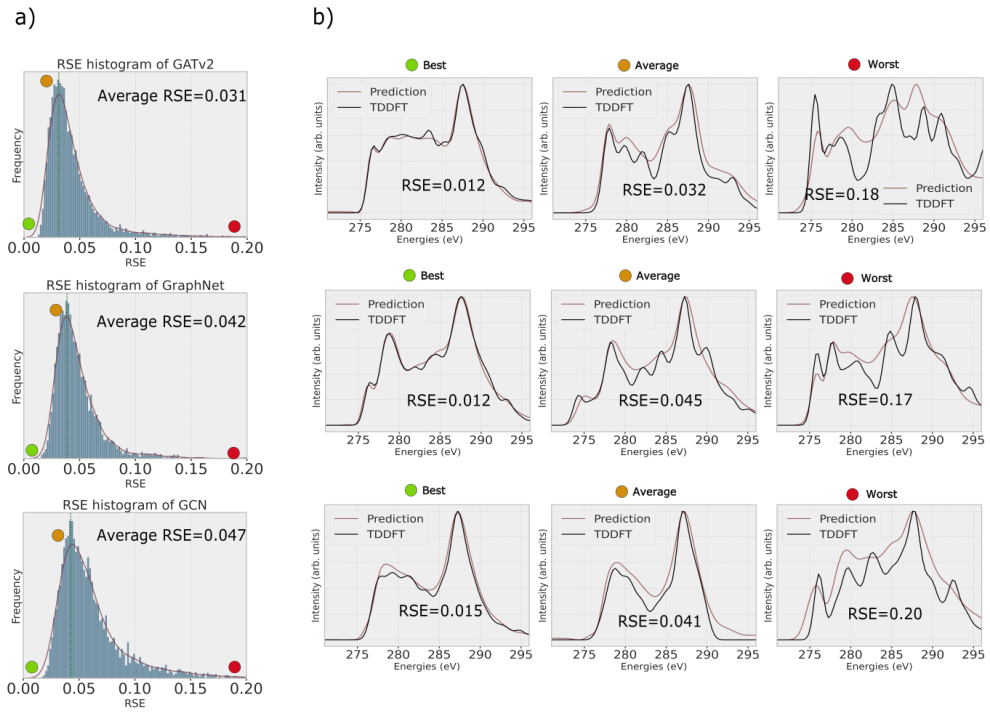

Figure 4.3: Evaluating the performance of various GNNs on the test dataset. RSE histogram for all GNN models (a). While average RSE performances are close, GATv2 has a more left-skewed histogram distribution, indicating better performance over large portions of the data. Best, worst, and average predictions of the three GNN models with their respective RSE values (b).

molecule, while the orbitals contributing are based only on two atoms of which one is not a part of the CAM at all.

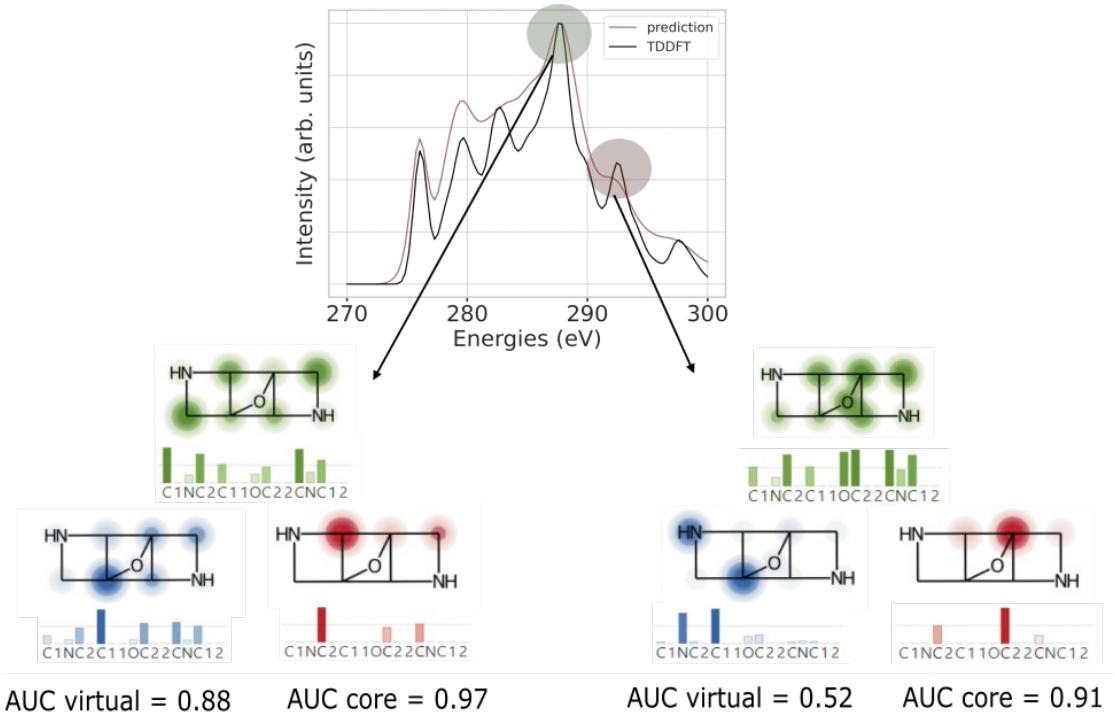

Figure 4.4: The attributions (green) are compared with the ground truth of core (red) and virtual (blue) orbitals via AUC values for two peaks of an XAS spectrum predicted by the GATv2 model. The model has higher AUC values when a peak in the predicted spectrum follows the TDDFT result.

Figure 4.5 gives a close-up visualization of the derivation of the CAM and the core and virtual orbital ground truth, by relating both to local excitations and the latter also to orbitals relevant to the respective excitation. This is done for the first three excitation states of the TDDFT calculation underlying the first signal of the broadened spectrum. Note that later signals are composed of a much larger number of transitions, making the visual comparison very cumbersome. We observe that the first two peaks originate from a transition of an electron on the cyano carbon atom to one of the π^* orbitals of the CN group. This is exactly reflected in the CAM weights obtained at exactly the transition energy. The CAM weights show a low contribution at other atoms, which is insignificant. The third peak belongs to the $s \rightarrow \pi^*$ transition on the amide group at the other end of the molecules, which is likewise highlighted by the local CAM. The total CAM overlays both transitions, and likewise does the ground truth of the contributing core (red) and virtual (blue) orbitals highlight the two C atoms or multiple bonds, respectively.

To further analyse the explainability of the best performing GNN model (i.e. GATv2), we performed TDDFT calculation of local atom XAS spectra of individual carbon atoms of a sample

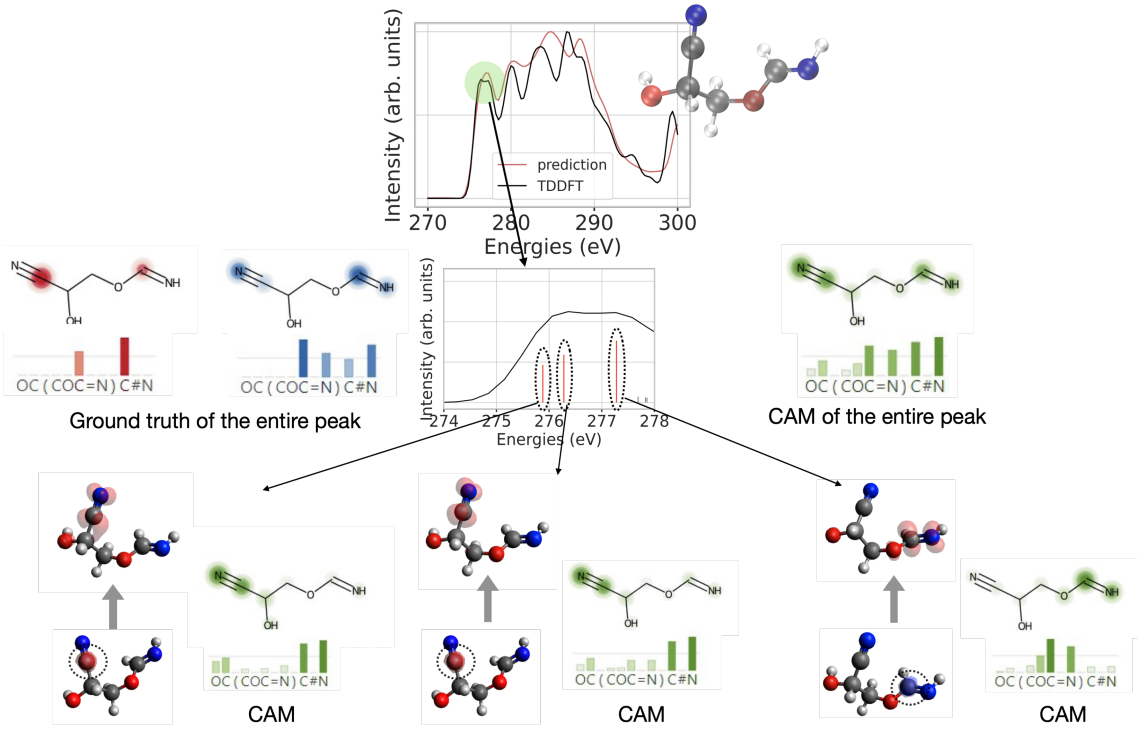

Figure 4.5: Exploring the correlation between CAM attributions of atoms and transition densities of a peak in XAS spectrum. CAM attributions (green) and transition densities of three excitation states are visualised for a sample molecule in the test dataset in the bottom part of the figure. The transition densities highlight the starting C core orbital, which is encircled for better visibility), in the bottom, and above the virtual orbital on the cyanide group for the two lower-energy peaks and the amide group for the third peak. The overlay of the three transition densities for the core (red) and the virtual (blue) states are shown on the left side of the close-up spectrum, while on the right side, the CAM of the entire peak is shown.

molecule in the test dataset with the CAM attribution weights assigned to these carbon atoms for which the comparison is displayed in Figure 4.6. The CAM attribution weights, which are X-ray photon energy dependent and hence appear as spectra in themselves, exhibit a reasonably accurate alignment with the main features of localized XAS spectra, although they do not entirely replicate all the peaks. In particular, CAM attribution weights of the carbon C1 (shown in Figure 4.6)

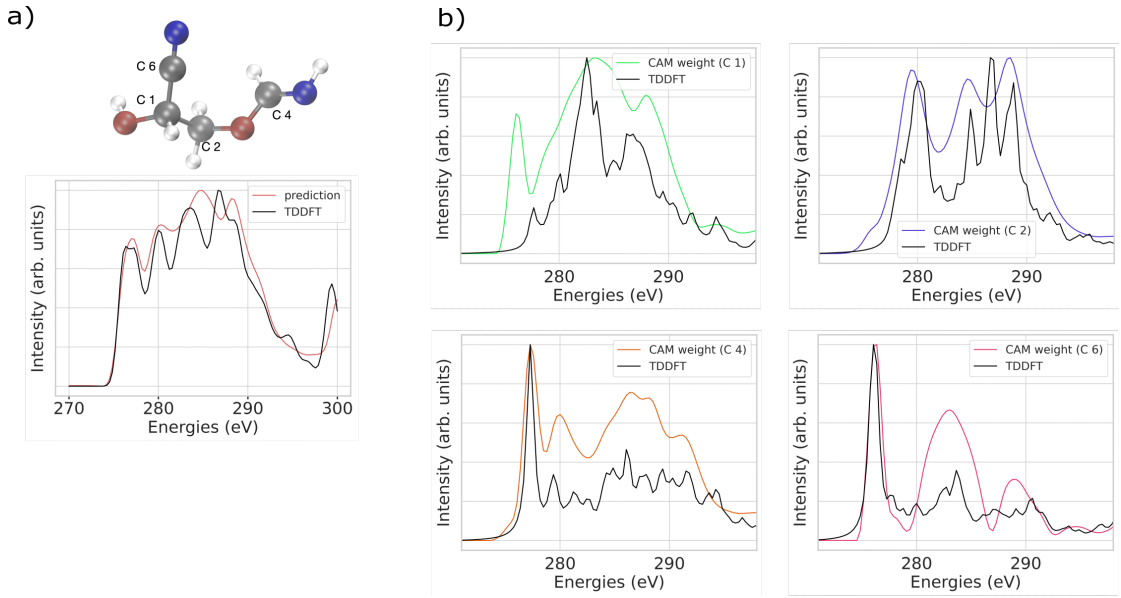

Figure 4.6: TDDFT(black) and GATv2(red) predicted C K edge XAS spectra for an entire sample molecule shown (a). Calculated local XAS spectra (black) and CAM attribution weights(multiple colours) of individual carbon atoms in the molecule (b).

next to the hydroxy group appear to show discrepancies, which can be due to the attribution technique or weaknesses in the model's explainability concerning this specific atom. Although training a GNN model using localized XAS spectra to predict the spectra of individual carbon atoms is achievable and could potentially enhance the alignment between TDDFT and ML in terms of spectral shape and CAM attribution, generating a dataset with atom-localized spectra through various methods require more computational resources. CAM attributions of atoms from a complete molecular spectrum can provide an opportunity for creating a dataset of localised spectra based on arbitrary XAS methods. Moreover, since the ultimate goal is to compare the predicted XAS to experimental spectra, training a model based on entire XAS spectra in certain energy ranges is more favorable.

With this rationalization, the next step is to evaluate the attribution quality over the entire dataset. Figure 4.7 shows box plots of the attribution AUC for core and virtual orbitals of the three GNNs evaluated over the full test dataset. As seen from the figure, the GCN model gives an average attribution AUC close to 0.5, which means the model barely outperforms a random classifier. This

combination of good spectra predictions on the test data, as shown in Figure 4.3, and low average attribution AUC value by the GCN model is in line with a previous study, suggesting that the combination of near-perfect model performance and low attribution AUC indicates that the model fails to learn the ground truth logic [173]. In contrast to this, the GNN models with multi-head GATv2 and GraphNet layers have a superior agreement with our developed ground truth logic, with median values greater than 0.7 for both virtual and core orbitals. As a general trend, it is shown that the spread of core AUC values is lower across all models, while the AUC values for virtual orbitals are more widely spread out, as indicated by the high variances in the figure. Nevertheless, it should be noted, that within the presented approach it is not possible to learn to distinguish between the more localized core orbitals and the more delocalized virtual orbitals, which could be useful information for the model to be included. Models that have higher attribution AUC values

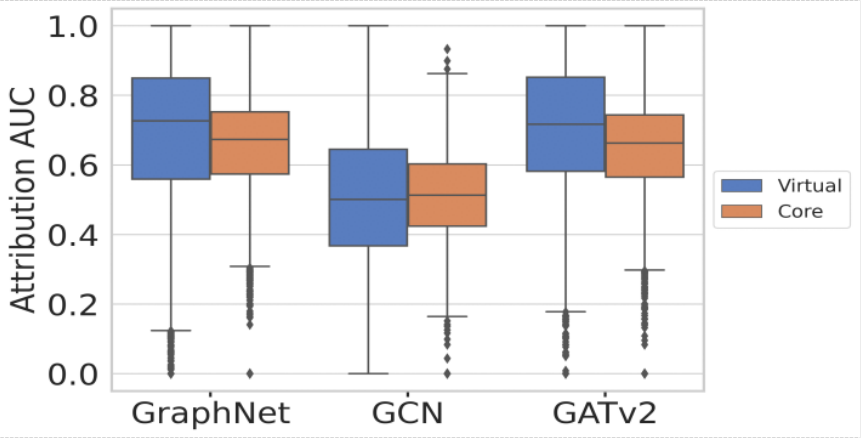

Figure 4.7: Attribution AUC score boxplots for the core and virtual orbitals of the three GNN models. The vertical line within the box indicates the median AUC value on the test data, while the length of the horizontal lines indicates the variance in AUC values for each model. Points beyond this range are considered outliers.

for core and virtual orbitals, i.e. GraphNet and GATv2, demonstrate a greater ability to learn the contribution of atoms to the excitation energies of the XAS spectrum. GraphNet models associate and encode global graph context in addition to the message-passing on node and edge level, and this perhaps positively influences CAM attributions giving them information beyond the local environment. For XAS analysis, the peaks in the spectra are strongly influenced by the local atomic arrangement and electronic configurations of the atoms [262, 263]. GraphNet models can effectively capture these complex relationships between atomic coordination and specific excitation states in the XAS spectrum by considering the interconnectedness of nodes and the overall structure of the molecular graph. It is expected that using multi-head GATv2 and GraphNet architectures as GNNs for learning XAS spectra aligns with the essential understanding of the delocalized nature of molecular orbitals, which is crucial for accurate XAS prediction. Vaswani, Shazeer, Parmar,

Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin [162] have shown previously that multi-head attention in multi-head GATv2 model can improve the performance of models by enabling them to attend to different parts of the input molecular graph simultaneously. Wiegreffe and Pinter [264] have additionally shown that models that use the attention mechanism can provide better interpretability compared to non-attention frameworks, since they allow the visualization of which parts of the input are being emphasized by each head, making it easier to understand how the model is making predictions.

Thus, when it comes to XAS analysis, it can be inferred that the attention framework, which dynamically assigns importance weights to nodes surrounding a target node, yields superior attribution values compared to the static node-weighting scheme employed by the GCN framework. Moreover, combinatorial generalization in GraphNets, which enhances their ability to generalize and perform well on new, unseen graph structures and tasks, is crucial for their applicability to XAS predictions in diverse molecular structures. On the other hand, robustness and generalization in GraphNet models, which incorporate relational inductive biases, have achieved improvement compared to traditional GNNs such as GCNs, over a range of graph classification and regression tasks [265, 266, 267]. Relational inductive biases in GraphNet models are assumptions about the relationships between entities in a graph that guide the learning process. These biases help GraphNet models to learn effective representations of graph-structured data and to generalize well to unseen data.

4.9 Robustness of the explainability

Having shown that CAM attributions allow the explanation of the individual peaks in predicted XAS, the next task is to determine how robust this explainability is with respect to the prediction accuracy itself and to the changes in the dataset. To address the influence of prediction quality on interpretability, the variation of attribution AUC scores across different RSE values for the three GNN models are explored. This is performed for each model by first subdividing the molecules of the test dataset into ten evenly-large groups based on their RSE values. For these RSE deciles the average attribution AUC scores are computed and plotted in Figure 4.8 for both the virtual (a) and the core (b) orbitals. For the multi-head GATv2 (red line) and the GraphNet (green line) models, the attribution AUC scores decline with increasing RSE values. The GATv2 sets on at the overall largest AUC of 0.83 (0.70) for the virtual (core) orbitals and then drops slightly below the GraphNet prediction to a value of about 0.6 (for both models). With the understanding that a larger counter of the RSE decile means a poorer prediction of the XAS spectrum, it becomes apparent that large AUC values are obtained when the overall spectrum prediction itself is reliable as well. Aligning this observation with the broader knowledge of quantum chemistry, we can

infer that if ML predicts the spectrum more accurately, its understanding of orbital contributions improves correspondingly. In contrast, the GCN model's average attribution AUC exhibits no variation across RSE deciles, staying close to the random baseline value of 0.5. This suggests that the model has a similar level of understanding of the ground truth for both strong and weak performances at XAS prediction, which is consistent with the attribution AUC scores of the GCN model shown in the last section.

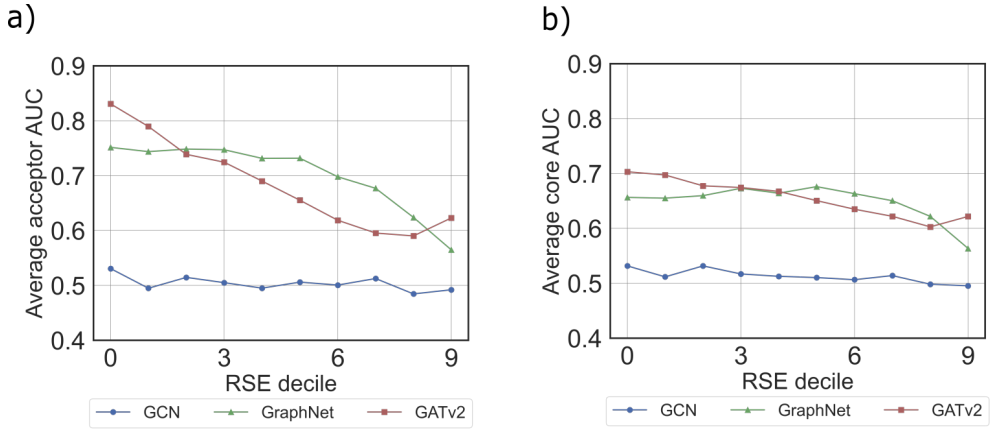

Figure 4.8: Variation of attribution AUC values for virtual (a) and core (b) orbitals with RSE decile values for three GNN models GraphNet (green triangles), GATv2 (red squares), and GCN (blue points).

The robustness of model predictions (and explainability) usually decreases when there are biases in the training dataset that the model erroneously learns [173]. The QM9 dataset is only a small representation of the vast chemical space of organic molecules, and as such is biased towards molecules with certain functional groups. Furthermore, choosing a random subset of structures from this dataset means that the resulting structures in the smaller QM9-XAS dataset could also be further biased toward one or several types of functional groups. To identify whether such biases are learned by the model, one approach is to analyze the attributions of the model's predictions and inspect whether CAM attributions are allocated to incorrect features of the input [268]. In this case, the robustness of model predictions is tested by looking at how the model performance varies for predictions across similar chemical environments. The simplest way of doing so is by perturbing the chemical space around a molecule, e.g. by adding one or several functional groups at different places. The impact of the addition of one methyl group was investigated on randomly selected molecules from the test dataset on both the attribution AUC and the RSE value obtained with the GNN models. For these novel 40 perturbed structures, XAS spectra were calculated as a reference for RSE determination using the same TDDFT method as above. Adding a methyl group at different positions in a molecule leads to changes in the TDDFT spectrum, as well as in the ML predictions as illustrated in the three right panels of Figure 4.9. The three GNN architectures respond differently to this change and give vastly varying predictions of the new spectrum as

indicated by their increased RSE values as well. Overall the ML spectra deviate significantly from the TDDFT spectra. This difference in predictions across all molecules is summarised in the left panel of Figure 4.9 which illustrates the change in the RSE performance of the models for the 40 selected structures before and after perturbation. The RSE distributions of the unperturbed set of molecules have slightly different shapes for the different models, but all give mostly the same average RSE value of approximately 0.03. With the perturbation, the RSE of the GCN and the GATv2 both shift to an RSE average of 0.18, while the GraphNet model gives about 0.13. The altered RSE distributions of the perturbations of these structures clearly indicate a decrease in model performance for perturbed molecules, with the GraphNet model demonstrating superior performance compared to the others. This difference indicates that the GraphNet model can generalize better to chemical environments that are rarely encountered in the dataset and are less susceptible to biases.

The changes in RSE are significant, even for the GraphNet model. This change can be attributed to the fact that when a methyl group is included and replaces a hydrogen atom, the size of the molecule increases. The largest molecules within the original QM9-XAS dataset consist of a maximum of nine heavy atoms (C, N, O, F), while the perturbed structures, on average, contain more than nine heavy atoms. This increase in molecular size potentially represents outliers to the trained model thereby leading to a decline in performance when predicting spectra.

Previous studies have demonstrated[173, 170] that when a model fails to learn the ground-truth logic, it can result in misplaced attributions and the misclassification of atoms within the molecule after perturbations. We therefore now look at how attribution AUC changes for the spectra of the perturbed structures when compared to the AUC values of the original molecules. Figure 4.10 shows the Δ-attribution AUC across all the models for the perturbed structures, where Δ-attribution AUC is the percentage-difference in the attributions of the 40 perturbed structures compared to the AUC values of the unperturbed molecules. While the multi-head GATv2 model shows a 30% decline in the attribution AUC of core orbitals after perturbation, GCN and GraphNet models experience over 40% change. In the case of virtual orbitals, GraphNet and multi-head GATv2 models decrease by 25% and 30% respectively, while the GCN model shows a 35% drop. The drop in relative attributions uniformly across all the models aligns with the increase in RSE values for these molecules, discussed in Figure 4.9. Such large changes in both core and virtual orbitals in all GNN models can originate from the effects of changes in both local and global molecular features on the spectrum after perturbations which results in changes in atomic contributions to the peaks in the spectrum. Hence, while the local environment of an atom, which refers to the atoms in close proximity to the absorbing atom, strongly affects the spectral features in the XAS spectrum, the global environment of that atom and changes caused by perturbations can also play a significant role in determining the electronic structure and thus the final XAS spectrum. This is also in line

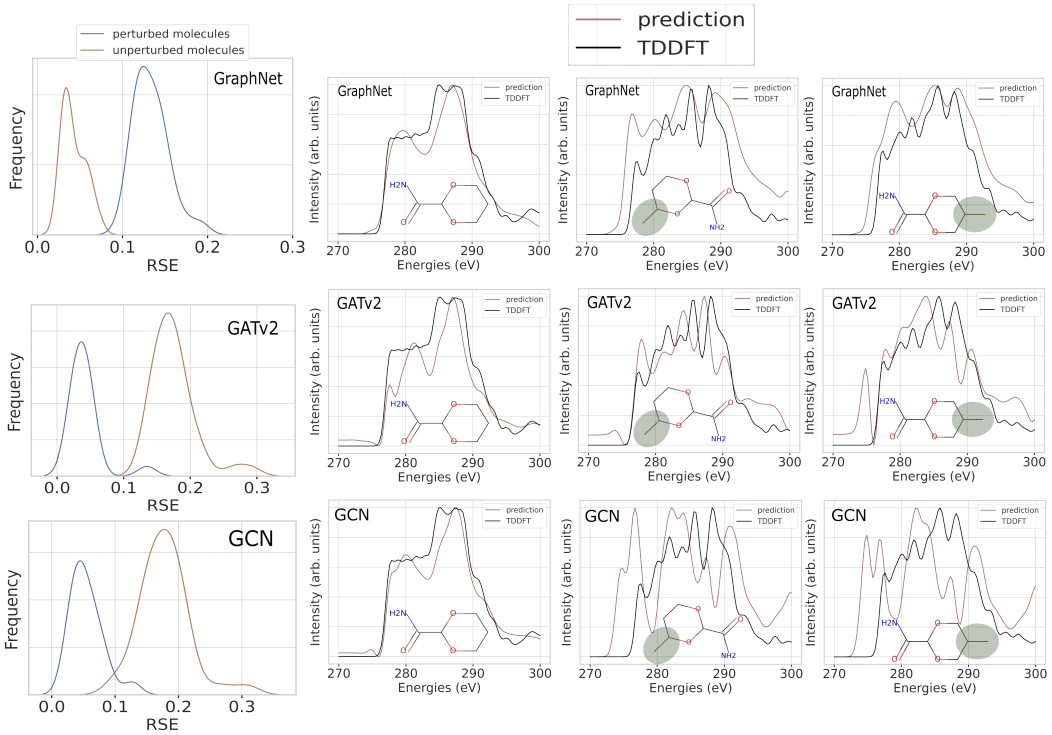

Figure 4.9: The impact of a perturbation through the replacement of functional groups. The left side of each row displays alterations in the RSE distribution for all GNN models when predicting the spectra for unperturbed structures selected from the dataset (blue) and the perturbations of these structures (orange). Additionally, XAS spectra for different exemplary perturbations are shown (right), where a methyl group (highlighted in the grey circle) is added at different positions. The changed TDDFT spectra are shown in black, and their ML predictions are in red.

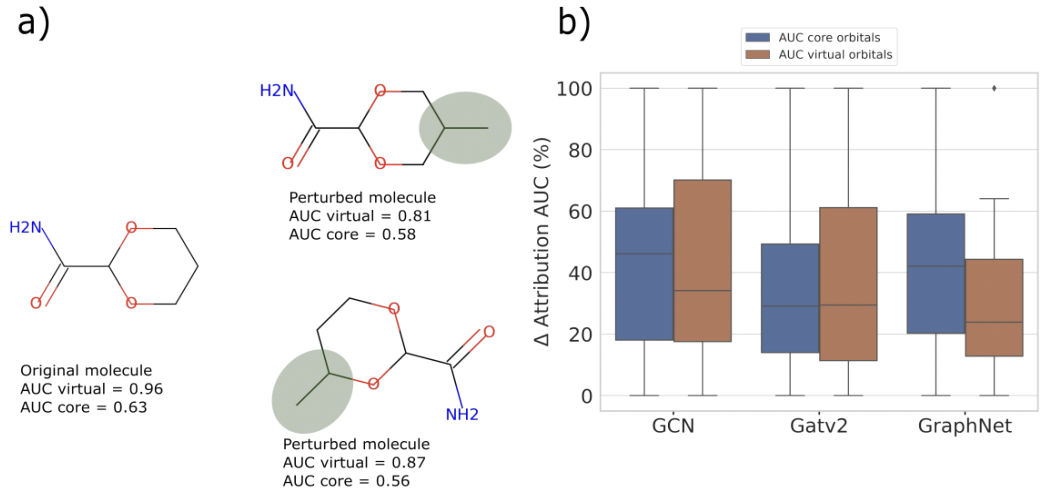

Figure 4.10: Attribution accuracy measured after perturbing random structures. (a) One specific molecular example to demonstrate the addition of $-CH_3$ group as perturbation along with the change of AUC values according to the GraphNet model. (b) Δ-AUC plots for the perturbed set of test molecules across the three GNNs.

with previous research which showed that the presence of long-range interactions between atoms, as well as the coordination number, chemical nature, and distance of these neighboring atoms, can have strong influences on the spectral features, such as the position and width of the XAS peaks [269, 236]. These findings demonstrate the importance of incorporating the local and global environment of nodes while learning structure-property relationships using GNNs [270].

In addition, we examine whether the GNN model with best performance (i.e. GATv2) mimics the changes expected with structural distortions. To obtain distorted molecules, we choose a distortion parameter $\sigma \in \{0.02, 0.05, 0.1\}$ Å to perturb randomly the atomic coordinates, i.e. in x-, y- and z-direction, in the respective molecule [271]. Figure 4.11 demonstrates how the TDDFT calculation and model's prediction change with respect to different distortion values. Upon

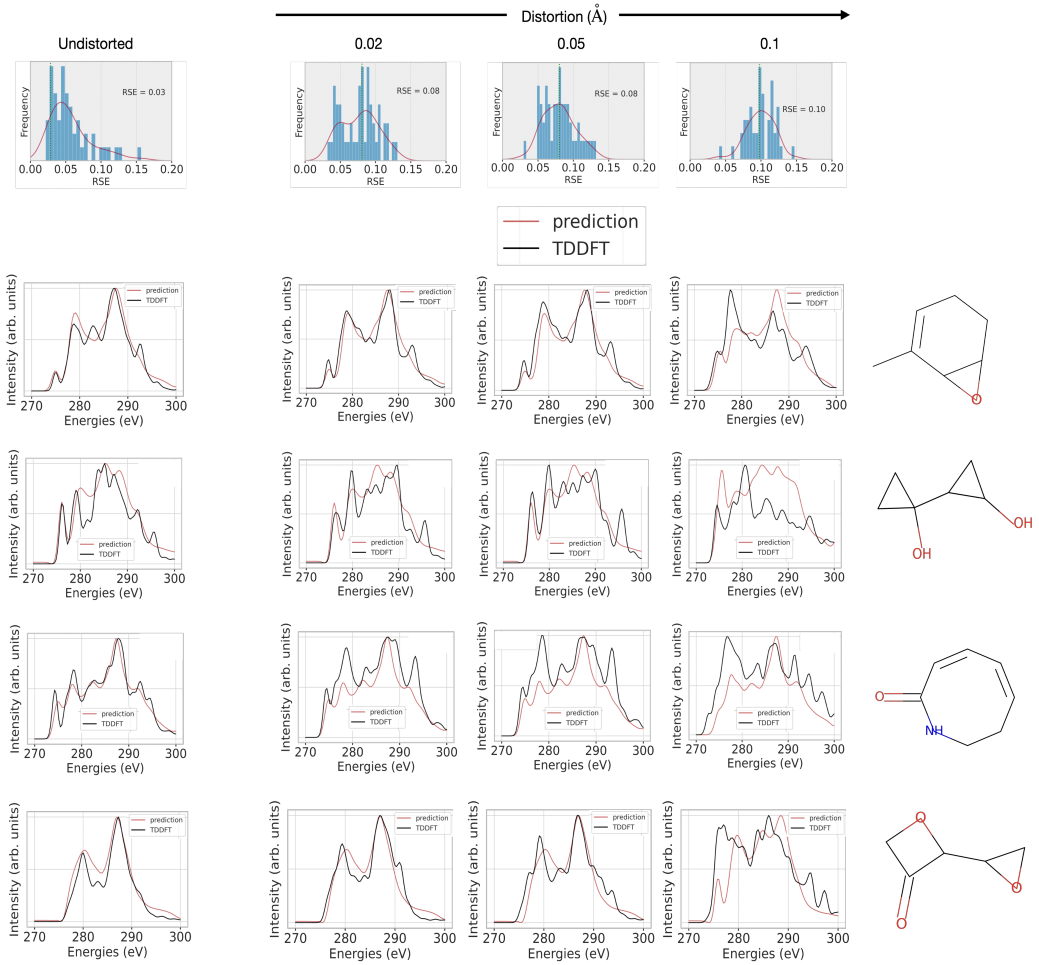

Figure 4.11: Evaluating the performance of the GATv2 model with structural distortions. The predicted spectra show that the model can learn the changes for larger distortions (i.e. 0.1 Å).

distortion of already only 0.02 Å, the TDDFT spectra change mostly in peak intensities and slightly in peak positions. These changes become more pronounced for a stronger distortion. The model's

prediction of small distortions looks similar to the undistorted one, i.e. predicting the general features of the spectrum which, however, results in an increasing RSE with increasing distortion. For the largest distortion of 0.1 Å, the models mimics more closely the changes of the TD-DFT spectra while the RSE values increase again. Such changes in XAS spectra prediction suggest that the representation of small molecular conformation changes would require more structural information in node and edge feature vectors beyond bond lengths. This could be the atomic pairwise distances, dihedral angles, etc. Additionally, training on datasets obtained with canonical ensemble at certain temperature, similar to approaches discussed in chapter 3, can perhaps include rich information about the representative conformers to enhance the robustness of the model to such changes in the conformation.

4.10 Conclusion and Discussion

The aim of this chapter is to assist in the interpretation of peaks in X-ray absorption spectra (XAS) using a black-box machine learning (ML) model, i.e. Graph Neural Networks (GNNs), as opposed to obtaining such information from purely conventional quantum chemical calculations. Yet, the underlying ground truth is based on the latter. In order to achieve this, the explainability technique is implemented on various architectures of GNNs trained on a custom-developed carbon K-edge XAS dataset of 56,000 small organic molecules, denoted as QM9-XAS which is a subset of the original QM9 dataset.

The main difficulty in explaining properties with GNN models, as complex as the physical origin of peaks in XAS spectra already is, is the inherent lack of knowledge about the internal mechanisms of the model and how to correlate the properties of the model with the knowledge gained from quantum chemical calculations. In this chapter, an approach was introduced that reflects a chemist's understanding of the XAS phenomenon as electronic excitations originating from individual atoms, which treats the underlying excitations of XAS peaks as a linear combination of core-to-valence orbital transitions and calculates the individual atom's contribution to the participating core and valence orbitals. This produces atom labels denoting whether a particular atom contributes to an XAS peak within a specified energy range, allowing the acquisition of the chemical ground truth and assessment of the extent to which an ML model comprehends the XAS spectra.

The rationale behind peaks observed in ML-predicted XAS spectra is unraveled via the so-called class activation map (CAM) attributions, highlighting the importance of individual nodes (atoms) in a molecular graph to the target signal of the spectrum. For a quantitative assessment of the graph attributions, the true and false positive rates of CAM attributions were characterised by calculating the area under the curve of the receiver operating characteristic (AUC-ROC), which

is effectively a measure of how well the node attributions match the atomic contributions from the ground truth. Through this comparison between the chemical ground truth, i.e. here the core-to-valence orbital transitions, and CAM attributions, it is shown that while it is important to consider the overall performance of the GNN model in accurately predicting XAS features, the degree of explainability of the different architectures of GNN models differentiates them. It is found that GNN models such as GraphNet and multi-head GAT layers, which are in principle able to capture both the local and the more global chemical environment of an atom in a molecule, not only perform well in their spectra predictions but the explanations obtained from these models are also consistent with the quantum chemical interpretation of XAS.

To examine model robustness, a methyl group is added as a perturbation to a random set of molecules of the test dataset of QM9-XAS. A decrease in performance is observed for all GNN models, with the GraphNet model showing the least decrease in performance, as assessed by the increase in Relative Spectral Error (RSE). Perhaps the differences in the learning mechanisms between the three GNN architectures used have a significant impact on the changes in the RSE distribution and AUC attributions. The observed changes in attribution AUC highlight the importance of integrating explainability into GNN models instead of relying only on the prediction accuracy obtained on a test dataset to evaluate the performance of a model.

In conclusion, the approach presented here provides a recipe for incorporating explainability into GNN models using custom-generated data which provides insight into the physical origin of spectroscopic predictions. Although the GNN models in this work are trained to predict the entire XAS spectrum, the model's attributions provide an opportunity to obtain some insights into local XAS spectra i.e. for individual carbon atoms, with cost-effective computational resources. While our framework was demonstrated for carbon K-edge XAS prediction, the approach can be easily extended to other energy regimes, such as nitrogen and oxygen edges of molecules and metal complexes, or even other spectroscopic techniques. Further, since this approach relies on theoretical data obtained from quantum chemical calculations, it can also be used to obtain ground truth data for models trained on experimental data.

Direct comparison of predictions made in this approach to experimental spectra is challenging due to several factors influencing the experimental observations including solvent effects, experimental conditions like temperature and pressure, and structure-determining factors such as coexistence of multiple meta-stable conformers contributing to the experimental spectra. Incorporating these effects is often not so trivial using the existing theoretical approaches, and thus corrections to theoretical spectra are necessary, often done on a case-by-case basis, depending on the molecular system and its environment. Considering the configurational phase space of the molecule in dataset generation for training the model is one of the ways one can improve the discrepancy between a model's prediction and experimental spectra. For large molecular structures

such as proteins and nano-particles, computation of spectra at *ab initio* level of theory is often a challenge, although their XAS spectra can give insights into their different local environments. While traditionally these have been tackled by the use of fingerprints determined on an ad-hoc basis, we believe that the development of more sophisticated and efficient machine learning frameworks, while maintaining explainability, offers a promising avenue for predicting spectra at low costs, as well as getting insights into local molecular environments.

5

Conclusions and outlook

High-throughput investigations of biomolecules' structure and their spectra prediction in gas phase have become essential components for comprehending the underlying biological processes of these systems. These investigations can be accelerated through the implementation of novel machine learning techniques.

The emergence of enhanced sampling techniques in molecular dynamics, coupled with unsupervised machine learning (ML) techniques, has made it possible to overcome the computational expenses associated with thoroughly investigating the intricate conformational space of biomolecules. Such investigations would otherwise be infeasible with expensive *ab initio* calculations. In particular, the approach presented in chapter 3 can provide experimentalist a fundamental starting point to take into account the macroscopic ensemble in which multiple conformers separated by low-energy barriers contribute to experimental observations at finite temperatures.

Implementing supervised ML such as graph neural networks (GNN) represents a key milestone in the roadmap toward predicting biomolecule's spectrocopic properties including X-ray absorption spectroscopy (XAS) and potentially infrared (IR) spectroscopy. Nevertheless, explaining properties as complex as the physical origin of XAS peaks predicted by GNN models benefits heavily the application of explainability techniques on these black-box models. This enables to shed light on the correlation of the model's predictions with the knowledge obtained from *ab initio* calculations.

In this thesis, we have introduced an approach to consider the realistic representation of canonical ensemble to reproduce and interpret experimental infrared (IR) spectrum of a peptide in gas phase using a combination of replica-exchange molecular dynamics simulations (REMD), unsupervised ML, and *ab initio* calculations. In gas-phase experiments conducted at room temperature, we demonstrate that considering the coexistence of recurring molecular conformations, including both stable and metastable states, along with the relative importance of each conformer can significantly improve the discrepancies between theoretical predictions and experimental results. Therefore, using the weighting factors calculated from relative populations of clusters, obtained by Probabilistic Analysis of Molecular Motifs (PAMM) technique, for averaging the calculated IR spectra of these recurring molecular conformations is a reasonable approach to capture the main features of the vibrational spectra. Enhancing the agreement between theory and experiment can be achieved by unraveling the non-Gaussian features present in the potential energy surface. This approach aids in explaining the IR fingerprints that are correlated with changes in the backbone conformation and H-bonding patterns.

Harmonic frequency calculations employed in this approach neglects the local anharmonic contributions from representative conformer of each energy basin in the conformational space. However, sampling the conformational space with REMD and decomposing the clusters containing similar conformations into even smaller subensemble, implicitly account for some of the anharmonicity.

Investigating the potential energy surface of biomolecules and considering it in interpretation of experimental observations is an active field of research, with strong effort to include more advance

enhanced sampling methods with novel ML techniques to this process. To further extend this study, the next step could involve exploring the canonical ensemble of the biomolecule under investigation to interpret more complex experiments at room temperature such as near-edge X-ray absorption mass spectrometry in the gas phase. Additionally, the approach presented here serves as an excellent starting point for extracting recurring molecular motifs from the conformational space and use these motifs to conduct costly *ab initio* calculations for generating a dataset suitable for supervised machine learning.

Furthermore, we have developed a supervised ML technique using Graph Neural Networks (GNNs) to predict X-ray absorption spectrum (XAS). As the explainability of these model's predictions remains a challenge, we introduced a framework to have a rigorous understanding of the predicted XAS generated by such ML models. This framework performs an in-depth investigation of the respective black-box ML model, here based on different architectures of GNN on custom-generated XAS dataset for small organic molecules. The result of the analysis shows that a thorough analysis of the different ML models with respect to the local and global environments of atoms considered in each ML model, is essential for the selection of an appropriate ML model which allows a robust XAS prediction. Furthermore, feature attribution is shown as a promising candidate to determine the respective contributions of various atoms in the molecules to the peaks observed in the XAS spectrum. This study demonstrate that it is possible to relate the atomic contributions via these orbitals to the XAS spectrum, by comparing this peak assignment to the core and virtual orbitals from the quantum chemical calculations underlying our dataset.

To summarise, the methods presented in this thesis offer a pathway to tackle the challenges associated with comprehending the behavior of biomolecules in the gas phase. They achieve a balance between costly *ab initio* calculations and the lower computational expenses of molecular dynamics by integrating both supervised and unsupervised machine learning techniques. Although the methods developed in this thesis were applied to small molecules, they can be extended to larger molecular structures like proteins and nanoparticles to reveal intricate structure-property relationships and identify metastable configurations that are crucial for comprehending and predicting the spectroscopic behavior of these systems.

Bibliography

[1] Clifford A. Hudis. “Trastuzumab — Mechanism of Action and Use in Clinical Practice”. In: *New England Journal of Medicine* 357 (1 July 2007), pp. 39–51.

[2] Vladimir P. Torchilin. “Targeted pharmaceutical nanocarriers for cancer therapy and imaging”. In: *The AAPS journal* 9 (2 May 2007).

[3] Chatterjee J, Giri S, Maity S, Sinha A, Ranjan A, Rajshekhar None, and Gupta S. “Production and characterization of thermostable alkaline protease of Bacillus subtilis (ATCC 6633) from optimized solid-state fermentation”. In: *Biotechnol Appl Biochem* 62 (5 Sept. 2015), pp. 709–718.

[4] Margit Mahlapuu, Joakim Håkansson, Lovisa Ringstad, and Camilla Björn. “Antimicrobial Peptides: An Emerging Category of Therapeutic Agents”. In: *Frontiers in Cellular and Infection Microbiology* 6 (Dec. 2016), p. 194.

[5] Harry Jubb, Alicia P. Higueruelo, Anja Winter, and Tom L. Blundell. “Structural biology and drug discovery for protein–protein interactions”. In: *Trends in Pharmacological Sciences* 33 (5 May 2012), pp. 241–248.

[6] Svetlana Gorina and Nikola P. Pavletich. “Structure of the p53 Tumor Suppressor Bound to the Ankyrin and SH3 Domains of 53BP2”. In: *Science* 274 (5289 Nov. 1996), pp. 1001–1005.

[7] Vladimir N. Uversky. “Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders”. In: *Frontiers in Aging Neuroscience* 7 (MAR Mar. 2015).

[8] Lucas Schwob, Simon Dörner, Kaan Atak, Kaja Schubert, Martin Timm, Christine Bülow, Vicente Zamudio-Bayer, Bernd Von Issendorff, J. Tobias Lau, Simone Techert, and Sadia Bari. “Site-Selective Dissociation upon Sulfur L-Edge X-ray Absorption in a Gas-Phase Protonated Peptide”. In: *Journal of Physical Chemistry Letters* 11 (4 2020), pp. 1215–1221.

[9] Simon Dörner, Lucas Schwob, Kaan Atak, Kaja Schubert, Rebecca Boll, Thomas Schlathölter, Martin Timm, Christine Bülow, Vicente Zamudio-Bayer, Bernd von Issendorff, J Tobias Lau, Simone Techert, and Sadia Bari. “Probing Structural Information of Gas-Phase Peptides by Near-Edge X-ray Absorption Mass Spectrometry”. In: *Journal of the American Society for Mass Spectrometry* 32 (3 Mar. 2021), pp. 670–684.

[10] Ron O. Dror, Albert C. Pan, Daniel H. Arlow, David W. Borhani, Paul Maragakis, Yibing Shan, Huafeng Xu, and David E. Shaw. "Pathway and mechanism of drug binding to G-protein-coupled receptors". In: *Proceedings of the National Academy of Sciences of the United States of America* 108 (32 Aug. 2011), pp. 13118–13123.

[11] Esther Marco and Federico Gago. "Overcoming the Inadequacies or Limitations of Experimental Structures as Drug Targets by Using Computational Modeling Tools and Molecular Dynamics Simulations". In: *ChemMedChem* 2 (10 Oct. 2007), pp. 1388–1401.

[12] Amir Kotobi, Lucas Schwob, Gregor B. Vonbun-Feldbauer, Mariana Rossi, Piero Gasparotto, Christian Feiler, Giel Berden, Jos Oomens, Bart Oostenrijk, Debora Scuderi, Sadia Bari, and Robert H. Meißner. "Reconstructing the infrared spectrum of a peptide from representative conformers of the full canonical ensemble". In: *Communications Chemistry* 6 (1 Mar. 2023), p. 46.

[13] Derek Jones, Jonathan E. Allen, Yue Yang, William F. Drew Bennett, Maya Gokhale, Niema Moshiri, and Tajana S. Rosing. "Accelerators for Classical Molecular Dynamics Simulations of Biomolecules". In: *Journal of Chemical Theory and Computation* 18 (7 July 2022), pp. 4047–4069.

[14] Pascal Parneix, Marie Basire, and Florent Calvo. "Accurate Modeling of Infrared Multiple Photon Dissociation Spectra: The Dynamical Role of Anharmonicities". In: *The Journal of Physical Chemistry A* 117.19 (May 2013), pp. 3954–3959.

[15] Trey Ideker, Vesteinn Thorsson, Jeffrey A. Ranish, Rowan Christmas, Jeremy Buhler, Jimmy K. Eng, Roger Bumgarner, David R. Goodlett, Ruedi Aebersold, and Leroy Hood. "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network". In: *Science* 292 (5518 May 2001), pp. 929–934.

[16] D.J. Wales. "Energy Landscapes: Applications to Clusters, Biomolecules and Glasses". In: *Cambridge University Press* (2003).

[17] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. "Simplifying the representation of complex free-energy landscapes using sketch-map". In: *Proceedings of the National Academy of Sciences* 108 (32 Aug. 2011), pp. 13023–13028.

[18] Florent Hédin, Nuria Plattner, J. D. Doll, and Markus Meuwly. "Spatial Averaging: Sampling Enhancement for Exploring Configurational Space of Atomic Clusters and Biomolecules". In: *Journal of Chemical Theory and Computation* 10 (10 Oct. 2014). doi: 10.1021/ct500529w, pp. 4284–4296.

[19] Hiroshi Fujisaki, Kei Moritsugu, and Yasuhiro Matsunaga. "Exploring Configuration Space and Path Space of Biomolecules Using Enhanced Sampling Techniques—Searching for Mechanism and Kinetics of Biomolecular Functions". In: *International Journal of Molecular Sciences* 19 (10 Oct. 2018), p. 3177.

[20] Andreas Vitalis and Rohit V. Pappu. "Chapter 3 Methods for Monte Carlo Simulations of Biomacromolecules". In: *Annual reports in computational chemistry* 5 (Jan. 2009), pp. 49–76.

[21] Denis Bucher, Levi C. T. Pierce, J. Andrew McCammon, and Phineus R. L. Markwick. "On the Use of Accelerated Molecular Dynamics to Enhance Configurational Sampling in Ab Initio Simulations". In: *Journal of Chemical Theory and Computation* 7 (4 Apr. 2011), pp. 890–897.

[22] Jim Pfaendtner. "Metadynamics to Enhance Sampling in Biomolecular Simulations". In: *Methods in molecular biology (Clifton, N.J.)* 2022 (2019), pp. 179–200.

[23] Hiraku Oshima, Suyong Re, and Yuji Sugita. "Replica-Exchange Umbrella Sampling Combined with Gaussian Accelerated Molecular Dynamics for Free-Energy Calculation of Biomolecules". In: *Journal of Chemical Theory and Computation* 15 (10 Oct. 2019), pp. 5199–5208.

[24] Omar Valsson and Michele Parrinello. "Variational Approach to Enhanced Sampling and Free Energy Calculations". In: *Phys. Rev. Lett.* 113 (9 Aug. 2014), p. 090601.

[25] Aleksandar R. Milosavljević, Christophe Nicolas, Miloš Lj. Ranković, Francis Canon, Catalin Miron, and Alexandre Giuliani. "K-Shell Excitation and Ionization of a Gas-Phase Protein: Interplay between Electronic Structure and Protein Folding". In: *The Journal of Physical Chemistry Letters* 6 (16 Aug. 2015), pp. 3132–3138.

[26] Annalisa Arcella, Guillem Portella, Maria Luz Ruiz, Ramon Eritja, Marta Vilaseca, Valérie Gabelica, and Modesto Orozco. "Structure of Triplex DNA in the Gas Phase". In: *Journal of the American Chemical Society* 134 (15 Apr. 2012), pp. 6596–6606.

[27] Nicole L. Burke, James G. Redwine, Jacob C. Dean, Scott A. Mcluckey, and Timothy S. Zwier. "UV and IR spectroscopy of cold protonated leucine enkephalin". In: *International Journal of Mass Spectrometry* 378 (2015), pp. 196–205.

[28] Nadja Heine and Knut R Asmis. "Cryogenic ion trap vibrational spectroscopy of hydrogen-bonded clusters relevant to atmospheric chemistry". In: *International Reviews in Physical Chemistry* 34.1 (Jan. 2015), pp. 1–34.

[29] O. González-Magaña, G. Reitsma, M. Tiemens, L. Boschman, R. Hoekstra, and T. Schlathölter. "Near-edge X-ray absorption mass spectrometry of a gas-phase peptide". In: *Journal of Physical Chemistry A* 116 (44 Nov. 2012), pp. 10745–10751.

[30] Corey N Stedwell, Johan F Galindo, Adrian E Roitberg, and Nicolas C Polfer. "Structures of Biomolecular Ions in the Gas Phase Probed by Infrared Light Sources". In: *Annual Review of Analytical Chemistry* 6.1 (June 2013), pp. 267–285.

[31] Stephan Warnke, Gert von Helden, and Kevin Pagel. "Protein Structure in the Gas Phase: The Influence of Side-Chain Microsolvation". In: *Journal of the American Chemical Society* 135.4 (Jan. 2013), pp. 1177–1180.

[32] Tim Meyer, Valérie Gabelica, Helmut Grubmüller, and Modesto Orozco. "Proteins in the gas phase". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3.4 (July 2013), pp. 408–425.

[33] Jaime A. Stearns Thomas R. Rizzo and Oleg V. Boyarkin. "Spectroscopic studies of cold, gas-phase biomolecular ions". In: *International Reviews in Physical Chemistry* 28.3 (2009), pp. 481–515.

[34] Ángela I López-Lorente and Boris Mizaikoff. "Mid-infrared spectroscopy for protein analysis: potential and challenges". In: *Analytical and Bioanalytical Chemistry* 408.11 (Apr. 2016), pp. 2875–2889.

[35] Anouk M Rijs and Jos Oomens. "Gas-Phase IR Spectroscopy and Structure of Biological Molecules". In: *Topics in Current Chemistry* 364 ().

[36] Jérôme Mahé, Sander Jaeqx, Anouk M Rijs, and Marie-Pierre Gaigeot. "Can far-IR action spectroscopy combined with BOMD simulations be conformation selective?" In: *Physical Chemistry Chemical Physics* 17.39 (2015), pp. 25905–25914.

[37] Audrius Doblies, Christian Feiler, Tim Würger, Eduard Schill, Robert H. Meißner, and Bodo Fiedler. "Mechanical degradation estimation of thermosets by peak shift assessment: General approach using infrared spectroscopy". In: *Polymer* 221 (2021), p. 123585.

[38] John B. Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M. Whitehouse. "Electrospray Ionization for Mass Spectrometry of Large Biomolecules". In: *Science* 246 (4926 Oct. 1989), pp. 64–71.

[39] Joseph A. Loo. "Mass spectrometry in biophysics: Conformation and dynamics of biomolecules". In: *Journal of the American Society for Mass Spectrometry* 16.12 (Dec. 2005), pp. 2064–2065.

[40] M Michaelis, N Hildebrand, R H Meißner, N Wurzler, Z Li, J D Hirst, A Micsonai, J Kardos, M Delle Piane, and L Colombi Ciacchi. "Impact of the Conformational Variability of Oligopeptides on the Computational Prediction of Their CD Spectra". In: *The Journal of Physical Chemistry B* 123.31 (Aug. 2019), pp. 6694–6704.

[41] Mariana Rossi, Volker Blum, Peter Kupser, Gert von Helden, Frauke Bierau, Kevin Pagel, Gerard Meijer, and Matthias Scheffler. "Secondary Structure of Ac-Ala n -LysH + Polyalanine Peptides (n = 5,10,15) in Vacuo: Helical or Not?" In: *The Journal of Physical Chemistry Letters* 1.24 (Dec. 2010), pp. 3465–3470.

[42] Nicole L Burke, Andrew F. DeBlase, James G Redwine, John R Hopkins, Scott A. McLuckey, and Timothy S Zwier. "Gas-Phase Folding of a Prototypical Protonated Pentapeptide: Spectroscopic Evidence for Formation of a Charge-Stabilized β-Hairpin". In: *Journal of the American Chemical Society* 138.8 (Mar. 2016), pp. 2849–2857.

[43] Tim Würger, Wolfgang Heckel, Kai Sellschopp, Stefan Müller, Andreas Stierle, Yuemin Wang, Heshmat Noei, and Gregor Feldbauer. "Adsorption of Acetone on Rutile TiO 2 : A DFT and FTIRS Study". In: *The Journal of Physical Chemistry C* 122.34 (Aug. 2018), pp. 19481–19490.

[44] Ronghu Wu and Terry B. McMahon. "Protonation sites and conformations of peptides of glycine (Gly 1-5H+) by IRMPD spectroscopy". In: *Journal of Physical Chemistry B* 113.25 (June 2009), pp. 8767–8775.

[45] Jonas Sjöqvist, Rafael C. González-Cano, Juan T. López Navarrete, Juan Casado, M. Carmen Ruiz Delgado, Mathieu Linares, and Patrick Norman. "A combined MD/QM and experimental exploration of conformational richness in branched oligothiophenes". In: *Phys. Chem. Chem. Phys.* 16.45 (2014), pp. 24841–24852.

[46] Tapta Kanchan Roy, Vladimir Kopysov, Aleksandr Pereverzev, Jiří Šebek, R. Benny Gerber, and Oleg V. Boyarkin. "Intrinsic structure of pentapeptide Leu-enkephalin: geometry optimization and validation by comparison of VSCF-PT2 calculations with cold ion spectroscopy". In: *Physical Chemistry Chemical Physics* 20.38 (2018), pp. 24894–24901.

[47] Kiyoshi Yagi, Kenta Yamada, Chigusa Kobayashi, and Yuji Sugita. "Anharmonic Vibrational Analysis of Biomolecules and Solvated Molecules Using Hybrid QM/MM Computations". In: *Journal of Chemical Theory and Computation* 15 (3 Mar. 2019), pp. 1924–1938.

[48] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong, and Chris Wolverton. "Recent advances and applications of deep learning methods in materials science". In: *npj Computational Materials* 8 (1 Apr. 2022), p. 59.

[49] Geemi P. Wellawatte, Heta A. Gandhi, Aditi Seshadri, and Andrew D. White. "A Perspective on Explanations of Molecular Prediction Models". In: *Journal of Chemical Theory and Computation* (Mar. 2023).

[50] Y. Mishin. "Machine-learning interatomic potentials for materials science". In: *Acta Materialia* 214 (Aug. 2021), p. 116980.

[51] Aldo Glielmo, Brooke E. Husic, Alex Rodriguez, Cecilia Clementi, Frank Noé, and Alessandro Laio. "Unsupervised Learning Methods for Molecular Simulation Data". In: *Chemical Reviews* 121 (16 Aug. 2021), pp. 9722–9758.

[52] Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, and Kristof T. Schütt. "Inverse design of 3d molecular structures with conditional generative neural networks". In: *Nature Communications* 13.1 (2022), p. 973.

[53] Niklas Gebauer, Michael Gastegger, and Kristof Schütt. "Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules". In: *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems* (2019), pp. 7566–7578.

[54] Kristof T. Schütt, Stefaan S. P. Hessmann, Niklas W. A. Gebauer, Jonas Lederer, and Michael Gastegger. "SchNetPack 2.0: A neural network toolbox for atomistic machine learning". In: *The Journal of Chemical Physics* 158.14 (Apr. 2023), p. 144801.

[55] Sheng Ye, Kai Zhong, Jinxiao Zhang, Wei Hu, Jonathan D. Hirst, Guozhen Zhang, Shaul Mukamel, and Jun Jiang. "A Machine Learning Protocol for Predicting Protein Infrared Spectra". In: *Journal of the American Chemical Society* 142 (45 Nov. 2020), pp. 19071–19077.

[56] Hao Ren, Hao Li, Qian Zhang, Lijun Liang, Wenyue Guo, Fang Huang, Yi Luo, and Jun Jiang. "A machine learning vibrational spectroscopy protocol for spectrum prediction and spectrum-based structure recognition". In: *Fundamental Research* 1 (4 July 2021), pp. 488–494.

[57] Fenris Lu, Lixue Cheng, Ryan J DiRisio, Jacob M Finney, Mark A Boyer, Pattarapon Moonkaen, Jiace Sun, Sebastian J R Lee, J Emiliano Deustua, Thomas F I I I Miller, and Anne B McCoy. "Fast Near Ab Initio Potential Energy Surfaces Using Machine Learning". In: *The Journal of Physical Chemistry A* 126 (25 June 2022), pp. 4013–4024.

[58] Li Zhu, Maximilian Amsler, Tobias Fuhrer, Bastian Schaefer, Somayeh Faraji, Samare Rostami, S Alireza Ghasemi, Ali Sadeghi, Migle Grauzinyte, Chris Wolverton, and Stefan Goedecker. "Unsupervised machine learning in atomistic simulations, between predictions and understanding". In: *The Journal of Chemical Physics* 144 (2016), p. 34203.

[59] Tsz Wai Ko, Jonas A. Finkler, Stefan Goedecker, and Jörg Behler. "A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer". In: *Nature Communications* 12 (1 Jan. 2021), p. 398.

[60] A. A. Guda, S. A. Guda, A. Martini, A. N. Kravtsova, A. Algasov, A. Bugaev, S. P. Kubrin, L. V. Guda, P. Šot, J. A. van Bokhoven, C. Copéret, and A. V. Soldatov. "Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms". In: *npj Computational Materials* 7 (1 Dec. 2021), p. 203.

[61] C. D. Rankine and T. J. Penfold. "Accurate, affordable, and generalizable machine learning simulations of transition metal x-ray absorption spectra using the XANESNET deep neural network". In: *The Journal of Chemical Physics* 156.16 (Apr. 2022). 164102.

[62] Kanishka Singh, Jannes Münchmeyer, Leon Weber, Ulf Leser, and Annika Bande. "Graph Neural Networks for Learning Molecular Excitation Spectra". In: *Journal of Chemical Theory and Computation* 18 (7 July 2022), pp. 4408–4417.

[63] Michael Gastegger, Jörg Behler, and Philipp Marquetand. "Machine learning molecular dynamics for the simulation of infrared spectra". In: *Chemical Science* 8.10 (2017), pp. 6924–6935.

[64] Anja Aarva, Volker L. Deringer, Sami Sainio, Tomi Laurila, and Miguel A. Caro. "Understanding X-ray Spectroscopy of Carbonaceous Materials by Combining Experiments, Density Functional Theory, and Machine Learning. Part I: Fingerprint Spectra". In: *Chemistry of Materials* 31.22 (Nov. 2019), pp. 9243–9255.

[65] Anja Aarva, Sami Sainio, Volker L. Deringer, Miguel A. Caro, and Tomi Laurila. "X-ray Spectroscopy Fingerprints of Pristine and Functionalized Graphene". In: *The Journal of Physical Chemistry C* 125.33 (Aug. 2021), pp. 18234–18246.

[66] Dorothea Golze, Markus Hirvensalo, Patricia Hernández-León, Anja Aarva, Jarkko Etula, Toma Susi, Patrick Rinke, Tomi Laurila, and Miguel A. Caro. "Accurate Computational Prediction of Core-Electron Binding Energies in Carbon-Based Materials: A Machine-Learning Model Combining Density-Functional Theory and *GW*". In: *Chemistry of Materials* 34.14 (July 2022), pp. 6240–6254.

[67] Philipp Schienbein. "Spectroscopy from Machine Learning by Accurately Representing the Atomic Polar Tensor". In: *Journal of Chemical Theory and Computation* 19.3 (Jan. 2023), pp. 705–712.

[68] Felix Faber, Luke Hutchinson, Huang Bing, Justin Gilmer, Sam Schoenholz, George Dahl, Oriol Vinyals, Steven Kearnes, Patrick Riley, and Anatole von Lilienfeld. "Prediction errors of molecular machine learning models lower than hybrid DFT error". In: *Journal of Chemical Theory and Computation* 13(11) (2017), pp. 5255–5264.

[69] Duch Wlodzislaw, Karthikeyan Swaminathan, and Jaroslaw Meller. "Artificial Intelligence Approaches for Rational Drug Design and Discovery". In: *Current pharmaceutical design* 13 (Feb. 2007), pp. 1497–508.

[70] Johann Lüder. "Determining electronic properties from *L*-edge x-ray absorption spectra of transition metal compounds with artificial neural networks". In: *Phys. Rev. B* 103 (4 Jan. 2021), p. 045140.

[71] Kunal Ghosh, Annika Stuke, Milica Todorović, Peter Bjørn Jørgensen, Mikkel N. Schmidt, Aki Vehtari, and Patrick Rinke. "Machine Learning: Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra (Adv. Sci. 9/2019)". In: *Advanced Science* 6.9 (May 2019), p. 1970053.

[72] Daniel P. Little, J. Paul. Speir, Michael W. Senko, Peter B. O'Connor, and Fred W. McLafferty. "Infrared Multiphoton Dissociation of Large Multiply Charged Ions for Biomolecule Sequencing". In: *Analytical Chemistry* 66 (1994), pp. 2809–2815.

[73] Nick C Polfer, Jos Oomens, Sándor Suhai, and Béla Paizs. "Infrared Spectroscopy and Theoretical Studies on Gas-Phase Protonated Leu-enkephalin and Its Fragments: Direct Experimental Evidence for the Mobile Proton". In: *Journal of the American Chemical Society* 129.18 (May 2007), pp. 5887–5897.

[74] Florian Schinle, Christoph R Jacob, Arron B Wolk, Jean-François Greisch, Matthias Vonderach, Patrick Weis, Oliver Hampe, Mark A Johnson, and Manfred M Kappes. "Ion Mobility Spectrometry, Infrared Dissociation Spectroscopy, and ab Initio Computations toward Structural Characterization of the Deprotonated Leucine-Enkephalin Peptide Anion in the Gas Phase". In: *The Journal of Physical Chemistry A* 118.37 (Sept. 2014), pp. 8453–8463.

[75] Ronghu Wu and Terry B. McMahon. "An Investigation of Protonation Sites and Conformations of Protonated Amino Acids by IRMPD Spectroscopy". In: *ChemPhysChem* 9.18 (2008), pp. 2826–2835.

[76] Sjors Bakels, Marie-Pierre Gaigeot, and Anouk M Rijs. "Gas-Phase Infrared Spectroscopy of Neutral Peptides: Insights from the Far-IR and THz Domain". In: *Chemical Reviews* 120.7 (Apr. 2020), pp. 3233–3260.

[77] Timothy D. Vaden, Tjalling S. J. A. de Boer, Neil A. MacLeod, Elaine M. Marzluff, John P. Simons, and Lavina C. Snoek. "Infrared spectroscopy and structure of photochemically protonated biomolecules in the gas phase: a noradrenaline analogue, lysine and alanyl alanine". In: *Phys. Chem. Chem. Phys.* 9 (20 2007), pp. 2549–2555.

[78] Jonathan Martens, Giel Berden, Christoph R Gebhardt, and Jos Oomens. "Infrared ion spectroscopy in a modified quadrupole ion trap mass spectrometer at the FELIX free electron laser laboratory". In: *Review of Scientific Instruments* 87.10 (Oct. 2016), p. 103108.

[79] Joost M Bakker, Thierry Besson, Joël Lemaire, Debora Scuderi, and Philippe Maître. "Gas-Phase Structure of a π-Allyl-Palladium Complex: Efficient Infrared Spectroscopy in a 7 T Fourier Transform Mass Spectrometer". In: *The Journal of Physical Chemistry A* 111.51 (Dec. 2007), pp. 13415–13424.

[80] Giel Berden, Mathijs Derksen, Kas J. Houthuijs, Jonathan Martens, and Jos Oomens. "An automatic variable laser attenuator for IRMPD spectroscopy and analysis of power-dependence in fragmentation spectra". In: *International Journal of Mass Spectrometry* 443 (2019), pp. 1–8.

[81] Ron O. Dror, Robert M. Dirks, J.P. Grossman, Huafeng Xu, and David E. Shaw. "Biomolecular Simulation: A Computational Microscope for Molecular Biology". In: *Annual Review of Biophysics* 41 (1 June 2012), pp. 429–452.

[82] Lili Duan, Xiaona Guo, Yalong Cong, Guoqiang Feng, Yuchen Li, and John Z. H. Zhang. "Accelerated Molecular Dynamics Simulation for Helical Proteins Folding in Explicit Water". In: *Frontiers in Chemistry* 7 (2019).

[83] Jacob D Durrant and J Andrew McCammon. "Molecular dynamics simulations and drug discovery". In: *BMC Biology* 9 (1 Dec. 2011), p. 71.

[84] Yoshinori Hirano, Noriaki Okimoto, Shigeo Fujita, and Makoto Taiji. "Molecular Dynamics Study of Conformational Changes of Tankyrase 2 Binding Subsites upon Ligand Binding". In: *ACS Omega* 6 (2021).

[85] Mariana Rossi Carvalho. "Ab initio study of alanine-based polypeptide secondary-structure motifs in the gas phase". In: *PhD thesis* (2011).

[86] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *J. Chem. Phys* 126 (2007), p. 14101.

[87] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. "Molecular dynamics with coupling to an external bath". In: *The Journal of Chemical Physics* 81.8 (Oct. 1984), pp. 3684–3690.

[88] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. "Towards exact molecular dynamics simulations with machine-learned force fields". In: *Nature Communications* 9 (1 2018), p. 3887.

[89] Mark Abraham, Andrey Alekseenko, Cathrine Bergh, Christian Blau, Eliane Briand, Mahesh Doijade, Stefan Fleischmann, Vytautas Gapsys, Gaurav Garg, Sergey Gorelov, Gilles Gouaillardet, Alan Gray, M. Eric Irrgang, Farzaneh Jalalypour, Joe Jordan, Christoph Junghans, Prashanth Kanduri, Sebastian Keller, Carsten Kutzner, Justin A. Lemkul, Magnus Lundborg, Pascal Merz, Vedran Miletić, Dmitry Morozov, Szilárd Páll, Roland Schulz, Michael Shirts, Alexey Shvetsov, Bálint Soproni, David van der Spoel, Philip Turner, Carsten Uphoff, Alessandra Villa, Sebastian Wingbermühle, Artem Zhmurov, Paul Bauer, Berk Hess, and Erik Lindahl. "GROMACS 2023.1 Manual". In: *Zenodo* (Apr. 2023).

[90] Lars Konermann, Haidy Metwally, Robert G. McAllister, and Vlad Popa. "How to run molecular dynamics simulations on electrospray droplets and gas phase proteins: Basic guidelines and selected applications". In: *Methods* 144 (July 2018), pp. 104–112.

[91] Lars Konermann. "Molecular Dynamics Simulations on Gas-Phase Proteins with Mobile Protons: Inclusion of All-Atom Charge Solvation". In: *The Journal of Physical Chemistry B* 121 (34 Aug. 2017), pp. 8102–8112.

[92] Jay W. Ponder and David A. Case. "Force fields for protein simulations". In: *Advances in protein chemistry* 66 (2003), pp. 27–85.

[93] George A. Kaminski, Harry A. Stern, B. J. Berne, Richard A. Friesner, Yixiang X. Cao, Robert B. Murphy, Ruhong Zhou, and Thomas A. Halgren. "Development of a polarizable force field for proteins via ab initio quantum chemistry: First generation model and gas phase tests". In: *Journal of Computational Chemistry* 23.16 (2002), pp. 1515–1531.

[94] Justin A. Lemkul, Jing Huang, Benoît Roux, and Alexander D. MacKerell. "An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications". In: *Chemical Reviews* 116 (9 May 2016), pp. 4983–5013.

[95] Michael D. Beachy, David Chasman, Robert B. Murphy, Thomas A. Halgren, and Richard A. Friesner. "Accurate ab Initio Quantum Chemical Determination of the Relative Energetics of Peptide Conformations and Assessment of Empirical Force Fields". In: *Journal of the American Chemical Society* 119 (25 June 1997), pp. 5908–5920.

[96] Cameron Abrams and Giovanni Bussi. "Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration". In: *Entropy* 16.1 (2014), pp. 163–199.

[97] Rafael C. Bernardi, Marcelo C.R. Melo, and Klaus Schulten. "Enhanced sampling techniques in molecular dynamics simulations of biological systems". In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1850.5 (2015). Recent developments of molecular dynamics, pp. 872–877.

[98] Yinglong Miao and J. Andrew McCammon. "Unconstrained Enhanced Sampling for Free Energy Calculations of Biomolecules: A Review". In: *Molecular simulation* 42 (13 Sept. 2016), pp. 1046–1055.

[99] Koji Hukushima and Koji Nemoto. "Exchange Monte Carlo Method and Application to Spin Glass Simulations". In: *Journal of the Physical Society of Japan* 65.6 (June 1996), pp. 1604–1608.

[100] Yuji Sugita, Motoshi Kamiya, Hiraku Oshima, and Suyong Re. "Replica-Exchange Methods for Biomolecular Simulations". In: *Methods in Molecular Biology* 2022 (2019), pp. 155–177.

[101] R.W Hockney, S.P Goel, and J.W Eastwood. "Quiet high-resolution computer models of a plasma". In: *Journal of Computational Physics* 14.2 (Feb. 1974), pp. 148–158.

[102] Herman Berendsen and Wilfred van Gunsteren. "H.J.C. Berendsen and W.F. van Gunsteren Practical Algorithms for Dynamic Simulation , in". In: *Molecular-Dynamics Simulation of Statistical Mechanical Systems* (1986), pp. 43–65.

[103] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". In: *Journal of Chemical Theory and Computation* 11.8 (Aug. 2015), pp. 3696–3713.

[104] Alexandra Patriksson and David Van Der Spoel. "A temperature predictor for parallel tempering simulations". In: *Physical Chemistry Chemical Physics* 10.15 (Apr. 2008), pp. 2073–2077.

[105] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J C Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". In: *Journal of Computational Physics* 23 (3 1977), pp. 327–341.

[106] Hans C Andersen. "Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations". In: *Journal of Computational Physics* 52 (1 1983), pp. 24–34.

[107] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas". In: *Phys. Rev.* 136 (3B Nov. 1964), B864–B871.

[108] W. Kohn and L. J. Sham. "Self-Consistent Equations Including Exchange and Correlation Effects". In: *Phys. Rev.* 140 (4A Nov. 1965), A1133–A1138.

[109] A. D. Becke. "Density-functional exchange-energy approximation with correct asymptotic behavior". In: *Phys. Rev. A* 38 (6 Sept. 1988), pp. 3098–3100.

[110] John P. Perdew. "Density-functional approximation for the correlation energy of the inhomogeneous electron gas". In: *Phys. Rev. B* 33 (12 June 1986), pp. 8822–8824.

[111] J. C. Slater. "The Self Consistent Field and the Structure of Atoms". In: *Phys. Rev.* 32 (3 Sept. 1928), pp. 339–348.

[112] E. B. Wilson, J. C. Decius, P. C. Cross, and Benson R. Sundheim. "Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra". In: *Journal of The Electrochemical Society* 102.9 (Sept. 1955), 235Ca.

[113] Y. Ozaki, C. Huck, S. Tsuchikawa, and S. B. Engelsen. *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*. Springer, 2021.

[114] "Time-dependent density functional theory". In: *Lecture notes in physics* 706 (2006). Ed. by Miguel A. L. Marques.

[115] John M. Herbert, Ying Zhu, Bushra Alam, and Avik Kumar Ojha. "Time-Dependent Density Functional Theory for X-ray Absorption Spectra: Comparing the Real-Time Approach to Linear Response". In: *Journal of Chemical Theory and Computation* 19 (19 Oct. 2023). doi: 10.1021/acs.jctc.3c00673, pp. 6745–6760.

[116] Nicholas A. Besley. "Density Functional Theory Based Methods for the Calculation of X-ray Spectroscopy". In: *Accounts of Chemical Research* 53 (7 July 2020), pp. 1306–1315.

[117] Ambuj Mehrish, Navonil Majumder, Rishabh Bhardwaj, Rada Mihalcea, and Soujanya Poria. "A Review of Deep Learning Techniques for Speech Processing". In: *Information Fusion* (2023).

[118] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. "Natural Language Processing (almost) from Scratch". In: *Journal of Machine Learning Research* (2011), pp. 2493–2537.

[119] Paul Covington, Jay Adams, and Emre Sargin. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. Association for Computing Machinery, 2016, pp. 191–198.

[120] Pavlo O. Dral. "Quantum Chemistry in the Age of Machine Learning". In: *The Journal of Physical Chemistry Letters* 11 (6 Mar. 2020), pp. 2336–2347.

[121] Manas Sajjan, Junxu Li, Raja Selvarajan, Shree Hari Sureshbabu, Sumit Suresh Kale, Rishabh Gupta, Vinit Singh, and Sabre Kais. "Quantum machine learning for chemistry and physics". In: *Chem. Soc. Rev.* 51 (15 2022), pp. 6475–6573.

[122] Matthew Turk and Alex Pentland. "Eigenfaces for Recognition". In: *Journal of Cognitive Neuroscience* 3.1 (Jan. 1991), pp. 71–86.

[123] Thomas K Landauer, Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis". In: *Discourse Processes* 25.2-3 (1998), pp. 259–284.

[124] Gareth A. Tribello and Piero Gasparotto. "Using Dimensionality Reduction to Analyze Protein Trajectories". In: *Frontiers in Molecular Biosciences* 6 (2019).

[125] Ian T. Jolliffe and Jorge Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (Apr. 2016), p. 20150202.

[126] Gareth A Tribello, C, and Michele Parrinello. "Using sketch-map coordinates to analyze and bias molecular dynamics simulations". In: *Proceedings of the National Academy of Sciences* 109 (14 Apr. 2012), pp. 5196–5201.

[127] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. "Demonstrating the Transferability and the Descriptive Power of Sketch-Map". In: *Journal of Chemical Theory and Computation* 9.3 (Mar. 2013), pp. 1521–1532.

[128] Piero Gasparotto, Robert Horst Meißner, and Michele Ceriotti. "Recognizing Local and Global Structural Motifs at the Atomic Scale". In: *Journal of Chemical Theory and Computation* 14 (2 Feb. 2018), pp. 486–498.

[129] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[130] Vinaitheerthan Renganathan. "Text Mining in Biomedical Domain with Emphasis on Document Clustering". In: *Healthcare Informatics Research* 23 (3 2017), p. 141.

[131] D. Comaniciu and P. Meer. "Mean shift: a robust approach toward feature space analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (2002), pp. 603–619.

[132] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: *ACM Comput. Surv.* 41.3 (July 2009).

[133] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137.

[134] D. Sculley. "Web-Scale k-Means Clustering". In: *Proceedings of the 19th International Conference on World Wide Web* (2010), pp. 1177–1178.

[135] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), pp. 226–231.

[136] Piero Gasparotto, Maria Fischer, Daniele Scopece, Maciej O Liedke, Maik Butterling, Andreas Wagner, Oguz Yildirim, Mathis Trant, Daniele Passerone, Hans J Hug, et al. "Mapping the Structure of Oxygen-Doped Wurtzite Aluminum Nitride Coatings from Ab Initio Random Structure Search and Experiments". In: *ACS Applied Materials & Interfaces* 13.4 (2021), pp. 5762–5771.

[137] Sandip De, Felix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti. "Mapping and classifying molecules from a high-throughput structural database". In: *Journal of cheminformatics* 9.1 (2017), pp. 1–14.

[138] Piero Gasparotto and Michele Ceriotti. "Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond". In: *The Journal of Chemical Physics* 141.17 (2014), p. 174110.

[139] Piero Gasparotto, Davide Bochicchio, Michele Ceriotti, and Giovanni M Pavan. "Identifying and tracking defects in dynamic supramolecular polymers". In: *The Journal of Physical Chemistry B* 124.3 (2019), pp. 589–599.

[140] Annie M Westerlund and Lucie Delemotte. "InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures". In: *Journal of Chemical Theory and Computation* 15.12 (Dec. 2019), pp. 6752–6759.

[141] B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26.

[142] Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. "Graph-based Semi-supervised Learning: A Comprehensive Review". In: *IEEE Transactions on Neural Networks and Learning Systems* (2021), pp. 8174–8194.

[143] C. Lemaréchal. *Cauchy and the Gradient Method*. Math Extra, 2012, pp. 251–254.

[144] Magnus R. Hestenes and Eduard Stiefel. "Methods of Conjugate Gradients for Solving Linear Systems". In: *Journal of Research of the National Bureau of Standards* (1952).

[145] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958.

[146] Ron Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* (1995), pp. 1137–1143.

[147] Andrey N. Tikhonov and Vasiliy Y. Arsenin. "Solutions of ill-posed problems". In: *Mathematics and Its Applications* (1977), pp. 258–270.

[148] Lutz Prechelt. "Early Stopping — But When?" In: *Neural Networks: Tricks of the Trade: Second Edition* (2012), pp. 53–67.

[149] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Deep Learning". In: *MIT Press* (2016).

[150] Henry J. Kelley. "Gradient Theory of Optimal Flight Paths". In: *ARS Journal* 30.10 (1960), pp. 947–954.

[151] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. "What is the best multi-stage architecture for object recognition?" In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 2146–2153.

[152] Vinod Nair and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 807–814.

[153] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Geoffrey Gordon, David Dunson, and Miroslav Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Nov. 2011, pp. 315–323.

[154] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80.

[155] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. "Graph Neural Networks for Social Recommendation". In: *The World Wide Web Conference* (2019), pp. 417–426.

[156] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge". In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (2008), pp. 1247–1250.

[157] Albert-László Barabási and Zoltán N. Oltvai. "Network biology: understanding the cell's functional organization". In: *Nature Reviews Genetics* 5 (2 Feb. 2004), pp. 101–113.

[158] Lingfei Wu, Peng Cui, Jian Pei, Liang Zhao, and Xiaojie Guo. "Graph Neural Networks: Foundation, Frontiers and Applications". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022), pp. 4840–4841.

[159] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. "Neural Message Passing for Quantum Chemistry". In: *Proceedings of the 34th International Conference on Machine Learning*. Proceedings of Machine Learning Research 70 (June 2017), pp. 1263–1272.

[160] Rianne van den Berg, Thomas N. Kipf, and Max Welling. "Graph Convolutional Matrix Completion". In: *ICSCA '21: Proceedings of the 2021 10th International Conference on Software and Computer Applications* (2017), pp. 51–56.

[161] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. "Graph Attention Networks". In: *International Conference on Learning Representations* (2018).

[162] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (June 2017), pp. 6000–6010.

[163] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local Neural Networks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[164] Andrew L. Maas. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: *Proceedings of the 30th International Conference on Machine Learning* (2013).

[165] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. "Relational inductive biases, deep learning, and graph networks". In: *arXiv* (June 2018).

[166] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *Artificial Intelligence* (2018), pp. 1–38.

[167] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. "Explainability in Graph Neural Networks: A Taxonomic Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 5782–5799.

[168] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128 (2 Feb. 2020), pp. 336–359.

[169] Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, and Heiko Hoffmann. "Explainability Methods for Graph Convolutional Neural Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10764–10773.

[170] Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wesley Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. "Evaluating Attribution for Graph Neural Networks". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* 33 (2020), pp. 5898–5910.

[171] Sam Sattarzadeh, Mahesh Sudhakar, Anthony Lem, Shervin Mehryar, K. N. Plataniotis, Jongseong Jang, Hyunwoo Kim, Yeonjeong Jeong, Sangmin Lee, and Kyunghoon Bae. "Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation". In: *arXiv* (2020).

[172] Pouya Pezeshkpour, Sarthak Jain, Byron C. Wallace, and Sameer Singh. "An Empirical Comparison of Instance Attribution Methods for NLP". In: *arXiv* (2021).

[173] Kevin McCloskey, Ankur Taly, Federico Monti, Michael P. Brenner, and Lucy J. Colwell. "Using attribution to decode binding mechanism in neural network models for chemistry". In: *Proceedings of the National Academy of Sciences* 116 (24 June 2019), pp. 11624–11629.

[174] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features Through Propagating Activation Differences". In: *ICML'17: Proceedings of the 34th International Conference on Machine Learning* (Apr. 2017), pp. 3145–3153.

[175] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning Deep Features for Discriminative Localization". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Dec. 2015), pp. 2921–2929.

[176] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* (Oct. 2016), pp. 618–626.

[177] Alok Sharma, Artem Lysenko, Keith A Boroevich, Edwin Vans, and Tatsuhiko Tsunoda. "DeepFeature: feature selection in nonimage data using convolutional neural network". In: *Briefings in Bioinformatics* 22 (6 Nov. 2021), pp. 1–12.

[178] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. "Evaluating explainability for graph neural networks". In: *Scientific Data* 10 (1 Mar. 2023), p. 144.

[179] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. "Machine learning for molecular and materials science". In: *Nature* (2018), pp. 547–555.

[180] Mirko Torrisi, Gianluca Pollastri, and Quan Le. "Deep learning methods in protein structure prediction". In: *Computational and structural biotechnology journal* 18 (Jan. 2020), pp. 1301–1310.

[181] Charles McGill, Michael Forsuelo, Yanfei Guan, and William H. Green. "Predicting Infrared Spectra with Message Passing Neural Networks". In: *Journal of chemical information and modeling* (2021), pp. 2594–2609.

[182] Frank Noé and Cecilia Clementi. "Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods". In: *Current Opinion in Structural Biology* 43 (2017), pp. 141–147.

[183] Peter G Bolhuis and Christoph Dellago. "Trajectory-Based Rare Event Simulations". In: *Reviews in Computational Chemistry* (Sept. 2010), pp. 111–210.

[184] Christoph Dellago, Peter G. Bolhuis, and P. L. Geissler. "Transition path sampling methods". In: *Advances in Chemical Physics* 703 (2006), pp. 349–391.

[185] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning". In: *Physical Review Letters* 108.5 (Jan. 2012), p. 058301.

[186] Ali Sadeghi, S Alireza Ghasemi, Bastian Schaefer, Stephan Mohr, Markus A Lill, and Stefan Goedecker. "Metrics for measuring distances in configuration spaces". In: *The Journal of Chemical Physics* 139.18 (Nov. 2013), p. 184118.

[187] Giacomo Fiorin, Michael L. Klein, and Jérôme Hénin. "Using collective variables to drive molecular dynamics simulations". In: *Molecular Physics* 111.22-23 (Dec. 2013), pp. 3345–3362.

[188] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. "Comparing molecules and solids across structural and alchemical space". In: *Physical Chemistry Chemical Physics* 18.20 (2016), pp. 13754–13769.

[189] Albert P Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments". In: *Physical Review B* 87.18 (May 2013), p. 184115.

[190] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. "Comparing molecules and solids across structural and alchemical space". In: *Physical Chemistry Chemical Physics* 18.20 (2016), pp. 13754–13769.

[191] Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. "Physics-inspired structural representations for molecules and materials". In: *Chemical Reviews* 121.16 (2021), pp. 9759–9815.

[192] Benjamin A Helfrecht, Rose K Cersonsky, Guillaume Fraux, and Michele Ceriotti. "Structure-property maps with Kernel principal covariates regression". In: *Machine Learning: Science and Technology* 1.4 (Nov. 2020), p. 045021.

[193] Benjamin A Helfrecht, Piero Gasparotto, Federico Giberti, and Michele Ceriotti. "Atomic motif recognition in (bio) polymers: Benchmarks from the protein data bank". In: *Frontiers in molecular biosciences* 6 (2019), p. 24.

[194] Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. "Physics-Inspired Structural Representations for Molecules and Materials". In: *Chemical Reviews* 121.16 (Aug. 2021), pp. 9759–9815.

[195] Félix Musil, Sandip De, Jack Yang, Joshua E. Campbell, Graeme M. Day, and Michele Ceriotti. "Machine learning for the structure–energy–property landscapes of molecular crystals". In: *Chemical Science* 9.5 (2018), pp. 1289–1300.

[196] Sandip De, Felix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti. "Mapping and classifying molecules from a high-throughput structural database". In: *Journal of Cheminformatics* 9.1 (Dec. 2017), p. 6.

[197] Jan G. Rittig, Qinghe Gao, Manuel Dahmen, Alexander Mitsos, and Artur M. Schweidtmann. "Graph neural networks for the prediction of molecular structure-property relationships". In: *arXiv* (2022).

[198] Yair Litman, Jeremy O Richardson, Takashi Kumagai, and Mariana Rossi. "Elucidating the Nuclear Quantum Dynamics of Intramolecular Double Hydrogen Transfer in Porphycene". In: *Journal of the American Chemical Society* 141.6 (Feb. 2019), pp. 2526–2534.

[199] Robert H Meißner, Julian Schneider, Peter Schiffels, and Lucio Colombi Ciacchi. "Computational Prediction of Circular Dichroism Spectra and Quantification of Helicity Loss upon Peptide Adsorption on Silica". In: *Langmuir* 30.12 (Apr. 2014), pp. 3487–3494.

[200] Carlos R. Baiz, Bartosz Błasiak, Jens Bredenbeck, Minhaeng Cho, Jun-Ho Choi, Steven A. Corcelli, Arend G. Dijkstra, Chi-Jui Feng, Sean Garrett-Roe, Nien-Hui Ge, Magnus W. D. Hanson-Heine, Jonathan D. Hirst, Thomas L. C. Jansen, Kijeong Kwac, Kevin J. Kubarych, Casey H. Londergan, Hiroaki Maekawa, Mike Reppert, Shinji Saito, Santanu Roy, James L. Skinner, Gerhard Stock, John E. Straub, Megan C. Thielges, Keisuke Tominaga, Andrei Tokmakoff, Hajime Torii, Lu Wang, Lauren J. Webb, and Martin T. Zanni. "Vibrational Spectroscopic Map, Vibrational Spectroscopy, and Intermolecular Interaction". In: *Chemical Reviews* 120.15 (June 2020), pp. 7152–7218.

[201] Kim E. van Adrichem and Thomas L. C. Jansen. "AIM: A Mapping Program for Infrared Spectroscopy of Proteins". In: *Journal of Chemical Theory and Computation* 18.5 (2022), pp. 3089–3098.

[202] Roman D. Gorbunov and Gerhard Stock. "Ab initio based building block model of amide I vibrations in peptides". In: *Chemical Physics Letters* 437.4-6 (Apr. 2007), pp. 272–276.

[203] Y.-S. Lin, J. M. Shorb, P. Mukherjee, M. T. Zanni, and J. L. Skinner. "Empirical Amide I Vibrational Frequency Map: Application to 2D-IR Line Shapes for Isotope-Edited Membrane Peptide Bundles". In: *The Journal of Physical Chemistry B* 113.3 (2009), pp. 592–602.

[204] Alexei A. Kananenka, Kun Yao, Steven A. Corcelli, and J. L. Skinner. "Machine Learning for Vibrational Spectroscopic Maps". In: *Journal of Chemical Theory and Computation* 15.12 (Oct. 2019), pp. 6850–6858.

[205] B. M. Auer and J. L. Skinner. "IR and Raman spectra of liquid water: Theory and interpretation". In: *The Journal of Chemical Physics* 128.22 (June 2008), p. 224511.

[206] Carsten Baldauf and Mariana Rossi. "Going clean: structure and dynamics of peptides in the gas phase and paths to solvation". In: *Journal of Physics: Condensed Matter* 27.49 (Nov. 2015), p. 493002.

[207] Judit Sztáray, Antony Memboeuf, László Drahos, and Károly Vékey. "Leucine enkephalin-A mass spectrometry standard". In: *Mass Spectrometry Reviews* 30.2 (July 2010), pp. 298–320.

[208] Xianmei Cai and Chhabil Dass. "Structural characterization of methionine and leucine enkephalins by hydrogen/deuterium exchange and electrospray ionization tandem mass spectrometry". In: *Rapid Communications in Mass Spectrometry* 19.1 (2004), pp. 1–8.

[209] Tobias N Wassermann, Oleg V Boyarkin, Béla Paizs, and Thomas R Rizzo. "Conformation-Specific Spectroscopy of Peptide Fragment Ions in a Low-Temperature Ion Trap". In: *Journal of the American Society for Mass Spectrometry* 23.6 (June 2012), pp. 1029–1045.

[210] David A Evans, David J Wales, Brian C Dian, and Timothy S Zwier. "The dynamics of conformational isomerization in flexible biomolecules. II. Simulating isomerizations in a supersonic free jet with master equation dynamics". In: *The Journal of Chemical Physics* 120.1 (Jan. 2004), pp. 148–157.

[211] Hongbao Li, Jun Jiang, and Yi Luo. "Identification of the protonation site of gaseous triglycine: the cis-peptide bond conformation as the global minimum". In: *Physical Chemistry Chemical Physics* 19.23 (2017), pp. 15030–15038.

[212] Franziska Schubert, Mariana Rossi, Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Frank Filsinger, Peter Kupser, Gerard Meijer, Mario Salwiczek, Beate Koksch, Matthias Scheffler, and Volker Blum. "Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala 19 -Lys + H + vs. Ac-Lys-Ala 19 + H + and the current reach of DFT". In: *Physical Chemistry Chemical Physics* 17.11 (Mar. 2015), pp. 7373–7385.

[213] Bijyalaxmi Athokpam, Sai G. Ramesh, and Ross H. McKenzie. "Effect of hydrogen bonding on the infrared absorption intensity of OH stretch vibrations". In: *Chemical Physics* 488-489 (May 2017), pp. 43–54.

[214] Nataliya S. Myshakina, Zeeshan Ahmed, and Sanford A. Asher. "Dependence of Amide Vibrations on Hydrogen Bonding". In: *The Journal of Physical Chemistry B* 112.38 (Sept. 2008), pp. 11873–11877.

[215] Jos Oomens, Boris G. Sartakov, Gerard Meijer, and Gert von Helden. "Gas-phase infrared multiple photon dissociation spectroscopy of mass-selected molecular ions". In: *International Journal of Mass Spectrometry* 254.1-2 (July 2006), pp. 1–19.

[216] Yan Ji, Xiaoliang Yang, Zhi Ji, Linhui Zhu, Nana Ma, Dejun Chen, Xianbin Jia, Junming Tang, and Yilin Cao. "DFT-Calculated IR Spectrum Amide I, II, and III Band Contributions of N -Methylacetamide Fine Components". In: *ACS Omega* 5.15 (Apr. 2020), pp. 8572–8578.

[217] Thomas Weymuth, Christoph R Jacob, and Markus Reiher. "A Local-Mode Model for Understanding the Dependence of the Extended Amide III Vibrations on Protein Secondary Structure". In: *The Journal of Physical Chemistry B* 114.32 (Aug. 2010), pp. 10649–10660.

[218] Andreas Barth. "Infrared spectroscopy of proteins". In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1767.9 (Sept. 2007), pp. 1073–1101.

[219] Evan G. Buchanan, William H. James, Soo Hyuk Choi, Li Guo, Samuel H. Gellman, Christian W. Müller, and Timothy S. Zwier. "Single-conformation infrared spectra of model peptides in the amide I and amide II regions: Experiment-based determination of local mode frequencies and inter-mode coupling". In: *The Journal of Chemical Physics* 137.9 (Sept. 2012), p. 094301.

[220] Kaustubh Joshi, David Semrouni, Gilles Ohanessian, and Carine Clavagu. "Structures and IR Spectra of the Gramicidin S Peptide: Pushing the Quest for Low-Energy Conformations". In: *J. Phys. Chem. B* 116 (2012), pp. 483–490.

[221] Ronghu Wu and Terry B. McMahon. "Infrared Multiple Photon Dissociation Spectroscopy as Structural Confirmation for GlyGlyGlyH + and AlaAlaAlaH + in the Gas Phase. Evidence for Amide Oxygen as the Protonation Site". In: *Journal of the American Chemical Society* 129.37 (Sept. 2007), pp. 11312–11313.

[222] Marie L. Laury, Scott E. Boesch, Ian Haken, Pankaj Sinha, Ralph A. Wheeler, and Angela K. Wilson. "Harmonic vibrational frequencies: Scale factors for pure, hybrid, hybrid meta, and double-hybrid functionals in conjunction with correlation consistent basis sets". In: *Journal of Computational Chemistry* 32.11 (Aug. 2011), pp. 2339–2347.

[223] Jacob C Dean, Evan G Buchanan, and Timothy S Zwier. "Mixed 14/16 Helices in the Gas Phase: Conformation-Specific Spectroscopy of Z-(Gly) n , n = 1, 3, 5". In: *Journal of the American Chemical Society* 134.41 (Oct. 2012), pp. 17186–17201.

[224] Patrick S Walsh, Ryoji Kusaka, Evan G Buchanan, William H. James, Brian F Fisher, Samuel H Gellman, and Timothy S Zwier. "Cyclic Constraints on Conformational Flexibility in γ-Peptides: Conformation Specific IR and UV Spectroscopy". In: *The Journal of Physical Chemistry A* 117.47 (Nov. 2013), pp. 12350–12362.

[225] Robert Horst Meißner, Gang Wei, and Lucio Colombi Ciacchi. "Estimation of the free energy of adsorption of a polypeptide on amorphous SiO 2 from molecular dynamics simulations and force spectroscopy experiments". In: *Soft Matter* 11.31 (2015), pp. 6254–6265.

[226] Tapta Kanchan Roy and R. Benny Gerber. "Vibrational self-consistent field calculations for spectroscopy of biological molecules: new algorithmic developments and applications". In: *Physical Chemistry Chemical Physics* 15.24 (2013), p. 9468.

[227] Alexander Goscinski, Guillaume Fraux, Giulio Imbalzano, and Michele Ceriotti. "The role of feature space in atomistic learning". In: *Machine Learning: Science and Technology* 2.2 (2021), p. 025028.

[228] Jaime A Stearns, Caroline Seaiby, Oleg V Boyarkin, and Thomas R Rizzo. "Spectroscopy and conformational preferences of gas-phase helices". In: *Phys. Chem. Chem. Phys.* 11.1 (2009), pp. 125–132.

[229] Jaime A. Stearns, Oleg V. Boyarkin, and Thomas R. Rizzo. "Effects of N-Terminus Substitution on the Structure and Spectroscopy of Gas-Phase Helices". In: *Chimia* 62.4 (Apr. 2008), p. 240.

[230] S. A. Krasnikov, A. B. Preobrajenski, N. N. Sergeeva, M. M. Brzhezinskaya, M. A. Nesterov, A. A. Cafolla, M. O. Senge, and A. S. Vinogradov. "Ni(II)-porphyrins with different ligands on the porphyrin ring, significant change of XAS spectra for different ligands on the ring". In: *Chemical Physics* 332 (2-3 2007), pp. 318–324.

[231] Ana Guilherme Buzanich. "Recent developments of X-ray absorption spectroscopy as analytical tool for biological and biomedical applications". In: *X-Ray Spectrometry* 51.3 (2022), pp. 294–303.

[232] Federico Frateloreto, Francesco Tavani, Marika Di Berto Mancini, Daniele Del Giudice, Giorgio Capocasa, Isabelle Kieffer, Osvaldo Lanzalunga, Stefano Di Stefano, and Paola D'Angelo. "Following a Silent Metal Ion: A Combined X-ray Absorption and Nuclear Magnetic Resonance Spectroscopic Study of the Zn2+ Cation Dissipative Translocation between Two Different Ligands". In: *The Journal of Physical Chemistry Letters* 13 (24 June 2022), pp. 5522–5529.

[233] P. Eisenberger and B. M. Kincaid. "EXAFS: New Horizons in Structure Determinations". In: *Science* 200 (4349 June 1978), pp. 1441–1447.

[234] Grant S. Henderson, Frank M.F. de Groot, and Benjamin J.A. Moulton. "X-ray Absorption Near-Edge Structure (XANES) Spectroscopy". In: *Reviews in Mineralogy and Geochemistry* 78.1 (Jan. 2014), pp. 75–138.

[235] George E. Cutsail Iii and Serena DeBeer. "Challenges and Opportunities for Applications of Advanced X-ray Spectroscopy in Catalysis Research". In: *ACS Catalysis* 12.10 (May 2022), pp. 5864–5886.

[236] J. J. Rehr and R. C. Albers. "Theoretical approaches to x-ray absorption fine structure". In: *Reviews of Modern Physics* 72 (3 July 2000), pp. 621–654.

[237] C. D. Rankine, M. M. M. Madkhali, and T. J. Penfold. "A Deep Neural Network for the Rapid Prediction of X-ray Absorption Spectra". In: *The Journal of Physical Chemistry A* 124 (21 May 2020), pp. 4263–4270.

[238] Matthew R. Carbone, Mehmet Topsakal, Deyu Lu, and Shinjae Yoo. "Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy". In: *Phys. Rev. Lett.* 124 (15 Apr. 2020), p. 156401.

[239] Hiram A. Castillo-Michel, Camille Larue, Ana E. Pradas del Real, Marine Cotte, and Geraldine Sarret. "Practical review on the use of synchrotron based micro- and nano- X-ray fluorescence mapping and X-ray absorption spectroscopy to investigate the interactions between plants and engineered nanomaterials". In: *Plant Physiology and Biochemistry.* Effects of Nanomaterials in Plants 110 (Jan. 2017), pp. 13–32.

[240] Patric Zimmermann, Sergey Peredkov, Paula Macarena Abdala, Serena DeBeer, Moniek Tromp, Christoph Müller, and Jeroen A. van Bokhoven. "Modern X-ray spectroscopy: XAS and XES in the laboratory". In: *Coordination Chemistry Reviews* 423 (Nov. 2020), p. 213466.

[241] Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning". In: *arXiv* (2017).

[242] D. Egorov, L. Schwob, M. Lalande, R. Hoekstra, and T. Schlathölter. "Near edge X-ray absorption mass spectrometry of gas phase proteins: the influence of protein size". In: *Phys. Chem. Chem. Phys.* 18 (37 2016), pp. 26213–26223.

[243] Stefan G. Minasian, Jason M. Keith, Enrique R. Batista, Kevin S. Boland, Stosh A. Kozimor, Richard L. Martin, David K. Shuh, Tolek Tyliszczak, and Louis J. Vernon. "Carbon K-Edge X-ray Absorption Spectroscopy and Time-Dependent Density Functional Theory Examination of Metal–Carbon Bonding in Metallocene Dichlorides". In: *Journal of the American Chemical Society* 135 (39 Oct. 2013), pp. 14731–14740.

[244] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. "Quantum chemistry structures and properties of 134 kilo molecules". In: *Scientific Data* 1.1 (Aug. 2014), p. 140022.

[245] M. Petersilka, U. J. Gossmann, and E. K. U. Gross. "Excitation Energies from Time-Dependent Density-Functional Theory". In: *Phys. Rev. Lett.* 76 (8 Feb. 1996), pp. 1212–1215.

[246] Frank Neese. "The ORCA program system". In: *WIREs Computational Molecular Science* 2.1 (Jan. 2012), pp. 73–78.

[247] Florian Weigend and Reinhart Ahlrichs. "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy". In: *Physical Chemistry Chemical Physics* 7 (18 Sept. 2005), p. 3297.

[248] "RDKit: Open-source cheminformatics". In: *http://www.rdkit.org* (Accessed April 16, 2022).

[249] Matthias Fey and Jan E. Lenssen. "Fast Graph Representation Learning with PyTorch Geometric". In: *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).

[250] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. "Convolutional Networks on Graphs for Learning Molecular Fingerprints". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15 (2015), pp. 2224–2232.

[251] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. "Graph Attention Networks". In: *International Conference on Learning Representations* (Oct. 2017).

[252] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations* (Nov. 2017).

[253] Jérémie Despraz, Stéphane Gomez, Héctor F. Satizábal, and Carlos Andrés Peña-Reyes. "Towards a Better Understanding of Deep Neural Networks Representations using Deep Generative Networks:" in: *Proceedings of the 9th International Joint Conference on Computational Intelligence* (2017), pp. 215–222.

[254] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. "Towards better understanding of gradient-based attribution methods for Deep Neural Networks". In: *International Conference on Learning Representations* (2018).

[255] José Jiménez-Luna, Miha Skalic, and Nils Weskamp. "Benchmarking Molecular Feature Attribution Methods with Activity Cliffs". In: *Journal of Chemical Information and Modeling* 62 (2 Jan. 2022), pp. 274–283.

[256] Bowen Tang, Skyler T. Kramer, Meijuan Fang, Yingkun Qiu, Zhen Wu, and Dong Xu. "A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility". In: *Journal of Cheminformatics* 12 (1 Dec. 2020), p. 15.

[257] Iulia Emilia Brumboiu and Thomas Fransson. "Core–hole delocalization for modeling x-ray spectroscopies: A cautionary tale". In: *The Journal of Chemical Physics* 156.21 (June 2022). 214109.

[258] Andrew P. Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern Recognition* 30.7 (1997), pp. 1145–1159.

[259] Henry Heberle, Linlin Zhao, Sebastian Schmidt, Thomas Wolf, and Julian Heinrich. "XSMILES: interactive visualization for molecules, SMILES and XAI attribution scores". In: *Journal of Cheminformatics* 15.1 (Jan. 2023), p. 2.

[260] Sangho Lee, Hyunwoo Park, Chihyeon Choi, Wonjoon Kim, Ki Kang Kim, Young-Kyu Han, Joohoon Kang, Chang-Jong Kang, and Youngdoo Son. "Multi-order graph attention network for water solubility prediction and interpretation". In: *Scientific Reports* 13 (1 Mar. 2023), p. 957.

[261] M. Withnall, E. Lindelöf, O. Engkvist, and H. Chen. "Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction". In: *Journal of Cheminformatics* 12 (1 Jan. 2020), pp. 1–18.

[262] Federica Frati, Myrtille O. J. Y. Hunault, and Frank M. F. de Groot. "Oxygen K-edge X-ray Absorption Spectra". In: *Chemical Reviews* 120 (9 May 2020), pp. 4056–4110.

[263] Marcel Risch, Dulce M. Morales, Javier Villalobos, and Denis Antipin. "What X-Ray Absorption Spectroscopy Can Tell Us About the Active State of Earth-Abundant Electrocatalysts for the Oxygen Evolution Reaction". In: *Angewandte Chemie International Edition* 61 (50 Dec. 2022), e202211949.

[264] Sarah Wiegreffe and Yuval Pinter. "Attention is not not explanation". In: *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2019), pp. 11–20.

[265] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. *Modeling Relational Data with Graph Convolutional Networks*. Cham, 2018.

[266] Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y. Hammerla. "Relational Graph Attention Networks". In: *arXiv* (2019).

[267] Seongjun Yun, Minbyul Jeong, Sungdong Yoo, Seunghun Lee, Sean S. Yi, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. "Graph Transformer Networks: Learning meta-path graphs to improve GNNs". In: *Neural Networks* 153 (2022), pp. 104–119.

[268] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning* 70 (June 2017), pp. 3319–3328.

[269] Frank de Groot. "X-ray absorption and dichroism of transition metal compounds". In: *AIP Conference Proceedings* 389.1 (Jan. 1997), pp. 497–520.

[270] Zhantao Chen, Nina Andrejevic, Nathan C. Drucker, Thanh Nguyen, R. Patrick Xian, Tess Smidt, Yao Wang, Ralph Ernstorfer, D. Alan Tennant, Maria Chan, and Mingda Li. "Machine learning on neutron and x-ray scattering and spectroscopies". In: *Chemical Physics Reviews* 2.3 (July 2021). 031301.

[271] Animesh Ghose, Mikhail Segal, Fanchen Meng, Zhu Liang, Mark S. Hybertsen, Xiaohui Qu, Eli Stavitski, Shinjae Yoo, Deyu Lu, and Matthew R. Carbone. "Uncertainty-aware predictions of molecular x-ray absorption spectra using neural network ensembles". In: *Physical Review Research* 5 (1 Mar. 2023), p. 013180.

[272] M. Petersilka, U. J. Gossmann, and E. K. U. Gross. "Excitation Energies from Time-Dependent Density-Functional Theory". In: *Phys. Rev. Lett.* 76 (8 Feb. 1996), pp. 1212–1215.

[273] J. F. Dobson E. K. U. Gross and Petersilka M. *Density functional theory of time-dependent phenomena*. Vol. 181(81). 1996.

[274] Bryan Edman Sundahl. "Time Dependent Density-Functional Theory - Linear Response". In: *master thesis* (2013).

[275] Robert van Leeuwen. "Mapping from Densities to Potentials in Time-Dependent Density-Functional Theory". In: *Phys. Rev. Lett.* 82 (19 May 1999), pp. 3863–3866.

[276] Mark E. Casida. "Time-Dependent Density Functional Response Theory for Molecules". In: *Recent Advances in Density Functional Methods* (Nov. 1995), pp. 155–192.

A

A.1 Harmonic approximation for IR spectra calculation

As mentioned in section 2.2.3.1 in chapter 2, the geometry optimisation is first performed using DFT in order to calculate the harmonic approximation for IR spectra and obtain the vibrational frequencies. The following details of harmonic approximation is written based on ref [113]. The molecule's potential energy $V(\mathbf{Q})$ is a many-parameter function of its atomic coordinates, represented as the vector $\mathbf{Q} = \{q_1, q_2, \ldots, q_{3N-N_{inv}}\}$ where $3N$ is the total number of degrees of freedom and N is the number of atoms. N_{inv} represents the translational and rotational degrees of freedom which are set apart from the vibrational degrees of freedom (vibrational modes). Geometry optimisation is then a purely mathematical optimisation problem of finding $\mathbf{Q}$ that minimizes $V(\mathbf{Q})$. For a stationary point on the PES, the energy gradient which is the derivative of the energy with respect to all atomic coordinates $\frac{\partial V}{\partial q_i}$, is zero. In case of a polyatomic oscillator, the vibrational Hamiltonian can be written as,

$$H = -\frac{1}{2}\sum_i \frac{1}{m_i}\frac{\partial^2}{\partial q_i^2} + \frac{1}{2}\sum_i m_i\omega_{0i}^2 q_i^2 + \sum_{i\leq j\leq k} k_{ijk}q_iq_jq_k + \sum_{i\leq j\leq k\leq l} k_{ijkl}q_iq_jq_kq_l + \cdots \tag{A.1}$$

where m_i is the reduced mass of the i-th normal mode and ω_{0i} is the corresponding harmonic frequency which is given as,

$$\omega_{0i} = \sqrt{\frac{k_i}{m_i}} \tag{A.2}$$

where k_i is the harmonic force constant. The third and higher terms in the expansion describe the anharmonic contributions to the vibrational Hamiltonian via the associated cubic and quartic force

constants, k_{ijk} and k_{ijkl}, respectively. Taking into account anharmonic contributions increases the complexity of obtaining vibrational frequencies. However, the contributions from the anharmonic terms in equation A.1 is relatively low for a number of molecular vibrations and hence harmonic approximation is constructed. Based on this approximation, no coupling between modes is permitted. Therefore, k_{ijk}, k_{ijkl}, and higher-order constants are set to zero. In other words, the normal vibrations of harmonic oscillator are entirely independent [113]. As the next step, the potential energy in the vicinity of the equilibrium is approximated as a Taylor series,

$$V(\mathbf{Q}) = V_0(\mathbf{Q}) + \Delta\mathbf{Q}^T \cdot g(\mathbf{Q}) + \frac{1}{2}\Delta\mathbf{Q}^T\mathbf{H}\Delta\mathbf{Q} + \cdots \tag{A.3}$$

with the higher terms in the expansion being neglected. At a stationary point on the PES (minima and transition states), the gradient $g(\mathbf{Q})$ and hence the second term of equation A.3 is equal to zero. This results in a quadratic function as the approximation of the potential which corresponds to a harmonic potential. The mass-weighted second-derivative matrix of the potential, or mass-weighted Hessian matrix $\mathbf{H}$ is introduced, which elements are written as,

$$H^{mw}_{i,j} = \frac{1}{\sqrt{m_i m_j}} \frac{\partial^2 V(\mathbf{Q})}{\partial q_i \partial q_j}. \tag{A.4}$$

Diagonalization of the mass-weighted Hessian matrix yields a matrix with $3N - N_{inv}$ columns which consists of orthonormal eigenvectors describing the vibrational motion of the system within the harmonic approximation, the so-called mass-weighted normal modes. The $3N - N_{inv}$ diagonal elements of the eigenvalue matrix $\mathbf{h}$ are proportional to the square frequency of the associated normal mode.

$$\mathbf{h} = \mathbf{U}^T\mathbf{H}\mathbf{U} \tag{A.5}$$

A.2 Linear-response time-dependent density functional theory

In linear-response TDDFT, the external perturbation is small enough that the system's response can be described in terms of linear response functions. All the properties of DFT can be used to calculate the ground-state density of the system. The linear response function is then constructed from the ground-state density and the exchange-correlation functional used in TDDFT.

A system of N electrons with coordinates $\underline{\mathbf{r}} = (\mathbf{r}_1 \ldots \mathbf{r}_N)$ is known to obey the time-dependent Schrödinger equation;

$$\mathrm{i}\frac{\partial}{\partial t}\Psi(\underline{\mathbf{r}}, t) = \hat{H}(\underline{\mathbf{r}}, t)\Psi(\underline{\mathbf{r}}, t), \tag{A.6}$$

$|\Psi(\underline{\mathbf{r}}, t)|^2$ is interpreted as the probability of finding the electrons in positions $\mathbf{r}$. The Hamiltonian can be written in the form;

$$\hat{T}(\mathbf{r}) + \hat{W}(\mathbf{r}) + \hat{V}_{ext}(\mathbf{r}, t). \tag{A.7}$$

in which the kinetic energy of the electrons is written as;

$$\hat{T}(\mathbf{r}) = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2, \tag{A.8}$$

$\hat{W}$ accounts for the Coulomb repulsion between electrons;

$$\hat{W}(\mathbf{r}) = -\frac{1}{2}\sum_{i,j=1(i\neq j)}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{A.9}$$

$\hat{V}_{ext}(\mathbf{r},t)$ represents the influence of a generic, time-dependent potential on the electrons.

Full solutions to the time-dependent Kohn-Sham (KS) equation can be expensive to calculate for even medium-sized systems. Linear response of the system is considered as the first approximation to the full solution to alleviate this cost. It has been shown that linear response can also produce exact excitation energies in the limit of the exact exchange-correlation kernel [272]. The linear response can be calculated using the perturbation theory. Following the work of [273] and according to [274], this can be shown as follows. Response of the system to a small perturbation $v(\boldsymbol{r},t)$ can be written as a Taylor series;

$$\rho(\mathbf{r},t) - \rho_0(\mathbf{r},t) = \rho_1(\mathbf{r},t) + \rho_2(\mathbf{r},t) + \rho_3(\mathbf{r},t) + \cdots \tag{A.10}$$

where the subscripts indicate the order of the external perturbation and $\rho_0(\mathbf{r},t)$ is the ground-state density of the unperturbed system. The first order response can then be written as;

$$\rho_1(\mathbf{r},t) = \iint \chi\left(\mathbf{r},t,\mathbf{r}',t\right) v_1\left(\mathbf{r}',t\right) d^3r'dt \tag{A.11}$$

where χ is the density response of the interacting system;

$$\chi\left(\mathbf{r},t,\mathbf{r}',t'\right) = \left.\frac{\delta\rho\left[v_{ext}\right](\mathbf{r},t)}{\delta v_{ext}\left(\mathbf{r}',t\right)}\right|_{v_0}. \tag{A.12}$$

The time-dependent KS equation has the form of;

$$i\frac{\partial\varphi_j(\mathbf{r},t)}{\partial t} = \left[-\frac{\nabla^2}{2} + v_{\mathrm{KS}}[n](\mathbf{r},t)\right]\varphi_j(\mathbf{r},t) \tag{A.13}$$

in which $n(\mathbf{r},t)$ is the density of both the fictitious system and the physical system and is written as;

$$n(\mathbf{r},t) = \sum_{j=1}^{N}\left|\varphi_j(\mathbf{r},t)\right|^2. \tag{A.14}$$

The one-to-one mapping between densities and potentials guarantees a unique local effective potential $v_{\mathrm{KS}}[\rho](\mathbf{r},\mathbf{t})$ for the non-interacting system which generates the same density as the

interacting system [275]. Therefore, the potential $v_{\mathrm{KS}}(\mathbf{r},\mathbf{t})$ is uniquely determined from this density and it can be defined as;

$$v_{\mathrm{KS}} = v_{\mathrm{ext}}(\mathbf{r},t) + v_{\mathrm{H}}(\mathbf{r},t) + f_{\mathrm{Xc}}(\mathbf{r},t). \tag{A.15}$$

$v_{\mathrm{H}}(\mathbf{r},t)$ is the time-dependent Hartree potential;

$$v_{\mathrm{H}}[\rho](\mathbf{r},t) = \int \frac{\rho\left(\mathbf{r}',t\right)}{|\mathbf{r}-\mathbf{r}'|} \tag{A.16}$$

And $f_{\mathrm{Xc}}(\mathbf{r},t)$ is the exchange-correlation kernel. The exchange-correlation kernel is unknown (similar to the exchange-correlation potential in ground state DFT) and the TDDFT equation will yield exact results in the limit that the exchange-correlation kernel becomes exactly known. By applying the chain rule for the functional;

$$\chi\left(\mathbf{r},t,\mathbf{r}',t'\right) = \iint \frac{\delta\rho(\mathbf{r},t)}{\delta v_{\mathrm{KS}}(\mathbf{y},\tau)} \frac{\delta v_{\mathrm{KS}}(\mathbf{y},\tau)}{\delta v_{ext}\left(\mathbf{r}',t'\right)}\bigg|_{v_0} d^3y d\tau. \tag{A.17}$$

As the next step, we take the functional derivative of equation A.15 with respect to the external potential;

$$\frac{\delta v_{\mathrm{KS}}(\mathbf{r},t)}{\delta v_{ext}\left(\mathbf{r}',t'\right)} = \delta\left(\mathbf{r}-\mathbf{r}'\right)\delta\left(t-t'\right) + \iint \left(\frac{\delta(t-\tau)}{|\mathbf{r}-\mathbf{y}|} + \frac{\delta v_{xc}(\mathbf{r},t)}{\delta\rho(\mathbf{y},\tau)}\right) \frac{\delta\rho(\mathbf{y},\tau)}{\delta v_{ext}\left(\mathbf{r}',t'\right)} d^3y d\tau \tag{A.18}$$

inserting equation A.18 into equation A.17 gives;

$$\begin{aligned}\chi\left(\mathbf{r},t,\mathbf{r}',t'\right) = \chi_{\mathrm{KS}}\left(\mathbf{r},t,\mathbf{r}',t'\right) + \int d^3y \int d\tau \int d^3y' \int d\tau' \chi_{\mathrm{KS}}\left(\mathbf{r},t,\mathbf{r}',t'\right) \\ \times \left(\frac{\delta\left(\tau-\tau'\right)}{|\mathbf{y}-\mathbf{y}'|} + f_{xc}\left[\rho_0\right]\left(\mathbf{y},\tau,\mathbf{y}',\tau'\right)\right) \chi\left(\mathbf{y}',\tau',\mathbf{r}',t'\right)\end{aligned} \tag{A.19}$$

in which $\chi_{\mathrm{KS}}\left(\mathbf{r},t,\mathbf{r}',t'\right)$ is the Kohn-Sham response function;

$$\chi_{\mathrm{KS}}\left(\mathbf{r},t,\mathbf{r}',t\right) := \frac{\delta\rho\left[v_{\mathrm{KS}}\right](\mathbf{r},t)}{\delta v_{\mathrm{KS}}\left(\mathbf{r}',t'\right)}\bigg|_{v_{\mathrm{KS}}[\rho_0]} \tag{A.20}$$

and $f_{xc}\left[\rho_0\right](\mathbf{r},t,\mathbf{r}',t')$ is the exchange-correlation kernel;

$$f_{xc}\left[\rho_0\right](\mathbf{r},t,\mathbf{r}',t') := \frac{\delta v_{xc}[\rho](\mathbf{r},t)}{\delta\rho\left(\mathbf{r}',t'\right)}\bigg|_{\rho_0} \tag{A.21}$$

Equation A.19 relates the fictitious non-interacting system to the physically relevant interacting system and it is the key equation in TDDFT theory. By inserting equation A.19 into equation A.11, the linear response of the density can be written as;

$$\rho_1(\mathbf{r},t) = \iint \chi_{\mathrm{KS}}\left(\mathbf{r},t,\mathbf{r}',t]\right) v_{\mathrm{KS},1}\left(\mathbf{r}',t\right) d^3r' dt' \tag{A.22}$$

where the effective potential is;

$$v_{\mathrm{KS},1}\left(\mathbf{r}',t\right)=v_1(\mathbf{r},t)+\int\frac{\rho_1\left(\mathbf{r}',t\right)}{|\mathbf{r}-\mathbf{r}'|}d^3r'+\iint f_{xc}\left[\rho_0\right]\left(\mathbf{r},t,\mathbf{r}',t'\right)\rho_1(\mathbf{r};,t)d^3r'dt' \tag{A.23}$$

and it holds the external perturbation $v_1(\mathbf{r},t)$, the Hartree Coulomb potential and the unknown exchange-correlation potential. To this point, all equations were considered in real-space. For calculating properties such as polarizabilities or excitation energies, the same equations should be considered in frequency space. Therefore, a Fourier transform of equations A.22 and A.23 must be performed for transition into frequency space. The frequency-dependent linear response equation can be written as;

$$\begin{aligned}\rho_1(\mathbf{r},\omega)=&\int\chi_{\mathrm{KS}}(\mathbf{r},\mathbf{y};\omega)v_1(\mathbf{y},\omega)d^3y\\&+\iint\chi_{\mathrm{KS}}(\mathbf{r},\mathbf{y};\omega)\left(\frac{1}{|\mathbf{y}-\mathbf{y}'|}+f_{xc}\left[\rho_0\right]\left(\mathbf{y},\mathbf{y}';\omega\right)\right)\rho_1\left(\mathbf{y}',\omega\right)d^3yd^3y'\end{aligned} \tag{A.24}$$

the frequency-dependent Kohn-Sham response function χ_{KS} can also be expressed in terms of its sum over states, which is

$$\chi_{\mathrm{KS}}\left(\mathbf{r},\mathbf{r}';\omega\right)=\sum_{j,k}\left(f_k-f_j\right)\frac{\psi_j(\mathbf{r})\psi_k^*(\mathbf{r})\psi_k^*(\mathbf{r})\psi_j\left(\mathbf{r}'\right)}{\omega-\left(\epsilon_j-\epsilon_k\right)+i\eta} \tag{A.25}$$

in which f_k is the occupation number of ground state Kohn-Sham orbital $\psi_k(\mathbf{r})$ with orbital energy ϵ_k.

Equation A.25 can be transformed into a matrix representation according to the work of Casida [276]. For this aim, the equation A.25 can be expanded as;

$$\begin{aligned}\chi_{\mathrm{KS}}\left(\mathbf{r},\mathbf{r}';\omega\right)&=\sum_{j,k}\left(f_k-f_j\right)\frac{\psi_j(\mathbf{r})\psi_k\left(\mathbf{r}'\right)\psi_k^*(\mathbf{r})\psi_j^*\left(\mathbf{r}'\right)}{\omega-\left(\epsilon_j-\epsilon_k\right)+i\eta}\\&=\sum_{k=1}^{N}\sum_{j=1}^{\infty}\frac{\psi_j(\mathbf{r})\psi_k\left(\mathbf{r}'\right)\psi_k^*(\mathbf{r})\psi_j^*\left(\mathbf{r}'\right)}{\omega-\left(\epsilon_j-\epsilon_k\right)}-\sum_{k=1}^{N}\sum_{j=1}^{\infty}\frac{\psi_k(\mathbf{r})\psi_j\left(\mathbf{r}'\right)\psi_j^*(\mathbf{r})\psi_k^*\left(\mathbf{r}'\right)}{\omega+\left(\epsilon_j-\epsilon_k\right)}\\&=\sum_{i,a}\left(\frac{\psi_a(\mathbf{r})\psi_i\left(\mathbf{r}'\right)\psi_i^*(\mathbf{r})\psi_a^*\left(\mathbf{r}'\right)}{\omega-\left(\epsilon_a-\epsilon_i\right)}-\frac{\psi_i(\mathbf{r})\psi_a\left(\mathbf{r}'\right)\psi_a^*(\mathbf{r})\psi_i\left(\mathbf{r}'\right)}{\omega+\left(\epsilon_a-\epsilon_i\right)}\right)\end{aligned} \tag{A.26}$$

where the subscript i takes values 1 through N, representing the occupied orbitals, and a takes values $N+1$ through ∞, representing the virtual orbitals of a complete basis set. Let;

$$P_{ai}(\omega)=\frac{\int\psi_i\left(\mathbf{r}'\right)\psi_a^*\left(\mathbf{r}'\right)v_{\mathrm{KS},1}\left(\mathbf{r}',\omega\right)d^3r'}{\omega-\left(\epsilon_a-\epsilon_i\right)} \tag{A.27}$$

and;

$$P_{ia}(\omega)=\frac{\int\psi_a\left(\mathbf{r}'\right)\psi_i^*\left(\mathbf{r}'\right)v_{\mathrm{KS},1}\left(\mathbf{r}',\omega\right)d^3r'}{-\left(\omega+\left(\epsilon_a-\epsilon_i\right)\right)} \tag{A.28}$$

then the linear density response can be shown as;

$$\rho_1(\mathbf{r},\omega) = \sum_{i,a} \psi_a(\mathbf{r})\psi_i^*(\mathbf{r})P_{ai}(\omega) + \psi_i(\mathbf{r})\psi_a^*(\mathbf{r})P_{ia}(\omega). \tag{A.29}$$

A small rearrangement of equations A.27 and A.28 gives;

$$\left(\omega - (\epsilon_a - \epsilon_i)\right) P_{ai}(\omega) = \int \psi_i\left(\mathbf{r}'\right) \psi_a^*\left(\mathbf{r}'\right) v_{\mathrm{KS},1}\left(\mathbf{r}',\omega\right) d^3r' \tag{A.30}$$

and;

$$\left(\omega + (\epsilon_a - \epsilon_i)\right) P_{ia}(\omega) = -\int \psi_a\left(\mathbf{r}'\right) \psi_i^*\left(\mathbf{r}'\right) v_{\mathrm{KS},1}\left(\mathbf{r}',\omega\right) d^3r'. \tag{A.31}$$

By writing the Hartree and exchange-correlation potentials as;

$$f_{H_{xc}}\left(\mathbf{r},\mathbf{r}',\omega\right) = \frac{1}{|\mathbf{r}-\mathbf{r}'|} + f_{xc}\left(\mathbf{r},\mathbf{r}',\omega\right) \tag{A.32}$$

the matrix elements $v_{ai}(\omega)$ can be defined as;

$$v_{ai}(\omega) := \int \psi_i(\mathbf{r}) v_1(\mathbf{r},\omega) \psi_a^*(\mathbf{r}) d^3r \tag{A.33}$$

and;

$$K_{kl,mn}(\omega) = \iint \psi_k(\mathbf{r})\psi_l^*(\mathbf{r}) f_{H_{xc}}\left(\mathbf{r},\mathbf{r}',\omega\right) \psi_m\left(\mathbf{r}'\right) \psi_n^*\left(\mathbf{r}'\right) d^3r d^3r'. \tag{A.34}$$

These, along with the equations A.29 and A.30 gives the matrix form of the frequency dependent linear response of the density as;

$$\left(\omega - (\epsilon_a - \epsilon_i)\right) P_{ai}(\omega) = v_{ai}(\omega) + \sum_{j,b} \left(P_{bj}(\omega)K_{ai,bj}(\omega) + P_{jb}(\omega)K_{ai,bj}(\omega)\right) \tag{A.35}$$

which is quivalent to;

$$\sum_{j,b} \left\{\left[\delta_{ij}\delta_{ab}\left(\epsilon_a - \epsilon_i - \omega\right) + K_{ai,bj}(\omega)\right] P_{bj}(\omega) + K_{ai,bj}(\omega)P_{jb}(\omega)\right\} = -v_{ai}(\omega). \tag{A.36}$$

Using equation A.31 instead of A.30 will give;

$$\sum_{j,b} \left\{\left[\delta_{ij}\delta_{ab}\left(\epsilon_a - \epsilon_i + \omega\right) + K_{ai,jb}(\omega)\right] P_{jb}(\omega) + K_{ai,bj}(\omega)P_{bj}(\omega)\right\} = -v_{ia}(\omega). \tag{A.37}$$

By defining;

$$X_{jb}(\omega) = P_{jb}(\omega) \tag{A.38}$$

$$Y_{jb}(\omega) = P_{bj}(\omega) \tag{A.39}$$

$$A_{ia,jb}(\omega) = \delta_{ij}\delta_{ab}\left(\epsilon_a - \epsilon_i\right) + K_{ai,jb}(\omega) \tag{A.40}$$

$$B_{ia,jb}(\omega) = K_{ia,bj}(\omega) \tag{A.41}$$

$$Q_{ia}(\omega) = -v_{ai}(\omega) \tag{A.42}$$

$$R_{ia}(\omega) = -v_{ia}(\omega) \tag{A.43}$$

we can write a very compact notation of equations A.36 and A.37 in matrix form;

$$\left[\begin{pmatrix} A(\omega) & B(\omega) \\ B^*(\omega) & A^*(\omega) \end{pmatrix} - \omega \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\right] \begin{pmatrix} X(\omega) \\ Y(\omega) \end{pmatrix} = \begin{pmatrix} Q(\omega) \\ R(\omega) \end{pmatrix}. \tag{A.44}$$

The equation A.44 can be turned into a pseudo-eigenvalue problem if the orbitals are real valued and f_{xc} is independent of the incident frequency. To see this, we first take the sum and difference respectively of each of the equations in A.44. Doing so gives;

$$(A+B)(Y+X)_q = \Omega_q (Y-X)_q \tag{A.45}$$

$$(A+B)(Y-X)_q = \Omega_q (Y+X)_q \tag{A.46}$$

where the q index indicates which eigenvector is being considered. Solving the equation A.46 and putting that into equation A.45 gives;

$$(A-B)(A+B)(X+Y)_q = \Omega_q^2 (X+Y)_q \tag{A.47}$$

the matrix $(A-B)$ has only positive values on its diagonal as it is positive definite. Therefore, equation A.47 can be written as;

$$(A-B)^{1/2}(A+B)(A-B)^{1/2}(A-B)^{-1/2}(X+Y)_q = \Omega_q^2 (A-B)^{-1/2}(X+Y)_q \tag{A.48}$$

which is usually written as;

$$WF_q = \Omega_q F_q \tag{A.49}$$

where;

$$F_q = (A-B)^{-1/2}(X+Y)_q \tag{A.50}$$

$$W = (A-B)^{1/2}(A+B)(A-B)^{1/2}. \tag{A.51}$$

As noted in the work of Casida [276], the eigenvalues of W are the squares of the excitation energies.