TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Data-driven Spectrum Interpretation in the Liquid Phase

Eemeli A. Eronen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 760 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2026

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

DATA-DRIVEN SPECTRUM INTERPRETATION IN THE LIQUID PHASE

Eemeli A. Eronen

TURUN YLIOPISTON JULKAISUJA ANNALES UNIVERSITATIS TURKUENSIS
SARJA SER. AI OSA TOM. 760 | ASTRONOMICA CHEMICA PHYSICA MATHEMATICA | TURKU 2026

University of Turku

Faculty of Science
Department of Physics and Astronomy
Physics
Doctoral Programme in Exact Sciences

Supervised by

Docent Johannes Niskanen
University of Turku
Turku, Finland

Dr. Anton Vladyka
University of Turku
Turku, Finland

Reviewed by

Professor Dr. Patrick Rinke
Technical University of Munich
Munich, Germany

Dr. Michael Meyer
European XFEL
Schenefeld, Germany

Opponent

Associate Professor Hannu-Pekka Komsa
University of Oulu
Oulu, Finland

In accordance with the guidelines of the University of Turku, the author declares that large language models (Microsoft Copilot, ChatGPT, and Writefull) were utilized in the preparation of this dissertation for searches, discussion of ideas, and language editing.

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-952-02-0676-5 (PRINT)
ISBN 978-952-02-0677-2 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)
Painosalama, Turku, Finland, 2026

UNIVERSITY OF TURKU
Faculty of Science
Department of Physics and Astronomy
Physics
ERONEN, EEMELI A.: Data-driven Spectrum Interpretation in the Liquid Phase
Doctoral dissertation, 115 pp.
Doctoral Programme in Exact Sciences
May 2026

ABSTRACT

Interpreting liquid-phase electronic spectra is challenging due to the complex interplay of atomic motion and intermolecular interactions. These effects create a wide distribution of local atomistic environments, each of which can yield a significantly different spectrum. Because experiments only capture ensemble averages, any observed spectral change reflects a latent shift in the underlying distribution of spectrally relevant structural features. In this thesis, I introduce a data-driven framework for disentangling the structure–spectrum relationship by adapting the recently proposed emulator-based component analysis. Applied to extensive simulated datasets, this spectrum-variance-driven dimensionality reduction method transforms the high-dimensional problem into a low-dimensional form, uncovering structural trends behind spectral changes. Across the studied cases, most of the spectral variation is explained by a small fraction of structural variance, represented by only a few latent structural degrees of freedom. This emerging pattern may indicate a general principle in spectroscopy of liquids.

KEYWORDS: Spectroscopy, Liquids, Simulations, Machine learning, Data analysis, Spectrum interpretation

TURUN YLIOPISTO
Matemaattis-luonnontieteellinen tiedekunta
Fysiikan ja tähtitieteen laitos
Fysiikka
ERONEN, EEMELI A.: Data-driven Spectrum Interpretation in the Liquid Phase
Väitöskirja, 115 s.
Eksaktien tieteiden tohtoriohjelma
Toukokuu 2026

TIIVISTELMÄ

Nesteiden elektronisten spektrien tulkinta on haastavaa. Atomien liike ja molekyylienväliset vuorovaikutukset johtavat laajaan kirjoon paikallisia atomistisia ympäristöjä, joista kullakin voi olla hyvin erilainen spektri. Kokeellisesti havaitaan vain näiden joukkokeskiarvoja, joten mitattu muutos spektrissä heijastaa piilevää muutosta sille olennaisten rakenteellisten ominaisuuksien jakaumassa. Tässä väitöskirjassa esittelen datalähtöisen viitekehyksen rakenne–spektri-riippuvuuden tulkintaan. Sovellan hiljattain julkaistua emulaattoripohjaista komponenttianalyysia, joka perustuu spektrivarianssiohjattuun rakenneavaruuden dimensionreduktioon. Menetelmä muuntaa korkeaulotteisen tulkintaongelman matalaulotteiseksi ja paljastaa spektrimuutoksen aiheuttaneet rakennekehityssuunnat laajoja simuloituja datajoukkoja hyödyntäen. Kaikissa tässä työssä tutkituissa tapauksissa valtaosa spektrivarianssista on selitettävissä pienellä osalla rakennevarianssia, joka lisäksi tiivistyy vain muutamaan piilevään rakennevapausasteeseen. Tämä tulos voi olla yleinen periaate nestesysteemien spektroskopiassa.

ASIASANAT: Spektroskopia, Nesteet, Simulaatiot, Koneoppiminen, Data-analyysi, Spektritulkinta

Acknowledgements

This dissertation was carried out in the Femto group of the Department of Physics and Astronomy, University of Turku. Although I am the author of this thesis, its completion relied on the support of many people and institutions.

I first wish to thank my primary supervisor, Docent Johannes Niskanen, for his exceptionally dedicated supervision throughout my PhD and his genuine commitment to producing the best possible research. I also thank my second supervisor, Dr. Anton Vladyka, for his expert advice, especially in advanced data analysis. I am grateful to Prof. Edwin Kukk for his scientific and academic guidance, as well as for his roles as my research director and custos. In addition, I thank Dr. Sari Granroth, Dr. Lassi Pihlava, Dr. Mikael Santonen, Kerttu Pusa, Md Golam Moctader, Hemmo Hartikainen, and Jalmari Passilahti for their contributions to a supportive and enjoyable work environment within the group.

Regarding the thesis manuscript, I thank Docent Johannes Niskanen, Dr. Anton Vladyka, Prof. Edwin Kukk, and Dr. Mikael Santonen for their insightful and constructive comments, which significantly improved the final version. I am grateful to the pre-examiners, Prof. Dr. Patrick Rinke and Dr. Michael Meyer, for their careful evaluation of the thesis and their encouraging feedback. I also thank Assoc. Prof. Hannu-Pekka Komsa for agreeing to act as the opponent at the public defence.

I acknowledge the European Synchrotron Radiation Facility for providing the synchrotron radiation infrastructure, and I especially thank Dr. Christoph Sahle for enabling my extended research visit and for his contributions to the manuscripts. I also acknowledge the significant computational resources provided by CSC – IT Center for Science and the Finnish Grid and Cloud Infrastructure. I am grateful for financial support from the Research Council of Finland, the Jenny and Antti Wihuri Foundation, the Doctoral Programme in Exact Sciences, the Magnus Ehrnrooth Foundation, and the Turku University Foundation.

Finally, I express my deepest gratitude to my friends and family for their constant love and support throughout my PhD journey.

27.04.2026
Eemeli A. Eronen

Table of Contents

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I E. A. Eronen, A. Vladyka, F. Gerbon, Ch. J. Sahle and J. Niskanen: Information bottleneck in peptide conformation determination by x-ray absorption spectroscopy. Journal of Physics Communications, 2024; 8: 025001.
Data and scripts available in Zenodo, doi: 10.5281/zenodo.8239300.

II E. A. Eronen, A. Vladyka, Ch. J. Sahle and J. Niskanen: Structural descriptors and information extraction from X-ray emission spectra: aqueous sulfuric acid. Physical Chemistry Chemical Physics, 2024; 26: 22752.
Data and scripts available in Zenodo, doi: 10.5281/zenodo.10650121.

III E. A. Eronen, A. Vladyka, Ch. J. Sahle and J. Niskanen: Structural Sensitivity of N 1s Excitations in *N*-Methylacetamide Solutions. The Journal of Physical Chemistry Letters, 2025; 16: 1666-1672.
Data and scripts available in Zenodo, doi: 10.5281/zenodo.14216288.

IV E. A. Eronen and J. Niskanen: Structural Decomposition of UV–Visible Spectral Variation: Azobenzene in Ethanol Solution. Manuscript, 2026; preprint available at arxiv.org/abs/2505.02610.
Data and scripts available in Zenodo, doi: 10.5281/zenodo.15349624.

V A. Vladyka, E. A. Eronen and J. Niskanen. Implementation of the emulator-based component analysis. Journal of Computational Science, 2024; 83: 102437.
Data and scripts available in Zenodo, doi: 10.5281/zenodo.10405685.

The publications were completed in collaboration with a team of researchers. For Publications **I**–**IV**, the author was responsible for the following: in Publication **I**, all simulations, machine learning, analysis, and writing; in Publication **II**,

machine learning, analysis, and writing; in Publication **III**, spectrum simulations, machine learning, analysis, and writing; and in Publication **IV**, machine learning, analysis, and writing. Additionally, the author contributed to the experimental work for Publications **I** and **III** and prepared the open data for Publications **I**–**IV**. In Publication **V**, the author contributed to code development, testing, and writing. Across all publications, the author contributed to the development of the analysis method.

1 Introduction

As direct observation of atomistic structures is difficult, indirect methods such as photon absorption [1, 2] and emission [3] spectroscopies are widely used. In the UV–visible and X-ray regimes, these techniques probe how a sample interacts with photons of different energies, producing a spectrum that reflects electronic transitions of the system. The discrete energies and transition probabilities depend strongly on the underlying atomistic configuration, allowing structural information to be deduced from the spectral response. However, in most cases, only a partial picture can be inferred. This occurs because only certain structural changes noticeably influence the spectrum, and because a given spectrum is not unique to a single configuration.

Spectrum interpretation becomes increasingly challenging as the system grows in complexity. Liquids are a prime example, as the phase allows for relatively strong intermolecular interactions, turning the spectrum into a high-dimensional function of both the target molecule and its surroundings. An additional challenge arises from the extensive motion of molecules caused by a rugged potential energy surface with numerous local minima. This prompts the sample to exhibit a broad distribution of local atomistic structures, which can yield significantly different spectra [4–11]. Experiments, however, capture only the ensemble-average spectrum, masking the underlying diversity. Consequently, any change in the observed spectrum reflects not a single structural alteration, but a shift in the distribution of structural properties that affect the spectrum. Identifying the spectrally decisive structural subspace and using it to interpret spectral effects is nontrivial but of considerable interest, as spectroscopy remains the most accessible tool for structural characterization of materials.

Computational spectroscopy combined with advanced analysis procedures can help disentangle the complex structure–spectrum relationship [12]. Unlike experiments, statistical simulations provide access to spectra of individual local structures, with the full dataset approximating the distributions and averages of the macroscopic ensemble. Recent advances in computational resources and simulation tools now allow the generation of large numbers of structure–spectrum pairs, enabling data-driven analysis with machine learning (ML). One such approach is the recently proposed emulator-based component analysis [13] (ECA), a spectrum-variance-driven dimensionality reduction algorithm. The method aims to find a few orthogonal vectors in the structural space that maximize the proportion of explained spectral variance when the structural data are projected onto them. This enables the identification

of the relevant structural changes behind spectral effects. Unlike methods that require prior hypotheses on the influential structural features, ECA operates in the full high-dimensional structural space, thus reducing possible bias. The iterative ECA fitting process requires a computationally fast spectrum emulator, such as a neural network trained to predict spectrum intensities from structures.

In this thesis, I adopt and extend the ECA method to develop a general framework for interpreting spectra of complex systems. After introducing the theoretical background, I discuss practical aspects of generating, validating and preprocessing extensive structure–spectrum datasets. Then, I show how the prepared data enable the training of a neural network and the fitting of an ECA model, reducing the high-dimensional problem into an interpretable form. Finally, I apply the framework to four liquid benchmark systems, uncovering the structural origins of several spectral phenomena. A consistent trend across all systems suggests a broader principle governing the structural information content of spectra in the liquid phase.

2 Molecular Spectroscopy

The analysis procedure of this thesis relies heavily on vast simulations of atomistic structures and their corresponding spectra. This chapter presents the theoretical framework later used to produce the necessary datasets, and briefly explains experimental measurements relevant for the work.

2.1 Molecular Dynamics

At any given time, an atomistic system is specified by the positions and momenta of all its atoms. The set of all possible values of these variables defines the phase space of the system. For a macroscopic sample, the thermodynamic ensemble, characterized by quantities such as temperature, pressure, and volume, defines the probability distribution over the phase space. Experimental observables correspond to ensemble averages with respect to this distribution.

Because macroscopic samples are impractical to model at the atomistic scale, molecular dynamics [14] (MD) simulations are commonly used. The method explores the accessible phase space by following the time evolution of a smaller, computationally feasible atomistic system. A sufficiently long trajectory should sample the probability distribution of the ensemble, with time averages over the trajectory approximating the ensemble averages.

The core principle of MD involves propagating the positions and velocities of each atom over time by integrating Newton's equations of motion

$$\mathbf{F}_i = m_i \frac{\mathrm{d}^2 \mathbf{R}_i}{\mathrm{d}t^2} = m_i \mathbf{a}_i \tag{1}$$

for each atom i where t is time, m_i is the mass of the atom, $\mathbf{a}_i$ its acceleration, $\mathbf{R}_i$ its position, and $\mathbf{F}_i$ the total force acting on it. Due to the complex nature of $\mathbf{F}_i$, Equation (1) is typically solved numerically using a scheme such as the Velocity Verlet [15]

$$\mathbf{R}_i(t + \Delta t) = \mathbf{R}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{1}{2}\mathbf{a}_i(t)\Delta t^2 \tag{2}$$

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + \frac{1}{2}\left[\mathbf{a}_i(t) + \mathbf{a}_i(t + \Delta t)\right]\Delta t, \tag{3}$$

where the position at time $t + \Delta t$ is calculated using the current acceleration and velocity $\mathbf{v}_i$ at time $t + \Delta t$ is calculated using the mean acceleration between the

current and new time steps. The time step Δt should be chosen to be short enough to capture all relevant vibrations of the system and to not allow atoms to get too close to each other, causing instabilities. The step length should also be long enough to ensure that the simulation progresses efficiently.

The approach for calculating the total force acting on each atom depends on the level of theory. In classical MD, the forces for an M-atom system are obtained from the gradient of a predefined empirical potential U as

$$\mathbf{F}_i = -\nabla_{\mathbf{R}_i} U(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M), \tag{4}$$

where the gradient $\nabla_{\mathbf{R}_i}$ is taken with respect to the coordinates of atom i. The functions U are known as force fields and typically include contributions from bond stretching, angle bending, torsional rotations, electrostatic forces, and van der Waals interactions. As the latter two decay slowly, a cutoff distance is often applied to van der Waals terms for efficiency, while long-range electrostatics are typically treated using methods such as Ewald summation [16]. While the individual contributions of a force field are based on physically motivated properties (*e.g.* bond lengths or angles), their numerical values are often optimized empirically. Numerous different force fields have been developed, each aiming to mimic experimental and higher-level computational results while allowing for fast calculations of relatively large systems. In Publication **I**, we compared three widely used force fields: AMBER03 [17] Charmm27 [18], and OPLS-AA [19], using aqueous triglycine as the test system. Following common practice, we paired the first two with TIP3P [20] water model and the last one with TIP4P [20]. We found that although each method produced slightly different structural distributions, the subsequently derived X-ray absorption spectra were similar.

Unlike in classical MD, the interatomic forces in *ab initio* and semiempirical MD approaches are computed on-the-fly from the quantum mechanical electron structure corresponding to the current nuclear configuration. These methods do not rely on a system-specific force-field parameterization. Consequently, they are, at least in principle, widely applicable. While these quantum mechanical methods can provide more accurate results, they are typically limited to smaller systems and shorter timescales due to significantly higher computational demands and reduced scalability [21]. The next section will provide a more detailed discussion of the electronic structure calculations that enable these methods.

Regardless of the method or computational resources, the number of atoms in a liquid MD simulation will always be small compared to a macroscopic sample. To approximate bulk behavior and to minimize surface effects, periodic boundary conditions are applied. The simulation cell, as shown in Figure 1, is conceptually replicated in all directions, and atoms near one edge of the cell can interact with atoms on the opposite edge. When an atom moves out of the cell on one side, it re-enters from the opposite side. The size of the cell must be large enough so that the

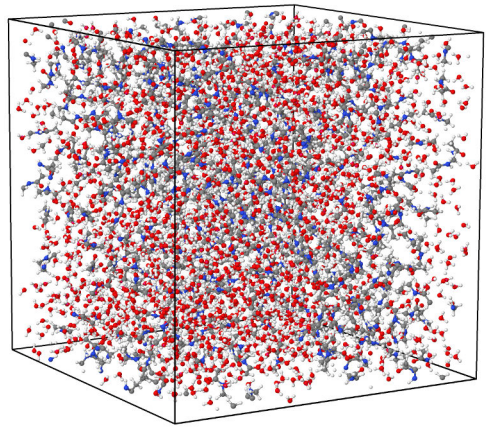

Figure 1. Example snapshot from a classical molecular dynamics production run of aqueous N-methylacetamide. The simulation cell is a cubic box of dimensions (50 Å)^3 with periodic boundary conditions. The plot was generated with the Jmol software [22] using data from Publication **III**.

molecules do not interact with their own images, and to ensure that relevant structural fluctuations are properly captured within the cell.

MD simulations are performed under a chosen statistical ensemble to mimic real-world conditions. Canonical (constant NVT), microcanonical (constant NVE), and isothermal-isobaric (constant NPT) are the most typical ones. Here N is the number of particles, V the volume, E the total energy, T the temperature, and P the pressure of the system. The choice of the ensemble depends on the physical conditions being modeled. Maintaining T and P in the vicinity of the chosen reference values is particularly important, as these quantities tend to fluctuate significantly due to numerical approximations and finite size of the simulation.

To control temperature, atoms are coupled to a thermostat. For example, the Berendsen thermostat applies a velocity scaling factor to drive the system toward the target temperature. A more sophisticated Nosé–Hoover thermostat [23, 24] aims to maintain the target temperature by introducing an additional degree of freedom representing a thermal reservoir. In these approaches, Equation (1) is modified to allow energy exchange between the atoms and the reservoir. Analogously, constant pressure is emulated by coupling the system to a barostat that modifies the volume of the simulation cell. As an example, the Parrinello-Rahman barostat [25] adjusts the size and shape of the cell to move towards the target pressure.

The first step in running an MD simulation is selecting the appropriate level of theory. For example, in Publications **I** and **III** we employed classical MD to generate large datasets, taking advantage of the availability of predefined force-field parametrizations. To accurately capture proton transfer between an acid molecule and its environment, *ab initio* MD was necessary for Publication **II**. Given the computational constraints, in Publication **IV** we adopted a semiempirical extended tight-

binding MD. Once the level of theory and relevant simulation parameters are selected, the MD run begins with an initial guess for atomic positions, with velocities typically assigned from the Maxwell-Boltzmann distribution. This is followed by an equilibration phase, allowing the system to relax and adjust to the chosen ensemble. After properties such as temperature and pressure have stabilized, the production run can begin. From this trajectory, snapshots (see Figure 1) can be extracted for spectrum calculations and subsequent analysis.

While methods such as classical MD rely on Newtonian mechanics of nuclei, the underlying laws arise from quantum mechanics. The next section will introduce the principles of a commonly used quantum mechanical simulation framework for a more complete view of the atomistic world.

2.2 Electronic Structure Calculations

Electronic structure calculations model the behavior of electrons explicitly. Compared to classical models, these methods can yield more accurate results and enable the study of properties that depend on the electron configuration of the system. Such properties include light–matter interactions and chemical reactivity.

At the heart of the methods lies the Schrödinger equation [26]

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N,t) = \hat{H}\,\Psi(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N,t), \tag{5}$$

where i is the imaginary unit, $\hbar$ the reduced Planck's constant, Ψ the N-electron wavefunction, $\mathbf{r}_j$ the position vector of electron j, t time, and $\hat{H}$ the Hamiltonian operator. For the purposes of this thesis, it is sufficient to consider the time-independent form

$$\hat{H}\Psi = E\Psi, \tag{6}$$

in which E is the energy eigenvalue of the system. An exact solution to Equation (6) is impossible for molecular systems, and even numerical approaches are typically nontrivial. Several approximations, introduced throughout this chapter, are required to transform the problem into a more manageable form.

A common starting point is to apply the Born–Oppenheimer approximation [27], which decouples electronic and nuclear motion and enables the calculation of electronic structure at a fixed nuclear geometry. This separation is justified by the large mass difference between nuclei and electrons, allowing the electronic density to adjust rapidly to nuclear motion. Under this assumption, the Hamiltonian can be expressed as the sum of the kinetic energy of the electrons, the electron–nuclear attraction, the electron–electron repulsion, and the nuclear–nuclear repulsion. External fields or other perturbations can be included by adding an additional potential to the Hamiltonian. Even with this simplified electronic Hamiltonian, substantial complex-

ity stems from the electron–electron interaction term, as it couples all electrons to one another.

The N-electron wavefunction is often modeled as a product of single-electron spinorbitals

$$\psi_I(\mathbf{r}_j) = \phi_i(\mathbf{r}_j)\,\alpha(j) \quad \text{or} \quad \psi_I(\mathbf{r}_j) = \phi_i(\mathbf{r}_j)\,\beta(j) \tag{7}$$

for spin up and spin down, respectively. Here $\phi_i(\mathbf{r}_j)$ is the spatial orbital which is a function of position vector $\mathbf{r}_j$ of electron j. Any such construction must satisfy the Pauli exclusion principle and the antisymmetry requirement: a spinorbital can only be occupied by a single electron, and exchanging any two electrons changes the sign of the total wavefunction. These conditions are fulfilled by expressing the approximate N-electron wavefunction as a Slater determinant [28]

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\mathbf{r}_1) & \psi_2(\mathbf{r}_1) & \ldots & \psi_N(\mathbf{r}_1) \\ \psi_1(\mathbf{r}_2) & \psi_2(\mathbf{r}_2) & \ldots & \psi_N(\mathbf{r}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\mathbf{r}_N) & \psi_2(\mathbf{r}_N) & \ldots & \psi_N(\mathbf{r}_N) \end{vmatrix}. \tag{8}$$

In many cases, a single Slater determinant of provides a sufficiently accurate description of the wavefunction.

The exact form of the spatial molecular orbitals $\phi_i(\mathbf{r})$ is unknown. An approximate solution can be obtained by representing each molecular orbital as a linear combination of a set of K basis functions $\chi_\mu(\mathbf{r})$:

$$\phi_i(\mathbf{r}) = \sum_{\mu=1}^{K} c_{\mu i}\,\chi_\mu(\mathbf{r}). \tag{9}$$

This representation, called linear combination of atomic orbitals [29] (LCAO), transforms the complex task of determining the wavefunctions into a more manageable optimization problem of finding the values of coefficients $c_{\mu i}$. For computational efficiency, Gaussian basis functions are commonly employed, as they allow fast analytic evaluation of the required integrals. Several Gaussians, with an additional angular factor, each called a primitive basis function, can be combined with certain fixed coefficients to produce one basis function χ_μ. This results in a contracted basis set used in practically all applications. While a minimal basis set would have a single function χ for each occupied orbital, additional functions are typically introduced to increase flexibility. The cardinal number ζ denotes the number of basis functions used to represent each orbital relative to a minimal basis set (*e.g.* double-ζ, triple-ζ). Since valence orbitals often exhibit more complex spatial structure than inner-shell orbitals, and since they change more in chemical environments, split-valence basis sets increase the number of valence functions relative to the core. As the atomic "orbitals" are centered around the nuclei, polarization and diffuse functions can be added to further increase the flexibility of the approximation.

An alternative approach is to find coefficients $c_{\mu i}$ of K plane waves [30]

$$\phi_i(\mathbf{r}) = \sum_{\mu=1}^{K} c_{\mu i} e^{\mathrm{i}\mathbf{G}_\mu \cdot \mathbf{r}}, \tag{10}$$

where $\mathbf{G}_\mu$ are wavevectors of the reciprocal lattice. The number of functions can be defined by an energy cutoff value limiting the maximum frequency of oscillations. Unlike atomic orbital functions, plane waves are not centered around an atom but span the entire space. While designed for periodic systems, this type of basis can be applied to non-periodic systems by embedding the system in a vacuum box. We employed plane waves for the spectrum calculations in Publications **I**–**III**, and LCAO basis sets for the simulations in Publication **IV** as well as for the *ab initio* MD in Publication **II**.

Density Functional Theory

Density functional theory (DFT) is a widely used quantum–mechanical simulation method based on the two Hohenberg–Kohn theorems [31]. The first theorem states that the ground state properties of a many-electron system are uniquely defined by the electron density $\rho(\mathbf{r})$ of the system. The second theorem, the variational principle, states that for any candidate density ρ', the exact energy functional obeys $E[\rho'] \geq E[\rho_0]$, where ρ_0 denotes the true ground-state density. Thus, the ground-state energy can be obtained by finding the density that minimizes $E[\rho]$.

The Kohn–Sham (KS) formalism [32] of time-independent DFT enables the method to be computationally feasible by approximating the electronic wavefunction as a Slater determinant of orthonormal single-electron spinorbitals. The goal is to find a set of N occupied spinorbitals $\psi_I(\mathbf{r})$ so that the electron density

$$\rho(\mathbf{r}) = \sum_{I=1}^{N} |\psi_I(\mathbf{r})|^2 \tag{11}$$

minimizes the approximate Kohn–Sham energy functional

$$\begin{aligned} E_{\mathrm{KS}}[\rho(\mathbf{r})] = & -\frac{\hbar^2}{2m_e} \sum_{I=1}^{N} \int \psi_I^*(\mathbf{r}) \nabla^2 \psi_I(\mathbf{r}) \, \mathrm{d}\mathbf{r} \\ & + \frac{e^2}{4\pi\varepsilon_0} \left(\sum_{A=1}^{M} \int \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \rho(\mathbf{r}) \mathrm{d}\mathbf{r} \right. \\ & \left. + \frac{1}{2} \int \int \frac{\rho(\mathbf{r}) \rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, \mathrm{d}\mathbf{r} \, \mathrm{d}\mathbf{r}' \right) \\ & + E_{\mathrm{XC}}[\rho(\mathbf{r})]. \end{aligned} \tag{12}$$

Here m_e is the electron mass, e the elementary charge, ε_0 the permittivity of vacuum, $\mathrm{d}\mathbf{r}$ the volume element of integration, $\mathbf{R}_A$ the position and Z_A the charge (atomic number) of nucleus A in a system of M atoms, and $E_{\mathrm{XC}}[\rho(\mathbf{r})]$ the electron-electron exchange-correlation (XC) functional.

Applying the variational principle to E_{KS} yields the KS equations, which have a form analogous to the Schrödinger equation. In the spin-restricted formalism, the KS equation for a single orbital can be written as

$$h_i^{\mathrm{KS}}\phi_i^{\mathrm{KS}}(\mathbf{r}_j) = \epsilon_i^{\mathrm{KS}}\phi_i^{\mathrm{KS}}(\mathbf{r}_j), \tag{13}$$

where $\phi_i^{\mathrm{KS}}(\mathbf{r}_j)$ is the single-electron wavefunction, ϵ_i^{KS} is the orbital energy and h_i^{KS} is an effective single-electron Hamiltonian for the electron in a multi-electron system. The Hamiltonian

$$h_i^{\mathrm{KS}} = -\frac{\hbar^2}{2m_e}\nabla^2 + \frac{e^2}{4\pi\varepsilon_0}\left(-\sum_{A=1}^{M}\frac{Z_A}{|\mathbf{r}_j - \mathbf{R}_A|} + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}_j - \mathbf{r}'|}\,\mathrm{d}\mathbf{r}'\right) + V_{\mathrm{XC}}(\mathbf{r}) \tag{14}$$

consists of the electronic kinetic energy, electron–nuclear interaction, electron–electron electrostatic repulsion (Hartree term [33]), and electron-electron exchange-correlation potential

$$V_{\mathrm{XC}}(\mathbf{r}) = \frac{\delta E_{\mathrm{XC}}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})}. \tag{15}$$

The potential V_{XC} is commonly referred to as the exchange-correlation "functional".

The XC term is crucial for the performance of DFT calculations, as it aims to incorporate all electron–electron interactions not accounted for by the direct electron–electron repulsion in Equation (14). Specifically, it estimates exchange effects arising from the tendency of electrons with the same spin to avoid each other, and the effect from the electron correlation, *i.e.* how the motion of an electron is affected by the movement of all other electrons in the system. In addition, the XC term compensates for the approximations employed in the KS scheme.

A wide variety of different XC functionals have been proposed with varying levels of theoretical complexity [21]. The local density approximation uses only the local electron density of the system, while functionals within the generalized gradient approximation (GGA) incorporate both the electron density and its gradient. Meta-GGA functionals go further by including additional ingredients such as higher-order gradients and kinetic energy density. The specific mathematical form varies across different functional families. Many of them are also empirically parameterized, which can make DFT resemble a semi-empirical method rather than a purely *ab initio* approach.

In Publications **I–III**, we employed a GGA-type approximation called the Perdew-Burke-Ernzerhof (PBE) functional [34], which we found to have a good balance between computational speed, accuracy, generalizability and robustness. While more

sophisticated XC functionals, such as hybrid functionals, can offer improved accuracy for certain systems, they are computationally demanding due to the inclusion of the exact exchange term from the Hartree-Fock theory [33, 35]. Moreover, the generalizability of hybrid functionals must be carefully assessed, as they often rely on empirical parameters fitted to a limited set of molecular properties of a limited set of systems. In Publication **IV** we found that a hybrid functional, specifically B3LYP [36, 37], might reproduce details of the UV–visible spectra with higher accuracy than the GGA functional PBE. Despite this, the overall spectral trends were still captured equally well by the computationally less demanding PBE.

Using all the information above, the electronic structure can now be approximately solved after applying a basis set expansion to the Kohn–Sham orbitals ϕ_i^{KS}. The problem can be expressed in matrix form

$$\mathbf{HC} = \mathbf{SCE}, \tag{16}$$

where $\mathbf{H}$ is the Kohn–Sham matrix with elements of operator (14), $\mathbf{C}$ contains coefficients c as in Equations (9) or (10), $\mathbf{E}$ is a diagonal matrix of orbital energies ϵ as in Equation (13), and $\mathbf{S}$, with elements $S_{ij}=\int \chi_i \chi_j \mathrm{d}\mathbf{r}$ is the overlap matrix of the non-orthonormal basis functions. In this formulation, each molecular orbital is a linear combination of all basis functions.

Because the form of the Hamiltonian in Equation (14) depends on the orbitals, and because coefficients $\mathbf{C}$ are found on both sides of Equation (16), the solution has to be approached iteratively. This is performed using the self-consistent field method, which proceeds as follows:

1. Guess the initial coefficients $\mathbf{C}$.
2. Construct the Kohn–Sham matrix $\mathbf{H}$.
3. Take iterative step for new $\mathbf{C}$.
4. Calculate new electron density from the now solved orbitals.
5. Check for convergence; stop or repeat from point number 2.

Convergence is reached after the energy and the orbitals no longer vary significantly between consecutive iterations. The orbitals are filled with electrons starting from the lowest energy applying the Aufbau principle. In some cases, core orbitals are not treated explicitly. Instead, their effect is represented by predefined atom-centered pseudopotentials [30] to improve computational efficiency. This approximation is justified by the idea that the core orbitals are mostly localized around a single atom and do not contribute significantly to the formation of the molecular orbitals.

The now-solved electron structure can be used for various applications, such as geometry optimization of isolated molecules, force evaluation in molecular dynamics simulations, or the derivation of electronic properties of the system. In Publications **I**–**IV** we applied DFT to local atomistic structures derived from MD to investigate

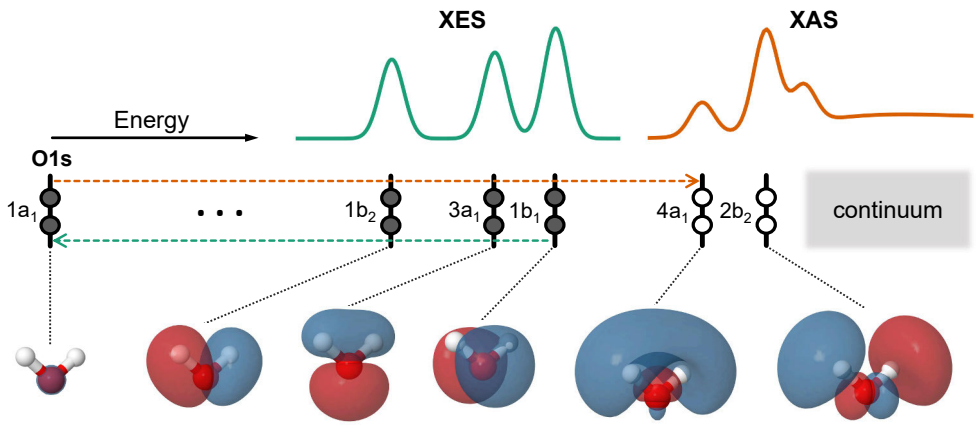

Figure 2. Schematic of the electronic structure and oxygen 1s (O1s) spectra of a single water molecule. The figure shows isosurfaces of the wavefunctions representing four of the five occupied molecular orbitals and two unoccupied bound orbitals, together with their energy levels and occupations. The continuum region is also indicated, representing states in which an excited electron is no longer bound to the molecule. Dashed arrows mark excitation from the O1s orbital to the lowest unoccupied molecular orbital and the highest-energy X-ray emission line that can appear when a core-excited state relaxes. The orbital labels follow the symmetry notation of an isolated water molecule, although this symmetry is broken in the liquid phase. The figure also includes an illustration of X-ray absorption (XAS) and X-ray emission (XES) spectra, derived from the electronic structure of the molecule. Only ground-state orbitals and energies are shown, whereas in reality each electronic transition involves a different set of orbitals and energies for the initial and final states. Orbital visualizations were generated using the Jmol software [22].

light–matter interaction of the system, specifically absorption or emission spectra as illustrated in Figure 2.

Spectrum Evaluation

To simulate how a system absorbs or emits a photon, the energy between the final and the initial electronic states $E_f - E_i$, as well as the transition cross-section σ, must be determined for each relevant electronic transition. The cross-section as a function of photon energy $\hbar\omega$ can be derived from Fermi's golden rule [38]. For X-ray absorption, this yields

$$\sigma(\hbar\omega) = 4\pi^2\alpha_0\hbar\omega \sum_{f,i} \left| \langle \phi_f | \, \hat{\epsilon} \cdot \mathbf{r} \, | \phi_i \rangle \right|^2 \delta(E_f - E_i - \hbar\omega). \tag{17}$$

Here α_0 is the fine structure constant, ω the angular frequency of the absorbed photon, ϕ_i the initial orbital of the electron, and ϕ_f its final orbital after transition. Dirac's δ-function denoted as δ ensures energy conservation. The transition operator is expressed in the dipole approximation form of $\hat{\epsilon} \cdot \mathbf{r}$ where $\hat{\epsilon}$ is the polarization vector of the incident photon and $\mathbf{r}$ is the position vector of the electron with respect

to the absorption site. For a randomly oriented sample, an angular average over all possible orientations introduces an additional factor of $1/3$ to the equation.

We carried out X-ray spectrum simulations using the GPAW software [39–41], employing its projector augmented wave (PAW) formalism [42] in combination with a plane-wave basis set. In the PAW approach, core orbitals are treated with the frozen-core approximation, and their contribution is included through an effective atomic potential. The valence orbitals are transformed into smooth pseudo-orbitals, removing rapid oscillations near the nuclei while maintaining the ability to reconstruct the original orbitals. This transformation enables the orbitals to be efficiently described by a moderate number of basis functions, while retaining the core–valence overlap needed for X-ray transition dipole moments $\langle \phi_f | \, \hat{\epsilon} \cdot \mathbf{r} \, | \phi_i \rangle$. Within this methodology, an expression for the absorption cross-section analogous to Equation (17) can be obtained [43].

The absorption spectra investigated in Publications **I** and **III** feature several bound excited states followed by a continuum, turning explicit computation of each possible excited state computationally laborious. One solution is to apply transition potential (TP) approximation, where the N-electron wavefunction is solved once with half of an electron removed from the core [44]. The resulting orbitals and their energies are then used for spectrum evaluation. Within this scheme, the continuum is approximated by simply including a large number of unoccupied (virtual) orbitals in the calculation. The energy scale of the half-core-hole calculation can be improved with the ΔKS scheme [45]. In the procedure, the energy axis is shifted so that the lowest-energy absorption line matches the energy obtained by subtracting the electronic energy of the ground state from the electronic energy of the explicitly simulated lowest-energy core-excited state.

An analogous formalism can be derived for X-ray emission studied in Publication **II**. In this case, the spectrum is modeled using a ground state calculation of filled orbitals. The energy axis is shifted so that the highest-energy emission line matches the energy obtained by subtracting the electronic energy of the cationic ground state from the electronic energy of a core hole state. Note that this calculation omits nuclear movement during the lifetime of the core hole, and is motivated by the significant additional computational burden required to include the according effects.

The aforementioned methodology applies to X-ray spectroscopy, where one of the involved orbitals is localized around a single nucleus. In contrast, UV–visible spectra, studied in Publication **IV**, involve excitations in which both orbitals are delocalized. For this purpose, we employed time-dependent density functional theory [46] (TDDFT), as implemented in the ORCA software [47, 48]. The program uses linear-response TDDFT [49], based on the time-dependent extension of the Kohn–Sham formalism. In the procedure, a standard DFT calculation is first performed to obtain the ground-state Kohn–Sham orbitals. Then, a linear-response eigenvalue problem is solved for a set of excitation energies and amplitudes. The former repre-

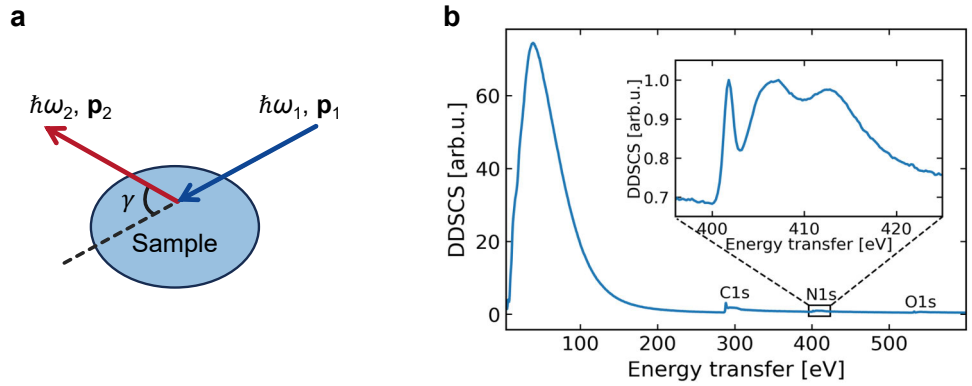

Figure 3. Overview of X-ray Raman scattering. (**a**) An incident photon with energy $\hbar\omega_1$ and momentum $\mathbf{p}_1$ interacts with a sample, producing an outgoing photon with lower energy $\hbar\omega_2$ and altered momentum $\mathbf{p}_2$. The scattering angle is denoted as γ. The illustration is adapted from Reference 51. (**b**) Inelastic scattering spectrum for liquid N-methylacetamide at 305 K studied in Publication **III**. The measurement was perfomed at a small scattering angle ($\mathbf{p}_1 \approx \mathbf{p}_2$). Starting from low energy transfer, the spectrum shows the valence Compton profile, followed by carbon (C1s), nitrogen (N1s), and oxygen (O1s) K-edges, with the near-edge spectrum of N1s highlighted.

sent the transition energies, and the latter estimate the contribution of each occupied orbital to a given transition. Combined with the ground-state calculation, the amplitudes can be used to obtain the transition dipole moments of Equation (17) [50].

The simulations produce a discrete stick spectrum, but in reality a smooth curve is expected due to factors such as instrumental and lifetime broadening, and vibrational effects. To account for these factors, we applied Gaussian convolution to the spectra. The width of the broadening is case-specific. We used either a constant width in energy across the entire spectrum, or a constant width followed by a linearly increasing one above the core-level binding energy of an absorption spectrum at a given atomic site.

2.3 Experiments

Although simulations are invaluable for understanding atomistic systems, they rely on a wide range of approximations. Therefore, the validity of the computational results should be assessed through a comparison with experimental observables. For each dataset in this thesis, we compared the calculated ensemble-averaged spectra with their experimental counterparts.

As part of Publications **I** and **III**, we performed inelastic X-ray Raman scattering (XRS) experiments [51] using the multi-element XRS endstation [52] of the beamline ID20 at the European Synchrotron Radiation Facility. During the scattering process, a hard X-ray photon with the energy $\hbar\omega_1$ and the momentum $\mathbf{p}_1$ is annihilated upon interaction with the sample, and subsequently the system emits a

new photon with reduced energy $\hbar\omega_2$ and altered momentum $\mathbf{p}_2$ as shown in Figure 3 a. The experiment measures the double differential scattering cross-section (DDSCS), dependent on both the outgoing photon energy and the solid angle of detection. At small scattering angles, that is when the momentum transfer $\mathbf{p}_1 - \mathbf{p}_2$ is small, the dipole operator dominates the interaction Hamiltonian. Under these conditions, the core-level XRS signal closely resembles that of X-ray absorption, with a cross-section analogous to Equation (17) [53]. Figure 3 b shows an example of such a spectrum over a wide energy range.

For static ensemble averages, the use of hard X-rays offers significant benefits arising from the high penetration depth of the photons. The liquid sample can be maintained at ambient pressure, and photons scattered from the bulk of the liquid can be reliably detected. Additionally, issues associated with microfluidic sample environments can be avoided by using relatively large sample volumes. On the other hand, the inelastic scattering cross-section is relatively small compared to resonant absorption, where the incident photon energy is tuned near the chosen core-level binding energy. This necessitates long acquisition times, with each XRS spectrum requiring hours to scan the core-level near-edge region. Such long measurements are made possible by a liquid flow cell [54], which continuously circulates the sample between a reservoir and a capillary tube. To allow precise temperature control, the cell incorporates a heating element and a thermostat.

We performed XRS measurements at the nitrogen K-edge for aqueous triglycine (Publication **I**), and for N-methylacetamide in its pure and aqueous forms (Publication **III**). In both cases, we carried out the experiment at two different temperatures. For the remaining computational datasets, we validated our data by comparisons with experimental results previously published in the literature.

3 Analysis Procedure

The complexity of the structure–spectrum relationship prompts the question of how to interpret the spectra. This chapter introduces a data-driven framework designed to address this challenge in a systematic and unbiased manner. First, Sections 3.1 and 3.2 describe the acquisition and preprocessing of large sets of structure–spectrum pairs required for the analysis. Section 3.3 then presents the neural network emulator, which serves as a central component of the analysis by enabling computationally efficient prediction of spectra for new structures. With these prerequisites in place, Section 3.4 explains the method itself.

3.1 Data Acquisition

Producing a dataset for the purposes presented in this thesis consists of two parts: the structural and spectral calculations, the principles of which were introduced in Chapter 2. The former begins by running an MD simulation and recording the atomic positions throughout the production run. The resulting trajectory can then be sampled to produce a set of snapshots (Figure 4 a). In order to reproduce an experimental result, these structures must be sampled in a way that yields a distribution representative of the statistical ensemble, and thus suitable for the evaluation of ensemble averages. Such a sampling can be achieved by selecting snapshots at specific time intervals or randomly from a sufficiently long standard MD run. The temporal independence of consecutive snapshots can then be assessed using autocorrelation functions.

The structural sampling is followed by spectrum calculations depicted in Figure 4 d. In a liquid, the valence molecular orbitals are not only bound to a single molecule, but span a vast spatial region covering the intermolecular local neighborhood. Consequently, the spectrum calculations should explicitly include both the target molecule and its neighbors. Ideally, the entire MD cell and periodic boundary conditions would be included in the calculation, as was the case in Publication **II**. However, in this case, the cell was relatively small. Computational demands quickly rise to unreasonable levels as the number of electrons increases. Therefore, we calculated the spectra of the rest of the data sets by extracting a local cluster around a molecule of interest, as shown in Figure 4 b. The size of the cluster is situational. In general, any molecule with an atom within a few Ångströms of the center molecule should be included. Additionally, for the UV–visible spectra of Publication **IV**, we

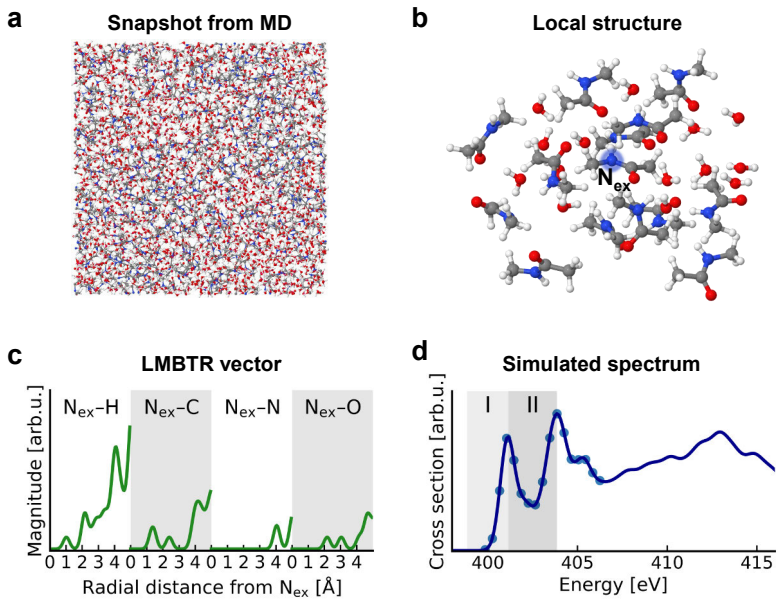

Figure 4. Steps for generating a dataset for the analysis procedure. The presented data originate from Publication **III**, and consist of computational nitrogen K-edge X-ray absorption spectra of aqueous N-methylacetamide (NMA). (**a**) Example snapshot from a molecular dynamics (MD) trajectory used to produce the raw structural data. (**b**) Example local atomistic structure extracted from the MD snapshot, with the nitrogen absorption site of the central NMA molecule denoted as $\mathrm{N_{ex}}$. (**c**) The same local structure encoded using the local version of the many-body tensor representation (LMBTR), centered on $\mathrm{N_{ex}}$. The standardized form of this vector can be used as $\tilde{\mathbf{x}}$ in the analysis. (**d**) Simulated spectrum corresponding to the local structure. The intensities within the shaded regions of interest (I and II) can be used as the target $\mathbf{y}$. A rebinned coarse spectrum, shown as circles, can alternatively be used as $\mathbf{y}$.

added a computationally inexpensive implicit solvent around the explicitly modelled cluster.

The ensemble averages usually converge within the inclusion of a few hundred independent snapshots. However, both ML and our statistical analysis requires a data set spanning all the available input and output space with enough data points to cover the nonlinear relation between them. In the context of this thesis, we have found that at least $\sim$10 000 structure–spectrum pairs must be obtained. For a liquid system of at minimum a hundred explicitly modeled atoms, quantum mechanical calculations are computationally heavy and require a large amount of memory. These factors create a strict limit for the level of theory that can be applied for the simulations. The calculations for each dataset in this thesis (Publications **I**–**IV**) consumed roughly between 100 000 and 1 000 000 processor hours on a supercomputer.

The aforementioned ensemble sampling of structures is a natural choice due to its connection with experimental results and its role as a traditional purpose of MD simulations. However, this approach can be suboptimal for producing a training

dataset for an ML model, as ensemble sampling tends to favor low-energy regions of the atomistic potential surface with less data from the rarely-visited high-energy regions. A better method would produce a data set spanning the entire accessible structural space with uniformly distributed data points. In Publication **III** we studied this alternative sampling strategy. We developed the disperse sampling algorithm, similar to the farthest point sampling [55], and chose a uniformly distributed subset of structural data from a large pool of local atomistic structures. We found that with this sampling method, an ML model could be trained with less data to consistently produce better results especially in sparse regions of the data cloud. This method is most effective when MD is computationally cheap but the consequent electron structure calculations are computationally demanding.

3.2 Data Preprocessing

The simulated structural data consists of cartesian coordinates and an atomic number for each atom. While this is a natural representation for simulations, it is suboptimal for ML of atomistic systems [56]. Consequently, numerous structural representation families, often referred to as atomistic descriptors, have been developed (see *e.g.* Publication **II**). Although designed to serve various different purposes, many descriptors share common properties, which we have also found valuable for our work. The descriptor should be invariant to translations and rotations, as the studied spectra are likewise unaffected by such transformations. Additionally, it is often beneficial if re-ordering atoms of the input does not affect the result. For ML models, it is important that the feature vector maintains constant length, even when the number of atoms varies between instances of the studied systems. Furthermore, smooth and continuous descriptor vectors tend to enable more stable model training than rough and discrete ones.

Selecting an appropriate descriptor for the purposes of this thesis was non-trivial. We outlined criteria necessary for analyzing the structure–spectrum relationship in Publication **I**. In essence, the descriptor must enable accurate spectrum predictions by an ML model and support the spectrum-variance-driven dimensionality reduction presented in Section 3.4. Moreover, the descriptor should consist of features that are interpretable by a human. We expanded this investigation in Publication **II** by comparing six different descriptor families. Although the ordered Coulomb matrix [57] has demonstrated strong performance for certain observables of an atomistic system [58], it proved less effective in our comparison. Instead, we found that the local version of the many-body tensor representation [59] (LMBTR), smooth overlap of atomic positions [60] (SOAP), and atom-centered symmetry functions [61] (ACSF) excelled in the comparison. Ultimately, we chose LMBTR as it outperformed the competition and also offered features with the most easily identifiable physical interpretation.

The two-body LMBTR descriptor is constructed by calculating the radial distance d between a center point and each atom of a given element. Gaussian functions centered at each d, are then added on a uniformly spaced distance grid. This process is repeated for each element of the system. Optionally, multiple centers may be included. The resulting vectors are concatenated into a single one, an example of which is presented in Figure 4 c. The Gaussian width, as well as the grid must be carefully tuned to fit the task at hand, which is discussed in the next section. While interpretable already in its native form, a division by the square of the distance from the center yields curves with an analogy to commonly used radial distribution functions.

LMBTR can also include three-body features based on broadened angular distributions. We incorporated these features in Publications **I** and **II**, but eventually excluded them from **III** and **IV** due to their lack of interpretability. Instead, we employed multiple system-specific center points. These can include the atoms of the molecule-of-interest and virtual sites defined by some geometric rule relative to a molecule. The selection can be guided by monitoring the ML performance as different sets of centers are included.

After constructing the descriptor vector $\mathbf{x}$ for a structure, each feature x_j covers a distinct range of values. Such variability can impede the performance of optimization algorithms, which often benefit from input features that are on a comparable scale and centered around zero. We address this by applying z-score standardization [62]. For each feature x_j with mean μ_{x_j} and standard deviation σ_{x_j}, its standardized counterpart $\tilde{x}_j$ is computed as

$$\tilde{x}_j = \frac{x_j - \mu_{x_j}}{\sigma_{x_j}}, \tag{18}$$

which yields a standardized descriptor vector $\tilde{\mathbf{x}}$. For any feature in this transformed space, a value of 0 corresponds to the mean, while a value of ± 1 represents one standard deviation above or below the mean. As the procedure is lossless, the original $\mathbf{x}$ can always be recovered from $\tilde{\mathbf{x}}$ via inverse transformation. Note that the values of μ_{x_j} and σ_{x_j} should be computed exclusively from the training data, after which the resulting standardization can be applied to all data. This prevents information leakage when later evaluating the generalizability of a predictive model on previously unseen data.

Another key preprocessing step is to select a representation for the spectrum intensities $\mathbf{y}$ corresponding to structure $\tilde{\mathbf{x}}$. As shown in Figure 4 d, we used two different approaches depending on the focus of the analysis. In Publication **II** we selected a relevant region of the spectrum and rebinned it to the coarsest resolution that preserves the essential spectral information, resulting in a 16-dimensional vector. In other publications, we simplified the presentation further by defining two or three regions of interest (ROIs) from the spectrum, and analyzing the total intensity

within each of them. These ROIs can be chosen by, for example, using a spectrum difference profile that highlights an effect between two enseble averages (Publications **I** and **III**), or by dividing a peak into halves (Publication **IV**). This coarsened representation enables the study of general trends in the spectrum while avoiding overinterpretation of potentially flawed details [63]. Finally, we applied z-score standardization to give each bin or ROI equal weight in the analysis. Spectral features with very small standard deviation were not included, as dividing by such small values can amplify noise and complicate ML processes. An alternative approach is to use absolute values, shifting the focus of the analysis to larger absolute changes in the spectrum.

3.3 Neural Networks

The analysis requires a computationally efficient emulator capable of approximating the intensities of a spectrum such that $\mathbf{y} \approx \texttt{Emulator}(\tilde{\mathbf{x}})$ for a given structural input $\tilde{\mathbf{x}}$. We have identified neural networks – a class of universal approximators [64] – well-suited for this regression task. Although the field of neural networks offers a wide range of architectures and techniques, here the scope will be limited to fully connected, multilayer feedforward neural networks [65] (NN). For Publication **I**, we implemented the NNs using scikit-learn [66], but in subsequent works switched to PyTorch [67] due to its greater flexibility, computational efficiency, and support for graphical processing units (GPUs).

An NN (Figure 5) consists of layers of units called neurons, each of which processes inputs from all neurons in the previous layer and passes the result forward. The structural input $\tilde{\mathbf{x}}$ (column vector) serves as the input layer $\mathbf{h}_0$ and propagates through $l-1$ hidden layers. Each layer $i = 1, 2, ..., l-1$ consists of a matrix calculation

$$\mathbf{h}_i = a_i(\mathbf{W}_i\mathbf{h}_{i-1} + \mathbf{b}_i), \tag{19}$$

where the weight matrix $\mathbf{W}_i$ acts on the previous output $\mathbf{h}_{i-1}$ after which a bias vector $\mathbf{b}_i$ is added. For a single neuron, this corresponds to computing a weighted sum of the outputs of all neurons in the previous layer and adding a scalar bias. The result is passed through an activation function a_i, applied elementwise on the vector $\mathbf{W}_i\mathbf{h}_{i-1} + \mathbf{b}_i$ to introduce nonlinearity to the output $\mathbf{h}_i$. This substantially enhances the flexibility of the model that would otherwise remain linear. After propagating through the hidden layers, the output layer yields a prediction for the spectrum

$$\mathbf{y}^{\text{pred}} = \mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{b}_l. \tag{20}$$

As is common in regression tasks, this layer does not include an activation function.

Several different activation functions have been developed. We mainly used the rectified linear unit (ReLU) [68], which sets negative values to zero while leaving

positive inputs unchanged. In many cases, ReLU is considered the standard choice due to strong empirical evidence for its good performance and its computational efficiency during gradient-based optimization [69]. We also tested scaled exponential linear units [70] and exponential linear units [71], but these offered only a small performance improvement at best while significantly increasing computational cost.

The objective is to train an NN model that generalizes well to unseen data. This requires optimizing the unknown model parameters in matrices $\mathbf{W}_i$ and vectors $\mathbf{b}_i$ through supervised learning. In this process, the parameters are iteratively adjusted to reduce the difference between a known training dataset and corresponding predictions. Next, I outline the steps involved in the training process.

Let us consider matrices $\tilde{\mathbf{X}}$, $\mathbf{Y}$, and $\mathbf{Y}^{\mathrm{pred}}$ obtained by stacking the vectors $\tilde{\mathbf{x}}$, $\mathbf{y}$, and $\mathbf{y}^{\mathrm{pred}}$ as rows, one for each data point. To compare the predicted and true outputs, a loss function L must be defined. The most common choice for regression is the mean squared error (MSE), which for n data points and o outputs (spectrum features) is defined as

$$L_{\mathrm{MSE}} = \frac{1}{n \cdot o} \sum_{i=1}^{n} \sum_{j=1}^{o} (y_{ij} - y_{ij}^{\mathrm{pred}})^2. \tag{21}$$

For Publications **II**–**V** we used the R^2 loss

$$L_{\mathrm{R}^2} = 1 - \mathrm{R}^2 \tag{22}$$

instead. Here, R^2 is the coefficient of determination (R^2 score), defined in the variance-weighted form as

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{o} (y_{ij} - y_{ij}^{\mathrm{pred}})^2}{\sum_{i=1}^{n} \sum_{j=1}^{o} (y_{ij} - \bar{y}_j)^2}, \tag{23}$$

where

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij} \tag{24}$$

is the mean value of the given dataset for output j. An R^2 score of 1 indicates a perfect fit, whereas a value of 0 corresponds to the performance of a model that always predicts the mean of the dataset. We chose this metric because it represents the proportion of variance explained by a predictive model. The importance of this will become more apparent in Section 3.4.

Before starting the training process, the model must be initialized to enable prediction. The values for the elements of the weight matrices $\mathbf{W}_i$ are typically drawn randomly from a distribution (see *e.g.* Ref. 72), while the elements of the bias vectors $\mathbf{b}_i$ are usually set to small positive numbers. Once initialized, predictions can be

computed through a forward pass, which involves applying Equation (19) layer by layer to all inputs until the corresponding outputs are obtained from Equation (20).

The learning process relies on the gradient of the loss function $\nabla L(\theta)$ with respect to all trainable model parameters θ, namely the elements of matrices $\mathbf{W}_i$ and vectors $\mathbf{b}_i$. This gradient is calculated using backpropagation [65, 73], which applies the chain rule in reverse direction by starting from the output and propagating backward through the layers to determine $\partial L/\partial\theta_i$ for each parameter θ_i. Because the loss function $L(\theta)$ is non-convex, the optimization must be performed approximately using iterative methods. Each iteration includes a forward pass to calculate the loss on the training data, followed by backpropagation to obtain $\nabla L(\theta)$.

The simplest optimization algorithm is gradient descent, which updates the values of each parameter by subtracting its partial derivative scaled by a factor called the learning rate. However, because the loss hypersurface contains numerous local minima and saddle points, this process can become trapped in a suboptimal region. For this reason, minibatch stochastic gradient descent [74] (SGD) is more commonly used for ML purposes. In SGD, the training dataset is split into randomly selected minibatches, each of which is used sequentially to update the model parameters. After one full sweep over the training set (one epoch), the process can be repeated with a newly shuffled set of minibatches. The slight variation in gradients across minibatches helps the optimization process to escape poor local minima and saddle points. Several modifications to standard SGD exist, the most popular being the adaptive moment estimation optimizer [75] (Adam). In addition to the gradient, the Adam algorithm incorporates momentum, an exponentially weighted average of past gradients, and their squared values to determine parameter updates. These additions yield a smoother and more efficient optimization path and allow automatic adjustment of the effective learning rate for each parameter individually. As an added benefit, default values for the internal parameters of Adam, such as the initial global learning rate, generally perform well for most tasks [76].

During training, the loss on the training set typically decreases until it eventually plateaus. In some cases, the training error can even approach zero. However, this can coincide with poor generalization, which is monitored using a validation set excluded from optimization. A small training loss combined with a large validation loss indicates overfitting: the model memorizes training data but produces erratic predictions for unseen inputs. To counteract this, regularization techniques can be employed. A common approach is to add a penalty term to the loss, encouraging the optimization algorithm to prefer simpler models. This can be achieved by, for example, adding a term penalizing the sum of the squares of all weights, resulting in a total loss of

$$L = L_{\mathrm{R}^2} + \lambda \sum_{i=1}^{l-1} \sum_{n,m} \widetilde{[\mathbf{W}_i]}_{nm}^2. \tag{25}$$

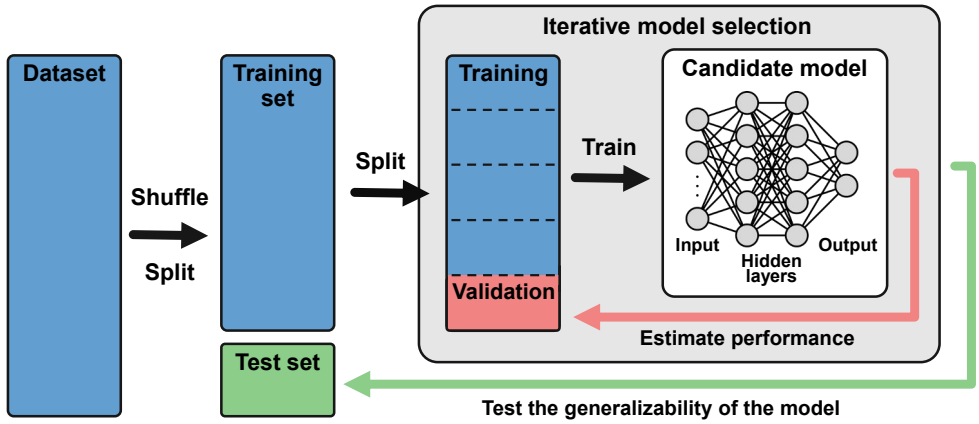

Figure 5. Schematic of the model selection process. Before any other steps, the full dataset is shuffled and split into training and test sets, with the latter set aside for final evaluation. The training set is used in an iterative model selection process, where candidate models with different neural network and descriptor hyperparameters are trained and their performance is measured. This evaluation can be performed using a single validation set separated from the training data or through cross-validation. The hyperparameter combination with the best validation score is then selected and retrained with the full training data. Finally, the test set is used to assess the generalizability of the chosen model. Figure adapted from a version provided by Dr. J. Niskanen.

The regularization strength λ must be carefully tuned. A value that is too small may permit overfitting, while an excessively large value leads to underfitting by preventing the model from capturing the underlying function. A properly chosen regularization strength reduces the variance of the predictions without introducing significant bias. Here, variance refers to the sensitivity of the predictions to small changes in the input, while bias denotes systematic deviation from the true output.

Another strategy to prevent overfitting is early stopping. Training is terminated and the best-performing model, based on validation loss, is saved if the validation loss does not improve for several consecutive epochs. Alternatively, training can be stopped after a fixed number of epochs or when the training loss plateaus, although these criteria provide less reliable control over overfitting.

Interestingly, overfitting is not necessarily always unfavorable. Models achieving very small training loss can sometimes generalize better than their less-overfitting counterparts. In such cases, the model has entered the interpolating regime with emergent self-regularization capability [77]. We observed possible evidence of this phenomenon in Publication **IV**. Ultimately, overfitting inferred from low training error may be acceptable if the model demonstrates good generalizability for its intended use.

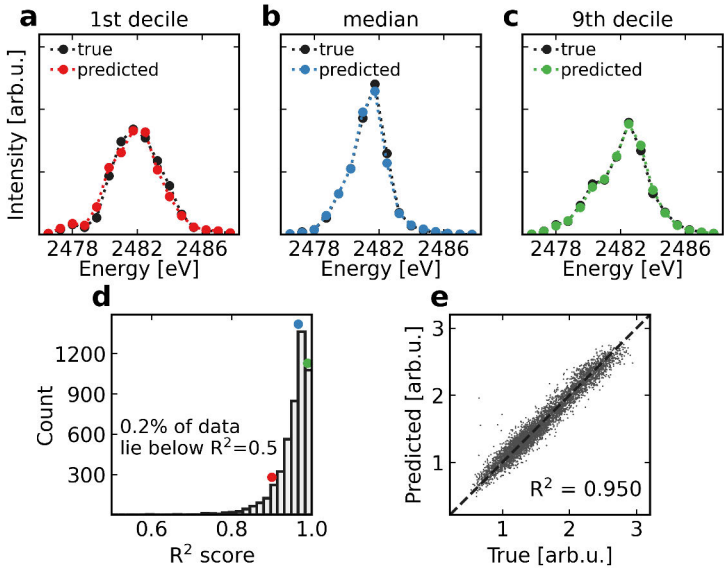

Figure 6. Demonstration of ML performance evaluation methods using test data and an ML model from Publication **II**. The data originates from simulations of sulfur Kβ X-ray emission spectra of aqueous sulfuric acid. (**a**)–(**c**) Examples of true and predicted spectra corresponding to the worst decile, the median, and the best decile of prediction quality, respectively. (**d**) Distribution of R^2 scores across the test set, with the scores of the three examples from panels (**a**)–(**c**) highlighted as dots. (**e**) Predicted versus true values for a single spectrum channel. For a perfect model, all points would lie on the one-to-one line. The panel also reports the average R^2 score for the entire test set.

Model Selection and Performance Evaluation

Both the NN and the atomistic descriptor involve multiple hyperparameters; properties without a single predefined value that universally guarantees optimal ML performance. For the NN, these include the number of layers (depth), the number of neurons in each hidden layer (width), and the regularization strength. Additional hyperparameters such as the initial learning rate, activation function, and minibatch size may also be relevant. Similarly, descriptors typically involve hyperparameters controlling the spatial range, the resolution of structural features, and the degree of smoothing via broadening functions. As an example, the performance of the LMBTR descriptor depends on the radial span of the grid, the number points within it, and the width of the Gaussian broadening. Depending on the system, different sets of centers may also need to be considered.

Finding a well-performing combination from the large pool of tunable hyperparameters is nontrivial. Therefore, a systematic model selection process, illustrated in Figure 5, is necessary. In Publication **II**, we introduced a joint hyperparameter search including both the neural network and the descriptor. We proposed that this avoids potential bias from considering their hyperparameters separately or from ne-

glecting descriptor optimization altogether. The procedure begins by constructing a vast grid containing all candidate hyperparameter values for both components. Then, a combination is selected, a candidate model is trained, and its performance is evaluated. This process is repeated hundreds or thousands of times, either systematically or randomly sampling combinations from the grid. In the end, the model with the best performance is chosen.

During the model selection phase, performance is evaluated using a validation set, which can be obtained by separating it from the training data beforehand. Alternatively, cross-validation can be used. In the latter approach, the training dataset is split into multiple subsets, one of which is used for validation and the others for training (Figure 5). The trial NN–descriptor model is trained multiple times so that each subset serves as validation once, and the average cross-validation loss is used to assess performance of the model. Finally, the best-performing model is chosen and typically retrained with the full training dataset.

After selecting and training the final ML model, its generalizability must be assessed to ensure that it approximates the function $\mathbf{y} = \mathbf{y}(\tilde{\mathbf{x}})$ for any $\tilde{\mathbf{x}}$ reasonably expected to occur. This requires a representative and statistically independent test dataset. A reliable approach is to define the test set already at an early stage of the project, during the data acquisition phase, keeping it strictly separate from the data used for training and validation. Alternatively, the test set may be obtained by randomly separating a portion from the full dataset, with the split performed prior to any preprocessing or training-related operations (see Figure 5).

The average performance of the model on the entire test set can be quantified by the R^2 score from Equation (23). As shown in Figure 6, additional insights can be obtained by visually comparing some individual predicted data points with their true counterparts, by analyzing the distribution of R^2 scores across the test set, or by examining how closely true and predicted values align on a line.

3.4 Emulator-Based Component Analysis

Due to the intra- and intermolecular interactions in a liquid, the structural descriptor will inevitably have at least hundreds of features (degrees of freedom), each exhibiting significant variation. Within this high-dimensional space lies an unknown subspace that explains most of the variation in the spectra. Finding this subspace is nontrivial yet fundamental to comprehensively analyze the structure–spectrum relationship. As the problem is impractical to address manually, it naturally motivates the development and application of data-driven statistical methods.

This thesis continues the development of a recently introduced technique called emulator-based component analysis (ECA). The method belongs to the family of projection pursuit algorithms [78], among which the most widely known is principal component analysis [79] (PCA). Both PCA and ECA perform dimensionality reduc-

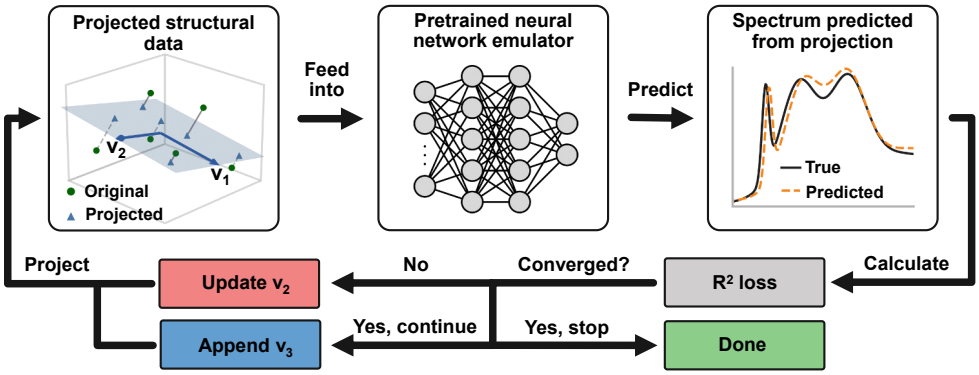

Figure 7. Schematic of the ECA fitting process. Structural data $\mathbf{X}$ are projected onto ECA component vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. The resulting projections $\mathbf{X}^{(2)}$ are passed through a pretrained neural network emulator, yielding predicted spectra $\mathbf{Y}^{(2)}$, which are then compared to the corresponding true spectra $\mathbf{Y}$ using the R^2 score. Since $\mathbf{v}_1$ is assumed to have undergone analogous optimization prior to considering $\mathbf{v}_2$, the second vector is iteratively tuned until R^2 score no longer improves. Once converged, the process can either stop or continue to the fitting of a third vector. Note that the structural data typically has far more than three dimensions.

tion by applying a projection to standardized input data $\tilde{\mathbf{X}}$ onto a set of orthonormal component vectors $\{\mathbf{v}_j\}_{j=1}^{k}$. For each data point $\tilde{\mathbf{x}}_i$, the rank k projection is defined as

$$\tilde{\mathbf{x}}_i^{(k)} = \sum_{j=1}^{k} \underbrace{(\mathbf{v}_j \cdot \tilde{\mathbf{x}}_i)}_{=:t_{i,j}} \mathbf{v}_j. \tag{26}$$

Here, t scores (t values) represent coordinates of the data points in the k-dimensional subspace spanned by $\{\mathbf{v}_j\}_{j=1}^{k}$. The key difference between the methods lies in the vector optimization criterion. In PCA, each successive vector is chosen so that the projections $\tilde{\mathbf{X}}^{(k)}$ capture the largest possible fraction of the total variance of data $\tilde{\mathbf{X}}$. In contrast, ECA aims to find projections $\tilde{\mathbf{X}}^{(k)}$, and vectors $\{\mathbf{v}_j\}_{j=1}^{k}$, that capture the largest possible proportion of variance of data $\mathbf{Y}$. In other words, ECA aims to highlight the spectrally decisive structural features while filtering out the irrelevant ones. Here, the use of R^2 score in the loss function and as a performance metric is natural, as the metric describes the proportion of variance explained by the ML and ECA models. Although applied to atomistic spectra in this thesis, the ECA method may be broadly applicable to the general problem of identifying features of any input $\tilde{\mathbf{x}}$ that are decisive for the corresponding outcome $\mathbf{y}$.

Finding the vectors for the target-variance-driven ECA decomposition is performed iteratively, one vector at a time. Either ML training data or a separate data set can be used for fitting, provided that the input data are standardized. The process begins with an initial guess for the first ECA component vector $\mathbf{v}_1$, onto which data $\tilde{\mathbf{X}}$ is projected. This yields a new set of projected structural data $\tilde{\mathbf{X}}^{(1)}$, for which

the corresponding spectrum must be calculated to evaluate the proportion of original spectral variance captured. Direct quantum mechanical simulations for these projections would be prohibitively slow, making optimization infeasible. Here, the numerically fast ML emulator shows its importance. Namely, a pre-trained NN provides an approximation for the spectra in milliseconds, producing predictions $\mathbf{Y}^{(1)}$ for projected data $\tilde{\mathbf{X}}^{(1)}$. This allows $\mathbf{v}_1$ to be tuned iteratively using gradient-based optimization of the R^2 loss of Equation (22), evaluated between $\mathbf{Y}^{(1)}$ and the corresponding true target values $\mathbf{Y}$.

After the rank-1 optimization converges, the above procedure can be repeated for the next component $\mathbf{v}_2$, constrained to remain orthogonal to the previous one. More generally, at rank k the procedure results in a set of orthonormal vectors $\{\mathbf{v}_j\}_{j=1}^{k}$, each optimized one-by-one. Due to non-linearity of $\mathbf{y}(\tilde{\mathbf{x}})$, this basis set must be considered as a whole in the optimization. Figure 7 shows a schematic of the entire fitting process. In practice, convergence near the optimal minimum is not guaranteed. Therefore, multiple initial guesses should be tested if suboptimal results are suspected. Note that the direction of a vector can flip depending on its initialization, as both $\mathbf{v}_j$ and $-\mathbf{v}_j$ yield the same R^2 score.

With the rank-k ECA decomposition completed, analysis of the results can begin using an independent test set. Equation (26) can be used to calculate projections $\tilde{\mathbf{X}}^{(k)}$ for the test data, which can then be fed through the emulator. Together with the true spectra $\mathbf{Y}$, the resulting predictions $\mathbf{Y}^{(k)}$ can be used for evaluation of the performance of the decomposition using the R^2 score. Since the ECA vectors are ordered by spectral relevance, the R^2 score as a function of ECA rank should initially rise quickly and then plateau near the emulation limit; the test set R^2 score of the emulator. For the applications of this thesis, just a few ECA vectors explained a majority of the target variance. In such cases, the analysis can be significantly simplified by focusing only on these components.

Both t values and vectors $\mathbf{v}_j$ of Equation (26) are highly useful for analyzing the structure–spectrum relationship. The former represents the coordinates of the structural data in the low-dimensional subspace spanned by the principal components of spectral response $\{\mathbf{v}_j\}_{j=1}^{k}$. They can be used to visually study how the data points are scattered across the spectrally relevant structural subspace. For example, a scatter plot can reveal clustering patterns (Figure 8 a), or distribution shifts between systems with different physical properties such as temperatures or concentrations. The latter can be plotted to reveal spectrally relevant and irrelevant structural features, as demonstrated in Figure 8 b. A non-zero value indicates that the structural feature affects the spectrum, whereas a zero suggests that it is spectrally irrelevant. Here, the interpretability of the atomistic descriptor is crucial, as each vector element corresponds to a specific feature in the descriptor. For two-body LMBTR, each feature corresponds to a point on the broadened radial distribution grid. On another note, in Publication **I** we showed that ECA vectors can measure feature importance for ML

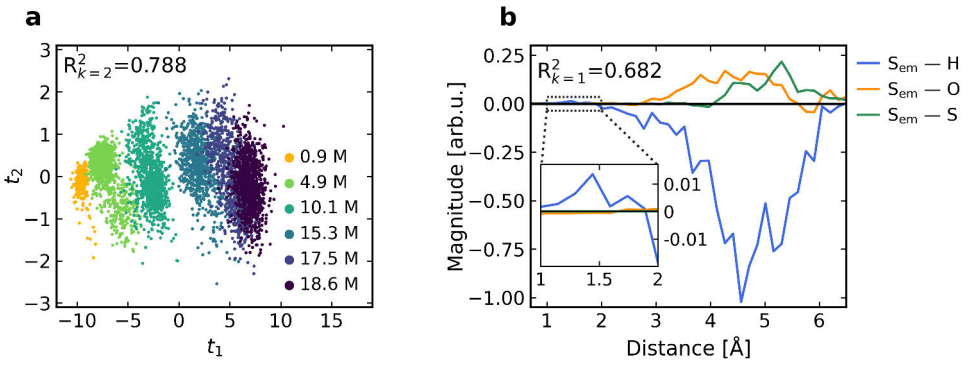

Figure 8. Example methods for analyzing ECA results using test data and results from Publication **II**. The data originates from simulations of sulfur Kβ X-ray emission spectra of aqueous sulfuric acid. (**a**) The two dimensional spectrally relevant structural subset shifts, revealing clustering when the sample molarity (M) changes. (**b**) The first ECA component vector. Non-zero features indicate a spectrally significant structural properties, whereas a zero suggests that the feature has no effect on the spectrum. With the local version of the many-body tensor representation descriptor, the x-axis of the curves correspond to element-wise radial distances from the chosen center atom; in this case the emitting sulfur atom S_{em}. The panels also highlight the test set R^2 score of rank 1 and 2 ECA decomposition.

performance, suggesting potential use in feature selection.

Furthermore, a deeper investigation is possible when two distinct ensemble-averaged spectra are observed. As the shift in the spectrum arises from a change in the underlying structural distribution, ECA can be used to identify two different data clusters in the spectrally decisive subspace. Afterwards, a representative structure from each case can be expanded using Equation (26). The difference between the two expansions reflects the structural trends behind the spectral shift. Chapter 4 presents a practical example of this type of analysis.

Our initial implementation of ECA relied on the original algorithm from Reference 13, which treated the vector fitting as a traditional optimization task. Later, we developed a new PyTorch-based implementation of ECA and published it as a ready-to-use package in Publication **V**. By treating ECA as an ML model, the new implementation achieved significantly faster optimization on high-dimensional data, which could be attributed to its efficient automated differentiation. It also demonstrated improved robustness in finding a good set of vectors, likely aided by the Adam optimizer and the use of minibatches. Additionally, the PyTorch implementation also offers flexibility through highly customizable PyTorch NN emulators and optional GPU acceleration.

To place the presented ECA in context, I conclude the chapter by briefly discussing some alternative methods. In Publication **IV**, considering UV–visible absorption spectra of ethanolic azobenzene, we compared the performance of ECA and

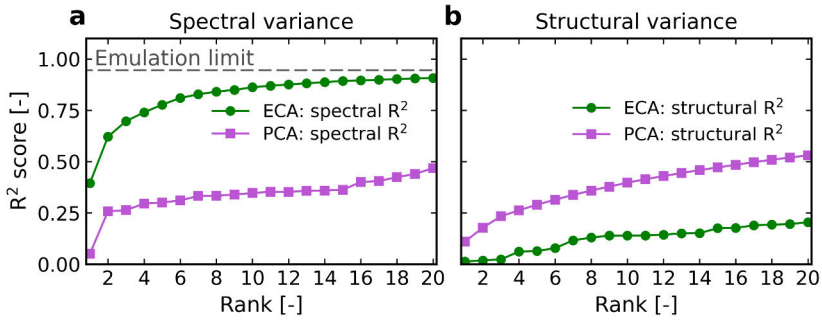

Figure 9. Comparison of ECA and structural PCA using test set results from Publication **IV**. The data originates from simulations of UV–visible absorption spectra of ethanolic azobenzene. (**a**) Spectral R^2 as a function of decomposition rank. ECA explains most of the spectral variance, with the R^2 score initially rising rapidly and then plateauing near the emulator performance level (emulation limit). Structural PCA, by contrast, accounts for significantly less spectral variance, with additional components after rank 2 providing only marginal improvement. (**b**) Proportion of explained structural variance. ECA explains only a small fraction of the structural variance, indicating that most of the spectral variance is explained by only a small portion of the structural variance. Structural PCA, in contrast, covers significantly more structural variance, as PCA by definition finds the linear projection that maximizes structural variance at a given rank.

structural PCA as a function of decomposition rank. Figure 9 a shows that ECA captures the majority of spectral variance with only a few structural degrees of freedom. In contrast, the most notable structural variation identified by PCA exhibits relatively low spectral significance when emulation is carried out for the projected data. Thus, structural PCA fails to capture spectrally relevant structural variance, while ECA excels in the task. This result can be explained by the observation that only a small fraction of the overall structural variance accounts for the majority of spectral variation, as shown in Figure 9 b.

Previous work has also shown that another decomposition tool, partial least squares fitting, is outperformed by ECA [13]. A simpler approach, correlation analysis, can be of use [9], but it inevitably overlooks potential interplay of several features and becomes tedious for complex systems. Tailored structural simulations allowing the movement of only some atoms could highlight the spectral impact of specific structural motifs. However, such constraints introduce selective bias and ignore possible contributions from the restricted degrees of freedom.

On another note, an analogous alternative to ECA, an encoder-decoder NN with an ECA-like bottleneck layer, has recently demonstrated the ability to capture most spectral variance using only a few structural degrees of freedom [80]. However, as found in the work, it is difficult to interpret the information encoded into the activation values of the neurons at the bottleneck. Judging by these findings, there appears to be noticeable benefits in applying data-driven dimensionality reduction methods specifically designed to focus on the target variance.

4 Results and Discussion

In this thesis, I present a new framework for investigating the relationship between atomistic structure and spectrum in liquids. Although spectrum is a function of structure, fully defined by the electronic Hamiltonian under the Born–Oppenheimer approximation, its interpretation remains challenging. This is due to the complexity of structural contributions and the inherent ensemble averaging in experiments. For these reasons, rather than trying to analyze detailed spectral impact of individual structural features, I aim to uncover general trends that describe how the distribution of spectrally relevant structural features changes when an observable spectrum shifts. Together with a team of researchers, I documented the results in Publications **I**–**V**. Four of these studies address the interpretation problem through benchmark systems, while Publication **V** introduced an improved implementation of the ECA algorithm applied in Publications **II**–**IV**.

Each benchmark study began with the generation of an extensive simulated dataset, validated against corresponding experimental spectra. Publications **I** and **III** included original measurements, while the remaining studies compared simulations to previously published spectra. The large sets of structure–spectrum pairs exhibited broad spectral variation depending on the local atomistic environment, with ensemble averages reproducing the experimental observations. Using the simulated datasets, the spectrum emulator achieved R^2 scores between 0.81 and 0.94, while a two-dimensional ECA decomposition retained the majority of the spectral variance with R^2 score of approximately 0.62 – 0.79. Across these studies, ECA on several different well-performing neural network–descriptor hyperparameter combinations yielded notably consistent results, highlighting the general robustness of the method. Next, I present the main case-specific results and discuss their broader implications.

In publication **I**, we studied the nitrogen K-edge X-ray absorption spectrum of aqueous triglycine. The molecule contains two backbone dihedral angles that describe the biologically relevant secondary structure of peptides and proteins. We examined whether these two Ramachandran angles [81] could be deduced from the spectrum. The results showed that the intensities in three main spectral regions, reproduced by simulations, are not sensitive to these structural properties, even though they dominate the overall structural variation. Instead, the spectrum is primarily influenced by the radial distance from the absorption site to neighboring atoms and by the water solvent environment. Additionally, we found that the choice of atomistic

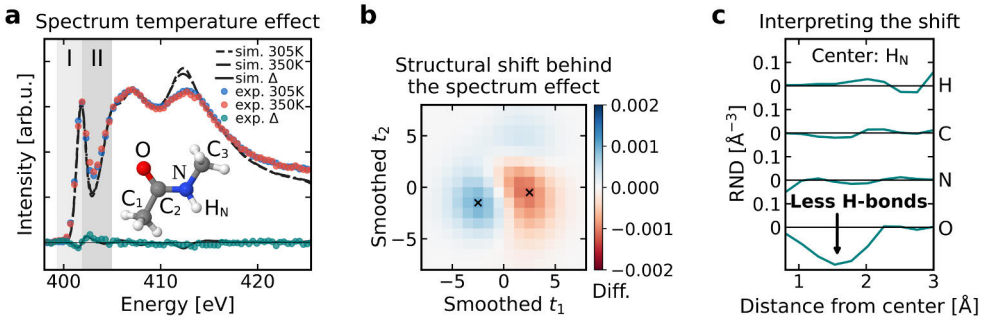

Figure 10. Selection of results from Publication **III** studying temperature dependency of nitrogen K-edge absorption spectra of liquid N-methylactamide (NMA). (**a**) Both experimental and computational spectra exhibit a pre-edge shift as temperature is increased from 305 K to 350 K. We used the difference curve Δ (350 K$-$305 K) to define two regions of interest I and II for analysis of the simulated dataset. The panel also shows an example NMA molecule with all selected centers for the LMBTR decriptor marked. (**b**) Smoothed difference between spectrally decisive structural distributions of systems at 350 K and 305 K, found using ECA. The displayed shift reflects the structural change behind the spectral effect. Crosses highlight two representative structures chosen from the maximum and minimum points of the histogram. (**c**) Difference between ECA expansions of the representative structures shows that the most significant phenomenon behind the spectral effect is weakening hydrogen bonds as temperature increases. The original curves of the LMBTR descriptor have been divided by the squared distance to the corresponding center to produce profiles representing changes in radial number density (RND). We observed similar effects and conclusions between spectra of pure and aqueous NMA. For full results, see Publication **III**.

descriptor plays a critical role in both ECA performance and interpretability.

Motivated by the previous finding, in Publication **II** we examined the ECA performance of six different descriptor families. We introduced a joint hyperparameter search that included both the descriptor and neural network in the same model selection process, avoiding bias from separate or manual tuning. As each descriptor had a different computational cost and number of tunable hyperparameters, we proposed that for a fair comparison, each should receive equal computational time for model selection. Among the studied schemes, LMBTR emerged as the best-performing descriptor, with the added benefit of including interpretable features. For this study, we used sulfur Kβ X-ray emission spectra of sulfuric acid at different concentrations. In the two-dimensional ECA space, clustering between concentrations appears along the first latent coordinate (Figure 8 a), while the protonation state of the emitting acid molecule is visible along the second coordinate. These two properties therefore contribute significantly to the spectral variation.

In Publication **III**, we measured an effect in the nitrogen K-edge X-ray Raman scattering spectra of pure and aqueous *N*-methylacetamide (NMA). As demonstrated in Figure 10 a, the prepeak of the spectrum shifts to higher energies when temperature increases or water solvent is replaced with pure NMA. Simulations reproduced this effect, enabling us to investigate its origins. We selected two regions of in-

terest from the spectrum, representing the below-zero and above-zero portions of the difference curve as seen in Figure 10 a. ECA revealed that the entire spectrally decisive structural distribution shifts as the temperature or concentration changes (Figure 10 b). This observation reflects the concept that spectral changes in liquids originate from changes in the distribution of local atomistic structures instead of a single structural alteration. Within the broader structure–spectrum relationship, the most significant single structural property behind the spectral shifts is the strength of hydrogen bonding with the absorbing NMA molecule (Figure 10 c). From a machine learning perspective, we found that wide and uniformly sampled training data improves model performance compared to data sampled to represent ensemble averages (see Section 3.1).

Finally, in Publication **IV** we extended our scope from X-ray spectra to the widely adopted UV–visible regime by studying the S_2 absorption peak of *trans*-azobenzene in ethanol solution. We observed substantial – even greater – spectral variation than that recorded in the X-ray regime. Despite the increased complexity arising from both the initial and final orbitals being highly delocalized, the system exhibited ECA behavior similar to the X-ray studies. Specifically, a few ECA vectors still capture most of the variation in the studied spectral regions. Further analysis revealed that a blueshift of the spectrum favors structures with less hydrogen bonding between azobenzene and ethanol, along with a contracted nitrogen–nitrogen bond at the azobenzene bridge. This suggests that certain structures may be overrepresented in photodynamics studies depending on the chosen incident photon energy within the S_2 peak. We performed the analysis using spectra calculated with both a GGA functional and a more computationally heavy hybrid functional, and found the main conclusions to remain consistent. While more advanced simulation methods may predict spectral details more accurately, the general structural dependencies are already captured by the GGA functional. Lastly, we also highlighted the advantage of the spectrum-variance-driven ECA compared to the commonly used PCA, as discussed in Section 3.4 and illustrated in Figure 9. In short, ECA captures a majority of spectral variance with just a few structural degrees of freedom, while PCA, focusing on the most notable structural variation, explains significantly less spectral variance at the same rank.

Overall, the presented analysis procedure proves to be remarkably effective in disentangling the complex structure–spectrum relationship in liquids. When combined with an extensive computational dataset, an accurate spectrum emulator, and a suitable atomistic descriptor, ECA consistently uncovers physically interpretable structural shifts behind the studied spectral effects. Across all benchmark systems, spectral changes appeared to be caused by only a small fraction of the overall structural variance. Moreover, most of the spectral variance is explained by just a few latent structural degrees of freedom, possibly indicating an underlying principle of liquid-phase spectroscopy.

List of References

[1] C. T. Chantler, G. Bunker, P. D'Angelo, and S. Diaz-Moreno. X-ray absorption spectroscopy. *Nature Reviews Methods Primers*, 4(1):89, 2024. doi:10.1038/s43586-024-00366-8.

[2] B. M. Tissue. Ultraviolet and Visible Absorption Spectroscopy. In *Characterization of Materials*, pages 1–13. John Wiley & Sons, Ltd, 2012. doi:10.1002/0471266965.com059.pub2.

[3] U. Bergmann and P. Glatzel. X-ray emission spectroscopy. *Photosynthesis Research*, 102(2): 255–266, 2009. doi:10.1007/s11120-009-9483-6.

[4] P. Wernet, D. Nordlund, U. Bergmann, M. Cavalleri, M. Odelius, H. Ogasawara, L. Å. Näslund, T. K. Hirsch, L. Ojamäe, P. Glatzel, L. G. M. Pettersson, and A. Nilsson. The Structure of the First Coordination Shell in Liquid Water. *Science*, 304(5673):995–999, 2004. doi:10.1126/science.1096205.

[5] M. Leetmaa, M. P. Ljungberg, A. Lyubartsev, A. Nilsson, and L. G. M. Pettersson. Theoretical approximations to X-ray absorption spectroscopy of liquid water and ice. *Journal of Electron Spectroscopy and Related Phenomena*, 177(2–3):135–157, 2010. doi:10.1016/j.elspec.2010.02.004.

[6] N. Ottosson, K. J. Børve, D. Spångberg, H. Bergersen, L. J. Sæthre, M. Faubel, W. Pokapanich, G. Öhrwall, O. Björneholm, and B. Winter. On the Origins of Core–Electron Chemical Shifts of Small Biomolecules in Aqueous Solution: Insights from Photoemission and ab Initio Calculations of Glycineaq. *Journal of the American Chemical Society*, 133(9):3120–3130, 2011. doi:10.1021/ja110321q.

[7] C. J. Sahle, C. Sternemann, C. Schmidt, S. Lehtola, S. Jahn, L. Simonelli, S. Huotari, M. Hakala, T. Pylkkänen, A. Nyrow, K. Mende, M. Tolan, K. Hämäläinen, and M. Wilke. Microscopic structure of water at elevated pressures and temperatures. *Proceedings of the National Academy of Sciences*, 110(16):6301–6306, 2013. doi:10.1073/pnas.1220301110.

[8] J. Niskanen, C. J. Sahle, K. O. Ruotsalainen, H. Müller, M. Kavčič, M. Žitnik, K. Bučar, M. Petric, M. Hakala, and S. Huotari. Sulphur Kβ emission spectra reveal protonation states of aqueous sulfuric acid. *Scientific Reports*, 6:21012, 2016. doi:10.1038/srep21012.

[9] J. Niskanen, C. J. Sahle, K. Gilmore, F. Uhlig, J. Smiatek, and A. Föhlisch. Disentangling Structural Information From Core-level Excitation Spectra. *Physical Review E*, 96:013319, 2017. doi:10.1103/PhysRevE.96.013319.

[10] J. Niskanen, M. Fondell, C. J. Sahle, S. Eckert, R. M. Jay, K. Gilmore, A. Pietzsch, M. Dantz, X. Lu, D. McNally, T. Schmitt, V. Vaz da Cruz, V. Kimberg, F. Gel'mukhanov, and A. Föhlisch. Compatibility of quantitative X-ray spectroscopy with continuous distribution models of water at ambient conditions. *Proceedings of the National Academy of Science of USA*, 116(10):4058–4063, 2019. doi:10.1073/pnas.1815701116.

[11] V. Vaz da Cruz, F. Gel'mukhanov, S. Eckert, M. Iannuzzi, E. Ertan, A. Pietzsch, R. C. Couto, J. Niskanen, M. Fondell, M. Dantz, T. Schmitt, X. Lu, D. McNally, R. M. Jay, V. Kimberg, A. Föhlisch, and M. Odelius. Probing hydrogen bond strength in liquid water by resonant inelastic X-ray scattering. *Nature Communications*, 10(1):1013, 2019. doi:10.1038/s41467-019-08979-4.

[12] T. Penfold, L. Watson, C. Middleton, T. David, S. Verma, T. Pope, J. Kaczmarek, and C. Rankine. Machine-learning strategies for the accurate and efficient analysis of x-ray spectroscopy. *Machine Learning: Science and Technology*, 5(2):021001, 2024. doi:10.1088/2632-2153/ad5074.

[13] J. Niskanen, A. Vladyka, J. Niemi, and C. J. Sahle. Emulator-based decomposition for

structural sensitivity of core-level spectra. *Royal Society Open Science*, 9:220093, 2022. doi:10.1098/rsos.220093.

[14] B. J. Alder and T. E. Wainwright. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, 31(2):459–466, 1959. doi:10.1063/1.1730376.

[15] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982. doi:10.1063/1.442716.

[16] P. P. Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921. doi:10.1002/andp.19213690304.

[17] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, 2003. doi:10.1002/jcc.10349.

[18] N. Foloppe and A. D. MacKerell, Jr. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104, 2000. doi:10.1002/(SICI)1096-987X(20000130)21:2<86::AID-JCC2>3.0.CO;2-G.

[19] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996. doi:10.1021/ja9621760.

[20] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2): 926–935, 1983. doi:10.1063/1.445869.

[21] A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Ltd, 2 edition, 2001.

[22] Jmol development team. Jmol: an open-source Java viewer for chemical structures in 3D, 2026. URL `http://www.jmol.org/`. Accessed: February 10, 2026.

[23] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, 1984. doi:10.1080/00268978400101201.

[24] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695–1697, 1985. doi:10.1103/PhysRevA.31.1695.

[25] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981. doi:10.1063/1.328693.

[26] E. Schrödinger. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Physical Review*, 28:1049–1070, 1926. doi:10.1103/PhysRev.28.1049.

[27] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20): 457–484, 1927. doi:10.1002/andp.19273892002.

[28] J. C. Slater. The Theory of Complex Spectra. *Physical Review*, 34:1293–1322, 1929. doi:10.1103/PhysRev.34.1293.

[29] R. S. Mulliken. Spectroscopy, Molecular Orbitals, and Chemical Bonding. *Science*, 157(3784): 13–24, 1967. doi:10.1126/science.157.3784.13.

[30] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54:11169–11186, 1996. doi:10.1103/PhysRevB.54.11169.

[31] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136:B864–B871, 1964. doi:10.1103/PhysRev.136.B864.

[32] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140:A1133–A1138, 1965. doi:10.1103/PhysRev.140.A1133.

[33] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24 (1):111–132, 1928. doi:10.1017/S0305004100011920.

[34] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77:3865–3868, 1996. doi:10.1103/PhysRevLett.77.3865.

[35] V. Fock. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Zeitschrift für Physik*, 61(1):126–148, 1930. doi:10.1007/BF01340294.

[36] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37:785–789, 1988. doi:10.1103/PhysRevB.37.785.

[37] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. doi:10.1063/1.464913.

[38] P. A. M. Dirac. The quantum theory of the emission and absorption of radiation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 114(767):243–265, 1927. doi:10.1098/rspa.1927.0039.

[39] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen. Real-space grid implementation of the projector augmented wave method. *Physical Review B*, 71(3):035109, 2005. doi:10.1103/PhysRevB.71.035109.

[40] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen, and K. W. Jacobsen. Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *Journal of Physics: Condensed Matter*, 22(25):253202, 2010. doi:10.1088/0953-8984/22/25/253202.

[41] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. doi:10.1088/1361-648X/aa680e.

[42] P. E. Blöchl. Projector augmented-wave method. *Physical Review B*, 50:17953–17979, 1994. doi:10.1103/PhysRevB.50.17953.

[43] M. Taillefumier, D. Cabaret, A.-M. Flank, and F. Mauri. X-ray absorption near-edge structure calculations with the pseudopotentials: Application to the K edge in diamond and α-quartz. *Physical Review B*, 66:195107, 2002. doi:10.1103/PhysRevB.66.195107.

[44] L. Triguero, L. G. M. Pettersson, and H. Ågren. Calculations of near-edge x-ray-absorption spectra of gas-phase and chemisorbed molecules by means of density-functional and transition-potential theory. *Physical Review B*, 58:8097–8110, 1998. doi:10.1103/PhysRevB.58.8097.

[45] M. P. Ljungberg, J. J. Mortensen, and L. G. M. Pettersson. An implementation of core level spectroscopies in a real space Projector Augmented Wave density functional theory code. *Journal of Electron Spectroscopy and Related Phenomena*, 184(8):427–439, 2011. doi:10.1016/j.elspec.2011.05.004.

[46] E. Runge and E. K. U. Gross. Density-Functional Theory for Time-Dependent Systems. *Physical Review Letter*, 52:997–1000, 1984. doi:10.1103/PhysRevLett.52.997.

[47] F. Neese. The ORCA program system. *WIREs Computational Molecular Science*, 2(1):73–78, 2012. doi:10.1002/wcms.81.

[48] F. Neese. Software update: The ORCA program system—Version 5.0. *WIREs Computational Molecular Science*, 12(5):e1606, 2022. doi:10.1002/wcms.1606.

[49] M. E. Casida. Time-dependent density functional response theory for molecules. In *Recent Advances in Density Functional Methods*, pages 155–192. World Scientific, 1995. doi:10.1142/9789812830586_0005.

[50] F. Neese. Prediction of molecular properties and molecular spectroscopy with density functional theory: From fundamental theory to exchange-coupling. *Coordination Chemistry Reviews*, 253 (5):526–563, 2009. doi:10.1016/j.ccr.2008.05.014.

[51] C. J. Sahle, A. Mirone, J. Niskanen, J. Inkinen, M. Krisch, and S. Huotari. Planning, performing and analyzing X-ray Raman scattering experiments. *Journal of Synchrotron Radiation*, 22(2): 400–409, 2015. doi:10.1107/S1600577514027581.

[52] S. Huotari, C. J. Sahle, C. Henriquet, A. Al-Zein, K. Martel, L. Simonelli, R. Verbeni, H. Gonzalez, M.-C. Lagier, C. Ponchut, M. Moretti Sala, M. Krisch, and G. Monaco. A large-solid-angle X-ray Raman scattering spectrometer at ID20 of the European Synchrotron Radiation Facility. *Journal of Synchrotron Radiation*, 24(2):521–530, 2017. doi:10.1107/S1600577516020579.

[53] Y. Mizuno and Y. Ohmura. Theory of X-Ray Raman Scattering. *Journal of the Physical Society of Japan*, 22(2):445–449, 1967. doi:10.1143/JPSJ.22.445.

[54] C. J. Sahle, C. Henriquet, M. Schroer, I. Juurinen, J. Niskanen, and M. Krisch. A miniature closed-circle flow cell for high photon flux X-ray scattering experiments. *Journal of synchrotron radiation*, 22(6):1555–1558, 2015. doi:10.1107/S1600577515016331.

[55] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. doi:10.1109/83.623193.

[56] J. Behler. Perspective: Machine learning potentials for atomistic simulations. *The Journal of Chemical Physics*, 145(17):170901, 2016. doi:10.1063/1.4966192.

[57] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters*, 108:058301, 2012. doi:10.1103/PhysRevLett.108.058301.

[58] A. Vladyka, C. J. Sahle, and J. Niskanen. Towards structural reconstruction from X-ray spectra. *Physical Chemistry Chemical Physics*, 25(9):6707–6713, 2023. doi:10.1039/D2CP05420E.

[59] H. Huo and M. Rupp. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, 3(4):045017, 2022. doi:10.1088/2632-2153/aca005.

[60] A. P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. *Physical Review B*, 87:184115, 2013. doi:10.1103/PhysRevB.87.184115.

[61] J. Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, 2011. doi:10.1063/1.3553717.

[62] J. M. H. Pinheiro, S. V. B. d. Oliveira, T. H. S. Silva, P. A. R. Saraiva, E. F. d. Souza, R. V. Godoy, L. A. Ambrosio, and M. Becker. The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks. *IEEE Access*, 13:199903–199931, 2025. doi:10.1109/ACCESS.2025.3635541.

[63] J. Niskanen, A. Vladyka, J. A. Kettunen, and C. J. Sahle. Machine learning in interpretation of electronic core-level spectra. *Journal of Electron Spectroscopy and Related Phenomena*, 260: 147243, 2022. doi:10.1016/j.elspec.2022.147243.

[64] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi:10.1016/0893-6080(89)90020-8.

[65] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi:10.1038/323533a0.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[67] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8026–8037, 2019.

[68] A. S. Householder. A theory of steady-state activity in nerve-fiber networks: I. Definitions and preliminary lemmas. *The bulletin of mathematical biophysics*, 3(2):63–69, 1941. doi:10.1007/BF02478220.

[69] X. Glorot, A. Bordes, and Y. Bengio. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 315–323, 2011.

[70] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-Normalizing Neural Networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 972–981, 2017.

[71] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv:1511.07289 [cs.LG]*, 2016. doi:10.48550/arXiv.1511.07289.

[72] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi:10.1109/ICCV.2015.123.

[73] S. Linnainmaa. Algoritmin kumulatiivinen pyöristysvirhe yksittäisten pyöristysvirheiden Taylor-kehitelmänä. Master's thesis. University of Helsinki, 1970.

[74] J. Bilmes, K. Asanovic, C.-W. Chin, and J. Demmel. Using PHiPAC to speed error back-propagation learning. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 4153–4156 vol.5, 1997. doi:10.1109/ICASSP.1997.604861.

[75] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs.LG]*, 2017. doi:10.48550/arXiv.1412.6980.

[76] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[77] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, 2019. doi:10.1073/pnas.1903070116.

[78] J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, 23(09):881–890, 1974. doi:10.1109/T-C.1974.224051.

[79] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi:10.1080/14786440109462720.

[80] J. Passilahti, A. Vladyka, and J. Niskanen. Encoder–decoder neural networks in interpretation of X-ray spectra. *Journal of Electron Spectroscopy and Related Phenomena*, 277:147498, 2024. doi:10.1016/j.elspec.2024.147498.

[81] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963. doi:10.1016/S0022-2836(63)80023-6.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2026

ISBN 978-952-02-0676-5 (PRINT)
ISBN 978-952-02-0677-2 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)