UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2025-038

Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 24 March 2025 in Luxembourg

to obtain the degree of

DOCTEUR DE L UNIVERSITE DU LUXEMBOURG

EN PHYSIQUE

by

Alessio Fallani

Born on 2 August 1995 in Florence (Italy)

Synergies between Quantum Mechanics and Machine Learning for Advancing Pharmaceutical Research

Dissertation Defense Committee:

Dr. Aurélia Chenu, Committee President

Professor, Université du Luxembourg

Dr. Alexandre Tkatchenko, Supervisor

Professor, Université du Luxembourg

Dr. Kostyantin Chernichenko, Co-supervisor

Senior Data Scientist, Janssen Pharma NV

Dr. Massimiliano Esposito

Professor, Université du Luxembourg

Dr. Pavlo O. Dral

Professor, Xiamen University

Affidavit

I hereby confirm that the PhD thesis entitled “Synergies between Quantum Mechanics and Machine Learning for Advancing Pharmaceutical Research” has been written independently and without any other sources than cited.

Luxembourg, _____ _____

Name

Abstract

The drug development process is resource-intensive, often costing billions and taking over a decade, yet many candidates still fail in late-stage trials. This thesis addresses key bottlenecks in early-stage drug discovery—such as navigating chemical space, modeling molecular interactions, and predicting biological properties—by integrating quantum chemistry and machine learning to develop more accurate and scalable computational methodologies. The analysis of the Aquamarine (AQM) dataset, designed to capture the interplay between molecular conformations, solvation effects, and non-covalent interactions, is presented as a key milestone for future machine learning models dealing with solvation effects for relevant molecules in medicinal chemistry. The results of the analysis reveal that many-body dispersion effects and implicit solvation significantly influence molecular geometries, reinforcing the necessity of accurate modeling for reliable predictions in biological environments. In a similar direction, the thesis introduces also a photonic quantum simulation framework for studying full Coulomb interactions between quantum Drude oscillators as a way to study dispersion beyond the dipole approximation typical of current models. This study uncovers nontrivial quantum effects, including the formation of entangled Schrödinger cat states during binding and offering insights into the fundamental nature of dispersion interactions. Moving from fundamental problems to more practical applications, the Quantum Inverse Mapping (QIM) framework is introduced to establish a direct, differentiable connection between quantum mechanical properties and molecular structures. This enables multi-objective molecular design and generation of transition path initializations, demonstrating its utility in navigating chemical spaces for different tasks. Finally, the thesis explores the role of quantum chemistry data in enhancing deep learning models for ADMET property modeling. A systematic study on Graph Transformer reveals that pretraining on atom-level quantum properties improves the model's representation, leading to superior performance. Collectively, these contributions bridge quantum chemistry with machine learning to address key challenges in molecular exploration, electronic structure calculation, and biological property modeling, advancing computational methodologies for rational drug discovery.

Preface

The content of this thesis is partly based on the following papers:

- Fallani, A.; Medrano Sandonas, L.; Tkatchenko, A. Inverse Mapping of Quantum Properties to Structures for Chemical Space of Small Organic Molecules. *Nat. Commun.* **2024**, *15*, 6061.

  Contribution: I conceived and designed this work together with L.M.S., with contributions from A.T. I developed the machine learning code, performed model training, and analyzed the model's performance across various applications in collaboration with L.M.S. A.T. supervised the project and provided feedback throughout. All authors contributed to discussions and the final manuscript.

  License: The content of this paper was used in Chapter 4 as allowed by Creative Commons Attribution 4.0 International License (CC BY 4.0). For a copy of the license visit `https://creativecommons.org/licenses/by/4.0/`.

- Fallani, A.; Nugmanov, R.; Arjona-Medina, J.; Wegner, J. K.; Tkatchenko, A.; Chernichenko, K. Pretraining Graph Transformers with Atom-in-a-Molecule Quantum Properties for Improved ADMET Modeling. *arXiv* **2024**, arXiv:2410.08024.

  Contribution: I wrote the code for the paper's results, conducted model training and evaluation, and devised and performed all representation analyses. R.N. conceived the idea of atom-level quantum mechanical properties pretraining, developed the Chytorch version of Graphormer, curated datasets, and provided code support and deep learning advice. J.A.-M. proposed the use of the TDC benchmark, contributed to the preliminary pretraining and fine-tuning results, and offered deep learning and code guidance. K.C. contributed to the attention rollout spectral analysis and supervised all stages of the project. All authors actively discussed the results and contributed to the final manuscript.

  License: The content of this preprint was used in Chapter 5 as allowed by Creative Commons Attribution 4.0 International License (CC BY 4.0). For a copy of the license

visit https://creativecommons.org/licenses/by/4.0/. A peer-reviewed version has been published in the *Journal of Cheminformatics* and is available at https://doi.org/10.1186/s13321-025-00970-0.

- Sarkis, M.; Fallani, A.; Tkatchenko, A. Modeling Noncovalent Interatomic Interactions on a Photonic Quantum Computer. *Phys. Rev. Res.* **2023**, *5*, 043072.

  Contribution: I contributed to the general discussions on the design of the model. More specifically, I performed the analysis of the cat state part and developed the corresponding visualizations presented in the paper. I worked closely with the co-authors to ensure the accuracy of the results and visualizations.

  License: The content of this paper was used in Chapter 3 as allowed by Creative Commons Attribution 4.0 International License (CC BY 4.0). For a copy of the license visit https://creativecommons.org/licenses/by/4.0/.

- Medrano Sandonas, L.; Van Rompaey, D.; Fallani, A.; Hilfiker, M.; Hahn, D.; Pérez-Benito, L.; Verhoeven, J.; Tresadern, G.; Wegner, J. K.; Ceulemans, H.; Tkatchenko, A. Dataset for Quantum-Mechanical Exploration of Conformers and Solvent Effects in Large Drug-Like Molecules. *Sci. Data* **2024**, *11*, 742.

  Contribution: I contributed to the curation and technical validation of the dataset, ensuring its quality and consistency, alongside M.H.. D.V.R. and J.V. selected relevant compounds from public datasets, with input from G.T. L.M.S. generated the 3D molecular structures using CREST/xTB and DFTB3+MBD, while D.V.R., L.P.B., and D.H. generated additional molecular structures with RDKit, Maestro, and Omega. L.M.S. performed the PBE0+MBD calculations in gas phase and implicit water for all structures. L.M.S. and D.V.R. designed and wrote the manuscript. A.T., J.K.W., and H.C. supervised and revised all stages of the work. All authors discussed the results and contributed to the final manuscript.

  License: The content of this paper was used in Chapter 3 as allowed by Creative Commons Attribution 4.0 International License (CC BY 4.0). For a copy of the license visit https://creativecommons.org/licenses/by/4.0/.

Acknowledgements

Completing this Ph.D. thesis has been an incredible journey, and I am deeply grateful to the many individuals and institutions that have supported me along the way. Given that this Ph.D. involved work across two locations, multiple secondments, and considerable travel, this section reflects the breadth of contributions that made it possible.

First and foremost, I would like to express my heartfelt thanks to my advisor, Prof. Alexandre Tkatchenko, for his invaluable guidance and encouragement throughout this process, and especially for the freedom and trust he granted me as a researcher. His expertise and thoughtful feedback have been instrumental in shaping not only this work but also the way I approach new challenges.

I would also like to extend my gratitude to my industrial supervisor, Dr. Kostiantyn Chernichenko, for his insightful guidance and practical perspective, which significantly enriched the applied aspects of my research. His support was essential in bridging the gap between academia and industry, and I deeply appreciate his time and expertise.

In addition, I would like to thank Dr. Ramil Nugmanov, with whom I had the privilege of collaborating closely during my time at Janssen. His expertise in cheminformatics, collaborative spirit, and thoughtful advice were pivotal in shaping key aspects of my work. I greatly value the opportunity to learn from his experience and I will always cherish the rewarding nature of our collaboration.

A special thanks goes to Dr. Leonardo Medrano Sandonas, whose guidance and support were particularly crucial during the early stages of my Ph.D. He generously shared his expertise, provided thoughtful advice and offered unwavering help, not only with academic matters but also with administrative challenges and personal guidance. His mentorship and camaraderie have had a lasting impact on both this work and my personal growth.

I am also thankful to the members of my Ph.D. defense committee for their insightful suggestions, constructive feedback, and critical perspectives, which have played a significant role in improving this thesis.

A heartfelt thanks goes to the friends and the colleagues I met along the way, both in Luxembourg during my time at the University and in Antwerp during my industrial second-

ment, as well as my fellow AIDD students from the European project. Although I wish I could name everyone individually along with their contributions, doing so would greatly extend this section. The stimulating conversations, shared challenges, and camaraderie have enriched my academic journey and made this experience much more rewarding.

I also thank my friend, Dr. Matthieu Sarkis, with whom I had the privilege of collaborating and discussing almost daily during my time at the University. He has been my go-to person for clarifying any doubts related to mathematics or theoretical physics, which he seems to know everything about. His generosity in sharing knowledge and his enthusiasm for problem solving have greatly enriched my understanding of many aspects of my work.

I am deeply grateful for the financial and institutional support provided by the European Union's Horizon 2020 research and innovation program under Marie Skodowska-Curie grant agreement No. 956832, and the partners of this project. This support allowed me to focus on my research and reach this important milestone.

Lastly and most importantly, I would like to thank my parents and my sister, without whom none of my academic efforts would have been remotely possible. Their unwavering support, encouragement, and love have been my foundation. I am also very grateful to my friends for their understanding and emotional support. Although I really enjoyed my time living abroad and traveling across Europe, I am truly happy to be back in Florence, where I can spend more time with them. No academic honor, career step, or qualification can outweigh the value of their company.

Finally, I would like to thank everyone who contributed, in large or small ways, to my growth as a scholar but more importantly as a person. Your kindness, encouragement, and wisdom have left an indelible mark on this work and on me.

Summary

Nowadays, the drug development process represents a massive investment of time and resources—often costing billions of dollars and taking up to 15 years. Despite this, a significant number of drug candidates still fail in the later stages of development. These challenges have made the development of accurate and efficient computational methodologies—particularly in the early drug discovery stages for modelling molecular behaviour and interactions—more important than ever. Addressing this class of problems means managing various forms of complexity: navigating the astronomically large chemical compound space, understanding molecular interactions with biological targets, and estimating complex biological properties linked to a molecule's journey through the human body, such as absorption, distribution, metabolism, excretion, and toxicity (ADMET). Quantum chemistry provides a rigorous foundation for computing molecular properties with high accuracy, particularly electronic structure and intermolecular interactions. However, these methods are computationally expensive and scale poorly with system size, making them impractical for high-throughput screening. Conversely, machine learning models—such as neural networks—are highly scalable and can approximate virtually any function, but their success depends critically on the availability of high-quality training data and physically grounded architectures. The contribution of this work lies in developing methods that strategically combine these two paradigms: leveraging the accuracy of quantum chemical data and the approximation power of neural networks to tackle the core complexities of early-stage drug discovery. After a chapter of introduction, a second chapter is used to lay the theoretical foundation for the rest of the work. This is done by reviewing the theoretical underpinnings of both quantum chemistry and machine learning. Starting from a brief historical introduction of quantum mechanics, the chapter outlines the construction of electronic structure methods—and in particular of density functional theory (DFT)—which enable the computation of molecular electronic properties at different levels of approximation and accuracy. Particular attention is given to the challenge of modelling non-local correlation effects, which are typically lacking in standard DFT methods and are critical for describing non-covalent interactions. The main approximations used to make the treatment of these effects computationally feasible are

introduced incrementally from the Adiabatic Connection Fluctuation Dissipation Theorem (ACFDT). Foundational aspects of machine learning are then covered, starting from neural networks as universal function approximators and essential training techniques such as backpropagation and stochastic gradient descent. A discussion of the bias–variance tradeoff highlights the key requirements for successful deep learning applications: large, diverse, and high-quality datasets, alongside expressive model architectures. The representation of molecules in machine-readable formats is explored, including text-based, graph-based, and 3D representations, along with the main neural network architectures developed for molecular modeling, with attention to symmetries, invariances, attention mechanisms, and equivariant message-passing updates. Following this theoretical foundation, the third chapter addresses the role of quantum chemistry data with a broad overview of datasets containing 3D molecular structures and their quantum mechanical properties. Landmark datasets of small gas-phase organic molecules—such as QM7, QM7-X, QM9, and ANI—are introduced, with a discussion of their scope, structure optimization methods, and quantum properties. While acknowledging their foundational value, their limitations in terms of chemical diversity, solvation modeling, system size, and relevance to pharmaceutical applications are outlined. More recent datasets attempt to address some of these gaps, but none fully captures the interplay between conformational variability, solvation, and quantum properties while providing broad coverage of the chemical space relevant to medicinal chemistry. The Aquamarine (AQM) dataset is then introduced as a resource specifically designed to address these limitations. Derived from public compounds selected to approximate a typical corporate chemical library—including molecules composed of H, C, N, O, F, P, Cl, and S atoms—AQM contains multiple conformers per molecule, with properties and geometries obtained both in vacuum and under implicit solvent conditions. After reviewing the data generation methodology—including the selection of relevant conformers—an extensive analysis of the dataset is presented. Special attention is given to the non-electrostatic components of solvation energy and its relation to dispersion energy computed with both pairwise Tkatchenko-Scheffler (TS) and Many-Body Dispersion (MBD) methods, highlighting the differences between these approaches across conformer ensembles. The extensive property and geometry coverage

emerging from this analysis place AQM as a valuable resource for training and benchmarking machine learning models beyond gas-phase approximations. The discrepancy between pairwise and many-body methods of treating non-covalent interactions just outlined serves as a clear illustration of how the level of physical approximation impacts dataset quality and, consequently, machine learning models. To further investigate these limitations, the following section introduces a proof-of-concept study exploring a modeling strategy beyond the dipole approximation inherent to standard dispersion models. This approach uses two Quantum Drude Oscillators interacting through full Coulomb potential (cQDOs), simulated on photonic quantum hardware. The bosonic nature of the cQDOs allows mapping onto the Fock space of optical modes, enabling a variational approach to finding the system's energy curve. The resulting binding curves exhibit features consistent with realistic molecular interactions and show excellent agreement with Morse-like potentials. Analysis in phase space reveals that the system transitions from vacuum states in the non-interacting regime to coherent states in the bonded regime, while at intermediate distances, it becomes an entangled Schrödinger cat state. This phenomenon is corroborated by entanglement entropy tracking and fitting an entangled cat-state ansatz to the quantum states at transition points. Overall, the methodology and the presented results offer insights into the quantum nature of dispersion and represents a first step toward a full-Coulomb treatment of non-covalent interactions on quantum devices. In the fourth chapter the focus is shifted towards developing tools for efficiently exploring chemical compound space (CCS). Starting from the observation that trends in quantum mechanical properties revealed meaningful structural patterns in AQM, this section starts by questioning if molecular properties can serve as coordinates for CCS navigation. This idea leads to the development of Quantum Inverse Mapping (QIM), a framework that modifies the Variational Autoencoder to learn a compressed latent representation enabling a differentiable mapping between quantum mechanical properties and molecular structures. QIM extends the standard VAE by incorporating an additional neural network that encodes quantum mechanical properties alongside the molecular encoder and decoder. A modified Evidence Lower Bound loss enforces alignment between latent representations of molecular structures and their associated properties, ensuring a shared

space where both can be seamlessly connected. Trained on the QM7-X dataset, the model demonstrates the ability to reconstruct molecular structures from quantum properties with good accuracy. A gradient analysis is then performed, identifying which quantum properties are most relevant for reconstruction, and noting that those properties tend to group molecules with the same chemical composition (isomers). It is then shown that QIM can be used for multi-objective targeted design of molecular structures, with its limited extrapolation capabilities addressed through training modifications that allow generation of structures with more heavy atoms than those in the training dataset. As a final application, the model's internal representation is exploited to obtain initial guesses for transition paths between molecular geometries. Applied to conformational isomerization cases, the interpolated geometries provide reasonable initial guesses for transition paths, demonstrating that the model's latent space captures essential physical aspects of molecular conformational changes, despite being trained exclusively on equilibrium structures. What is shown in this chapter validate the QIM framework as a significant conceptual advance by establishing quantum mechanical properties as natural coordinates for navigating chemical space, but also as a model that allows for many concrete applications. Finally, the fifth chapter explores how quantum chemistry data can enhance deep learning models for predicting ADMET properties. This chapter is presented as a natural extension of the previous ones, aiming at a way to leverage quantum mechanical properties, which can be computed from first principles, to improve the machine learning modelling of ADMET properties which instead must be measured experimentally. With data scarcity and noise posing a fundamental challenge for machine learning approaches, the proposed approach uses quantum chemistry data in a pretraining stage to obtain more robust representations for ADMET modeling. The investigation centers around the Graph Transformer architecture, comparing three pretraining strategies: atom-level quantum mechanical properties (charges, NMR shifts, Fukui indices), molecular-level quantum mechanical properties (HOMO-LUMO gaps), and self-supervised masking. Models were fine-tuned and tested on diverse ADMET tasks from public benchmarks and on a larger internal company dataset of microsomal clearance. Atom-level quantum pretraining showed the strongest improvements overall, while masking pretraining performed

well on public benchmarks but poorly on internal data. The chapter then continues with an in-depth analysis of the models' representations under multiple aspects. Firstly, it is found that pretraining information is preserved after fine-tuning and that all pretraining strategies mitigate the collapse in latent expressivity across layers. A novel spectral analysis is then introduced, showing that models pretrained on atom-level quantum properties develop attention patterns that mimic the low-frequency eigenmodes of molecular graph Laplacians. Finally, a gradient analysis demonstrates that atomic quantum mechanical pretraining produces representations more sensitive to local atomic environments. Overall, the results presented in this chapter confirm that this methodology is an effective way to leverage in-silico quantum chemistry data for improving models for experimental ADMET properties. Furthermore, the study reveals that the in-depth representation analysis provides insights that align more strongly with performance on high-quality internal datasets than with public benchmark results, suggesting that representation quality is a more reliable indicator of model performance in real-world pharmaceutical applications. The work is concluded with a chapter of discussion and perspectives. Overall, the work presented in this thesis outlines multiple contributions, spanning dataset development, computational methods, and representation learning techniques that together demonstrate how to leverage quantum chemistry data and machine learning methods in order to overcome the intractability issues inherent in chemical space exploration, electronic structure calculation, and biological property prediction. While each implementation has its limitations, the consistent success across applications validates this synthesis of physical understanding and data-driven methods as a promising direction for advancing molecular science and drug discovery. As computational capabilities continue to evolve, this integration of mechanistic understanding with machine learning models will play an increasingly central role in rational molecular design.

Index

Chapter 1

Introduction

Drug and material discovery has long been among the most challenging and resource-intensive tasks in science. Until the early 20th century, this process relied heavily on intuition and a trial-and-error approach, often referred to as the "Edisonian approach." One of the most notable examples is the discovery of penicillin in 1928, not through a deep understanding of molecular interactions but by accident—a fortunate event that ushered in the age of antibiotics. While Alexander Fleming's discovery was serendipitous, it took years of systematic research during the 1940s to fully realize penicillin's therapeutic potential. This reliance on empirical knowledge and manual experimentation made progress slow, as researchers navigated the vast and uncharted territory of chemical space without a fundamental understanding of the mechanisms governing molecular behavior.

The advent of quantum mechanics in the early 20th century marked a turning point in our understanding of matter at the subatomic level. It provided a framework for explaining the chemical processes that underpin molecular interactions. By the 1960s, methods like Density Functional Theory (DFT) were developed, allowing researchers to computationally predict electronic structures and molecular properties with unprecedented precision. These breakthroughs in quantum chemistry rapidly became integral to drug and material discovery, providing insights into reactivity, molecular stability, and electronic behavior. However, quantum methods posed significant computational challenges. Scaling these techniques to handle larger, more complex systems—such as modeling the biochemical processes in

drug interactions—remained a major hurdle, and the vastness of chemical space rendered comprehensive exploration impractical.

Concurrently, a revolution in machine learning (ML) was gaining momentum. Although the mathematical foundations of neural networks had existed for decades, it was not until the development of modern computing hardware—especially the GPU—that these models could be trained effectively. In 2012, AlexNet demonstrated the power of deep learning by training an over-parameterized model with millions of parameters on large datasets. This is seen by many as the turning point where deep learning started following the paradigm it keeps following today. This approach brought about many revolutions that are actually used in technology every day, with neural networks being able to track the motion of cars and pedestrians, to understand and mimic human language flawlessly and to generate any kind of absurd or realistic image or even video based on a text prompt. While this breakthrough revolutionized the field of artificial intelligence, it also had profound implications for many scientific disciplines, including chemistry and materials science. Deep learning's ability to approximate any function, given sufficient data, offered a new way to tackle the complexity of molecular systems.

The applications of machine learning in drug and materials discovery have, in fact, spurred an entire field of research, producing models that can be leveraged in numerous ways. One prominent application is virtual screening, where models are trained to predict specific properties of chemical structures, enabling the rapid evaluation of vast libraries of molecules. These models allow researchers to efficiently screen large numbers of compounds and select candidates with the most promising characteristics for further testing. Machine learning has also been employed to learn the complex energy landscape of a molecule and produce learned molecular force fields. These models are trained to predict interatomic forces and energies with great accuracy, significantly speeding up molecular simulations and structure optimization. Beyond virtual screening, machine learning models can also be trained for generative tasks, such as designing new molecules with targeted properties. By learning how to transform a known probability distribution into one that spans the space of compounds, these models generate molecules that align with desired chemical

or biological characteristics, expanding the toolkit available for drug discovery and material design. This dual capability—of both screening and generating novel compounds, together with the possibility to study their behaviour with accurate learned force fields highlights the transformative potential of machine learning in accelerating the discovery process, improving molecular simulations, and optimizing outcomes across both fields.

This new approach, however, comes with its own challenges. Deep learning models, while powerful, are notoriously data-hungry and prone to overfitting, especially in fields like drug discovery, where experimental data is often scarce and noisy. Furthermore, the dynamic nature of chemical space means that models trained on one region may not transfer well to others, necessitating continual retraining and data collection. In this context it becomes essential being able to leverage the data that you can actually produce by means of computation as is the case for quantum mechanical properties. This can be produced at scale and leveraged to provide these neural network based models with some degree of fundamental understanding of the underlying physics and chemistry governing molecular interactions. By training models on quantum mechanical properties, researchers can imbue these models with a more grounded representation of chemical systems. This mitigates the risk of overfitting to small, noisy datasets by anchoring predictions in fundamental physical principles. However, even with computational data, challenges remain in ensuring that these models generalize well across different chemical spaces, especially as quantum mechanical calculations become increasingly expensive for larger systems. Thus, the integration of quantum chemistry data with machine learning continues to evolve, seeking a balance between accuracy, computational cost, and scalability for meaningful predictions in drug discovery and materials science.

However, the relationship between quantum mechanics and machine learning is not unidirectional. While quantum mechanical data helps improve machine learning models, the opposite is also being actively explored. For instance, neural networks are being integrated into quantum chemistry workflows to create more expressive ansatzes for the electron density both in DFT-like methods or for variational calculations on actual quantum devices. This interplay between quantum mechanics and neural networks further pushes the boundaries

of what traditional computational methods can achieve.

Through this thesis, we explore how these revolutionary advancements—both in quantum mechanics and machine learning—are converging to address the complexity of molecular design. This is achieved by focusing on key aspects of data generation, exemplified by contributions to the analysis of the Aquamarine dataset, containing high quality data from molecules of interest for medicinal chemistry with a focus on dispersion interactions, solvation effects and conformational landscapes, and non-covalent interaction modeling on a photonic quantum simulator. We also investigate chemical space exploration through the inverse mapping from properties to molecules and representation learning by examining the effects of pretraining a graph-based model on quantum mechanical data. Importantly, explainability methods are integrated throughout this work, providing valuable insights into the knowledge encoded within these black-box models. These methods help uncover features inherent in the underlying data distributions and, more importantly, clarify the mechanisms that drive molecular behavior.

Chapter 2

Background

2.1 From the hydrogen atom to DFT and non-covalent interactions

This section delves into the evolution of atomic models, starting from the classical understanding of the hydrogen atom to the development of Density Functional Theory (DFT) and the modeling of non-covalent interactions. It begins with the limitations of classical atomic models and progresses through the advancements and intuitions that lead to the introduction of the Schrödinger equation, which revolutionized our understanding of atomic structure. The section then explores how quantum theory can be used to study molecular systems up to the principles of DFT. Finally, it addresses the challenges of modeling non-covalent interactions, such as van der Waals forces, and introduces the theory leading to advanced methods like Tkatchenko-Scheffler (TS) and Many Body Dispersion (MBD).

2.1.1 Limitations of Classical Atomic Models and the Schrödinger equation

When most people think of an atom, they imagine a small, dense nucleus surrounded by orbiting electrons. However, this understanding did not exist prior to 1911. This model was proposed by Ernest Rutherford [1] that year and experimentally validated through his famous gold foil experiment, in which ionized helium atoms (alpha particles) were fired at a

thin sheet of gold. The results showed that while most alpha particles passed through the foil with minimal deflection, a small fraction were scattered at large angles. This unexpected behavior could only be explained by the presence of a dense, positively charged nucleus at the center of the atom. This revelation fundamentally altered the prevailing understanding of atomic structure, replacing Thomson's Plum Pudding Model, which had pictured the atom as a diffuse "soup" of positive charge with negatively charged electrons embedded within it.

While Rutherford's model was able to explain numerous experimental observations, it still presented significant challenges. According to classical electrodynamics, an accelerated charged particle must radiate energy, which would cause the orbiting electrons to lose energy and spiral into the nucleus. This effect would lead to the collapse of the atom within a time frame on the order of $10^{-10}\mathrm{s}$. Since both experimental evidence and common sense dictate that matter exists stably, this poses a major flaw in Rutherford's atomic model. Around the same time, the idea that energy could be exchanged in discrete amounts, rather than continuously, was gaining traction. This concept was first introduced by Max Planck in his work on black body radiation [2] and later extended by Albert Einstein in explaining the photoelectric effect[3]. In both cases, the experimental evidence was explained by considering energy transfers as ensembles of small, discrete quantities called quanta, proportional to the frequency ν of the radiation. This relationship is expressed as $\epsilon_{\mathrm{q}} = h\nu$, where h is Planck's constant. This concept was used by Bohr in 1913 [4], who proposed a model which was a major step towards a more modern understanding of atomic structure. Bohr postulated that the possible orbits of electrons around their nuclei are only a countable number. These orbits are called stationary states and that electrons are only allowed to jump from one orbit to the next by losing or gaining an amount of energy that is the energy "quanta" associated to the energy difference. Following this model, an electron can jump between an orbit a and an orbit b if it radiates or absorbs a photon with frequency ν, where $h\nu = E_a - E_b$ with E_a energy associated to the stationary state a and E_b energy associated with the stationary state b. Albeit with some postulates, this model does not present the issue of charge collapse and furthermore was able to predict remarkably well the experimentally measured spectral lines of the hydrogen atom and explained the principle behind atoms emitting only at specific

wavelengths. While this was a major breakthrough, this model introduced the key mechanism behind its success via postulate. A good explanation for this came with De Broglie, who in 1923 [5] proposed a new and bold hypothesis. Stemming from the dual nature of light, behaving classically like a wave but also transferring energy like a discrete entity (photons), De Broglie hypothesized that matter might behave in the same way, behaving in some limit as a wave. In what became known as the "matter-wave hypothesis", De Broglie suggested that particles such as electrons could behave as waves and, borrowing the momentum formula $p = \frac{hv}{c} = h\lambda$ from photons, that their wavelengths would be inversely proportional to their momentum following the equation $\lambda = \frac{h}{p}$. This groundbreaking idea provided immediately a basis for Bohr's quantization of electron orbits as being stationary states, electrons could only exist in the form of standing waves, hence in the orbit of length $2\pi r$ should be divisible by the wavelength λ associated with the electron. This condition can be stated as $r = \frac{n\lambda}{2\pi} = \frac{nh}{2\pi p}$, with $n \in \mathbb{N}$.

While De Broglie's hypothesis provided a theoretical explanation for the quantization of orbits, a full understanding of electron behavior required the development of a more comprehensive mathematical framework. This came in 1926 [6], when Erwin Schrödinger formulated his now-famous wave equation, which described the electron as a wave, rather than a point particle. As a qualitative derivation of this equation one can start from considering in 1D a typical plane wave such as $\psi = Ae^{i(kx-\omega t)}$ with k being the wave number and ω being the frequency in radians per second. If one considers the derivatives in time t and space x, it is easy to see that:

$$\frac{\partial \psi}{\partial t} = -i\omega A e^{i(kx-\omega t)} \tag{2.1}$$

$$\frac{\partial \psi}{\partial x} = ikA e^{i(kx-\omega t)}. \tag{2.2}$$

Now, for a free particle with mass m and momentum p we know that $E = \frac{p^2}{2m}$, furthermore

$k = \frac{2\pi}{\lambda} = \frac{p}{\hbar}$ and $E = \hbar\omega$ where $\hbar = \frac{h}{2\pi}$. From 2.1 and 2.2 we can now find:

$$i\hbar\frac{\partial\psi}{\partial t} = EAe^{i(kx-\omega t)} = E\psi \tag{2.3}$$

$$-i\hbar\frac{\partial\psi}{\partial x} = pAe^{i(kx-\omega t)} = p\psi. \tag{2.4}$$

By noticing that $E\psi = \frac{p^2}{2m}\psi = \frac{1}{2m}(-i\frac{\partial}{\partial x})^2\psi$, it is easy to see that for a 1D free particle following a wave-like behaviour according to De Broglie's hypothesis:

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi, \tag{2.5}$$

which is indeed the celebrated Schrödinger equation. In a 3D setting, adding a potential $V(\mathbf{x})$ we obtain the more usual form:

$$i\hbar\frac{\partial\psi}{\partial t} = \left(-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x})\right)\psi. \tag{2.6}$$

The Schrödinger equation not only provided a rigorous mathematical framework for describing the wave-like behavior of particles, but also marked a fundamental shift in our understanding of the atom and matter itself. Under some assumption it now possible to obtain a time independent equation written as:

$$\hat{H}\left|\psi\right\rangle = E\left|\psi\right\rangle, \tag{2.7}$$

that is effectively an eigenvalue problem where the eigenfunctions of the Hamiltonian operator $\hat{H}$ are the stationary states of the system. In the case of the hydrogen atom this means being able to find the standing waves from the previous models as solution of this equation whose eigenvalues are the energy levels (spectrum) of the atom.

Albeit this framework is really effective, the interpretation of the wave function here introduced is not trivial at all. What does $\psi(\mathbf{x}, t)$ represent for the system? How should we use this to make experimental predictions and how should we interpret it? According to the Born rule, $|\psi(\mathbf{x}, t)|^2$ represents the probability density of finding the particle at position $\mathbf{x}$ at time

t. This means that quantum mechanics, in contrast to classical physics, does not provide deterministic outcomes, but rather predicts the likelihood of different outcomes upon measurement. In particular, in this framework observables are hermitian linear operators whose expectation value for the system can be computed from the wave function $\psi(\mathbf{x}, t)$. Namely for an observable $\hat{O}$ we will write the expectation value at time t as $\langle \hat{O} \rangle = \int \psi(\mathbf{x}, t)^* \hat{O} \psi(\mathbf{x}, t) \mathrm{d}\mathbf{x}$. The wave function $\psi(\mathbf{x}, t)$ thus encodes all possible information about the system and its observable behavior, but in the statistical sense. At its core it can actually be shown that quantum mechanics is in fact a probability theory [7].

Being that observables are hermitian linear operators, that their expectation values are essentially an inner product, and that solving a system in most cases is reduced to solving an eigenvalue problem, a simplified abstract notation that allows us to use all the tools of linear algebra comes naturally. This is known as braket notation, first introduced by Paul Dirac [8], and allows to write quantum states $|\psi\rangle$ as vectors in a generic Hilbert space $\mathcal{H}$. In this notation the wave function earlier mentioned is nothing but the projection of this vector on the eigenbasis of the position operator $\hat{\mathbf{x}}$, namely we have that $\psi(\mathbf{x}, t) = \langle \mathbf{x} | \psi(t) \rangle$. Under this light, it is possible to write the whole Eq. 2.6 as:

$$i\hbar \frac{\partial}{\partial t} |\psi\rangle = \left(\frac{\hat{\mathbf{p}}^2}{2m} + \hat{V} \right) |\psi\rangle = \hat{H} |\psi\rangle \tag{2.8}$$

where $\hat{H} = \frac{\hat{\mathbf{p}}^2}{2m} + \hat{V}$ is the generic Hamiltonian operator. This is the more familiar form of Schrödinger equation which will be used in the rest of the text.

The mathematical structure of the theory has a number of consequences. For example if we consider the operators associated to position and momentum, respectively $\hat{\mathbf{x}}$ and $\hat{\mathbf{p}}$, we find that they do not commute, in fact:

$$[\hat{\mathbf{x}}_i, \hat{\mathbf{p}}_j] = -i\hbar \delta_{ij}, \tag{2.9}$$

which can be easily shown to lead to the famous Heisenberg uncertainty principle:

$$\Delta \mathbf{x}_i \Delta \mathbf{p}_j \geq \frac{\hbar}{2} \delta_{ij} \tag{2.10}$$

which is in general valid in similar forms for non commuting observables and states that one cannot measure two such quantities with arbitrary precision at the same time.

Considering now functions of the position and momentum operators, also angular momentum can be introduced as $\hat{\mathbf{L}} = -i\hbar(\mathbf{r} \times \nabla)$, which can easily be shown to be Hermitian and satisfy the commutation relations:

$$[\hat{L}_i, \hat{L}_j] = i\hbar\epsilon_{ijk}\hat{L}_k \tag{2.11}$$

$$[\hat{L^2}, \hat{L}_j] = 0. \tag{2.12}$$

This set of operators is of great relevance as their spectrum and the set of eigenvalues and eigenfunctions of $\hat{L^2}$ and $\hat{L}_z$ define atomic and molecular orbitals. To summarize this, considering $\hat{L}_\pm = \hat{L}_x \pm i\hat{L}_y$ we have:

$$\hat{L^2}\,|l,m\rangle = l(l+1)\hbar^2\,|l,m\rangle \tag{2.13}$$

$$\hat{L}_z\,|l,m\rangle = m\hbar\,|l,m\rangle \tag{2.14}$$

$$\hat{L}_\pm\,|l,m\rangle = \hbar\sqrt{l(l+1) - m(m\pm1)}\,|l,m\pm1\rangle \tag{2.15}$$

where $l \in \{0,1,2,\dots\}$ is called orbital angular momentum quantum number and $m \in \{-l,-(l-1),\dots,0,\dots,l-1,l\}$ is called magnetic quantum number. For a classical intuition of how these make up atomic orbitals one can think of an orbital with fixed energy as also having fixed radius, meaning also fixed magnitude of the angular moment. An orbital will be hence defined by the magnitude of the angular momentum l, while the direction of the angular momentum in space will be given by m.

As these operators were not enough to explain further experimental effects related to the interaction with magnetic fields, such as the results from the Stern-Gerlach experiment [9], a new set of operators were introduced. These are associated with an intrinsic angular momentum of particles and is known as spin. Spin operators $\{\hat{S}_x, \hat{S}_y, \hat{S}_z\}$ follow the same algebra and rules of an angular momentum, having spectrum $\hat{S^2}\,|s,m\rangle = s(s+1)\hbar^2\,|s,m\rangle$ and $\hat{S_z}\,|s,m\rangle = m\hbar\,|s,m\rangle$ with the crucial difference that also half integer values are here

admitted, namely: $s \in \{0, \frac{1}{2}, 1, \dots\}$ and $m \in \{-s, -(s-1), \dots, \dots, s-\ s\}$. The nature of spin leads to significant physical implications. For fermions, the Pauli exclusion principle [10] states that no two identical fermions can occupy the same quantum state simultaneously, giving rise to the electronic structure of atoms and the stability of matter. Conversely, bosons do not adhere to such restrictions, allowing multiple particles to occupy the same state, which facilitates phenomena like Bose-Einstein condensation [11]. Since electrons are fermions with spin $\frac{1}{2}$, for what concerns atomic orbitals we now obtain an additional degree of freedom for which the previously defined orbitals can now be occupied by up to two electrons (spin $\frac{1}{2}$ and $-\frac{1}{2}$).

To close this section, a result that can be trivially derived from the linear character of this theory, is the variational theorem. This states that if a physical system admits a ground state $|\psi_{GS}\rangle$ with minimal energy $E_{GS} = \langle\psi_{GS}|\,\hat{H}\,|\psi_{GS}\rangle$, then for any other state $|\psi\rangle$ with energy $E = \langle\psi|\,\hat{H}\,|\psi\rangle$ one has $E_{GS} \leq E$. This can be shown by decomposing $|\psi\rangle$ on the eigenbasis given by the spectrum of $\hat{H}$, resulting in $\langle\psi|\,\hat{H}\,|\psi\rangle = \sum_n \sum_k a_n a_k^* \langle\psi_k|\,\hat{H}\,|\psi_n\rangle = \sum_n |a_n|^2 E_n \geq E_0 \sum_n |a_n|^2 = E_0 = E_{GS}$. As we will see in the next section, this result is at the core of quantum chemistry since a lot of the statistical properties and room temperature properties of molecules and materials depend on the ground state.

2.1.2 Density Functional Theory (DFT) and Its Role in Computational Chemistry

Quantum mechanics, with the Schrödinger equation at its core, allows predictions that went far beyond the possibilities of previous atomic models. With this theory one can try to tackle the generic electronic problem, being hence able, in principle, to compute the properties of any molecule or material. In practice, as mentioned in the previous section, one proceeds by starting from an eigenvalue problem given by the time independent Schrödinger equation (Eq. 2.7) where, if we consider a system of M nuclei and N electrons described by the position vectors $\mathbf{R}_A$ and $\mathbf{r}_i$, and consider the distances $r_{iA} = |\mathbf{r}_i - \mathbf{R}_A|$, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ and

$R_{AB} = |\mathbf{R}_A - \mathbf{R}_B|$, then the Hamiltonian operator $\hat{H}$ will be:

$$\hat{H} = \sum_{i=1}^{N} \frac{1}{2}\hat{\mathbf{p}}_i^2 + \sum_{A=1}^{M} \frac{1}{2}\hat{\mathbf{p}}_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{\hat{r}_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{\hat{r}_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{\hat{R}_{AB}} \tag{2.16}$$

where Z_A is the atomic number of the A-th nucleus and where following atomic unit convention masses are in units of electron mass and energies in Hartree ($1\mathrm{E_h} = \frac{e^2}{4\pi\epsilon_0 a_0}$).

While Eq.2.16 appears to be very complicated considering also quantum effects at the nuclear level, a simplification can be obtained by realizing that the mass of protons is ~ 2000 times higher than the mass of the electrons. As they have the same charge in absolute value, they are subject to similar forces and hence electrons will move on a much lower timescale seeing protons and nuclei as essentially standing still. This is known as the Born-Oppenheimer approximation[12] and plays a pivotal role in quantum chemistry as it is used in the majority of cases. Under this approximation we can consider in Eq.2.16 only the terms involving electrons, hence:

$$\begin{aligned} \hat{H}_{elec} &= \sum_{i=1}^{N} \frac{1}{2}\hat{\mathbf{p}}_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{\hat{r}_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{\hat{r}_{ij}} \\ &= \hat{T} + \hat{V}_{ext} + \hat{V}_{ee}. \end{aligned} \tag{2.17}$$

This leaves us with "only" the problem of solving Eq.2.7 for the electronic states:

$$\hat{H}_{elec} \left|\psi_{elec}(\{\mathbf{R}_A\})\right\rangle = E_{elec} \left|\psi_{elec}(\{\mathbf{R}_A\})\right\rangle \tag{2.18}$$

where now $|\psi_{elec}(\{\mathbf{R}_A\})\rangle$ depends on the positions of the nuclei $\{\mathbf{R}_A\}$ purely parametrically. The total energy of the system will now be given by $E_{tot} = E_{elec} + \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{\hat{R}_{AB}}$ and the solution to the dynamics of such systems and their equilibrium states can be obtained by iteratively computing the forces resulting from the equilibrium electronic state for a fixed nuclear configuration, and then taking one classical integration step in the dynamics of the nuclei. Hence, from now on we will solely focus on the electronic part of the problem. As mentioned in the previous section, the main interest when dealing with this kind of systems

is to find their ground state, and this problem is tackled relying on the variational theorem previously mentioned. The idea is in fact to find a state $|\psi\rangle$ for all of the electrons in the system by solving a problem of finding a functional minimum with the constraint that electrons will not be in the same state for the Pauli exclusion principle. Provided a functional form for the energy of the system as $\langle\psi|\,\hat{H}_{elec}\,|\psi\rangle = E[\psi]$, then the problem is to solve the following in $|\psi\rangle$:

$$\frac{\delta\tilde{E}[\psi]}{\delta\psi} = 0 \tag{2.19}$$

$$\tilde{E}[\psi] = E[\psi] - \sum_{ij}\epsilon_{ij}(\langle\psi_i|\psi_j\rangle - \delta_{ij}) \tag{2.20}$$

where $|\psi_i\rangle$ are the single electron states and where ϵ_{ij} are the Lagrange multiplier from the orthonormality constraint imposed by Pauli exclusion principle. To attempt a solution we start from guessing a wave function, which in order to respect fermionic rules will have to be antisymmetric w.r.t. the exchange in pairs of electrons so that the probability of finding two electrons in the same state is 0 by construction. To obtain this constraint it is customary to use the so-called Slater determinant [13] defined in position representation as:

$$\Psi(x_1, x_2, \ldots, x_N) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(x_1) & \ldots & \psi_N(x_1) \\ \ldots & \ddots & \ldots \\ \psi_1(x_N) & \ldots & \psi_N(x_N) \end{vmatrix} \tag{2.21}$$

which can be written in state vector form as $|\Psi\rangle = \mathcal{S}_-(N)\,|\psi_1, \psi_2, \ldots, \psi_N\rangle$ where $\mathcal{S}_-(N) = \sum_\sigma (-1)^\sigma P_\sigma$ is the sum over the electrons permutations σ where odd permutation give a negative sign thereby obtaining antisymmetry w.r.t. exchange. Note that in general one should consider that there are an infinite number of wave functions in the basis set used $\{\psi\}$. In practice the choice for this usually falls on a finite set of $2K > N$ orthogonal spin-orbitals (K orbitals with spin). The problem of assigning $2K$ spin-orbitals to N electrons produces $\binom{2K}{N}$ possible combinations which is also the number of possible Slater determinants. Using only one Slater determinant and not a linear combination thereof is hence a simplification.

By substituting this in the functional form of the energy $E[\psi]$ and considering how one- and two-body operators from Eq. 2.17 operate on a Slater determinant, it is possible to find that:

$$\tilde{E}[\Psi] = \sum_i \langle\psi_i| \frac{\hat{\mathbf{p}}^2}{2} |\psi_i\rangle - \sum_i \sum_A \langle\psi_i| \frac{Z_A}{\hat{r}_{iA}} |\psi_i\rangle + \frac{1}{2} \sum_{ij} \langle\psi_i\psi_j| \frac{1}{\hat{r}_{ij}} |\psi_i\psi_j - \psi_j\psi_i\rangle - \sum_{ij} \epsilon_{ij}(\langle\psi_i|\psi_j\rangle - \delta_{ij}). \tag{2.22}$$

The condition on the functional derivative can then be shown to yield the following:

$$\frac{\delta\tilde{E}[\Psi]}{\delta\psi_k^*} = \frac{\hat{\mathbf{p}}^2}{2} |\psi_k\rangle - \sum_A \frac{Z_A}{\hat{r}_{kA}} |\psi_k\rangle + \sum_j \langle\cdot\ \psi_j| \frac{1}{\hat{r}_{kj}} |\psi_k\psi_j - \psi_j\psi_k\rangle - \sum_j \epsilon_{kj} |\psi_j\rangle = 0. \tag{2.23}$$

With an appropriate choice of set for the electron states $\{\psi_j\}$, ϵ_{kj} is diagonal (if the wave-functions are already orthogonal then off-diagonal constraints are not needed and set to zero) and we obtain:

$$\hat{\mathcal{F}}[\{\psi_j\}] |\psi_k\rangle = \epsilon_k |\psi_k\rangle \tag{2.24}$$

which is a set of equations where the solution for the energy and state of each electron is dependent on the solution of all the others via the electron-electron interaction term of the Hamiltonian. Notice that, the only two terms coupling to other electrons can be made explicit as follows:

$$\sum_j 2\hat{J}_j |\psi_k\rangle = \sum_j \langle\cdot\ \psi_j| \frac{1}{\hat{r}_{kj}} |\psi_k\psi_j\rangle = \psi_k(\mathbf{x}_k) \sum_j \int \psi_j^*(\mathbf{x}_j) \frac{1}{r_{kj}} \psi_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j \tag{2.25}$$

$$\sum_j \hat{K}_j |\psi_k\rangle = -\sum_j \langle\cdot\ \psi_j| \frac{1}{\hat{r}_{kj}} |\psi_j\psi_k\rangle = -\sum_j \psi_j(\mathbf{x}_k) \int \psi_j^*(\mathbf{x}_j) \frac{1}{r_{kj}} \psi_k(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j \tag{2.26}$$

where $\hat{J}_j$ and $\hat{K}_j$ are known respectively as Coulomb and exchange operators. These operators both act on the i-th electron state considering the action of the other electrons as a mean field, which is evident from the summation and integration operations. This approximation, which is a consequence of the choice of single Slater determinant ansatz, is known as Hartree-Fock (HF) method [14, 15]. The solution of this systems is achieved by first guessing a set states for each electron, usually assigning each one to a different orbital in the set, and then solving the eigenvalue problem from Eq. 2.24 with the exchange operator

computed using the current set (it is an integral operator that depends on its own basis set). The solution of this is a set of eigenvalues and eigenvectors, where each electron will now in general be assigned to a new orbital which will be some linear combination of the basis different from the previous one. This new set of electron states can be used to solve again the eigenvalue problem and this procedure is repeated iteratively until the new electron states are the same as in the previous iteration. Similar methods are known as Self Consistent Field (SCF) methods. While approaches that consider more Slater determinants, such as Full Configuration Interaction (Full CI), are possible, the computational cost quickly becomes prohibitive.

In the 1960s, a new approach to address the complexities coming from these electron-electron interaction terms came alive. Hohenberg and Kohn in their 1964 paper [16] shifted the focus from treating the many-body electron wavefunction $\psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ which depends on 3N coordinates to treating the electron density defined as:

$$\rho(\mathbf{x}) = N \int \mathrm{d}\mathbf{x}_2 \ldots \mathrm{d}\mathbf{x}_N |\Psi(\mathbf{x}, \mathbf{x}_2, \ldots, \mathbf{x}_N)|^2, \tag{2.27}$$

which only depends on the 3 coordinates $\mathbf{x}$ and represents the probability of finding one electron in that point in space. This shift was possible because they derived two main results: (i) the electron density uniquely defines the hamiltonian operator of the system, hence its properties and (ii) the electron density that minimizes the energy of the system is the electron density obtained from the ground state of the system. For a proof of these theorems we refer to the original paper. This allows for a natural correspondence $\hat{V}_{ext} \leftrightarrow \rho_{GS}$ and hence $\hat{V}_{ext} \rightarrow \psi_{GS}(\hat{V}_{ext}) \rightarrow E_{GS}[\hat{V}_{ext}] \rightarrow E_{GS}[\rho]$. This means that the new variational principle will be:

$$E_{GS}[\rho_{GS}] = \langle \psi_{GS}[\rho_{GS}] | \hat{H} | \psi_{GS}[\rho_{GS}] \rangle \leq E_{GS}[\rho] \quad \forall \rho, \tag{2.28}$$

and the generic energy functional will be of the form:

$$E[\rho] = T[\rho] + E_{ext}[\rho] + E_{ee}[\rho], \tag{2.29}$$

where E_{ext} is the energy contribution coming from the interaction with the nuclei and E_{ee} contains the contribution stemming from the electron-electron interaction terms. While by itself it is not trivial to understand how to proceed to solve the electronic problem operatively, it is via the so-called Kohn-Sham method that DFT turns the theoretical results into practical calculations. Following this method we consider a system of N non-interacting independent electrons $\{|\phi_i\rangle\}$ that produces the same electron density as the original system, namely $\rho(\mathbf{r}) = \sum_i \phi_i(\mathbf{r})^* \phi_i(\mathbf{r})$, by moving in an effective potential. This results in a system of independent Shrödinger equations of the form:

$$\left(\frac{\hat{\mathbf{p}}_i^2}{2} + \hat{V}_i^{eff}\right)|\phi_i\rangle = \epsilon_i |\phi_i\rangle . \tag{2.30}$$

To define what is this effective potential, we need some manipulation to the energy functional contributions. Namely we can write:

$$\begin{aligned} E[\rho] &= T[\rho] + E_{ext}[\rho] + E_{ee}[\rho] + T_0[\rho] - T_0[\rho] + E_H[\rho] - E_H[\rho] \\ &= T_0[\rho] + E_H[\rho] + E_{ext}[\rho] + E_{XC}[\rho] \end{aligned} \tag{2.31}$$

where $T_0[\rho] = \sum_i \langle\phi_i|\frac{\hat{\mathbf{p}}_i^2}{2}|\phi_i\rangle$ is the kinetic energy of independent electrons, where $E_H[\rho] = \frac{1}{2}\iint \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \frac{\rho(\mathbf{x})\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}$ is known as Hartree energy and is the classical Coulomb repulsion energy, and where $E_{XC}[\rho] = T[\rho] - T_0[\rho] + E_{ee}[\rho] - E_H[\rho]$ contains all the other effects from the electron-electron interactions. This allows us to write:

$$\hat{V}_{eff} = \hat{V}_H + \hat{V}_{ext} + \hat{V}_{XC} \tag{2.32}$$

where $\hat{V}_{XC} = \frac{\delta E_{XC}[\rho]}{\delta\rho}$. It is easy to understand that now all the complexity of the modeling of the electron-electron interaction is not solved but only hidden in this exchange and correlation potential. This trick, though, is the key to the success of this method as it allows for a lot of freedom for modeling this in different ways depending on the level of computational complexity, approximation and on which system is under investigation while maintaining a system of equations that is iteratively solvable. For example, one can use highly costly

wave-function based methods as a reference to fit some parameters in this exchange and correlation functional. As one can imagine, this freedom spanned a plethora of computational methods, a few examples are:

- **Local Density Approximation (LDA)**: A foundational method in DFT that approximates the exchange-correlation energy as a function of local electron density alone. It belongs to the *semi-local functional family* and is particularly effective for systems with uniform electron density but can struggle with inhomogeneous systems.
- **Generalized Gradient Approximation (GGA)**: This method improves upon LDA by incorporating the gradient of the electron density, allowing for better handling of non-uniform systems. GGA also falls under the *semi-local functional family* and is widely used for various materials due to its improved accuracy [17].
- **Hybrid Functionals (e.g., PBE0)**: These functionals blend DFT with Hartree-Fock theory by including a portion of exact exchange from Hartree-Fock. Hybrid functionals are part of the *hybrid functional family* and offer enhanced accuracy, especially for systems with strong electron correlation effects [18].
- **Density Functional Tight Binding (DFTB)**: A computationally efficient method that simplifies DFT using a tight-binding approximation to treat electron interactions. The core of this method consists in assuming that the actual density is close to the one of the isolated atoms. After expanding at the second order in the difference between actual density and isolated atom density a number of simplifications are possible, among which the repulsive potential only depending on the isolated atom density. DFTB, together with other methods such as GTN-xTB, is considered part of the *tight-binding family* and is particularly useful for studying larger systems where full DFT would be computationally prohibitive [19].

All these methods work well for electronic structure calculations and found their way into materials and drug discovery pipelines. The common shortcoming is the lack of van der Waals interactions in these models due to their reliance on local approximations and limi-

tations in handling non-local electron correlation effects. In the next section we will discuss further how to overcome this issue.

2.1.3 Modeling Non-Covalent Interactions

Van der Waals interactions, or non-covalent interactions, are related to long range effects in the electron density and are the result of correlated charge fluctuations[20, 21]. The energy term that is involved in these forces is the correlation component of the exchange-correlation contribution to the total energy. In order to build on the more local commonly used DFT methods, we start from the Adiabatic Connection Fluctuation Dissipation Theorem (ACFDT). The ACFDT provides a formal framework for connecting the correlation energy of a quantum many-body system to the fluctuations of the electron density. It describes how the correlation energy can be understood in terms of the dissipation of the system to perturbations in the electron density, by saying that an electron system dissipates internal charge fluctuations in the same way as it dissipates fluctuations from changes in the external electric field. This effectively provides an expression for the correlation energy as:

$$E_{\text{corr}} = \frac{1}{2\pi}\int_0^\infty d\omega \int_0^1 d\lambda \int\int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \left[\chi_\lambda(\mathbf{r},\mathbf{r}';i\omega) - \chi_{\lambda=0}(\mathbf{r},\mathbf{r}';i\omega)\right] \frac{1}{|\mathbf{r}-\mathbf{r}'|} \tag{2.33}$$

where λ is the coupling parameter with $\lambda = 1$ being the fully correlated system Hamiltonian and $\lambda = 0$ being the independent particle picture such as the Kohn-Sham one, and where $\chi_\lambda(\mathbf{r},\mathbf{r}';i\omega)$ is the fourier transform of the response function obtained by considering the charge density fluctuation in response to a change in the external field $\Delta\phi(\mathbf{r},t)$:

$$\Delta\rho(\mathbf{r},t) = \int \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \chi(\mathbf{r},\mathbf{r}';t-t')\Delta\phi(\mathbf{r}',t'). \tag{2.34}$$

Considering now that at a charge fluctuation is associated a dipole density one can also relate it to a polarization density as $\Delta\rho(\mathbf{r},t) = -\nabla\cdot\mathbf{P}(\mathbf{r},t)$. This quantity is related to the

external electric field as:

$$\mathbf{P}(\mathbf{r},t) = \int \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\alpha}(\mathbf{r},\mathbf{r}',t,t')\mathbf{E}(\mathbf{r}',t'), \tag{2.35}$$

where $\boldsymbol{\alpha}(\mathbf{r},\mathbf{r}',t,t')$ is the polarizability tensor. Considering now that $\mathbf{E}(\mathbf{r},t) = -\nabla(\Delta\phi(\mathbf{r},t))$, together with the previous relations, one can find that:

$$\chi(\mathbf{r},\mathbf{r}',t,t') = \nabla_{\mathbf{r}} \cdot \nabla_{\mathbf{r}'} \cdot \boldsymbol{\alpha}(\mathbf{r},\mathbf{r}',t,t'). \tag{2.36}$$

This can be used in Eq. 2.33, together with integration by parts, to obtain:

$$E_{corr} = \frac{1}{2\pi} \int_0^\infty d\omega \int_0^1 d\lambda \int\int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \,\mathrm{Tr}\left[(\boldsymbol{\alpha}_\lambda(\mathbf{r},\mathbf{r}';i\omega) - \boldsymbol{\alpha}_{\lambda=0}(\mathbf{r},\mathbf{r}';i\omega))\mathbf{T}(\mathbf{r},\mathbf{r}')\right], \tag{2.37}$$

where $\mathbf{T}(\mathbf{r},\mathbf{r}') = \nabla_{\mathbf{r}}\nabla_{\mathbf{r}'}\frac{1}{|\mathbf{r}-\mathbf{r}'|} = \frac{3(\mathbf{r}-\mathbf{r}')\times(\mathbf{r}-\mathbf{r}')-|\mathbf{r}-\mathbf{r}'|^2\mathbb{I}}{|\mathbf{r}-\mathbf{r}'|^2}$ is the dipole potential tensor, and where $\mathrm{Tr}\left[\boldsymbol{\alpha}\mathbf{T}\right] = \sum_{ij}\alpha_{ij}T_{ij}$. For what concerns polarizability, $\boldsymbol{\alpha}_0$ can be obtained easily for the system of independent particles using Kohn Sham orbitals, while for $\boldsymbol{\alpha}_\lambda$ one can show that the following Dyson equation holds:

$$\begin{aligned}\boldsymbol{\alpha}_\lambda(r,r') &= \boldsymbol{\alpha}_0(r,r') - \iint \mathrm{d}\boldsymbol{r}''\mathrm{d}\boldsymbol{r}'''\boldsymbol{\alpha}_0(r,r'')\mathbf{T}_{xc,\lambda}(r'',r''')\boldsymbol{\alpha}_\lambda(r''',r') \\ &= \sum_{n=0}^{\infty}\langle\boldsymbol{\alpha}_0(-\lambda\mathbf{T}_{xc,\lambda}\boldsymbol{\alpha}_0)^n\rangle(\boldsymbol{r},\boldsymbol{r}')\end{aligned} \tag{2.38}$$

where $\mathbf{T}_{xc,\lambda}(\mathbf{r},\mathbf{r}',\omega) = \mathbf{T}(\mathbf{r},\mathbf{r}') - \lambda^{-1}\nabla_{\mathbf{r}}\nabla_{\mathbf{r}'}f_{xc,\lambda}(\mathbf{r},\mathbf{r}',\omega)$ is the dipole potential together with the exchange-correlation kernel $f_{xc,\lambda}(\mathbf{r},\mathbf{r}',\omega) = \frac{\delta V_{xc,\lambda}(\mathbf{r},\omega)}{\delta\rho(\mathbf{r}',\omega)}$ and where $\langle\cdot\rangle$ was used as shorthand for the integration over spatial coordinates. This kernel is notoriously harder to approximate than the original problem of approximating $\hat{V}_{xc}$ itself, but a different approach is here taken. Starting from the obvious consideration that using this formula for correlation energy on top of a usual DFT method would count short range effects twice, we separate the space integrals by using a range separation like:

$$\iint \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' = \iint \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'(1-f(|\mathbf{r}-\mathbf{r}'|)) + \iint \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' f(|\mathbf{r}-\mathbf{r}'|), \tag{2.39}$$

where f is one at long range and zero at short range, allowing to consider $E_{corr} = E_{corr,sr} + E_{corr,lr}$. This way one can proceed with a DFT approach for the short range component while focusing on the long range one for the addition to the correlation component. By splitting $\mathbf{T}_{xc,\lambda}(\mathbf{r},\mathbf{r}') \simeq (1-f(|\mathbf{r}-\mathbf{r}'|))\mathbf{T}_{xc,\lambda}(\mathbf{r},\mathbf{r}') + f(|\mathbf{r}-\mathbf{r}'|)\mathbf{T}(\mathbf{r},\mathbf{r}') \equiv \mathbf{T}_{sr,\lambda}(\mathbf{r},\mathbf{r}') + \mathbf{T}_{lr}(\mathbf{r},\mathbf{r}')$, one can obtain $\boldsymbol{\alpha}_\lambda(\mathbf{r},\mathbf{r}') = \sum_{n=0}^{\infty}\langle \boldsymbol{\alpha}_{sr}(-\lambda \mathbf{T}_{lr}\boldsymbol{\alpha}_{sr})^n\rangle(\mathbf{r},\mathbf{r}')$ for the polarizability where we introduced the short range polarizability as $\boldsymbol{\alpha_{sr}}$ by contracting all the short range screening effects like $\ldots\boldsymbol{\alpha}_0\mathbf{T}_{sr}\boldsymbol{\alpha}_0\ldots$. This separation results in a long range contribution to the correlation energy of the form:

$$E_{corr,lr} = -\sum_{n=2}^{\infty}\frac{(-1)^n}{n}\int_0^\infty \frac{\mathrm{d}\omega}{2\pi}\iint \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'\mathrm{Tr}\left[\langle(\boldsymbol{\alpha}_{sr}\mathbf{T}_{lr})^n\rangle(\mathrm{r},\mathrm{r}',i\omega)\right]. \tag{2.40}$$

A plethora of methods are possible using the theory laid out so far, and in general can involve truncation of the many body sum or a coarse graining of the system for polarizability approximation or both. For example considering both a truncation to second order and an effective localized isotropic polarizability α_{eff} based on a partitioning of the electron density (such as the Hirschfeld partitioning) in place of $\boldsymbol{\alpha}_{sr}$, leads to:

$$\begin{aligned} E^{(2)}_{corr,\mathsf{lr}} &= -\frac{1}{2}\int_0^\infty \frac{du}{2\pi}\iint \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'\times \mathrm{Tr}[\alpha_{\mathsf{eff}}(\mathbf{r},iu)\mathbf{T}_{\mathsf{lr}}(\mathbf{r},\mathbf{r}')\alpha_{\mathsf{eff}}(\mathbf{r}',iu)\mathbf{T}_{\mathsf{lr}}(\mathbf{r}',\mathbf{r})] \\ &= -\frac{1}{2}\iint \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'\times\left(\frac{3}{\pi}\int_0^\infty du\,\alpha_{\mathsf{eff}}(\mathbf{r},iu)\alpha_{\mathsf{eff}}(\mathbf{r}',iu)\right)\mathrm{Tr}\left[\frac{1}{6}\mathbf{T}_{\mathsf{lr}}(\mathbf{r},\mathbf{r}')^2\right] \\ &= -\frac{1}{2}\iint \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'\,C_6(\mathbf{r},\mathbf{r}')\frac{f(|\mathbf{r}-\mathbf{r}'|)^2}{|\mathbf{r}-\mathbf{r}'|^6}, \end{aligned} \tag{2.41}$$

where we defined $C_6(\mathbf{r},\mathbf{r}') = \frac{3}{\pi}\int_0^\infty du\,\alpha_{\mathsf{eff}}(\mathbf{r},iu)\alpha_{\mathsf{eff}}(\mathbf{r}',iu)$. Another commonly used approximation is then aimed at modeling the frequency response of the α_{eff}. While this can in principle be very complex, it turns out that a good proxy for this is given by the well known model of Quantum Drude Oscillators. This is essentially a charged harmonic oscillator that has polarizability given by:

$$\alpha^{QDO}(\omega) = \frac{q^2}{m(\tilde{\omega}^2 - \omega^2 - i\delta\omega)} \tag{2.42}$$

where q is the charge of the oscillator, m is its mass, $\tilde{\omega}$ is the characteristic frequency and δ is an infinitesimally small number meaning that the absorption spectrum is a Dirac delta located in $\tilde{\omega}$. Thanks to this simplification we obtain an analytical form for the C_6 coefficients in the case of identical oscillator pairs as:

$$C_6^{QDO} = \frac{3}{\pi}\int_0^\infty d\omega\, \alpha^{QDO}(i\omega)\alpha^{QDO}(i\omega) = \frac{3}{4}\left(\alpha^{QDO}(0)\right)^2 \tilde{\omega}, \tag{2.43}$$

while a convenient composition formula for pairs of oscillators (A, B) with different sets of parameters:

$$\begin{aligned} C_6^{A,B} &= \frac{3}{\pi}\int_0^\infty d\omega\, \alpha_A(i\omega)\alpha_B(i\omega) \\ &= \frac{2C_6^{AA}C_6^{BB}}{C_6^{AA}\frac{\alpha_B(0)}{\alpha_A(0)} + C_6^{BB}\frac{\alpha_A(0)}{\alpha_B(0)}}, \end{aligned} \tag{2.44}$$

albeit using this formula introduces a problem with the oscillator approximation as the characteristic frequency needed to reproduce the results of the integral without the oscillator approximation are different in the two integrations 2.42 and 2.44. This problem, though, only leads to a 3% error on the heteroatomic C_6 coefficients[22]. This results in a simplified formula for the correlation energy at long range as:

$$E_{corr,lr} = \sum_{A<B} C_6^{A,B}\frac{f(|\mathbf{R}_A - \mathbf{R}_B|)^2}{|\mathbf{R}_A - \mathbf{R}_B|^6}. \tag{2.45}$$

An example of a method leveraging on this level of approximation is the Tkatchenko-Sheffler (TS) method[22], which is a pairwise coarse grained model for van der Waals interactions. This relies on the observation that atomic polarizabilities of the free atoms linearly scale with the available volume from the partitioning used for the coarse graining, yielding the relations:

$$\alpha_{eff,A}(0) = \alpha_A(0)\frac{V_{eff}[\rho]}{V_{free}}, \quad C_{6,AA} = C_{6,AA}^{free}\left(\frac{V_{eff}[\rho]}{V_{free}}\right)^2 \tag{2.46}$$

which allows us to compute the long range correlation energy based on the current ρ and by only referring to a table of precomputed quantities per each atomic species. If we now want

to maintain this advantage which comes from the coarse-graining choice, but also consider higher order in the many body summation, we can shift to the Many Body Dispersion (MBD) framework[23]. In this framework we start from considering the atoms as three-dimensional QDOs, with polarizability obtained by accounting for the short range dipole interaction with a Dyson equation, considering the initial polarizability as obtained from the scaling law used also in the TS method. After tracing over Cartesian components of the resulting polarizabilities, and hence restoring isotropy, one can write an Hamiltonian for the interaction between these charged harmonic oscillators as:

$$\begin{aligned} H_{MBD} &= \sum_{A=1}^{N} -\frac{1}{2}\nabla^2_{\boldsymbol{\zeta}_A} + \sum_{A=1}^{N} \frac{\tilde{\omega}_A}{2}||\boldsymbol{\zeta}_A||^2 + \sum_{AB} \frac{\tilde{\omega}_A\tilde{\omega}_B}{2}\sqrt{\tilde{\alpha}_A(0)\tilde{\alpha}_B(0)}\boldsymbol{\zeta}_A^T\mathbf{T}_{lr}^{AB}\boldsymbol{\zeta}_B \\ &= \mathcal{T}_{\boldsymbol{\zeta}} + \frac{1}{2}\boldsymbol{\zeta}^T\mathcal{V}\boldsymbol{\zeta}, \end{aligned} \tag{2.47}$$

where $\boldsymbol{\zeta}_A = \sqrt{m}_A(\mathbf{r}_A - \mathbf{R}_A)$ are the relative displacements and where $\mathcal{V}_{AB}^{ij} = \tilde{\omega}_A\tilde{\omega}_B(\delta_{ij} + \sqrt{\tilde{\alpha}_A(0)\tilde{\alpha}_B(0)}\mathbf{T}_{AB}^{ij})$ with $\boldsymbol{\zeta}$ being the direct sum of all the $\boldsymbol{\zeta}_A$ and (i,j) running over the cartesian components of the displacements. Diagonalizing this Hamiltonian, one obtains a system of non interacting oscillators made of collective displacements with a new set of characteristic frequencies $\{\eta_i\}$. This new system has an energy of half the sum of its characteristic frequency. The interaction energy of this system is the equivalent of the long range correlation energy that we are trying to approximate in the electron system and is then trivially:

$$E_{MBD} = \frac{1}{2}\sum_{i=A}^{3N}\eta_i - \frac{3}{2}\sum_{A=1}^{N}\tilde{\omega}_A. \tag{2.48}$$

This method offers a significant improvement over pairwise-additive models by capturing collective many-body effects. The MBD approach provides a balance between computational feasibility and accuracy, making it particularly useful for modeling vdW interactions in large, complex systems such as molecular crystals[24], layered materials[25], and biological macromolecules[26].

2.2 Neural Networks in Chemistry: Can They Understand Molecular Systems?

This section lays out the theory and current practice of the application of neural networks in chemistry, focusing on their potential to understand and predict molecular systems. It starts with a short discussion on classical regression methods leading to a definition of neural network and an explanation of the Universal Approximation Theorem with a recipe for the success of the current deep learning paradigm based on the bias-variance dilemma. Various molecular representations are then discussed as inputs for machine learning models, including fingerprints, text-based representations, 2D graphs, and 3D structures. This is followed by a final section where the main model architectures are introduced together with the methods for molecular representation learning.

2.2.1 The Universal Approximation Theorem and Its Implications

In physics and data analysis, a common starting point is linear regression, where we aim to model a system with a linear equation that best fits observed data. Given a dataset with N samples, each having d features, we represent the input data as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top$ and the output as $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$. The linear regression model can then be written as $y_i \approx \mathbf{w}^\top \mathbf{x}_i + b$ for each sample i, where $\mathbf{w}$ is a vector of weights and b is the bias term. In matrix notation, this becomes $\mathbf{y} \approx \mathbf{X}\mathbf{w} + \mathbf{b1}$, where $\mathbf{b1}$ represents a vector where $\mathbf{1}$ is a vector of ones, allowing for the bias term to be included in the matrix formulation. To find the best fit, we minimize the Mean Squared Error (MSE), given by $\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right)^2$, resulting in a solution for $\mathbf{w}$ that can be expressed as $\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, assuming $\mathbf{X}^\top\mathbf{X}$ is invertible. This model, however, is limited to capturing only linear relationships, which are often insufficient for modeling real-world phenomena where complex, nonlinear relationships are common.

To capture more intricate patterns, we can expand the model by including polynomial terms of the features. For instance, a second-order polynomial regression in one dimension takes the form $y = w_0 + w_1 x + w_2 x^2$, while higher-dimensional data might include terms like

$x_i^j x_k^m$ to represent interactions between variables. With this approach, we extend our feature space by defining a new feature vector $\phi(\mathbf{x})$, which represents the polynomial expansion of the original features. This transforms the model to $y \approx \mathbf{w}^\top \phi(\mathbf{x}) + b$, effectively applying linear regression in a higher-dimensional space. While this approach can fit nearly any continuous function with sufficient polynomial terms, it quickly becomes computationally expensive, especially as the data's dimensionality increases.

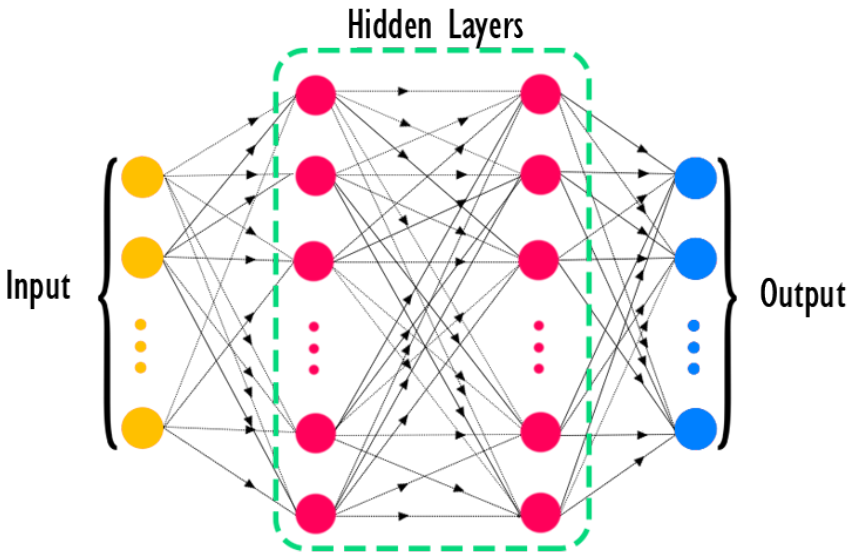


Figure 2.1: Schematic representation of a feed-forward neural network showing the structure from the input layer (yellow nodes), through multiple hidden layers (red nodes), to the output layer (blue nodes), emphasizing the unidirectional flow of information. Mathematically, the network's output $\mathbf{y}$ is given by $\mathbf{y} = \mathbf{W}_L \sigma(\mathbf{W}_{L-1} \ldots \sigma(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}) + \mathbf{b}_L$, where $\mathbf{x}$ is the input vector, $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weight matrices and bias vectors for each layer i, σ represents the activation function.

A natural evolution of this is kernel methods[27], where instead of computing the feature vector $\phi(\mathbf{x})$ and then finding the optimal weights $\mathbf{w}$ based on the training data $\{\mathbf{x}_i\}$, we directly use a kernel function $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$, which does not require the direct computation of ϕ while implicitly projecting data into a higher-dimensional space. A common example is the Gaussian or radial basis function (RBF) kernel, given by $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$, which represents an infinite-dimensional feature space and allows us to capture complex, nonlinear relationships. Prediction then follows by writing $\tilde{y} = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \tilde{\mathbf{x}})$ for any new data point $\tilde{\mathbf{x}}$, where α_i are the coefficients of the basis expansion determined

by solving a linear system involving the kernel matrix of the training data $\mathbf{K}$, with $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. This method thus provides a computationally efficient way to model nonlinear data without an explicit feature transformation.

While kernel methods are powerful, they require us to select a kernel function beforehand, which can be challenging for complex data where the optimal transformation is unknown and can limit the expressive power of the regression law. Neural networks address this challenge by learning the transformation directly from data. A neural network consists of a set of layers, each made of a linear transformation and a non linear function, that subsequently transform the input. Mathematically, this can be expressed for a single-layer neural network as $y = \mathbf{w}^\top \sigma(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+b$, where the matrix $\mathbf{W}^{(1)}$ serves as the weight matrix for the first layer of the neural network, mapping the input vector $\mathbf{x}$ to the hidden layer. Networks such as these ones are the simplest ones and are called feed-forward neural networks or alternatively Multi Layer Perceptrons (MLPs). Similarly, $\mathbf{b}^{(1)}$ is a bias vector for this layer. The term $\sigma(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})$ applies the nonlinear activation function $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ element-wise to the linear transformation of the input, yielding the hidden layer activations. The vector $\mathbf{w}$ represents weights applied to these hidden layer activations, while b is the output layer's bias term. The main result that makes neural networks such a powerful tool is the universal approximation theorem. In its initial form, published in a work by Cybenko in 1989 [28], it states:

Theorem 1 (Universal Approximation Theorem)**.** *Let σ be any continuous discriminatory function. Then finite sums of the form*

$$G(\mathbf{x}) = \mathbf{w}^\top \sigma(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + b \tag{2.49}$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\epsilon > 0$, there exists a sum $G(\mathbf{x})$ of the above form for which

$$|G(\mathbf{x}) - f(\mathbf{x})| < \epsilon \quad \textit{for all } \mathbf{x} \in I_n. \tag{2.50}$$

Here, $\mathbf{x} \in \mathbb{R}^n$ represents the input vector, and $I_n = [0, 1]^n$ is the n-dimensional unit in-

terval over which the approximation is valid. For the definition of discriminative function as well as the proof we refer to the original paper. In essence, this theorem states that a neural network with a single hidden layer and an appropriate choice of activation function can approximate any continuous function $f \in C(I_n)$ on a compact domain I_n to within any desired accuracy ϵ, provided the hidden layer has a sufficiently large number of neurons. While Cybenko's original theorem applies specifically to feedforward networks with a single hidden layer and continuous, sigmoid-like activation functions, later research generalized these results. Hornik, Stinchcombe, and White [29] showed that the theorem holds for networks with a variety of activation functions, not limited to sigmoids, as long as they are non-constant, bounded, and continuous. In deeper networks, these transformations are repeated across multiple layers, allowing the network to learn hierarchical, data-driven representations of the input. Physicists can in fact think of neural networks as analogous to a generalized basis expansion where each layer learns an adaptive basis, transforming inputs through nonlinear functions in ways that correspond to nested interactions among variables. Unlike kernel methods, neural networks do not rely on a pre-specified transformation; instead, they learn the transformation that best fits the data as part of the optimization process. This flexibility is why neural networks are so powerful in capturing high-dimensional, nonlinear relationships that are common in physics and other natural sciences.

While the theoretical power of neural networks as universal approximators is well-established, practical success has hinged on advances in optimization techniques, notably the backpropagation algorithm introduced by Rumelhart, Hinton, and Williams[30]. Backpropagation facilitates efficient computation of gradients for each network parameter through the application of the chain rule, enabling the iterative optimization of a loss function $L(\mathbf{w})$. The gradient descent update for weights $\mathbf{w}$ for a given dataset is represented as:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} L_i,$$

where L_i denotes the loss for a particular data sample, and η is the learning rate. This approach is complemented by Stochastic Gradient Descent (SGD) [31], which applies back-

propagation in practice. SGD operates by performing gradient descent steps based on a surrogate of the actual loss calculated as the average over smaller subsets of the training data, known as mini-batches.

In a more formal context, training a neural network can be framed within statistical learning theory. This framework defines a learning problem by the pair $(\mathcal{D}, L)$, where $\mathcal{D}$ represents the joint data distribution over the feature space $\mathcal{X}$ and the label space $\mathcal{Y}$, while L is the chosen loss function. Given the data, which will typically be a sample $\mathcal{S}$ from the distribution $\mathcal{D}$, the goal is to identify a function $h^* \in \mathcal{H}_m : \mathcal{X} \to \mathcal{Y}$ that minimizes the expected loss, satisfying:

$$\mathbb{E}[L(Y, f^*(\mathbf{X}))|\mathbf{X} = \mathbf{x}] \leq \mathbb{E}[L(Y, h^*(\mathbf{X}))|\mathbf{X} = \mathbf{x}] \leq \mathbb{E}[L(Y, h(\mathbf{X}))|\mathbf{X} = \mathbf{x}] \quad \forall h \in \mathcal{H}_m, \quad (2.51)$$

where f^* is the optimal function minimizing the expected loss for any $\mathbf{x} \in \mathcal{X}$, while $\mathcal{H}_m$ is the space of functions obtainable through algorithmic training on subsamples of $\mathcal{S}$ of size m. In practice, as noted for SGD, we approximate the function that minimizes the expected loss over mini-batches, with the Chernoff-Hoeffding lemma [32] guaranteeing that this approximation closely resembles the expectation over the actual distribution with high probability. Two critical aspects now emerge: (i) the distance between the best attainable function h^* and the actual optimal function f^* is dependent on the data sample $\mathcal{S}$ and is referred to as *bias*, and (ii) the function produced by a specific algorithm, denoted as h_{algo}, will not equate to h^*; this difference arises from the choice of model and optimization method, known as *variance*. The *bias-variance* dilemma [33] highlights the essential ingredients for successful deep learning: large, diverse, and high-quality datasets to reduce bias, alongside expressive architectures and optimization strategies tailored to specific problems to reduce the search space and facilitate convergence to the optimal result.

When these conditions are fulfilled, the outcomes can be remarkably impressive. A landmark example occurred in 2012 with the introduction of AlexNet [34], a neural network model with approximately 60 million parameters trained on the extensive ImageNet dataset. This dataset, which contained over 14 million labeled images, was the largest available at the

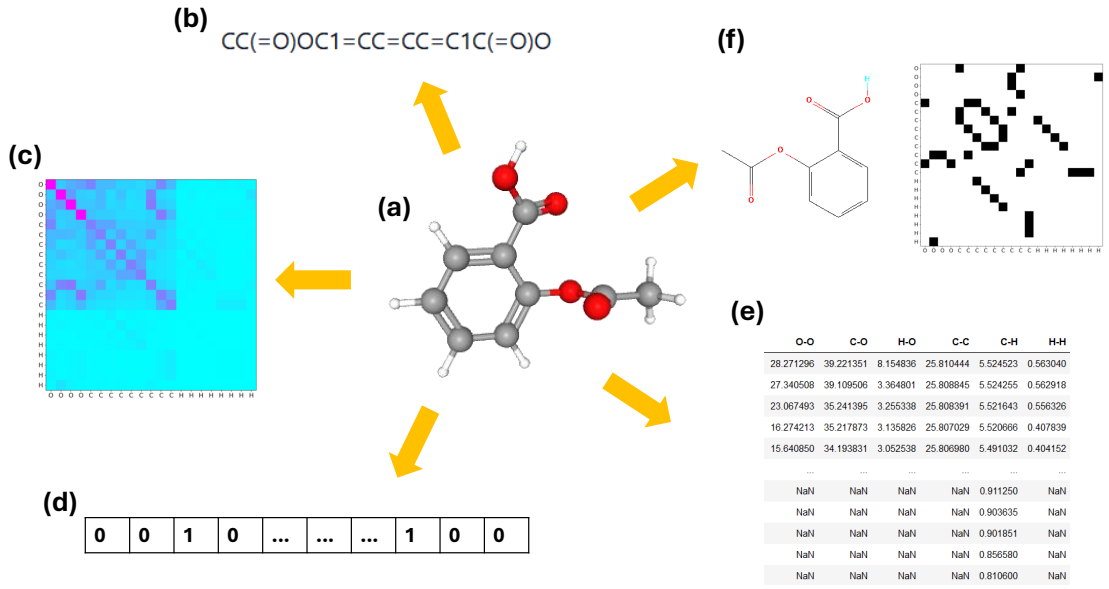
time. AlexNet's Convolutional Neural Networks (CNN) architecture, known for its effectiveness in image processing, achieved unprecedented performance, marking a pivotal moment in deep learning. Furthermore, the successful use of GPUs for training large, overparameterized models demonstrated the feasibility of this endeavours at scale, promising better and better results attainable scaling both model and dataset dimension.

This spurred a revolution in deep learning, which has led to a whole landscape of applications and models. Learning how to deal with specific data modalities lead to specialized architectures for diverse applications, especially with respect to inductive bias and symmetries. Convolutional Neural Networks (CNNs) excel in image processing [35] by leveraging spatial hierarchies through convolutional and pooling layers, enabling efficient feature extraction. Recurrent Neural Networks (RNNs), particularly with LSTM and GRU variants, are designed for sequential data [36], capturing temporal dependencies essential for tasks like language modeling. The Transformer architecture transformed Natural Language Processing (NLP) [37, 38] by using self-attention mechanisms to model relationships between input tokens, leading to breakthroughs in translation and text generation with models like BERT [39] and GPT [40]. The flexibility of deep learning enables various approaches, including pretraining and self-supervised learning, which are powerful strategies for enhancing model performance. Pretraining on large datasets allows models to learn general representations that can be fine-tuned for specific tasks, while self-supervised learning leverages unlabeled data to extract meaningful features without requiring extensive labeled datasets. Moreover, the possibility of learning arbitrary transformations from simple distributions, such as $N(0, 1)$, to complex data distributions facilitates innovative applications through generative models across fields like art creation and drug discovery [41, 42]. As we turn our focus to the realm of chemistry, we explore how these deep learning paradigms can be employed to define and represent molecules, paving the way for advanced machine learning applications in this field.

2.2.2 Defining Molecules as Input for Machine Learning Models

Preparing data for neural network models is often straightforward for many inputs, but finding optimal representations for molecules is complex and an active area of research. At their core, as seen in previous sections, molecules are defined by the distribution in space of positive and negative charges in a bound state. While this is true from the fundamental point of view, a chemist will likely think of a molecule in different ways. Taking the example of water, this can be simply called by the chemical formula H_2O, less commonly using the IUPAC-compliant name of *oxydane*, or again drawn as an undirected 2D graph using the chemical structure. As one can see, the way we represent molecules can vary and depend on what kind of information we need for the problem at hand as well as on other factors such as practicality, namely a chemist writing down a formula for balancing a reaction will not need to know the electronic ground state of the system. When addressing the problem of representing molecules in a machine-readable format, the choice of molecular representation is hence crucial, with each representation providing unique advantages depending on the applications [43]. For the purposes of this thesis, here we will discuss the most common representations considering the categories of fingerprints, text-based, 2D graphs, and 3D based, postponing learned representations to the following section. An overall illustration on the possible ways to see a molecule is reported in Fig. 2.2, where some of the representations mentioned in this section are represented.

Starting from the category of fingerprints, these are compact, binary representations encoding the presence or absence of specific structural fragments or substructures [44]. These binary vectors are highly effective in chemical similarity searches. Fingerprints like Morgan and ECFP (Extended-Connectivity Fingerprints) [45] provide substructure resolution by varying the "radius" parameter, which determines the atomic environment size around each atom. Fingerprints offer computational efficiency and robust performance for similarity-based searches and provide a highly detailed description of molecular substructures, which is particularly advantageous in machine learning and quantitative structure-activity relationship (QSAR) modeling [46, 47], albeit the lack of invertibility limits their utility in generative modeling applications.



O-O	C-O	H-O	C-C	C-H	H-H
28.271296	39.221351	8.154836	25.810444	5.524523	0.563040
27.340508	39.109506	3.364801	25.808845	5.524255	0.562918
23.067493	35.241395	3.255338	25.808391	5.521643	0.556326
16.274213	35.217873	3.135826	25.807029	5.520666	0.407839
15.640850	34.193831	3.052538	25.806980	5.491032	0.404152
...	...	...	...	...	...
NaN	NaN	NaN	NaN	0.911250	NaN
NaN	NaN	NaN	NaN	0.903635	NaN
NaN	NaN	NaN	NaN	0.901851	NaN
NaN	NaN	NaN	NaN	0.856580	NaN
NaN	NaN	NaN	NaN	0.810600	NaN

Figure 2.2: Overview of molecular representations for aspirin. (a) 3D ball-and-stick model of the aspirin molecule, showcasing its atomic structure. (b) SMILES representation providing a textual description of the molecule's connectivity. (c) Coulomb matrix visualized in color, emphasizing atomic interactions based on electrostatic potential. (d) Binary fingerprint indicating the presence or absence of specific molecular features. (e) Bag-of-bonds matrix highlighting interatomic bond information and corresponding features. (f) Two-dimensional molecular graph structure alongside the adjacency matrix, depicting atomic bonds in a binary format.

For what concerns text-based representation, the representation that due to its efficient storage and parsing capabilities is currently the standard in any cheminformatics application is the Simplified Molecular Input Line Entry System (SMILES) [48]. This linear notation simplifies molecular structures by encoding them as strings using specific symbols to denote atoms and bonds. SMILES employs atomic symbols and additional symbols to capture structure. Aspirin ($C_9H_8O_4$) is represented as `CC(=O)OC1=CC=CC=C1C(=O)O`. In this case, `CC(=O)O` encodes the acetyl group, while `C1=CC=CC=C1` describes a benzene ring, with the `1`s marking where the ring begins and ends. However powerful, SMILES strings have the disadvantage that they can be redundant, with multiple SMILES strings potentially representing the same molecule, complicating structural comparisons. While the success of this representation in the world of cheminformatics predates the advent of deep learning, this

representation is well suited for deep learning models for sequential data like LSTMs and GRUs[49], finding application in property prediction tasks as well as generative modeling [50, 51, 52].

2D Graph-based representations, instead, revolve around the fact that a molecule can be seen as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where in most cases the set of nodes (vertices) $\mathcal{V}$ is identified with the atoms and the set of edges $\mathcal{E}$ with the chemical bonds. These representations are very powerful in capturing a lot of the properties of the molecule and have spanned a whole array of works which develop and apply a whole zoo of models from the category of architectures known as Graph Neural Networks (GNN) [53]. A prominent approach within GNNs is the Message Passing Neural Network (MPNN) [54] which will be discussed in the next section. MPNNs, together with other variants such as Graph Convolutional Neural Networks (GCNN) [55] and Spectral Graph Neural Networks (SGNN) [56], form the basis for many models with very successful applications, such as for example the Chemprop model [57]. A more flexible way of using connectivity information is instead found in the Graph Transformer (GT) architecture, where the graph topology is used as a bias in self-attention mechanisms.

3D representations of molecules, finally, are particularly significant as they capture the spatial arrangement of atoms in three-dimensional space, which is crucial for understanding a number of molecular interactions and behaviors. Starting from the set of positions $\{\mathbf{R}_i\}$ and atomic numbers $\{Z_i\}$, a number of possible ways to represent molecules become available depending on our requirements. The idea behind these can be summarized into finding a suitable transformation $\mathcal{T} : \mathbb{R}^3 \times \mathbb{N} \to \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$, where $\{d_1, d_2, \ldots, d_N\}$ are the dimensionalities of the output tensor space, so that for any molecule $\mathcal{T}(\{\mathbf{R}_i\}, \{Z_i\})$ respects a number of properties such as for example translation and permutation invariance, and rotational invariance or equivariance. Following this definition a whole zoo of precomputed representations are available here which are usually based on some physical intuition. Coulomb Matrix (CM) [58] for example represents molecules using the interaction matrix as-

sociated with the Coulomb potential, and is defined as:

$$CM_{ij} = \begin{cases} 0.5\ Z_i^{2.4}, & \text{if } i = j \\ \frac{Z_i Z_i}{|\mathbf{R}_i - \mathbf{R}_j|}, & \text{if } i \neq j \end{cases} \tag{2.52}$$

This representation has the advantage of being rotationally invariant, and allows to retrieve both the original atomic species and the atomic positions up to an affine transformation with the drawback of not being permutation invariant. Bag of Bonds (BoB) [59] is an improvement on this, and attains also permutational invariance by collects all the pairwise interactions used in CMs in *bags* made of all bonds in which a certain atom is involved (excluding repetitions) with some predefined sorting criterion. A more sophisticated representation that also accounts for three-body interactions is the Spectrum of London and Axilrod-Teller-Muto potential (SLATM) representation [60]. This method extends the ideas behind the Coulomb Matrix and Bag of Bonds by incorporating the contributions from triplet interactions, capturing more complex relational data among atoms, at the cost of a much higher dimensional representation to handle. SLATM is an instance of a set of more sophisticated representations, which are based on the idea of representing the chemical environment around each atom by leveraging on symmetry properties of specific sets of functions. We report here the famous example of the symmetry functions put forth by Behler [61], which are radial and angular symmetry functions that describe the local environment around atoms. The radial symmetry function G_i^1 for atom i is given by

$$G_i^1 = \sum_{j \in \mathcal{N}_i} e^{-\eta(r_{ij} - R_s)^2} \tag{2.53}$$

where $\mathcal{N}_i$ is the set of neighboring atoms, $r_{ij} = |\mathbf{R}_i - \mathbf{R}_j|$ is the distance between atoms i and j, and η and R_s are parameters that control the width and position of the Gaussian function, respectively. The angular symmetry function G_{ijk}^2 involving atoms i, j, and k is defined as

$$G_{ijk}^2 = \sum_{j \in \mathcal{N}_i} \sum_{k \in \mathcal{N}_i} f_c(r_{ij}) f_c(r_{ik}) \cos(\theta_{ijk}) e^{-\eta(r_{jk} - R_s)^2} \tag{2.54}$$

where θ_{ijk} is the angle formed by the bonds ij and ik, and $f_c(r)$ is a cutoff function. The three-body angular symmetry function G^3_{ijk} is defined similarly, incorporating interactions among triplets of atoms. The result is a set of numbers that are expressive representation of the atomic environments and are differentiable w.r.t. atomic coordinates and invariant under global rotations and permutations of identical atoms, making them suitable for neural network potentials [62]. Building on these predefined representations, more recent advances have shifted focus towards learnable representations that adaptively capture molecular features while maintaining some requested invariance properties by construction as will be outlined in the next section. These representation are more commonly used in applications where the 3D structure of the molecules and its physical properties become relevant, examples of this are the prediction of HOMO-LUMO gap, dipole moment and atomization energy, as well as the already mentioned neural network potentials and others [63, 64, 65, 66, 67, 68].

2.2.3 Architecture Search and Representation Learning in Neural Networks for Chemistry

As the quest for deeper insights into molecular properties intensifies, representing complex chemical structures in formats suited to computational analysis has become paramount. Traditional representation methods, discussed earlier, have laid essential groundwork in the application of machine learning in computational chemistry. These methods remain competitive and, depending on the task, are still state-of-the-art [69, 70]. However, with recent advancements in computational power and data availability, there is increased focus on optimizing how neural network architectures perceive molecular inputs at their inner layers—a process termed *molecular representation learning* [71, 72]. This section explores the key points of the most widely used model architectures, discusses the most common representation learning techniques used for enhancing model accuracy, generalization, and interpretability in molecular applications.

A widely adopted model is the already mentioned MPNN [54], which is a good starting point since, as we will see, other approaches can be reduced to a similar form. MPNNs begin

with an initial feature representation of nodes (atoms), denoted $\mathbf{e} \in \mathbb{R}^{n \times d}$, where n represents the number of nodes and d the dimensionality of features. Each node's representation is iteratively updated based on "messages" pooled from neighboring nodes. Using an adjacency matrix $\mathcal{A} \in \{0,1\}^{n \times n}$, a transformation matrix $W \in \mathbb{R}^{d \times d}$ with learnable weights, and a non-linear activation f, we define initial and update steps as follows:

$$\mathbf{h}_{ij}^{0} = f\left(\sum_{l} \mathcal{A}_{il} \sum_{k} \mathbf{e}_{lk} W_{kj}\right) \tag{2.55}$$

$$\mathbf{h}_{ij}^{l+1} = \mathbf{h}_{ij}^{l} + f\left(\sum_{l} \mathcal{A}_{il} \sum_{k} \mathbf{h}_{lk}^{l} W_{kj}\right), \tag{2.56}$$

where we used the sum over messages as pooling. The final receptive field depends on the number of update layers, yet increasing layers can lead to *oversmoothing* [73], where node representations converge into indistinguishable states. Similarly, *oversquashing* [74] occurs when pooling messages from numerous neighbors compresses important information into a single representation, limiting distinctive features.

To mitigate these issues, researchers are exploring transformer architectures for graph data [75, 76]. Transformers leverage *self-attention* mechanisms [38], where, given a latent representation $\mathbf{X}_l$ at layer l, the update rule is (for simplicity of notation we do not consider bias terms):

$$\mathbf{X}_{l+1} = \mathbf{X}_l + \mathrm{MLP}\left(\mathrm{softmax}\left(\frac{\mathbf{X}_l W_Q W_K^T \mathbf{X}_l^T}{\sqrt{d}}\right) \mathbf{X}_l W_V\right) \tag{2.57}$$

Here, $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are linear transformations for query, key, and value, respectively, while MLP is a neural network applied uniformly across nodes. In graph-based adaptations, a learnable encoding of topological information (such as an adjacency or topological distance matrix) C biases the attention mechanism, enabling the attention matrix $A = \mathrm{softmax}\left(\frac{\mathbf{X}_l W_Q W_K^T \mathbf{X}_l^T}{\sqrt{d}} + C\right)$ to play the role of $\mathcal{A}$ in MPNNs but with adaptive capabilities, capturing non-local interactions and reducing oversmoothing and oversquashing risks [77].

Interestingly, the first applications of transformers as molecule encoders emerged from text-based representations like SMILES [78, 79, 80]. Here, each SMILES character is to-

kenized, allowing transformers to process molecules as sequential data, similar to natural language sequences. By adjusting attention weights, transformers can learn chemical "grammar" from SMILES strings, yielding impressive results in property prediction, compound generation, and reaction mapping [81, 82, 83].

When handling 3D molecular data, additional constraints like *rotation invariance* and *rotation equivariance* are essential. For scalar quantities (e.g., atomization energy), the learned representation should remain constant regardless of molecular orientation, while tensorial properties (e.g., dipole moments) should vary consistently with rotation. Models like SchNet [63] exemplify 3D molecular MPNNs, where after an initial learnable embedding based on atom type, continuous filter convolution updates modify the atomic representations based on atomic distances. Specifically:

$$\mathbf{x}_i^{l+1} = \mathrm{MLP}\left(\sum_j \mathbf{x}_j^l \circ W(||\mathbf{R}_i - \mathbf{R}_j||)\right) \tag{2.58}$$

Here, $\circ$ denotes element-wise multiplication, and $W(||\mathbf{R}_i - \mathbf{R}_j||)$ is a filter which is learned with a neural network applied to an RBF expansion of the interatomic distance. This approach inherently supports rotation and permutation invariance due to the filter properties and the fact that the $\mathrm{MLP}(\cdot)$ is an atom-wise transformation with shared weights. By adding one extra dimension to the tensors involved, it is easy to take a step forward and achieve equivariance. If we now consider for atom i at layer l scalar features $\mathbf{s}_i^l \in \mathbb{R}^{d\times 1}$ and vector features $\mathbf{s}_i \in \mathbb{R}^{d\times 3}$ with d being the feature dimension, one can for example write the following convolutions:

$$\mathbf{s}_i^{l+1} = \sum_j \mathrm{MLP}(\mathbf{s}_j^l) \circ W(||\mathbf{R}_i - \mathbf{R}_j||) \tag{2.59}$$

$$\begin{aligned}\mathbf{v}_i^{l+1} = &\sum_j \mathbf{v}_j^l \circ \mathrm{MLP}(\mathbf{s}_j^l) \circ W(||\mathbf{R}_i - \mathbf{R}_j||) \\ &+ \sum_j \mathrm{MLP}(\mathbf{s}_j^l) \circ W'(||\mathbf{R}_i - \mathbf{R}_j||)\frac{\mathbf{R}_i - \mathbf{R}_j}{||\mathbf{R}_i - \mathbf{R}_j||}.\end{aligned} \tag{2.60}$$

It is possible to see how here Eq.2.59 is rather similar to what encountered in SchNet,

while in Eq. 2.60 the first sum is a way to introduce nonlinearity in the equivariant vector features with a convolution of equivariant features with an invariant filter, while the second sum is where the propagation of the directional information from the system happens via a convolution of invariant features with an equivariant filter. It is easy to see that any rotation applied to the input system results here in a rotation in the space of features. While this is the updated used in PaiNN [68] and a similar method is used also in EGNN [84], more complex and computationally heavy networks such as Nequip [64] make use of spherical harmonics basis functions and an update rule based on Clebsch-Gordan coefficients.

This non-exhaustive list of network structures highlights powerful representation learners and molecule encoders that can be used in various ways to distill molecular representations optimal for specific tasks. Techniques such as masking [85, 86, 87], most commonly used on text-based representations, allow models to learn the structure of the input data by obscuring certain input features, improving generalization. Pretraining on large datasets [88, 82, 89], equips models with rich feature representations before fine-tuning on specific tasks, effectively addressing the challenge of limited labeled data. Additionally, contrastive learning [90, 91, 92, 93] enhances the ability to differentiate between similar and dissimilar molecular samples, as well as allowing for the combination of multiple data modality for the same molecule fostering robust embeddings that capture multiple aspects of similarities and dissimilarities between input samples. Other approaches involving simpler architectures and precomputed representations can also be leveraged to learn guided latent representations of molecular structures. Successful examples of this are all the representations learned from autoencoders [94, 50], where a network is used to compress the input information into a small dimensional representation that contains enough information for a second network to be able to reconstruct the original input. This, usually coupled with some other operation to guide the shape of the compressed latent space, provides an efficient way to extract meaningful features useful for prediction tasks as well as data analysis.

Chapter 3

The Role of Quantum Chemistry Datasets

The success of machine learning in various domains has fueled great optimism about its potential to accelerate progress in chemistry and materials science. However, realizing this potential requires access to high-quality datasets that can capture the complex relationships between molecular structure and function. This chapter explores the role of quantum chemistry datasets in enabling the development of advanced machine learning models. The chapter begins by providing an overview of some key datasets that have become foundational resources in the field. These datasets offer diverse chemical spaces and a wealth of computed quantum mechanical properties, serving as crucial testbeds for evaluating machine

learning models. Building on this foundation, the Aquamarine dataset, which was designed to address gaps in existing resources by incorporating many-body dispersion interactions and solvation effects via implicit models, is introduced and analysed. The analysis of this dataset reveals that these collective effects become increasingly significant for large, flexible drug-like molecules, particularly when considering solvation. This underscores the importance of accurately modeling non-covalent interactions, which are critical for understanding molecular behavior in biologically relevant conditions. Motivated by these findings, the chapter then explores how quantum computing can be leveraged to develop new approaches for modeling non-covalent interactions, going beyond current approximations coming from computational limitations. Harnessing photonic quantum simulation, a proof-of-concept study is then presented, that maintains the full Coulomb interaction between QDOs, without the need for truncated multipole expansions as commonly considered for most current computational methods. This study, which features a quantum neural network ansatz, uncovers intriguing emergent phenomena such as the formation of entangled "cat states" during the binding process, and provides unique insights into the quantum mechanical underpinnings of molecular interactions.

3.1 Overview of Key Datasets: QM7 and Beyond

As mentioned in the previous sections, the key ingredient of any machine learning application is high-quality and extensive datasets. This enables models to learn effectively from diverse examples and perform well on new, unseen inputs. In the context of computational chemistry, this is indeed a problem as the complexity of the space of combination of atoms is huge. The chemical compound space (CCS) is in fact estimated to contain $\sim 10^{60}$ chemical structures [95], and while enumeration of a high number of structures following chemical rules is feasible, obtaining good quality quantum chemistry calculations for a single molecule can take a non negligible amount of time depending on number of atoms and level of approximation [96], making the time required for calculations diverge quickly with the number of molecules.

Despite the effort to produce good enough datasets appears a gargantuan one, researchers over time started producing and collecting quantum mechanical calculations focusing their interest on covering specific regions of this vast CCS. The generation of a quantum chemistry dataset usually starts with the enumeration of structures based on some chemical rules and constraints dependent on the application focus. For example, for applications in pharma and life sciences, one might want to cover parts of the CCS that are of interest for medicinal chemistry producing synthesizable drug-like molecules. Options are parsing public datasets of chemical structures like PubChem [97], the Cambridge Structural Dataset (CSD) [98] or ZINC [99], or else the enumeration following chemical rules. Using this last method, a strong structure enumeration effort comes from the whole GDB series of datasets [100, 101, 102], which produced comprehensive libraries of chemical structures through systematic techniques. The GDB datasets, such as GDB-13 and GDB-17, utilize algorithms to explore and generate possible organic compounds based on predefined valence rules, connectivity patterns and synthesizability constraints. By employing such methodologies, the GDB series has made a significant contribution to the field by generating billions of unique molecular structures. The following step is to produce, from the chosen subset of chemical structures, the set of 3D structures. Depending on whether conformers or non-equilibrium geometries are needed, the correspondence can be one-to-one or one-to-many. For conformers, a single molecular structure can yield multiple equilibrium spatial arrangements, while non-equilibrium geometries typically correspond to specific states that can be obtained via some sampling method of the PES of the molecule. The level of theory used to produce these arrangements can vary and in general is chosen to be computationally lighter than the one used in the final stages for computing the quantum mechanical properties. The final step consists in running single-point quantum mechanical calculations, that is to say for fixed atom arrangements the electron density is optimized following methods similar to the ones outlined in Section 2.1 and molecular and atomic quantum mechanical properties can be parsed and collected. The choice of approximation for these calculations can be dictated by a number of factors, usually scientists try to find a trade off between computational cost and their necessities in terms of accuracy and system under study.

Following a similar procedure, among the first and most influential datasets, QM7 and QM7b [101, 58, 103] have played a pivotal role in this field. Derived from the GDB-13 database, QM7 includes approximately 7k small organic molecules, exhaustively covering the space of molecules with up to seven heavy atoms from the set (C, N, O, S) and with atomization energies calculated using DFT at the PBE0 level. QM7b extends this dataset adding Cl atoms to the set and by incorporating 13 additional quantum-chemical properties, such as polarizabilities, HOMO-LUMO gaps, and dipole moments, computed using different methods (ZINDO, SCS, PBE0, GW).

Building upon the foundation laid by QM7 and QM7b, QM7x [104] further extends these datasets by offering more molecular diversity and including dispersion energy via MBD corrections both at the stage of structures relaxation and for the final single point calculations. QM7x includes approximately 42k equilibrium geometries resulting from a thorough search of conformers per each molecule considered in QM7b relaxed using DFTB3+MBD. Additionally, QM7x includes multiple non-equilibrium geometries per each molecular structure, obtained via normal modes sampling, resulting in $\sim$ 4.2M structures and making it especially useful for studies that require an understanding of molecular flexibility and the impact of conformational diversity on molecular properties. The final single point calculations are carried out using PBE0+MBD with tightly converged numeric atom centered orbitals [105] and yield both global and atomic molecular properties for a total of 42 physicochemical properties.

QM9 [106], another foundational dataset, expands the scope of molecular diversity with over 134k stable organic molecules containing up to nine heavy atoms in the set of (C, N, O, F). Albeit a much smaller number than what found in QM7x, QM9 includes some quantum-chemical properties like HOMO and LUMO energies, dipole moments, and polarizabilities calculated using DFT at the B3LYP [107, 108] level (hybrid functional) with 6-31G* basis set [109] with no dispersion correction. Also differently from QM7x, this dataset does not consider conformers per each of the considered molecules. Together with the previously mentioned datasets, this is among the most common benchmarks for machine learning applications being employed in a number of works with different scopes.

The QMugs dataset is a more recent addition to quantum chemistry resources, designed specifically for drug discovery and pharmaceutical applications. It contains over 665k drug-like molecules selected from the ChEMBL [110] database, a manually curated database of bioactive molecules with drug-like properties. In this case three conformers per molecule are considered and relaxed at the GFN2-xTB level [111], followed by single point calculations at the ωB97X-D [112] level (hybrid functional with empirical dispersion correction) using the def2-SVP basis set [113]. The properties considered here are 33, with 9 properties reported both at GFN2-xTB and ωB97X-D/def2-SVP level and including vibrational frequencies and thermodynamic data at the GFN2-xTB level. Also, wavefunctions data is included in the form of densities and orbital matrices. Albeit the quantum chemistry methods used here are a compromise between computational cost and accuracy when compared to the more precise calculations found in QM7x, this dataset contains much bigger molecules with up to 100 heavy atoms. Furthermore, being sampled from ChEMBL, it is of particular interest for applications in pharma and life sciences.

With another application in mind, the ANI (Atomic-Network-Inspired) datasets are a series of machine learning-ready datasets specifically curated for training neural network potentials in quantum chemistry. Developed to provide accurate molecular energies and forces, the ANI datasets include extensive collections of molecular conformations with associated quantum mechanically computed energies and force vectors, generated using different levels of density functional theory (DFT). Each dataset in the ANI family—such as ANI-1 [114], ANI-1x and -1ccx [115], and ANI-2x [116] has distinct features and levels of accuracy. For example, ANI-1 contains around 20 million configurations of small organic molecules (around 50k molecules with meta-stable geometries and extensive non-equilibrium sampling) with energies and forces computed with DFT at the B3LYP level of theory. ANI-1x is instead composed of 5M configurations obtained with an active learning approach and ANI-1ccx enhances this by reporting the results of coupled cluster calculations on a selection of 10% of the configurations in the ANI-1x dataset. ANI-2x is instead an extension of ANI-1x in the range of included elements. These datasets have proven highly valuable for developing transferable neural network potentials, enabling accurate energy predictions across a wide

chemical space with minimal computational cost, and are particularly useful for simulations requiring Ab Initio Molecular Dynamics (AIMD) at a fraction of its actual cost.

The GEOM dataset [117] provides a comprehensive collection of energy-annotated molecular conformations, supporting machine learning applications in molecular property prediction and generation. Compiled through advanced sampling and density functional theory (DFT) methods, it includes approximately 37 million conformations across 450,000 unique molecules. The dataset emphasizes the value of conformer ensembles rather than isolated molecular structures, addressing a gap in existing resources. It consists of 133,000 molecules from the QM9 dataset and 317,000 experimental compounds relevant to biophysics, physiology, and physical chemistry, including 1,511 BACE-1 [118] inhibition samples with high-quality energy annotations in an implicit water solvent. Additionally, the CENSO [119] refinement process further optimizes 534 of these species with DFT in a water solvent model, enhancing the dataset's utility for biologically relevant modeling. GEOM facilitates training models for predicting molecular properties based on accessible conformations and for generating realistic 3D structures, positioning it as a valuable tool in computational chemistry and molecular design research

Together with others, these datasets provide a foundation of quantum chemistry data, covering a wide array of molecular properties, conformational diversity, and, in datasets like ANI and QM7x, non-equilibrium geometries. These extensive resources enables the development of accurate, scalable, and interpretable machine learning models across applications in drug discovery, raging from molecular property prediction [120], to neural network potentials [121], generative models for docking and pose estimation [122, 123, 124] and can even serve as a pretraining for related endpoint tasks [125]. However, a crucial aspect for life sciences remains underexplored: the role of modern dispersion corrections for noncovalent interactions together with solvent effects on the molecular conformer ensembles and quantum mechanical properties. In the following section, I delve into these interactions, their implications in implicit solvent calculations, and introduce a new dataset specifically designed to address these critical aspects.

3.2 Non-Covalent Interactions and Solvation in Molecular Properties: the Aquamarine dataset

As seen in the previous section, with some exceptions, in the realm of existing quantum chemistry datasets the most common dispersion corrections are generally methods that considered some cut to the many body expansion, as mentioned in 2.1.3. This is justified by the computational cost and the regimes in which these datasets cover the CCS. Many body effects, though, can have significant repercussions on the potential energy surface (PES), as demonstrated by studies comparing pairwise methods like TS with MBD corrections. For example, research on polymers and polyalanine [126] has shown that MBD corrections produce a much smoother PES and a different force profile during MD when compared to pairwise corrections. This results in fundamentally distinct molecular conformations, especially in systems where cumulative long-range interactions drive folding or clustering behavior. In applications pertinent to the life sciences, understanding a molecule's configurational landscape is often crucial, and particularly so in a solvated environment. Biological molecules, including many drug-like compounds, interact within aqueous environments where the many body nature of non-covalent interactions together with solvation can significantly alter both conformational preferences and stability. Effectively modeling these two effects can help simulate the molecule's true functional shape in biologically relevant conditions by accounting for the electrostatic and dispersion forces in a more realistic environment, e.g. *in solutio*. These interactions also influence key properties like binding affinity, solubility, and overall reactivity [127, 128, 129, 130, 131]. Consequently, an accurate dataset that incorporates many-body dispersion alongside solvent effects is essential for realistic modeling of molecular properties in life science applications. Notable mentions in this direction are the GEOM dataset, which contains geometries generated considering molecule-solvent interactions with the implicit analytical linearized Poisson-Boltzmann (ALPB) [130] for 1511 molecules form the BACE dataset (dataset of candidate inhibitors for a specific enzyme), and the SPICE dataset [132], considering geometries for 26 aminoacids with an explicit solvation model based on the AMBER-14 force field. Despite these efforts, a more systematic

assessment of solvent effects as well as collective dispersion interactions from small to large drug-like molecules is still missing. In particular, such an effort would be useful: (i) study the combined effects of different dispersion corrections and solvation on the conformational landscape of molecules, covering different sizes from small to large drug-like molecules of interest for medicinal chemistry applications (ii) to offer a large set of molecular (global) and atom-in-a-molecule (local) physicochemical properties that would enable a comprehensive exploration and analysis of these interactions in structure-property and property-property relationships throughout chemical space, and (iii) to provide accurate and reliable QM data that will enable the construction of models for describing covalent and non-covalent vdW interactions in large (solvated) molecules.

Here the Aquamarine (AQM) dataset is discussed, as a dataset designed to address these challenges. AQM features extensive conformational sampling of molecules containing C, N, O, F, P, S, and Cl, spanning a wide range of sizes and compositions. Geometries are optimized in both gas-phase and implicit solvent environments to enable direct comparisons. Additionally, the dataset includes over 40 detailed physicochemical properties for gas-phase structures and for solvent-optimized conformations.

3.2.1 Selection of relevant molecular structures

The dataset's uniqueness originates already in the selection of molecules, which were specifically chosen to reflect compounds typical of a pharmaceutical corporate library. To achieve this, 5000 compounds were sampled from ChEMBL and compared with the Johnson & Johnson Innovative Medicines corporate database. Compounds were filtered by excluding those with molecular weight above 1200, more than 30 rotatable bonds, QED scores under 0.4, or heavy atom counts exceeding 200. After removing undesirable substructures through diversity selection and manual review, $2,635$ unique molecules remained (≤ 60 non-hydrogen atoms, $N \leq 116$). These included building blocks, lead-like compounds, protein degraders, and macrocycles. RDKit was used to generate and optimize all possible stereoisomers while maintaining connectivity. QM calculations assessed isomer stability, resulting in $\sim 10,000$ total structures. Initial 3D conformations were created using RDKit and optimized with

MMFF94 force field [133, 134, 135, 136, 137].

3.2.2 Conformational sampling

Conformational sampling is pivotal in generating the AQM dataset, as it ensures comprehensive exploration of the potential energy surface (PES) and molecular property space of large drug-like molecules. After evaluating various workflows, the CREST [138] code was adopted, which employs metadynamics (MTD) and the semi-empirical extended tight-binding method (GFN2-xTB) to generate 3D conformations. The MTD-based algorithm uses atomic root-mean-squared deviation (RMSD) values to guide sampling, generating conformers by an iterative process of exploration and semi-empirical optimization selecting conformers when they overcome specific energy (> 12.0 kcal/mol) and RMSD (> 0.1 Å) thresholds. The lowest-energy conformers undergo molecular dynamics simulations at 400K and 500K to explore low-energy barrier crossings, supplemented by a genetic Z-matrix crossing algorithm to further diversify the conformer ensemble. Calculations, including geometry optimization and conformational searches, were performed in implicit water using the GBSA model [139], which has been successfully used in the study of free solvation energies of neutral/ionic molecules and the folding of short peptides. This procedure yielded 2242490 conformers for 2635 molecules. Unlike other public datasets, here a method was devised in order to select representative conformers by clustering structures with RMSD values under 1.5 Å and performing energy-based filtering using DFTB3+MBD [140, 141, 142] single-point calculations to identify those with distinct total (E_{tot}) and many-body dispersion (E_{MBD}) energies, where MBD was considered given its relevance in large molecules and molecular crystals [143]. This method was applied on a subset of 1653 molecules with up to 54 heavy atoms, and reduced the number of conformers for those from 280182 to 59783, prioritizing structural and energetic diversity while maintaining stability assessments. The conformer search method based on CREST was also validated against other common codes, namely MAESTRO, Omega[144] and RDKit[145]. This was done by randomly selecting 18 chemical compositions containing approximately 50 atoms, and resulted in a different total number of conformers depending on the code employed for their generation, i.e., CREST $\rightarrow$ 3747

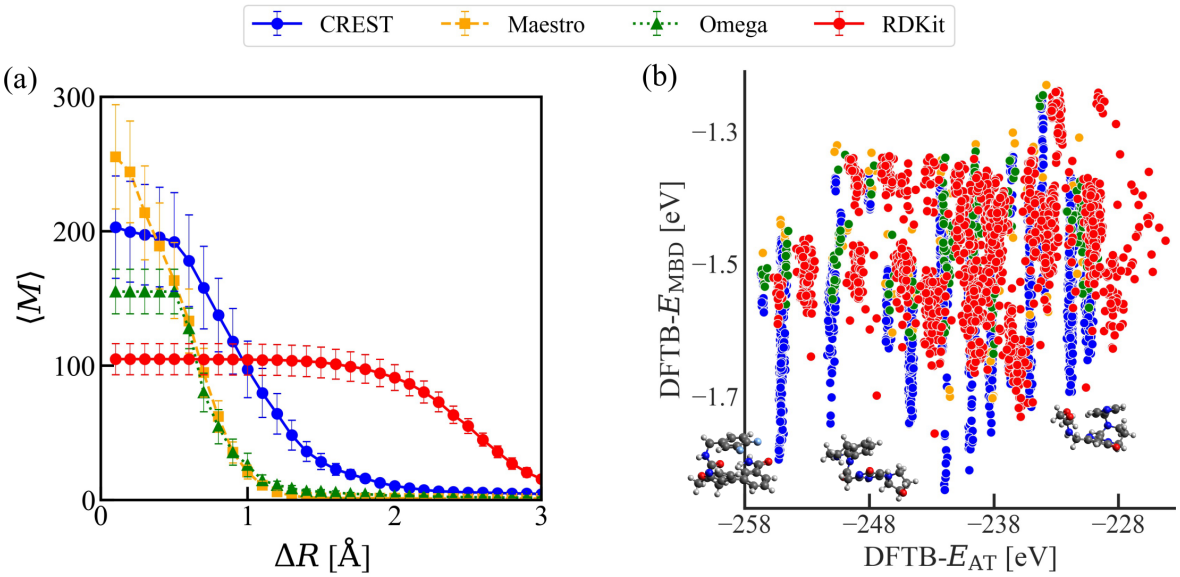


Figure 3.1: (a) Dependence of the average number of clusters, $\langle M \rangle$, on the root-mean-square deviation (ΔR) between conformers generated by the four conformational search workflows analyzed in this study (see legend above). The average is calculated over 18 randomly selected compositions, each containing approximately $N = 50$ atoms. (b) Two-dimensional property space defined by E_{AT} and E_{MBD} for the representative conformers obtained. Energy calculations were performed using the DFTB3+MBD method. Examples of structures uniquely generated by the CREST workflow are included in the graph.

conformers, Maestro $\rightarrow$ 100 conformers, Omega $\rightarrow$ 204 conformers and RDKit $\rightarrow$ 1872 conformers, confirming that CREST shows a more organized coverage of the energetic space defined by DFTB-E_{AT} and DFTB-E_{MBD} (see the well-defined cluster in Fig. 3.1). To enhance the dataset's utility in studying molecule-solvent interactions, structures were optimized in both gas phase and implicit water, resulting in the AQM-gas and AQM-sol subsets. Calculations were conducted at the DFTB3+MBD level, using the Atomic Simulation Environment (ASE)[146] interfaced with DFTB+ code [147] and using GBSA for solvated geometries. While most structures were identified as local minima, some remained at saddle points on the PES.

3.2.3 Analysis of solvent effects in property space

The DFTB-optimized structures were then subjected to more accurate quantum mechanical (QM) single-point calculations using dispersion-inclusive hybrid density functional the-

ory (DFT) to compute energies, forces, and various physicochemical properties. AQM-gas molecules were evaluated at the PBE0+MBD level, which has been chosen as baseline level of theory for property calculations due to its well-established accuracy and reliability in the description of intramolecular degrees of freedom as well as intermolecular interactions in organic molecular dimers, supramolecular complexes, and molecular crystals [148, 149, 150]. For AQM-sol molecules, instead, solvation effects were included on top of PBE0+MBD through the modified Poisson-Boltzmann (MPB) model[151]. The MPB model, which accounts for electrolytic solvation by solving the size-modified Poisson-Boltzmann equation and models the Stern layer to include non-mean-field ion-solute interactions, has been shown to provide an accurate description in the study of diverse electrochemical reactions [152, 153, 154, 155]. Calculations utilized the FHI-aims code (version 221103) with "tight" settings for basis functions and grids, achieving energy convergence to 10^{-6} eV and force accuracy to 10^{-4} eV/Å. MBD energies and atomic forces were computed using the range-separated self-consistent screening (rsSCS) approach [23], while molecular polarizabilities and C_6 coefficients (both atomic and molecular) were derived from the SCS approach [156]. TS dispersion energies, vdW radii, and Hirshfeld ratios were also calculated, the latter representing Hirshfeld volumes normalized by free atom volumes. Atomization energies were obtained by subtracting atomic PBE0 energies from the total molecular energy, and exact exchange energy reflected the Hartree-Fock exchange contribution within the exchange-correlation energy. These calculations provided a comprehensive dataset of electronic properties for both gas-phase and solvated conformers, covering both the intensive/extensive range and the global/atomic range, which, together with the respective geometries, allows us to visualise and analyse the effects of solvation and dispersion under a number of different points of view. Indeed, starting from the comparison of geometries from AQM-gas and AQM-sol one can already get a picture of the effects of solvation on geometry configuration by system size. Fig. 3.2(a) illustrates, in fact, the size dependence of the averaged RMSD deviation ΔR between geometries optimized in gas phase and in implicit water (blue dots), along with the total range of ΔR values observed across different molecular sizes (blue shaded region). The results indicate that the geometries coming from small

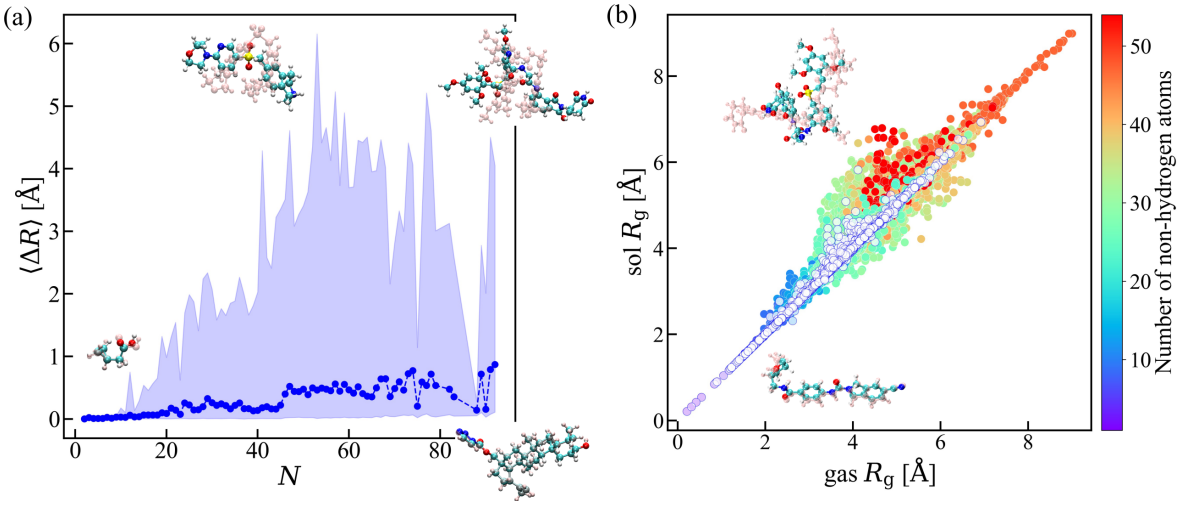


Figure 3.2: (a) Averaged root-mean-square deviation ($\langle \Delta R \rangle$) as a function of molecular size, comparing geometries optimized in the gas phase with those optimized in implicit water using the GBSA model and the DFTB3+MBD method. The blue shaded region indicates the range of ΔR values observed across different molecule sizes (N). (b) Relationship between the gyration radius (R_g) of conformations in the gas phase and in solution. Data points are color-coded by the number of heavy atoms (non-hydrogen). This analysis includes the 59,783 conformers spanning low- and high-energy states from the AQM-gas and AQM-sol datasets. Example conformations are included to demonstrate structural changes induced by solvation, with solvated configurations highlighted by pink spheres.

molecules (N$\leq$20 atoms) exhibit $\langle \Delta R \rangle < 0.1$ Å, implying that they are only minimally influenced by solvent interactions. In contrast, for molecules with N $>$ 40 atoms, solvent effects become more pronounced, leading to greater deviations in ΔR values (greater than 2.0 Å). A similar trend is observed when comparing the gyration radius R_g for gas-phase and solvated molecular structures, as shown in Fig. 3.2(b). Notably, larger compounds (e.g., N$\sim$90) exhibit extensively constrained structures that remain largely unaffected by the interaction with implicit water. These findings are of particular interest in pharmaceutical research, especially when generating non-equilibrium structures for ML force fields, as the differences just outlined imply that solvent effects significantly alter the potential energy surface (PES) and must be considered in dataset creation. The dataset can be further analyzed also from the point of view of electronic global properties. As an example, the 2D property space defined by isotropic molecular polarizability α and the HOMO-LUMO gap E_{gap} for the 59783 conformations in AQM-gas and AQM-sol was analysed, as well as the most stable conformer per

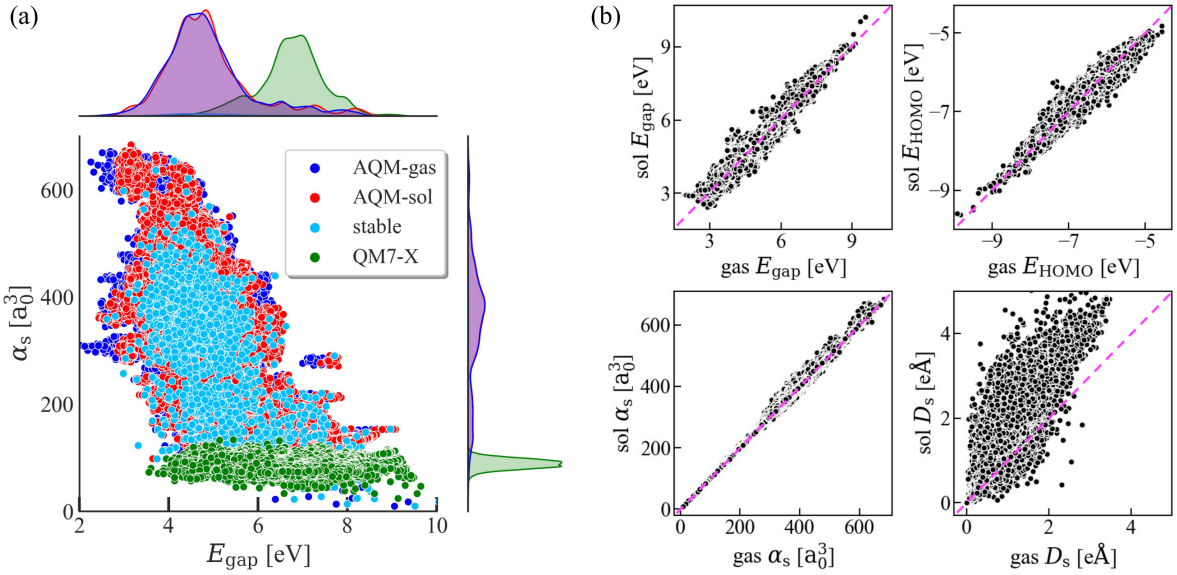


Figure 3.3: (a) Two-dimensional projection of the AQM property space defined by the isotropic molecular polarizability (α_s) and the HOMO-LUMO gap (E_{gap}). Property values are shown for the 59,783 low- and high-energy conformers from AQM-gas (blue) and AQM-sol (red), along with the most stable conformer of the 1,653 unique molecules in AQM-sol (cyan). For comparison, equilibrium structures from QM7-X are included (green). Distributions for each property and dataset are displayed in the marginal plots (top and right). (b) Correlation between gas-phase and solvated structures for selected QM properties: E_{gap}, HOMO energy (E_{HOMO}), isotropic polarizability (α_s), and scalar dipole moment (D_s).

molecule in AQM-sol (1,653 conformations), as shown in Fig. 3.3(a). For comparison, values for QM7-X equilibrium molecules are also plotted (green circles). The results show that AQM molecules cover a much broader range of α, surpassing QM7-X by a factor of 6, due to the extensive nature of α. The E_{gap} range spans about 2.5 eV, with the mean value dropping from 7.0 eV to 4.5 eV, as seen in the distribution plots in the top panel of Fig. 4(a). The slight differences between AQM-gas and AQM-sol in this space are likely due to compensation between fluctuations in α (mainly for $\alpha > 300a_0^3$) and the more sensitive behavior of E_{gap} to implicit water, as shown in the correlation plots in Fig. 3.3(b). These findings are especially relevant when considering the problem of tailored design of large drug-like molecules with specific (α,E_{gap}) values[157, 158]. Notably, the conformational sampling improved coverage of both properties, linking isolated regions tied to specific molecular sizes and compositions. Fig. 3.3(b) also shows the correlation between HOMO energy E_{HOMO} and dipole moment

D_s for the conformers in AQM-gas and AQM-sol, illustrating that intensive properties are more sensitive to solvent interactions in QM calculations compared to extensive properties.

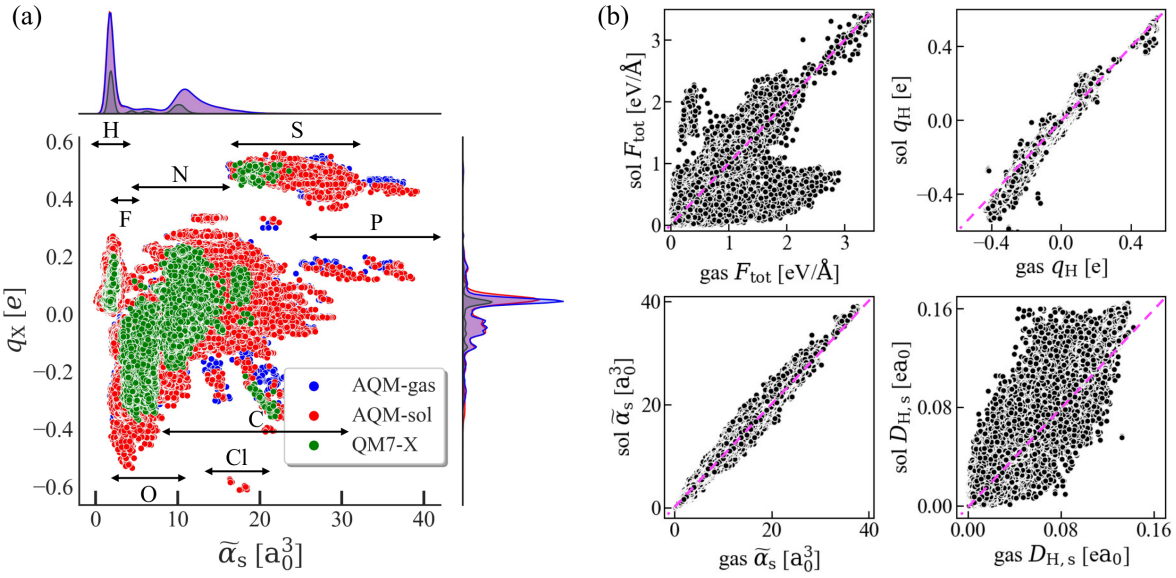


Figure 3.4: (a) Two-dimensional projection of the AQM property space defined by Hirshfeld charges (q_H) and atomic polarizabilities (α_s). Property values are displayed for 59,783 low- and high-energy conformers from AQM-gas (blue) and AQM-sol (red). For comparison, values from QM7-X equilibrium molecules are included (green). Arrows highlight the regions of the property space corresponding to specific local chemical environments for different atom types (X). Marginal frequency plots (top and right) show the distributions of each property for the datasets. (b) Correlation between gas-phase and solvated structures for selected local QM properties, including total atomic forces (F_{tot}), Hirshfeld charges (q_H), atomic polarizabilities (α_s), and atomic Hirshfeld dipole moments ($D_{H,s}$).

Focusing on atom-level properties, instead, enables a better understanding of how solvent interactions impact chemical environments and the local electron density. For example, Fig. 3.4(a) illustrates the 2D property space defined by Hirshfeld charges (q_H) and atomic polarizabilities (α_s) for all the conformations in AQM-gas and AQM-sol, alongside QM7-X equilibrium molecules (green circles). Distinct clusters emerge, primarily associated with specific atom types, with slight overlaps. This is especially interesting, and can potentially be useful information for representation learning purposes, as a number of works in chem-informatics use partial charges and similar information to induce better representations via chemical environment understanding. Notably, implicit solvation strongly influences q_H, especially for heavier atoms like P, S, and Cl—key elements in pharmaceutical compound

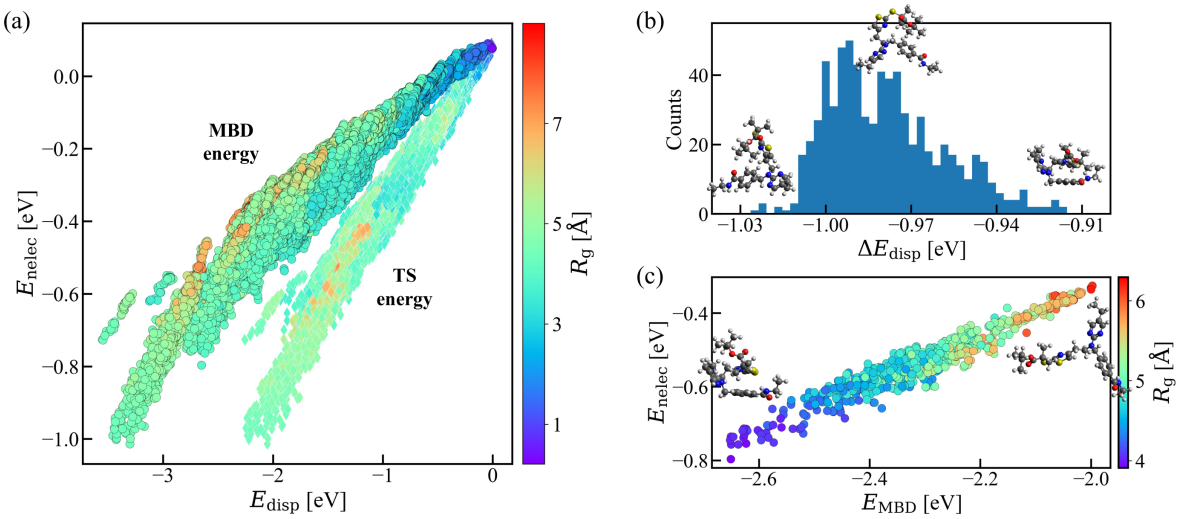


Figure 3.5: (a) Variation of the non-electrostatic free energy (E_{nelec}) as a function of the dispersion energy (E_{disp}) for all solvated structures in AQM-sol. Two different dispersion models are considered: Many-body dispersion (MBD) and Tkatchenko-Scheffler (TS). (b) Frequency distribution of the energy difference between MBD and TS models ($\Delta E_{\text{disp}} = E_{\text{MBD}} - E_{\text{TS}}$) for 720 low- and high-energy conformers of the molecule C29H39N5O3S2. (c) Variation of E_{nelec} as a function of E_{MBD} for all conformers of C29H39N5O3S2. In panels (a) and (c), the color coding corresponds to the gyration radius (R_g) of each solvated structure. Selected conformations are shown in panels (b) and (c) to highlight the impact of variations in these properties on the molecular structure.

design and property prediction. As shown in Fig. 3.4(b), molecule-solvent interactions exert a stronger influence on local properties than global ones, a trend further supported by variations in atomic forces (F_{tot}). This observation is particularly noteworthy, as many neural network potentials are trained on forces derived from gas-phase molecular configurations. It underscores the importance of datasets that account for solvation effects to better represent regions of phase space relevant to more realistic scenarios.

A final analysis focuses on the interplay between solvent effects and dispersion corrections, which is key to understanding the behavior of large drug-like molecules. To this end, in Fig. 3.5(a) a correlation plot between non-electrostatic component of solvation energy E_{nelec} and dispersion interaction energy, E_{disp} is reported, with the last one being calculated using two established methods: MBD and TS. The data points, colored by the gyration radius R_g of each solvated structure, reveal a strong correlation between these properties, emphasizing the importance of accounting for both molecule-solvent and dispersion interactions when

studying large molecules, as in the generation of the AQM dataset. Notably, the increasing divergence between energies obtained via the MBD and TS methods with growing system size confirms the significant impact of considering many-body interactions on the energetic description of these compounds. To explore this further, an analysis was conducted by selecting the molecule $C_{29}H_{39}N_5O_3S_2$ ($N = 78$ atoms) with 720 conformers, examining their respective energy values. These conformers exhibit a dispersion energy difference between the two considered methods, ΔE_{disp}, of up to ≈ 1.0 eV, with smaller values corresponding to compact molecular structures and larger values observed for extended ones (see Fig. 3.5(b)). In addition to its size dependence, Fig. 3.5(c) shows that, like E_{MBD}, E_{nelec} also varies with molecular conformation, further highlighting the interplay between solvation and structural effects.

3.3 Modeling Non-Covalent Interactions Using Photonic Quantum Simulation and Neural Networks

The analysis of the Aquamarine dataset in Section 3.2 demonstrated that many-body effects in dispersion interactions can be significant, especially for large drug-like molecules and particularly so when considering solvation effects. As discussed in Section 2.1.3, considering a pairwise model effectively truncates the series expansion of the correlation energy. While the MBD model advances beyond this limitation by treating the full series, it still relies on the approximation of localized dipoles for the QDOs. A more complete treatment considering the full Coulomb interaction between QDOs could theoretically provide higher accuracy. Such an approach was proposed by Whitfield and Martyna[159], who showed that by carefully selecting the charge, mass, and characteristic oscillation frequency of QDO pseudo-particles, one could enable the description of response properties, many-body induction, and dispersion interactions to infinite order. However, practical implementation of this model has been limited by the computational expense of classical methods such as imaginary-time path integration or Diffusion Monte Carlo.

In this section, we examine how recent advancements in quantum computing can ad-

dress the computational challenges inherent to classical approaches. Quantum computers, which harness quantum mechanical phenomena such as superposition and entanglement, hold significant promise for applications in chemistry for two main reasons. First, they can efficiently represent quantum states that would otherwise require exponential resources on classical hardware. Second, they enable the direct simulation of quantum systems by mapping the problem effectively onto quantum hardware. For a review on the topic see [160]. While the current state of hardware strongly limits quantum computing's applicability, applications on noisy hardware are an active field of study, especially in quantum chemistry, with the name of NISQ (Noisy Intermediate Scale Quantum) applications [161, 162, 163]. In this context, this section presents an approach to simulating non-covalent interactions with photonic quantum hardware. Simulating this kind of interaction is oddly something not very explored in the current state of research. One previous excellent work on the topic is [164], which simulates QDO interaction on NISQ hardware and goes beyond the dipole approximation, but still relies on a truncation in the multi-pole expansion, not treating a full-Coulomb interaction. Here, it is shown that photonic hardware is a more natural choice for this purpose. A novel method is proposed for simulating full Coulomb interactions in QDO systems, leveraging the observation that photons, like the modes of quantum harmonic oscillator states in the QDO model, exhibit a bosonic nature. This proof-of-concept implementation focuses on a system of two Coulomb-coupled QDOs (cQDOs), a simplified yet insightful scenario that allows to establish core principles while facilitating detailed analysis. Beyond demonstrating the viability of this quantum computing approach, these results unveil intriguing qualitative features, such as characteristic binding curves and the formation of quantum states known as entangled cat states at saddle points. This last point of particular interest, as entanglement is a purely quantum property of multipartite states that cannot be separated in tensor products of the single systems [165], while cat states are a family of quantum states that are superposition of classical states, and are extensively studied for their connection to the problem of obtaining macroscopic quantum superposition effects [166, 167].

3.3.1 Definition of the Model

The fundamental description of the considered system of cQDOs begins with the Hamiltonian for N quantum Drude oscillators in three dimensions:

$$\hat{H} = \sum_{i=1}^{N} \left(\frac{\hat{\mathbf{p}}_i^2}{2m_i} + \frac{1}{2} m_i \omega_i^2 \hat{\mathbf{x}}_i^2 \right) + \sum_{i<j} V_{\text{Coul}}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j). \tag{3.1}$$

This Hamiltonian is the one of a charged harmonic oscillator where the nuclei are considered positively charged particles with fixed position and the oscillators are negatively charged particles known as drudons. Here, $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{p}}_i$ represent the position and momentum operators for the i-th drudon relative to its nucleus, while m_i and ω_i denote the oscillator's mass and characteristic frequency respectively. The Coulomb interaction term $V_{\text{Coul}}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)$ accounts for all pairwise interactions between drudon-drudon and drudon-nucleus pairs:

$$V_{\text{Coul}}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = \frac{q_i q_j}{\mathbf{r}_{ij}} - \frac{1}{|\mathbf{r}_{ij} + \hat{\mathbf{x}}_i|} - \frac{1}{|\mathbf{r}_{ij} - \hat{\mathbf{x}}_j|} + \frac{1}{|\mathbf{r}_{ij} - \hat{\mathbf{x}}_j + \hat{\mathbf{x}}_i|} \tag{3.2}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ denotes the position vector between nuclei i and j. While this expression maintains the complete Coulomb interaction without truncation, to reconnect with the approximation used in models such as MBD, we remind that this usually involves using a multipolar expansion:

$$V_{\text{Coul}}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = \sum_{n=0}^{\infty} V_n(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \tag{3.3}$$

where each term V_n scales as $\mathbf{r}_{ij}^{-n-3}$. The leading term V_0 corresponds to the dipole-dipole interaction central to the MBD model, while V_1 and V_2 represent dipole-quadrupole and quadrupole-quadrupole interactions respectively. This expansion, while mathematically convenient, assumes the nuclear separation substantially exceeds the typical drudon-nucleus distance, imposing a lower bound on the interatomic distance. Moving forward, one can introduce dimensionless position and momentum operators:

$$\hat{\mathbf{X}}_i := \sqrt{\frac{m_i \omega_i}{\hbar}} \hat{\mathbf{x}}_i, \quad \hat{\mathbf{P}}_i := \sqrt{\frac{\hat{\mathbf{p}}_i}{\hbar m_i \omega_i}}. \tag{3.4}$$

These operators allow us to express the creation and annihilation operators for each oscillator:

$$\hat{\mathbf{a}}_i = \frac{\hat{\mathbf{X}}_i + i\hat{\mathbf{P}}_i}{\sqrt{2}}, \quad \hat{\mathbf{a}}_i^\dagger = \frac{\hat{\mathbf{X}}_i - i\hat{\mathbf{P}}_i}{\sqrt{2}}, \tag{3.5}$$

leading to a reformulation of the Hamiltonian particularly suitable for quantum simulation:

$$\begin{aligned} \hat{H} = \sum_{i=1}^{N} \hbar\omega_i \left(\hat{\mathbf{a}}_i^\dagger \hat{\mathbf{a}}_i + \frac{3}{2}\right) + \\ \sum_{i<j} V_{\text{Coul}}\left(\frac{\hbar}{m_i\omega_i}\left(\hat{\mathbf{a}}_i + \hat{\mathbf{a}}_i^\dagger\right), \frac{\hbar}{m_j\omega_j}\left(\hat{\mathbf{a}}_j + \hat{\mathbf{a}}_j^\dagger\right)\right). \end{aligned} \tag{3.6}$$

For this proof-of-concept implementation, a system of two QDOs ($N = 2$) separated by distance d is considered, and to reduce the computational complexity while maintaining the essential physics, the motion of the drudons is restricted to a common axis defined by unit vector $\hat{\mathbf{e}}_\theta$, forming an angle $\theta \in [0, \pi/2]$ with the internuclear axis. This geometric configuration allows the study of the interplay between binding and smoothness of the potential energy surface, with particularly interesting behavior emerging at intermediate values of θ. For a general angle θ and interatomic distance d, the one-dimensional Coulomb potential becomes:

$$\begin{aligned} V_{\theta,d}^{\text{Coul}}(\hat{x}_1, \hat{x}_2) = \frac{q_1 q_2}{d} - \frac{1}{\sqrt{d^2 + 2d\cos\theta\hat{x}_1 + \hat{x}_1^2}} - \\ \frac{1}{\sqrt{d^2 - 2d\cos\theta\hat{x}_2 + \hat{x}_2^2}} + \frac{1}{\sqrt{d^2 - 2d\cos\theta(\hat{x}_2 - \hat{x}_1) + (\hat{x}_2 - \hat{x}_1)^2}}, \end{aligned} \tag{3.7}$$

while the corresponding one-dimensional Hamiltonian in terms of creation and annihilation operators takes the form:

$$\begin{aligned} \hat{H}_{\theta,d} = \hbar\omega_1\left(\hat{a}_1^\dagger\hat{a}_1 + \frac{1}{2}\right) + \hbar\omega_2\left(\hat{a}_2^\dagger\hat{a}_2 + \frac{1}{2}\right) + \\ V_{\theta,d}^{\text{Coul}}\left(\frac{\hbar}{m_1\omega_1}\left(\hat{a}_1 + \hat{a}_1^\dagger\right), \frac{\hbar}{m_2\omega_2}\left(\hat{a}_2 + \hat{a}_2^\dagger\right)\right). \end{aligned} \tag{3.8}$$

For the numerical implementation, natural units are adopted where $\hbar = 4\pi\epsilon_0 = 1$ and set all oscillator parameters (masses, charges, and frequencies) to unity. This choice, while

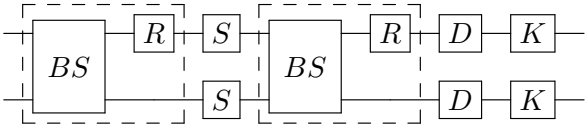


Figure 3.6: A single layer of the optical quantum circuit, consisting of various gates such as beamsplitters, rotation, squeezing, displacement, and Kerr gates, is collectively parameterized by ω.

simplifying the calculations, preserves the essential physical features of the system and allows focusing on the fundamental aspects of the quantum simulation.

3.3.2 Neural Network Ansatz, Photonic Circuit and Variational Algorithm

In order to map the physical model onto the photonic hardware, the intrinsic bosonic nature of QDOs is leveraged by mapping their Hilbert space directly onto that of photonic quantum hardware. In this framework, the Fock space of a single harmonic oscillator is identified with the Fock space of one mode of the quantum electromagnetic field, allowing a very natural correspondence between the position and momentum of the drudon particle and the quadratures of the electromagnetic field. Then, to find the ground state of the system, a continuous-variable version of the Variational Quantum Eigensolver (VQE) algorithm[168] is employed. Without a priori assumptions on the form of the ground state solution, the ansatz chosen for this variational approach will have to be as expressive as possible. Following previous work [169, 170], the already mentioned universal function approximation theorem is leveraged in the form of a quantum neural network ansatz. This ansatz is in fact easily attained in optical circuits despite the linear nature of quantum mechanics thanks to the possibility to introduce optical non-linearities in the form of Kerr gates[171]. Starting from encoding a multipartite state as $|\mathbf{x}\rangle = |x_1\rangle \otimes |x_1\rangle \otimes \cdots \otimes |x_1\rangle$, the idea is to encode the equation for an MLP as $|\sigma(\mathbf{W}\mathbf{x} + \mathbf{b})\rangle$. This is achieved by applying singular value decomposition to $\mathbf{W}$ as $\mathbf{W} = \mathbf{O}_1 \mathbf{\Sigma} \mathbf{O}_2$, where $\mathbf{O}_k$ are orthogonal matrices and $\mathbf{\Sigma} = \mathrm{diag}(\{c_i\})$ is a positive diagonal matrix. In this optical setting these transformations are easily achieved. Orthogonal transformations are unitary transformations obtained through phaseless interferometry, e.g. $\hat{U}(\theta_k, 0)\,|\mathbf{x}\rangle = |\mathbf{O}_k \mathbf{x}\rangle$, while the positive definite diagonal matrix can be obtained through

mode-wise squeezing operations such as $\otimes_{i=0}^{N}\hat{S}(\log(c_i))\,|\mathbf{x}\rangle = |\mathbf{\Sigma x}\rangle$. The bias term can then be introduced via the application of a displacement operator, namely $\otimes_{i=0}^{N}\hat{D}(b_i)\,|\mathbf{x}\rangle = |\mathbf{x}+\mathbf{b}\rangle$, and finally the non-linearity can be introduced via the mode-wise application of a Kerr gate such that $\otimes_{i=0}^{N}\hat{\Phi}(\lambda_i)\,|\mathbf{x}\rangle = |\sigma(\mathbf{x})\rangle$. Altogether, the application of one quantum neural network layer will read:

$$\hat{\Phi}\circ\hat{D}(\{b_i\})\circ\hat{U}(\theta_2,0)\circ\hat{S}(\{\log(c_i)\})\hat{U}(\theta_2,0)\,|\mathbf{x}\rangle = |\sigma(\mathbf{W}\mathbf{x}+\mathbf{b})\rangle\,, \tag{3.9}$$

where it is interesting to notice that Kerr gates are the only operations that allow the ansatz to escape the strictly Gaussian nature of the state, which would otherwise be preserved by Gaussian transformations such as the ones previously mentioned [172].

The resulting optical circuit for this case study, hence, implements a unitary transformation $\hat{U}(\omega)$ which will be the subsequent stacking of a number of layers such as the one in Eq.3.9 acting on an input reference state (see Fig. 3.6), which is taken to be the Fock vacuum state $|0\rangle\otimes|0\rangle$. The state prepared by the circuit is therefore given by $|\psi(\omega)\rangle = \hat{U}(\omega)|0\rangle\otimes|0\rangle$, where ω represents the set of all circuit parameters. The optimization objective for the variational algorithm is the expectation value of the Hamiltonian in this state:

$$\begin{aligned}\langle\hat{H}\rangle = \hbar\omega_1(\langle\hat{a}_1^\dagger\hat{a}_1\rangle+\frac{1}{2}) + \hbar\omega_2(\langle\hat{a}_2^\dagger\hat{a}_2\rangle+\frac{1}{2}) + \\ \left\langle V_{\theta,d}^{\text{Coul}}\left(\sqrt{\frac{\hbar}{m_1\omega_1}}\hat{X}_1, \sqrt{\frac{\hbar}{m_2\omega_2}}\hat{X}_2\right)\right\rangle.\end{aligned} \tag{3.10}$$

The evaluation of this expectation value requires computing terms of the form $\langle f(\hat{X}_1,\hat{X}_2)\rangle$, which is obtained by extracting the joint statistics of the position quadratures through preparation and measurement of the state $|\psi(\omega)\rangle$ in the quadrature basis. The joint probability density ρ of $(\hat{X}_1,\hat{X}_2)$ in the state $|\psi(\omega)\rangle$ allows then to compute:

$$\langle f(\hat{X}_1,\hat{X}_2)\rangle = \int_{\mathbb{R}^6} f(x_1,x_2)\rho(x_1,x_2)dx_1dx_2, \tag{3.11}$$

where the integral is implemented as a finite sum over a sufficiently refined grid in the position quadratures plane. For the numerical simulations, a linear grid of 500 points in the interval $[-6, 6]$ and using the Xanadu's STRAWBERRY FIELDS simulator, the output state vector in the Fock basis is directly accessed:

$$|\psi(\omega)\rangle = \sum_{n_1,n_2=0}^{\infty} \alpha_{n_1 n_2}(\omega)|n_1\rangle \otimes |n_2\rangle, \tag{3.12}$$

where a truncation at 5 modes per oscillator is applied, following the observation in previous full configuration interaction approaches[173]. The amplitude of a specific pair of quadratures $(\hat{X}_1, \hat{X}_2)$ is then given by:

$$\langle X_1, X_2|\psi(\omega)\rangle = \sum_{n_1,n_2=0}^{\infty} \alpha_{n_1 n_2}(\omega) \prod_{i=1}^{2} \frac{e^{-X_i^2/2} H_{n_i}(X_i)}{\sqrt{\pi^{1/2} 2^{n_i} n_i!}} \tag{3.13}$$

in terms of the Hermite polynomials, yielding the joint probability density:

$$\rho(X_1, X_2) = |\langle X_1, X_2|\psi(\omega)\rangle|^2. \tag{3.14}$$

The circuit parameters are finally optimized iteratively following 1 to minimize the cost function:

$$C(\omega) = \langle\psi(\omega)|\hat{H}|\psi(\omega)\rangle. \tag{3.15}$$

Through this variational optimization procedure, a close approximation to the ground state of the coupled QDO system is obtained, enabling the study of binding energies and other quantum mechanical properties that will be explored in subsequent sections.

3.3.3 Binding Energy Curves and Ground State Properties

For each configuration of the quantum system, characterized by angle θ and interatomic distance d, the continuous-variables VQE algorithm is used to extract properties of the ground state $|\psi_{\theta,d}\rangle$ of the Hamiltonian defined in Eq. 3.10. This is done for a grid $G_\theta \times G_d \subset [0, \pi/2] \times (0, 3.5]$, where the cardinality of G_θ is 20 and the cardinality of G_d is 200. The

Algorithm 1: Training of the parameterized photonic circuit

Parameters: Model (θ, d), $N_{\text{steps}} \in \mathbb{N}$, initial circuit parameters $\omega_0 \in \mathbb{R}^K$, learning rate $\eta \in \mathbb{R}_+$

Result: Optimized hyperparameters $\omega \in \mathbb{R}^K$

Initialize hyperparameters $\omega \leftarrow \omega_0$;
for $i = 1$ *to* N_{steps} **do**
  Compute the loss $\mathcal{C}$ according to eq. (3.15);
  Compute the gradient $\nabla_\omega \mathcal{C}$ with the shift rule;
  Update the parameters $\omega \leftarrow \omega - \eta \nabla_\omega \mathcal{C}$;
  end for
return ω.

binding energy, which quantifies the strength of interaction between the two QDOs, is then defined as the difference between the ground state energy of the interacting system and that of the non-interacting system:

$$E_\theta^b(d) = \langle \psi_{\theta,d} | \hat{H}_{\theta,d} | \psi_{\theta,d} \rangle - \langle \psi_0 | \hat{H}_0 | \psi_0 \rangle \tag{3.16}$$

where $\hat{H}_0$ represents the Hamiltonian with electric charges turned off, effectively describing two independent harmonic oscillators. The analysis of the results reveals a rich dependence of the binding behavior on the angle θ. For intermediate values around $\theta = 0.58$, binding curves exhibit remarkable agreement with a Morse potential [174, 175, 176, 177, 178], characterized by:

$$f(d) = E_b \left(e^{-2\frac{d-d_b}{s}} - 2e^{-\frac{d-d_b}{s}} \right), \tag{3.17}$$

where d_b represents the equilibrium distance, E_b the binding energy, and s the characteristic length scale of the interaction. For $\theta = 0.58$, it is found that $d_b \approx 0.54$ with binding energy $-E_b \approx 0.46$ and length scale $s \approx 2.75$ (see Fig. 3.7b). The quality of the Morse fit, in particular, reveals a fascinating interplay between the angle θ and the nature of the QDO interaction. As θ varies, we observe two distinct regimes with competing effects (see Fig. 3.7a). For small angles ($\theta \approx 0$), corresponding to drudons moving predominantly along the internuclear axis, the binding curve starts showing extreme behaviours, and it is found to strongly deviate from the Morse form, particularly at short distances. This deviation stems

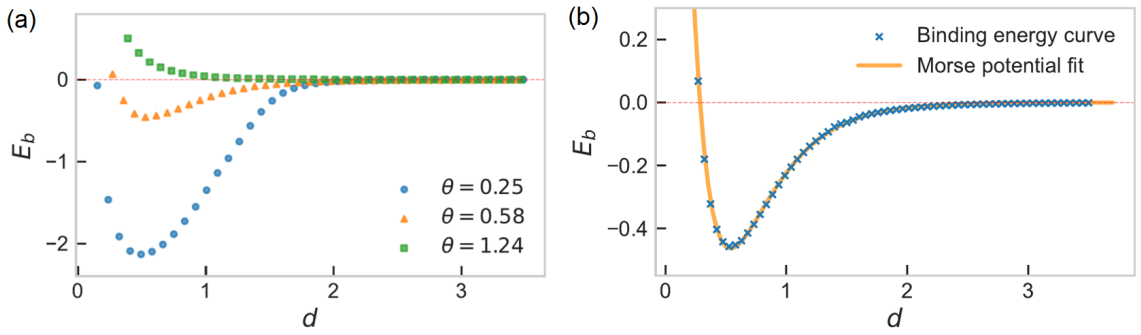


Figure 3.7: (a) The binding energy curve is shown for three different values of the angle θ, highlighting the tension between models near $\theta = 0$ and those near $\theta = \frac{\pi}{2}$. For small angles (blue curve), the steep curvature hinders an effective Morse fit. In contrast, for large angles (green curve), the transverse configuration of the drudons prevents the formation of a bound state. The orange curve represents an intermediate angle, demonstrating both binding and an excellent Morse fit. (b) Morse fit of the binding curve obtained for the value of $\theta = 0.58$ which provides the best Morse fit while still exhibiting binding behaviour.

from configurations where the drudons can approach arbitrarily close to each other, leading to divergent Coulomb repulsion in the one-dimensional geometry. The extreme case of $\theta = 0$ is in fact found to produce highly non-smooth behavior at short interatomic distances. Conversely, as θ increases toward $\pi/2$, the binding curves become progressively smoother and better described by the Morse potential. However, this improved smoothness comes at the cost of binding strength. Beyond a critical angle θ^*, the system transitions to an unbound state, characterized by the disappearance of the negative global minimum in the binding energy. This behavior can be understood physically by considering the transverse model ($\theta = \pi/2$), where two dominant configurations emerge: either the drudons align on opposite sides of the internuclear axis, experiencing primarily nuclear repulsion, or they align on the same side, experiencing both nuclear and drudon-drudon repulsion. Both scenarios preclude binding, and by continuity, this extends to a neighborhood of $\theta = \pi/2$.

These observations suggest a natural prescription for modeling molecular interactions: while the longitudinal model ($\theta = 0$) correctly predicts the existence of bound states, its instability due to overlapping drudon configurations makes it impractical. By introducing a finite angle θ, it is possible to regularize the model while maintaining binding behavior. The optimal angle should be large enough to ensure smoothness (quantified by the quality of the Morse

fit) but smaller than the critical angle θ^* where binding disappears. This analysis reveals a fundamental trade-off in quantum mechanical models of molecular binding: the competition between stability of the numerical description (favoring larger angles) and the strength of quantum mechanical binding (favoring smaller angles). The existence of an optimal intermediate regime suggests that this quantum simulation approach successfully captures essential features of molecular interactions while maintaining computational tractability. The emergence of the Morse potential form is particularly significant, as it reproduces a characteristic feature of covalent bonding in molecular systems. This suggests that the QDO model, despite being primarily motivated by dispersion interactions, could capture broader aspects of molecular binding. The binding curves obtained through the quantum simulation thus provide a valuable bridge between simplified classical models and full quantum mechanical treatments of molecular interactions.

3.3.4 Phase Space Analysis and Quantum Correlations

To gain deeper insight into the nature of the quantum mechanical ground state and the binding mechanism, one can analyze the system's representation in phase space at different interatomic separations. Focusing on the model at angle $\theta = 0.58$, which exhibits optimal binding behavior as discussed in the previous section, for each QDO, it is possible to compute the marginal Wigner function. This provides a quasi-probability distribution in phase space [179], alongside the joint probability distribution of the position quadratures that reveals spatial correlations between the oscillators. This phase space analysis reveals a rich evolution of the quantum state as the interatomic distance varies (see Fig. 3.8). At large separations ($d \approx 3.16$), the two QDOs behave essentially as independent systems, with their marginal Wigner functions displaying the characteristic Gaussian distributions of ground state harmonic oscillators (Fig. 3.8, top row). The joint position quadrature distribution shows a simple product state structure, confirming the absence of quantum correlations. As the QDOs approach each other ($d \approx 1.36$), their quantum states begin to deviate from this simple picture, with the marginal Wigner functions showing elongation primarily along the position quadrature axis, indicating spatial reorganization in response to the inter-oscillator

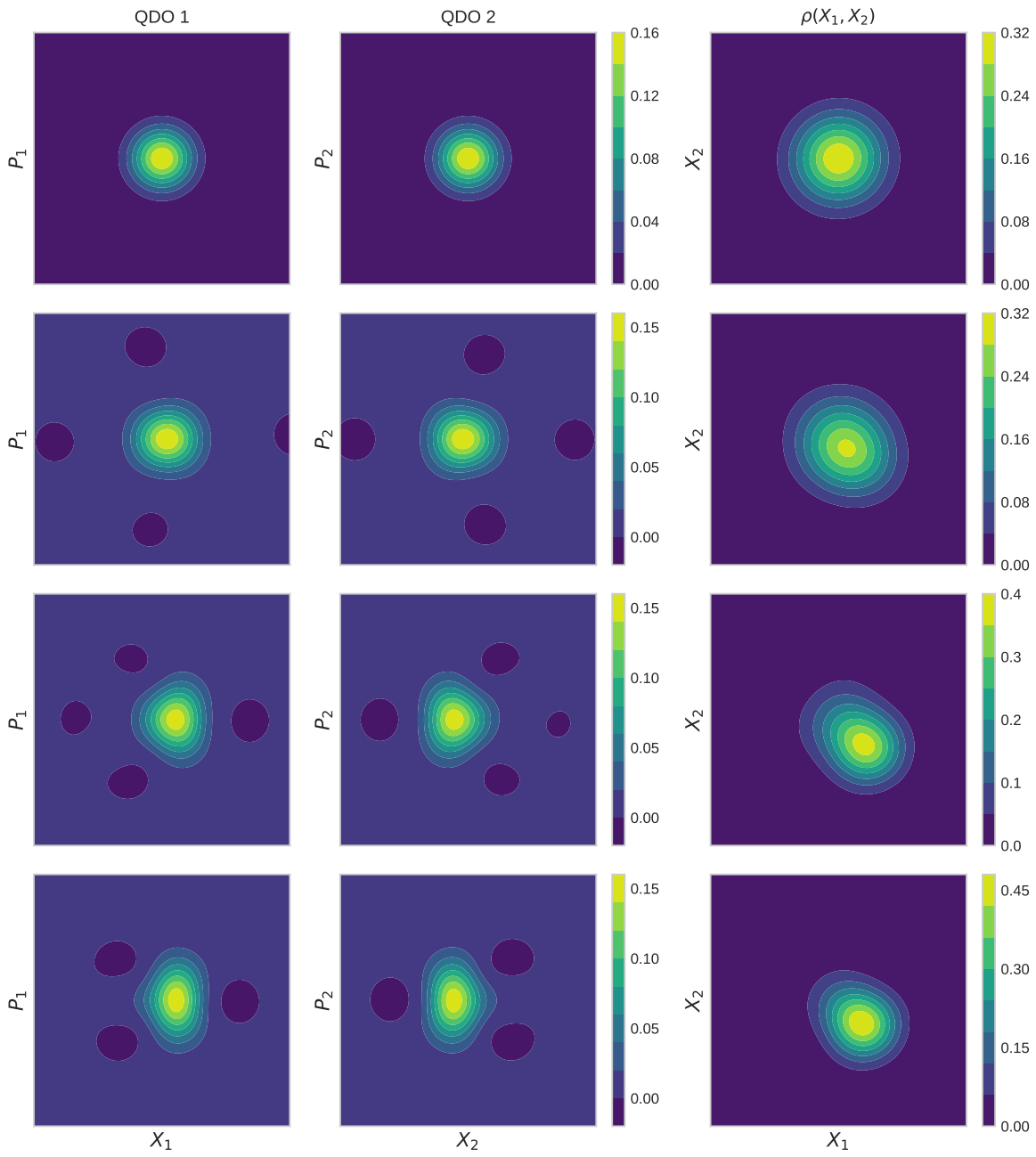


Figure 3.8: Visualization of the evolution of the quantum state across binding configurations. Top to bottom: separation distances $d \approx 3.16$, 1.36, 0.82, and 0.54. Left column shows marginal Wigner functions for QDO 1 (first) and QDO 2 (second), displaying phase space distributions of individual oscillators. Right column shows joint probability distributions of position quadratures, revealing spatial correlations between the oscillators. The progression demonstrates the transition from independent oscillators to the bound configuration.

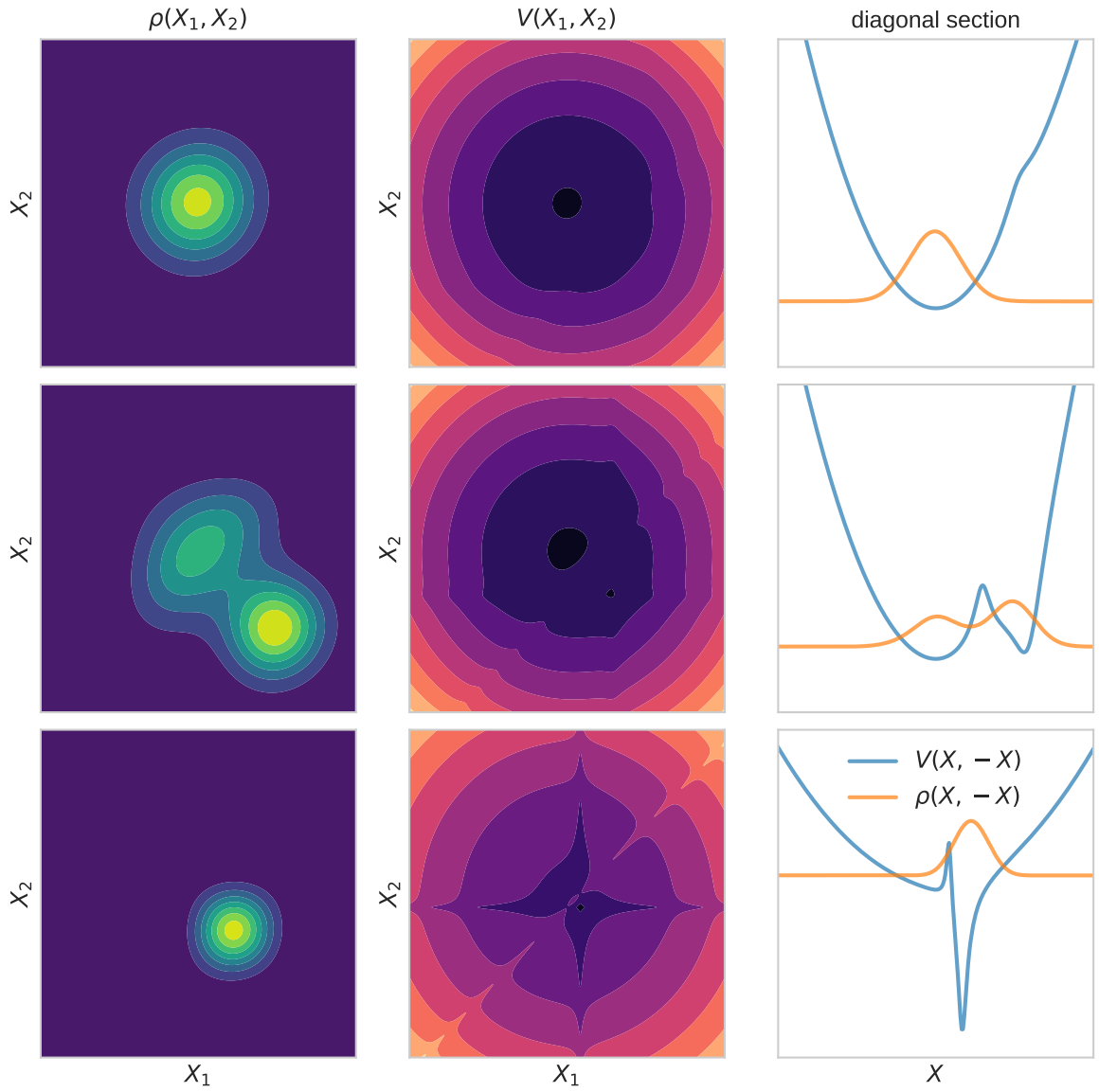


Figure 3.9: Top to bottom: different stages of the binding process from unbound at $d = 3.16$ where QDOs are far apart, then at $d = 1.75$ where the interaction between the QDOs is maximal, and finally at $d = 0.51$ where the bound state is achieved. Left to right: joint position quadrature distribution of the two drudons, compared to a scaled classical potential energy $V(X_1, X_2) = V_{\text{Coul}}(X_1, X_2) + \frac{X_1^2+X_2^2}{2}$, considering the diagonal section of both for better visualization of the binding mechanism.

interaction (Fig. 3.8, second row). The most interesting behavior emerges at intermediate distances ($d \approx 0.82$), where we observe maximal quantum correlations between the oscillators (Fig. 3.8, third row). Here, the marginal Wigner functions exhibit significant spread in both position and momentum quadratures, accompanied by regions of negative quasi-probability - a distinctive signature of non-classical behavior and a witness of entanglement in bipartite states [180, 181]. Finally, in the bound configuration ($d \approx 0.54$), the marginal Wigner functions show predominant elongation along the momentum quadrature axis, reflecting relatively well-localized drudons with increased momentum uncertainty due to their close proximity and strong interaction (Fig. 3.8, bottom row). This evolution becomes even more revealing when examining the system at smaller angles, where tunneling effects become more pronounced (see Fig. 3.9). For $\theta = 0.17$, one can clearly observe the interplay between the classical potential landscape and the quantum state distribution. At large separations ($d \approx 3.16$), the classical potential shows a single minimum and the joint distribution remains localized signaling the usual unbound case (Fig. 3.9, top row). The most intriguing behavior appears at intermediate distances ($d \approx 1.75$), where the classical potential $V(X_1, X_2) = V_{\text{Coul}}(X_1, X_2) + (X_1^2 + X_2^2)/2$ develops two distinct local minima. At this distance, the joint distribution becomes distinctly bimodal, providing clear evidence of tunneling between the quadratic and Coulomb wells (Fig. 3.9, middle row). The diagonal section of both the joint distribution and the potential further illuminates this tunneling phenomenon. Finally, in the deep bound state ($d \approx 0.51$), the system settles into a new configuration with the distribution centered away from the origin, indicating that the displacements of the drudons are locked in to form a bound state (Fig. 3.9, bottom row).

The findings just discussed, pertaining Fig. 3.8 and Fig. 3.9, gives us an intuition of what happens to the ground state of the system along the binding curve, allowing for an interpretation of the solution found via the quantum neural network ansatz. For $d \to \infty$ the two QDOs will not interact and hence the natural state for them will be the vacuum state $|0\rangle|0\rangle$ (ground state of independent harmonic oscillators). On the other hand when $d \sim d_{\text{bonding}}$ we see that the system is shifted towards an antisymmetric configuration where $\langle X_1\rangle = -\langle X_2\rangle$ and $\langle P_1\rangle = \langle P_2\rangle = 0$, which can be approximated by the bipartite coherent state $|\alpha\rangle|-\alpha\rangle$

with $\alpha = \langle X_1 \rangle$. In the transition region, instead, the system will pass through an intermediate state, which is reflected in the position joint probability by the transition from a single mode to a bimodal distribution, which is naturally represented as a superposition of the form:

$$\frac{1}{N}(|0\rangle \otimes |0\rangle + |\alpha\rangle \otimes |-\alpha\rangle) \tag{3.18}$$

where N is a normalization factor and α quantifies the displacement from the origin. This

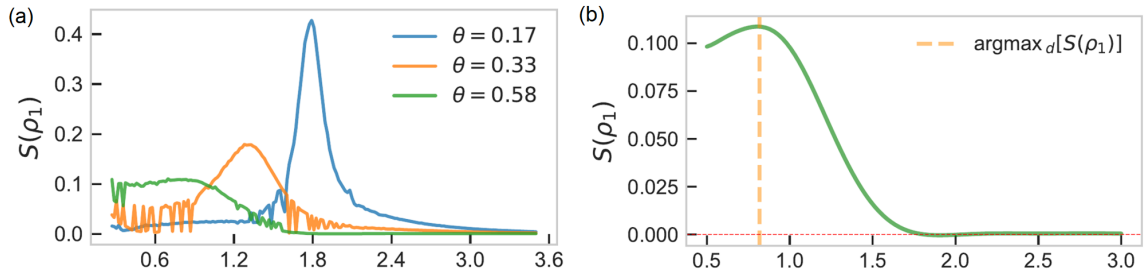


Figure 3.10: (a) Entanglement entropy plotted against the distance d between QDOs for various values of θ. (b) Smoothed entanglement entropy as a function of interatomic distance for $\theta = 0.58$.

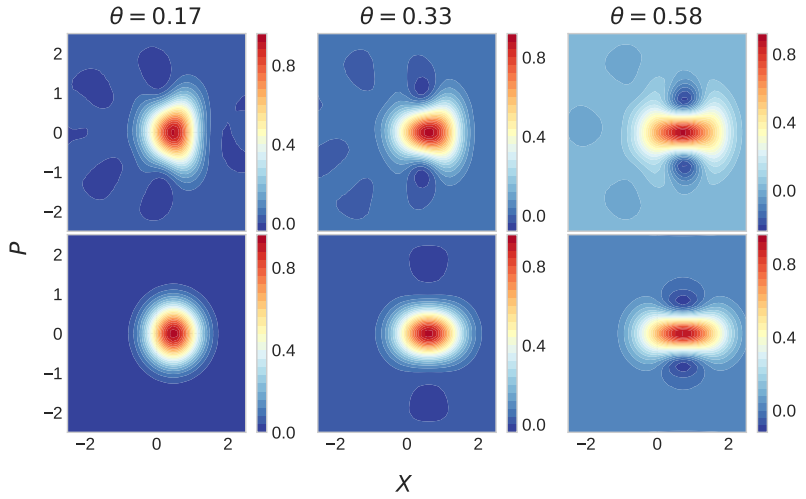


Figure 3.11: Comparison between simulated ground states and fitted cat state ansatz. Top row shows the Wigner functions of the ground states obtained from VQE at entropy peaks. Bottom row displays the Wigner functions of the fitted cat state ansatz $\frac{1}{N}(|0\rangle \otimes |0\rangle + |\alpha\rangle \otimes |-\alpha\rangle)$ at the same points. Wigner functions are sliced along the plane $(X = X_1 = -X_2, P = P_1 = -P_2)$ to obtain 2D visualizations. The high fidelity ($F \sim 0.96$) between simulated and fitted states validates the cat state description of the bonding mechanism.

state is indeed an entangled state, and also is of the form of a cat state. This simple expression for the state at the transition point of the binding curve elegantly explains the observed bimodality in the position distribution. Furthermore, this interpretation is supported by examining the von Neumann entropy, or entanglement entropy, through the binding curve. This is defined as $S(\rho_1) = -\mathrm{Tr}[\rho_1 \log(\rho_1)]$ where ρ_1 is the partial trace of the density matrix $\rho = \sum_{\substack{n_1,n_2 \\ m_1,m_2}} \alpha^*_{m_1 m_2} \alpha_{n_1 n_2} |n_1\rangle\langle m_1| \otimes |n_2\rangle\langle m_2|$ of the whole system over the Hilbert space of the second cQDO and is symmetric in the choice of cQDO to be traced out. In Fig. 3.10, it is indeed found that the entanglement entropy shows a characteristic peak that coincides with the maximal tunneling between the two configurations and converges to the inflection point of the binding energy curve for increasing values of θ in agreement with 3.18.

The quality of this entangled cat state description at the transition point can be tested. This is done by tuning the parameters defining the state in Eq. 3.18 for each of the transition points obtained at different values of θ in order to maximize quantum state fidelity $\mathcal{F} \in [0, 1]$ (overlap between two quantum states) with the actual state resulting from the neural network ansatz. The results are reported in Fig.3.11, and show that $\mathcal{F}$ remains high throughout the binding regime, with values of $\mathcal{F} \sim 0.96$ even as θ varies. The emergence of entangled cat states in this system is particularly noteworthy as it naturally explains the necessity of including Kerr gates in the quantum circuit - these non-Gaussian operations can indeed be used for cat state generation [182, 183]. Furthermore, the role played by such non-classical effects in the binding process suggests a fundamental connection between quantum correlations and the binding mechanism, providing an unexpected bridge between the disparate fields of quantum optics and molecular binding and suggests that quantum optical concepts and techniques might offer new perspectives on molecular interactions, particularly in regimes where quantum effects dominate.

Chapter 4

A Whole Chemical Space in a Set of Properties

A recurring theme across the work laid out so far has been the use of different ways to characterize molecular systems: from electronic densities to Cartesian coordinates, up to representations for machine learning and representation learning techniques. When discussing the analysis of the Aquamarine dataset, though, it became increasingly clear how the analysis of property-property relations could reveal substantial information about molecular structures. Together with the idea of 'freedom of design' in property space put forth in [158], this naturally leads us to a deeper question about the nature of chemical space itself: can we use molecular properties as coordinates to navigate the chemical compound space (CCS)? The idea is compelling: while structures encode "what a molecule is", properties encode "what a molecule does". A parameterization based on properties could therefore provide not only a more natural way to explore chemical space from the perspective of molecular function but also a much more efficient approach to exploring an otherwise

incredibly vast and high-dimensional space.

This chapter develops this idea by constructing a differentiable parameterization of CCS using quantum-mechanical properties as intrinsic coordinates. We demonstrate how combining variational auto-encoders with property encoders enables us to learn a common latent representation that bridges structures and properties in the case of the QM7x dataset. This representation reveals fundamental insights about the organization of chemical space - how certain properties naturally cluster molecules, how others provide local coordinates for exploring specific regions, and how the interplay between intensive and extensive properties shapes molecular diversity. Beyond these conceptual insights, we show how this framework can be used in principle also for applications like targeted exploration of chemical space and even prediction of transition pathways between conformers. The tools developed here thus provide both a new lens for understanding chemical space and practical methods for navigating it guided by molecular properties.

4.1 Compressing Chemical Space with Variational Auto-Encoders (VAE)

In chapter 3, we mentioned the challenges associated with the exploration of the vast and complex CCS. While exhaustive screening via quantum mechanical calculations is unfeasible regardless of computational power, there we saw that an effort for covering interesting regions of this space is ongoing and strong. From the ML side, in order to facilitate and give a direction to this exploration effort, different approaches have been developed. The main branch of these methods certainly regards generative models [184], that with the aim of generating molecules based on the correlations learned from the datasets, has emerged as a promising approach to navigate this vast space more efficiently. In the family of generative models used for drugs and material discovery, we can find approaches such as Generative Adversarial Networks (GANs)[185], which generate realistic molecular structures through adversarial training [186], Recurrent Neural Networks (RNNs) [36], which are often used to produce molecular SMILES strings or other sequential representations[51], diffusion models,

which transform random noise into structured molecular forms through iterative refinement [187, 188], and Variational Autoencoders (VAEs) [189], compressing molecular data into a low-dimensional latent space that facilitates the creation of novel compounds by sampling and decoding from a finite-size learned continuous latent space [50].

Focusing on this last family of models, VAEs consist of two neural networks: an encoder that compresses input data into a lower-dimensional latent space representation, and a decoder that attempts to reconstruct the original input from this compressed representation. Unlike traditional autoencoders, VAEs are probabilistic models - the encoder and decoder networks parameterize probability distributions rather than deterministic functions. Specifically, for an input molecule x, the encoder defines a distribution $q_\phi(z|x)$ over latent vectors z, while the decoder defines a distribution $p_\theta(x|z)$ over reconstructed molecules. VAEs are trained by optimizing the Evidence Lower BOund (ELBO):

$$\text{loss}_{\text{ELBO}} = D_{\text{KL}}[q_\phi(z|x)||p(z)] - \mathbb{E}_{q_\phi}[\log(p_\theta(x|z))] \tag{4.1}$$

where D_{KL} is the Kullback-Leibler divergence and $p(z)$ is typically chosen as a standard normal distribution $\mathcal{N}(0, I)$. The first term acts as a regularizer, ensuring the latent space remains compact and avoids sparsity and memorization by pushing the encoded distributions close to the prior. The second term instead has the function of encouraging accurate reconstructions by maximizing their log-likelihood.

Several landmark works have demonstrated the power of VAEs for molecular applications [190, 191, 192, 50]. Gómez-Bombarelli et al. showed that VAEs trained on SMILES strings and coupled with a prediction task from latent space could learn chemically meaningful latent spaces suitable for optimization of molecular properties [50]. Their success sparked numerous follow-up studies exploring different molecular representations and model architectures. These approaches have consistently shown that the discrete, variable-sized space of molecular structures can be effectively compressed into a continuous latent space of fixed dimensionality. This remarkable ability to compress the complex combinatorial space of molecules into a smooth, continuous representation of fixed size suggests an intriguing

possibility: that the underlying complexity of chemical space might be captured by a limited number of continuous degrees of freedom. Such an observation naturally leads to the hypothesis that molecular structures could be parameterized using a fixed set of physical descriptors - specifically, quantum mechanical properties that themselves arise from the fundamental laws governing molecular systems. This insight motivates the investigation into whether quantum mechanical properties can serve as intrinsic coordinates for navigating chemical space.

4.2 A Differentiable Mapping Between Properties and Molecules

To address this fundamental question, the QIM (Quantum Inverse Mapping) model is here presented, where the standard VAE framework is adapted to enable a differentiable mapping from molecular properties to structures. The key idea is to introduce a third network, the property encoder, that learns to map the molecular properties y to the same latent space variable z which the VAE uses to encode the correspondent molecular structure x, producing the probability distribution $p_\psi(z|y)$. During training, the VAE and the property encoder are jointly optimized by modifying the ELBO loss to include an additional likelihood term, so that the latent z sampled from the VAE training is also used to train the property encoder. The new loss reads:

$$\text{Loss} = \beta D_{\text{KL}}[q_\phi(z|x)||p(z)] - \mathbb{E}_{q\phi(z|x)}[\log p_\theta(x|z)] - \tau\mathbb{E}_{q\phi(z|x)}[\log p_\psi(z|y)] \tag{4.2}$$

where β and τ are adjustable coefficients introduced as hyperparameters to control the relative importance of the KL divergence and the property encoding. The training procedure involves sampling a molecular structure x and its corresponding properties y from the dataset, encoding x into the latent representation z using the VAE encoder $q_\phi(z|x)$, reconstructing x from z using the VAE decoder $p_\theta(x|z)$, encoding the properties y into the latent space probability distribution $p_\psi(z|y)$ using the property encoder, and finally computing the modified ELBO loss and updating the networks' parameters via backpropagation. By jointly

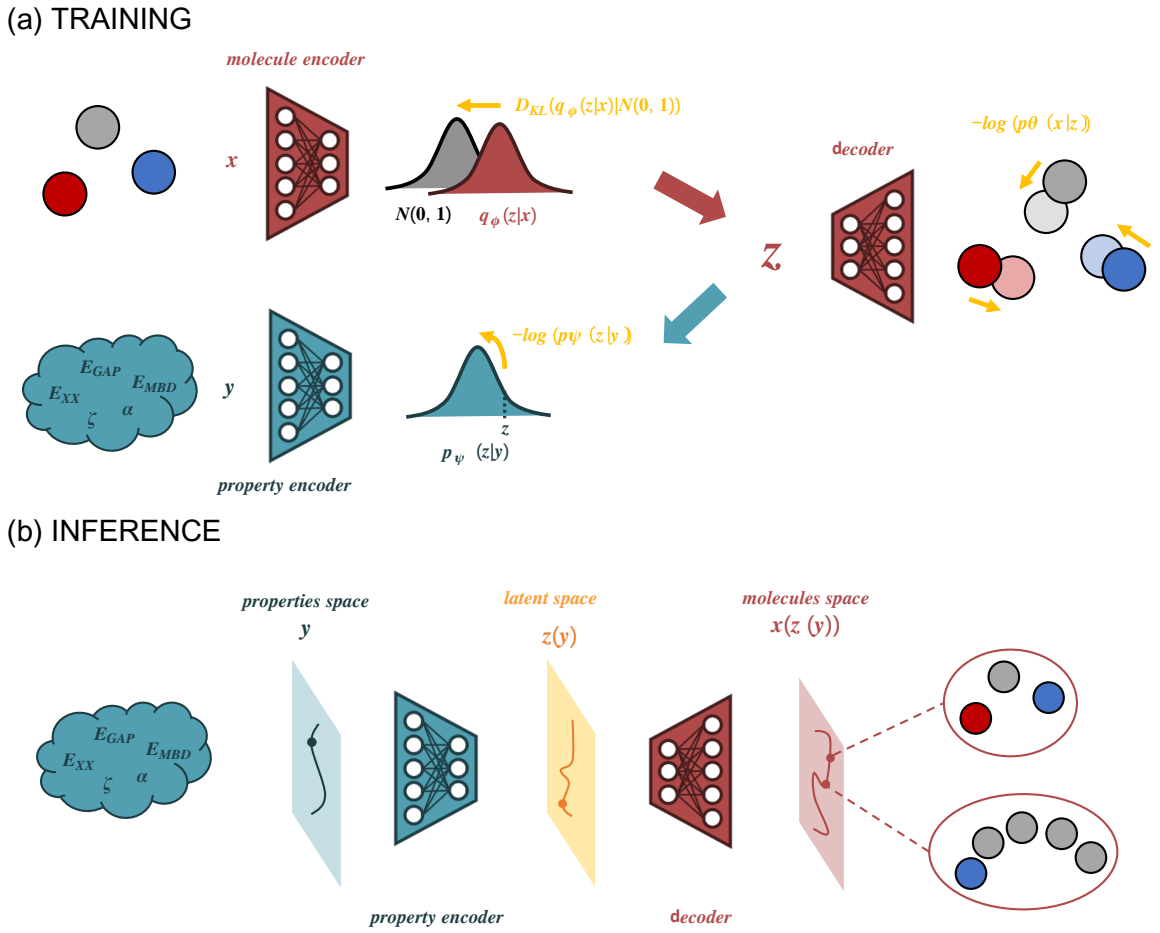


Figure 4.1: (a) The Quantum Inverse Mapping (QIM) model is based on a Variational Auto-Encoder (VAE) architecture. During training, molecular structures x (depicted with atoms as colored spheres) are encoded into a latent distribution $q_\phi(z \mid x)$. A latent variable z is sampled from this distribution and passed to a decoder to reconstruct the original structure. Simultaneously, a property encoder maps the associated quantum-mechanical (QM) properties y into a distribution $p_\psi(z \mid y)$. Both networks are trained jointly using the Evidence Lower Bound (ELBO) loss, which combines a molecular reconstruction term $\log(p_\theta(x \mid z))$, a Kullback-Leibler (KL) regularization term $D_{\mathrm{KL}}(q_\phi(z \mid x) \parallel N(0,1))$, and an additional loss term $\log(p_\psi(z \mid y))$ to link the latent z with the properties y. This results in a shared latent space representation that unifies molecular structures and QM properties. (b) During inference, the property encoder is coupled with the VAE decoder to approximate the parameterization of Chemical Compound Space (CCS) using QM properties as intrinsic coordinates. This differentiable CCS representation facilitates identifying key properties in molecular reconstruction and enables various molecular design tasks.

optimizing the VAE and the property encoder, a shared latent space is induced that captures the salient features of both molecular structures and their properties. After training, this al-

Symbol	Property Description	Units	Type	Class
E_{AT}	Atomization energy	eV	M,G	E
E_{MBD}	MBD energy	eV	M,G	E
E_{XX}	Exchange energy	eV	M,G	E
E_{NN}	Nuclear-nuclear energy	eV	M,G	E
E_{EE}	Electron-electron energy	eV	M,G	E
E_{KIN}	Kinetic energy	eV	M,G	E
E_{GAP}	HOMO-LUMO gap	eV	M,G	I
E^0_{HOMO}	HOMO energy	eV	M,G	I
E^0_{LUMO}	LUMO energy	eV	M,G	I
E^1_{HOMO}	HOMO-1 energy	eV	M,G	I
E^1_{LUMO}	LUMO+1 energy	eV	M,G	I
E^2_{HOMO}	HOMO-2 energy	eV	M,G	I
E^2_{LUMO}	LUMO+2 energy	eV	M,G	I
ζ	Total dipole moment	e·Å	M,G	I
α	Isotropic molecular polarizability	a_0^3	M,R	E
D_{MAX}	Maximum atom-atom distance	Å	S,G	I

Table 4.1: This table lists the QM properties (and their corresponding symbols) obtained from the QM7-X dataset [104], which were utilized during the training of our model (see Fig. 4.1). The units specified for these properties include a_0, the atomic unit of length (Bohr radius). All listed properties are scalars, meaning they have a dimensionality of 1. Properties were classified into the following categories: structural (S), global/molecular (M), ground-state (G), response (R), extensive (E), and intensive (I).

lows to combine the property encoder and the VAE decoder to establish a differentiable mapping from properties y to structures x:

$$y \to z \sim p_\psi(z|y) \to x \sim p_\theta(x|z). \tag{4.3}$$

where the mean of the parameterized distributions is used in place of the sampling operation as maximum likelyhood points.

For what concerns the molecular representation, the Coulomb Matrix (CM) is here used [58], which encodes both the atomic positions and species (see Sec. 2.2.2) considering implicit hydrogen atoms and hence treating directly only heavy atoms. The CM is invariant to translations and rotations, and allows for the retrieval of atomic positions and atomic species. While a graph-based representation would treat molecular fragments more robustly, it would introduce a number of complications in architecture design, hence a distance based repre-

sentation such as the CM is more lightweight and practical for this proof-of-concept implementation as it makes it easier to be associated with a fixed size vector of scalar molecular properties. Moreover, using an appropriately padded CM using fixed rules enables the QIM model to potentially generate molecules with more heavy atoms than the largest molecule in the QM7-X dataset.

The retrieval of Cartesian coordinates and chemical composition is then achieved with a two-step process. First, the chemical composition $\{Z_i\}$ is obtained from the diagonal elements using the inverse transformation $g = (\cdot)^{\frac{1}{2.4}}$, then the interatomic distances d_{ij} are computed from the off-diagonal elements as $d_{ij} = \left(\frac{CM_{ij}}{Z_i Z_j}\right)$, and classical multidimensional scaling (MDS) is applied to the resulting Euclidean Distance Matrix (EDM) to obtain the Cartesian coordinates [193]. Hydrogen atoms are finally added back to the reconstructed heavy-atom structures using the OpenBabel [194] software and their positions are optimized with the DFTB3+MBD method while keeping the heavy atoms fixed.

The performances of the QIM model in reconstructing molecular structures from their quantum mechanical properties are assessed by training and testing on a random split of the 40988 equilibrium structures of the QM7x dataset including only molecules with (C, N, O) heavy atoms and considering 17 QM global extensive and intensive properties (listed in Table 1). This is done with 28000 structures for training, 2000 for validation and the remaining used for testing. The testing phase consists into using the obtained map from Eq. 4.3 and comparing the original molecular structures to the ones generated from properties. The comparison is made first in terms of relative error on the reconstruction of the representation (CMs) for varying number of properties considered, and then for the full set considering the RMSD between reconstructed geometries. In Fig. 4.2a we present boxplots of the relative error in CM reconstruction as a function of the number of properties used to train the model. This is defined as $\Delta = \frac{|\tilde{C}-C|}{|C|} \times 100$ where C is the original CM, and $\tilde{C}$ is the reconstructed CM, both treated as vectors. The analysis focuses on the relative error in the CM rather than the root-mean-squared deviation (RMSD) between molecular structures since RMSD is only defined for molecules with correctly predicted compositions, leading to greater fluctuations when fewer properties are used—an outcome of higher chemical composition errors in such

cases. The results show that the QIM model achieves an average reconstruction error on the representation that stabilizes around 5% when more than seven properties are included. Beyond this threshold, the error distribution becomes increasingly skewed, as evidenced by the median value. While the mean and standard deviation of the CM reconstruction error are reported, it is important to note that this metric does not fully capture the quality of the mapping due to its noisy and nonlinear relationship with RMSD.

The distribution and the cumulative distribution of the RMSD over the test set is hence reported in Fig. 4.2b for the case of the full set of properties. The model achieves an average RMSD of 0.62 Å, with over 70% of the molecules reconstructed within an RMSD of 0.7 Å (see Fig. 4.2 in the main text). This threshold was empirically found to separate molecules with an acceptable reconstruction of the heavy-atom structure in terms of topology and orientation (in Fig 2c some examples of reconstructions as a reference). As a side note, albeit somewhat expected given the low diversity of the chemical compositions in the dataset, the model correctly predicts the heavy atom composition for 99.96% of the test set molecules. Interestingly, the QIM model's performance is enhanced when both extensive and intensive properties are used in the training process, compared to using either type of property alone. This highlights the complementary information captured by these two classes of properties. While extensive properties primarily govern the overall molecular size and composition, intensive properties refine the finer details of the 3D structure.

4.3 Scientific Insights from a Neural Network

To analyze how the QIM model functions and what insights it can provide about chemical space at hand, a gradient attribution map was implemented which enables the assessment of the individual contributions of each property to the output structures. Similar to standard approaches in machine learning applications with image inputs [195], one can compute an attribution map A for each property by calculating the gradient of the CM components with respect to that property. Specifically, the Jacobian matrix is here computed and its norm is taken over the output dimension of the CM, averaging over a subset of best reconstructed

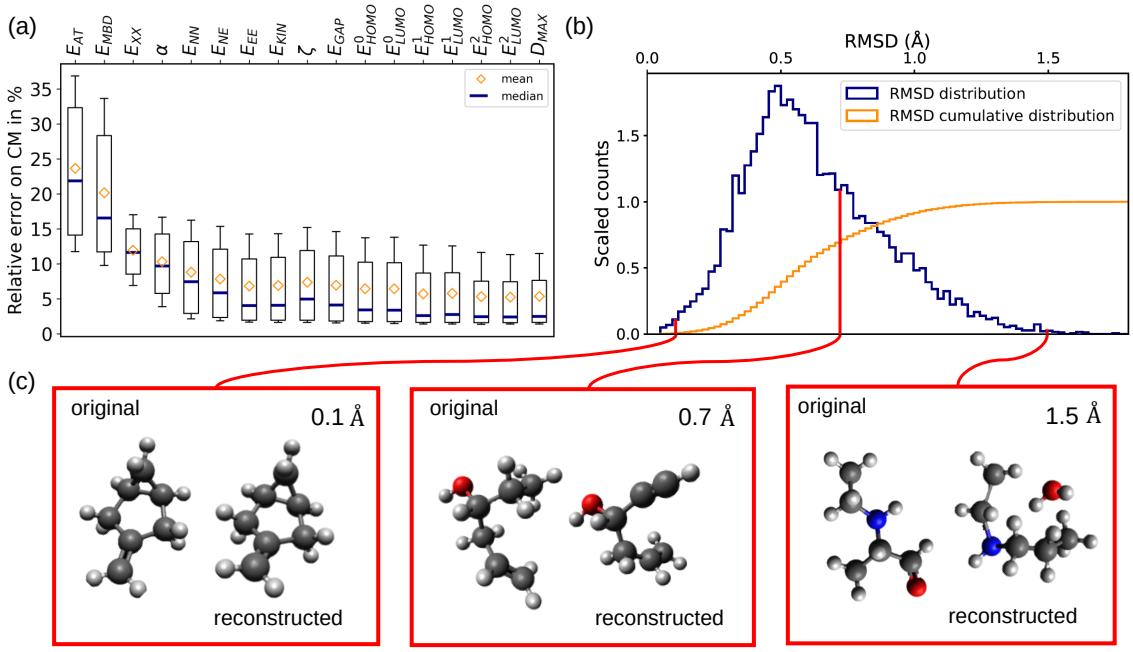


Figure 4.2: (a) Boxplots showing relative error in Coulomb matrix (CM) reconstruction from QM properties versus number of training properties. Analysis spans 10988 test molecules, with whiskers from 15th-85th percentiles. Property definitions in Table 4.1. (b) RMSD frequency distribution between original and reconstructed geometries, with cumulative distribution (orange). (c) Example molecules comparing originals to QM property-based reconstructions at various RMSD values. RMSD < 0.7 Å indicates acceptable heavy atom structure and topology reconstruction. Atoms: carbon (gray), oxygen (red), nitrogen (blue), hydrogen (white).

molecules. Namely, for a given property $\mathbf{p}_j$:

$$A_j = \frac{1}{N} \sum_{k \in \mathbb{B}} \| \frac{\partial \mathbf{CM}_i^k}{\partial \mathbf{p}_j} \|, \tag{4.4}$$

where $\mathbb{B}$ is of best reconstructed molecules (150 molecules with RMSD ≤ 0.2 Å). The normalized A values for each property reveal that extensive properties are more informative than intensive ones for molecular reconstruction (Fig. 4.3d). This can be explained by the fact that these extensive properties depend on crucial molecular features that are also considered in a 3D representation like the Coulomb matrix, such as number of atoms, number of electrons (related to chemical composition), and geometry. When comparing CMs, even a slight difference of one atom can significantly increase the loss, leading to a larger sensitivity

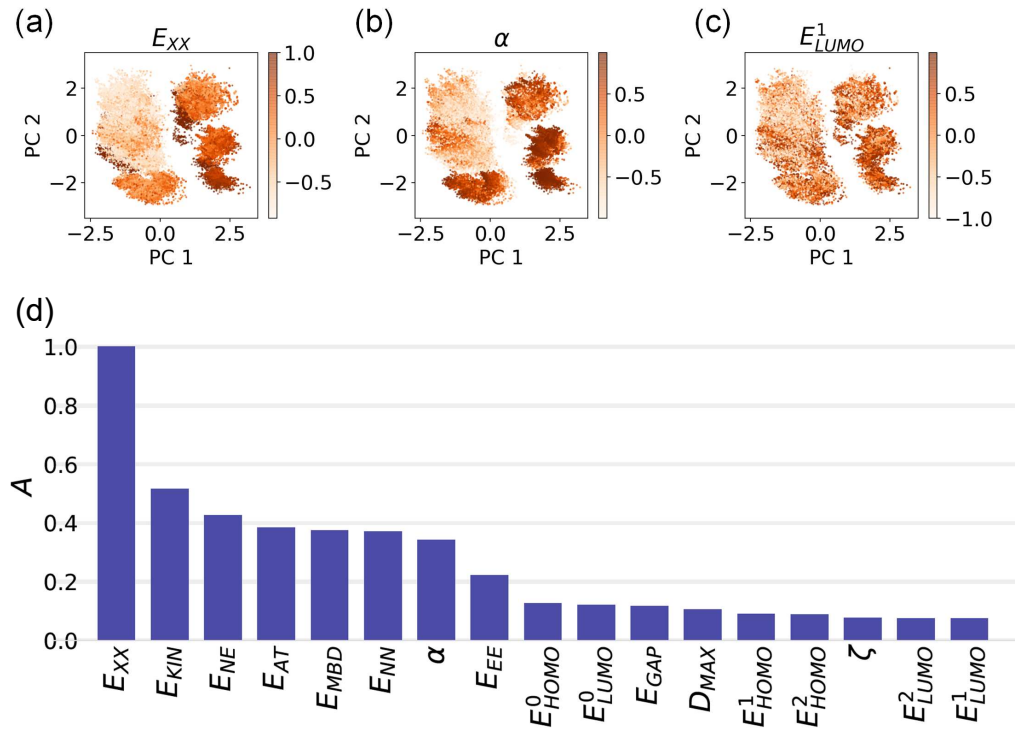


Figure 4.3: Two-dimensional PCA projection of the VAE latent space, with data points colored by (a) E_{XX}, (b) α, and (c) E^1_{LUMO} values (hyperbolic tangent scaling applied to prevent outlier dominance). (d) Attribution values A for each molecular property, obtained by computing the partial derivatives of the reconstructed Coulomb Matrix w.r.t. the selected property and taking the norm. Values are averaged over the set of best reconstructed molecules and normalized over the maximum value among the properties.

of the model to variations in system size and composition. Consequently, A values for the components of the total energy and molecular polarizability are higher compared to those for molecular orbital energies and dipole moment; in particular, E_{XX} and E_{KIN} present the largest A values. This finding correlates remarkably with the identification of molecular clusters in the two-dimensional principal component analysis (PCA) of the latent space of the VAE encoder, where the higher the A value of a property, the more correlated it is with respect to the PCA representation (Fig. 4.3a-c).

The hierarchy of QM properties revealed by the attribution analysis provides deeper insights into how they organize chemical space. Starting with the properties showing highest A values, E_{KIN} and E_{XX}, the two-dimensional projection of the QM7-X molecular prop-

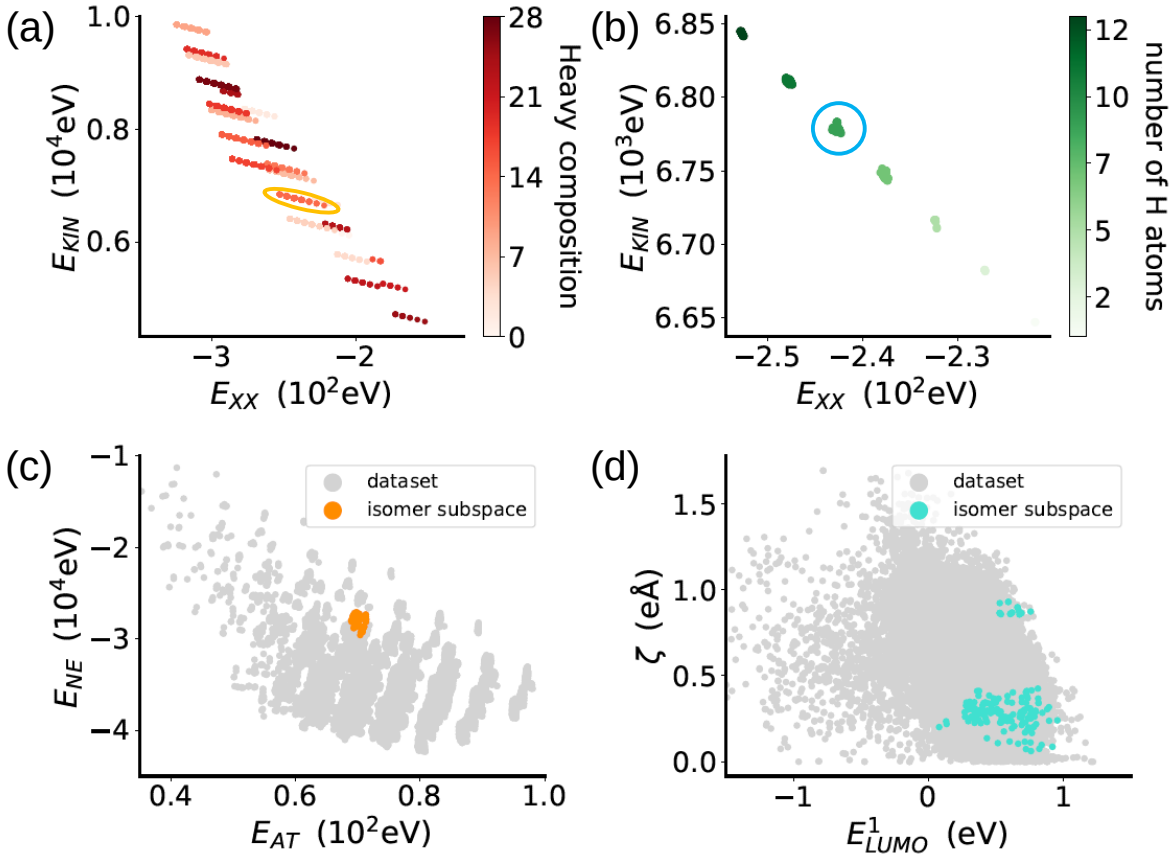


Figure 4.4: Analysis of QM7-X chemical space organization based on property hierarchies. (a) Distribution of molecules in (E_{XX}, E_{KIN}) space, with colors indicating distinct heavy atom compositions (each color in the colormap is a different composition), showing linear clustering by composition. (b) Detailed view of a high-density cluster (yellow ellipse in (a)), colored by hydrogen atom count, with a selected isomer subspace marked (blue ellipse). Comparison of molecular structure distribution within an isomer subspace using property pairs with (c) high and (d) low attribution values A.

erty space defined by these properties is considered (Fig. 4.4a). Despite their high inverse correlation (Pearson coefficient = -0.92), the molecules in the dataset organize into linear-shape clusters containing molecules with the same heavy atom composition. Upon closer examination, it becomes evident that E_{KIN} is mostly influenced by the heavy atom composition within a molecule, while E_{XX} is highly sensitive to the number of hydrogen atoms, indicating a dependence on particular bond types. A closer examination of one highly populated cluster reveals a finer local structure with almost perfect inverse correlation (Pearson coefficient = -0.99) as well as very compact clusters formed by isomers (Fig. 4.4b). This

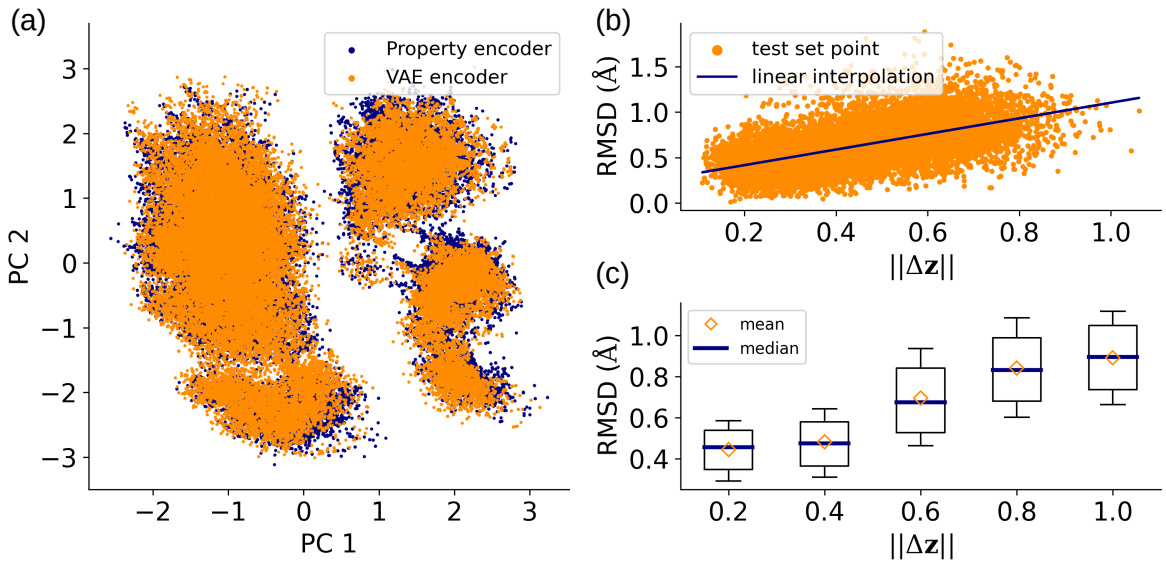


Figure 4.5: (a) Two-dimensional PCA projection showing overlap between property encoder and VAE latent representations. (b) RMSD between original and reconstructed structures versus latent space difference $\|\Delta z\| = \|z - \tilde{z}\|$, where z is the property encoder latent representation and $\tilde{z}$ is obtained by re-encoding reconstructed structures with the VAE encoder. (c) Distribution of RMSD values for different $\|\Delta z\|$ intervals (±0.1 Å width, whiskers from 15th to 85th percentile), leading to the definition of optimal generation criterion $\|\Delta z\| \in [0, 0.4]$.

behavior can be understood by considering the qualitative aspects of E_{KIN} and E_{XX}: the dominant contribution to E_{KIN} stems from the inner shell electrons (a trivial consequence of the virial theorem) while the primary influence on E_{XX} arises from the valence electrons. Exchange-related quantities have been found to play a significant role in characterizing bonds [196, 197], explaining their sensitivity to the number of hydrogen atoms in a molecule. Going further, when examining molecular isomer subspaces using other pairs of QM properties with high and low A values (Fig. 4.4c,d), we find that properties with high A values serve as better local coordinates for exploring these subspaces. These properties present relatively smaller changes in their values across related structures compared to properties with low A values. This demonstrates their efficiency in identifying molecular structures within a specific molecular isomer subspace while effectively distinguishing them from other structures spanning the entire property spectrum.

From the point of view of the latent space, in order to probe the hypothesis of having ob-

tained a common latent space representation for molecular structures and properties, one can look at the similarity between the latent representation of the properties and structures. This correspondence is indeed confirmed by the significant overlap between the PCA projections of both latent representations (Fig. 4.5a), verifying the initial assumption about the joint training procedure. Moreover, the correlation between the differences in latent space representation and the quality of the reconstructed molecules is analysed, aiming to obtain a self-consistent method for error estimation. To this end, if z is the latent representation from the property encoder, one can take the reconstructed representations and encode them again with the VAE encoder, producing a new latent representation $\tilde{z}$. Analysis of the correlation between the quantity $\|\Delta z\| = \|z - \tilde{z}\|$ and the RMSD between the original structures and those reconstructed from properties reveals an approximate bulk correlation with numerous outliers (Fig. 4.5b). Further investigation through boxplots of the RMSD for varying values of Δz shows a nonlinear but monotonic behavior for the relationships between these quantities, with a minimum for RMSD in the region of low $\|\Delta z\|$ values ($\in [0, 0.4]$) (Fig. 4.5c). Since the primary interest is having a low RMSD when reconstructing actual molecular structures, it is possible to use this interval to filter out generated structures, thereby enhancing the quality of molecules generated with a targeted set of QM properties.

Altogether, these insights demonstrate how this methodology is explainable and flexible enough to obtain fundamental insights int the structure-property relationship at hand, as well as an empirical estimation of the quality of the reconstructed geometry.

4.4 Multi-Objective Targeted Structure Generation

The QIM model, while not developed to be a generative model per se, can indeed be adapted to act as a conditional generative model for multi-objective targeted generation. For this purpose, it is necessary to define a way to sample the property space conditionally, namely obtaining the diversity that usual models obtain by sampling a known latent distribution by sampling a conditional distribution in property space for fixed values of the targeted properties. This procedure is implemented through a multi-Gaussian fitting of the distribution of

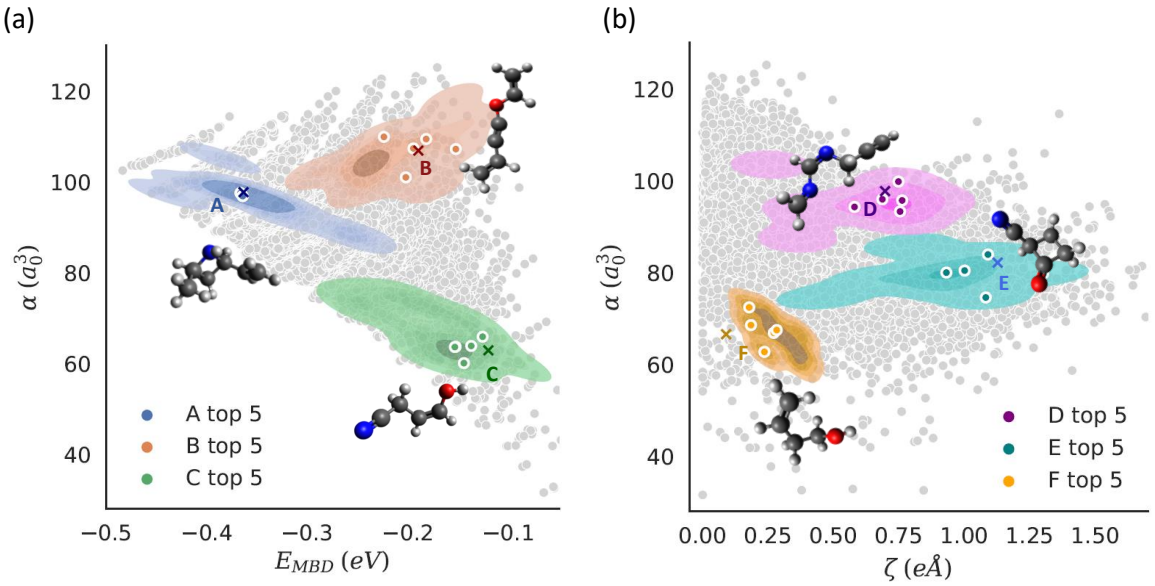


Figure 4.6: Generated molecular distributions around specified targets (A-F, colored crosses) in (a) (α, E_{MBD}) and (b) (α, ζ) spaces, with sample density indicated by shading intensity and QM7-X reference molecules shown as gray dots. Top 5 generated molecules per target highlighted as colored circles, with one representative structure shown per target. Atoms colored as: carbon (gray), oxygen (red), nitrogen (blue), hydrogen (white).

the QM7-X property space. This approach begins by constructing a model with 91 multivariate Gaussian distributions $N(\mu_k, \Sigma_k)$, with μ_k and Σ_k representing the mean value and covariance matrix of each Gaussian k (the choice of 91 distributions was determined through Bayesian information criterion analysis). When targeting specific property values m^*, one can select the most likely Gaussian k^* using the maximum likelihood criterion:

$$k^* = \arg\max_k N_m(m = m^* | \mu_k, \Sigma_k) \tag{4.5}$$

where N_m is the marginal distribution for the targeted properties $\{m\}$. The conditional probability formula for multivariate Gaussian distributions then allows the retrieval of the distribution of non-targeted properties $\{n\}$ for $m = m^*$:

$$p(n|m = m^*) = N(n|\tilde{\mu}, \tilde{\Sigma}) \tag{4.6}$$

where $\tilde{\mu}$ and $\tilde{\Sigma}$ are defined as:

$$\tilde{\mu} = \mu_n + \Sigma_{nm}\Sigma_{mm}^{-1}(m^* - \mu_m) \tag{4.7}$$

$$\tilde{\Sigma} = \Sigma_{nn} - \Sigma_{nm}\Sigma_{mm}^{-1}\Sigma_{mn}. \tag{4.8}$$

Multiple values for non-targeted properties are sampled from this distribution, with each sample corresponding to a complete set of properties that maintains a fitted consistency with the targeted values. These property sets are then passed through the QIM model to generate molecular structures. The generated samples undergo filtering based on the self-consistency criterion $\|\Delta z\| \in [0, 0.4]$, followed by reconstruction of Cartesian coordinates and optimization of hydrogen atom positions using DFTB3+MBD. The model's targeted generation capabilities are then evaluated across two distinct property spaces: (α, E_{MBD}) and (α, ζ), as shown in Fig. 4.6. For the (α, E_{MBD}) space, which pairs two properties with high attribution values, fifteen samples were generated per target, achieving mean relative errors of approximately 3.2% for E_{MBD} and 1.3% for α in the optimal 5-molecule sets. The moderate correlation between these properties (Pearson coefficient = 0.60) enabled exploration of diverse molecular structures while maintaining target accuracy. When targeting the more weakly correlated (α, ζ) space (Pearson coefficient = 0.44), combining extensive and intensive properties, the model achieved 2.8% error for α but higher 5.9% error for ζ across ten samples per target. This aligns with the attribution analysis findings regarding the model's reduced accuracy for intensive properties. Furthermore, it is found that the spread of generated molecules correlates with the relative variance of non-targeted extensive properties, offering a hint into the mechanism that controls generation diversity. Targets with lower negative log-likelihood values and small relative variances in extensive properties showed more precise generation, demonstrating the model's ability to balance accuracy and molecular diversity in targeted regions of chemical space.

As it is, the current model did not generate any molecule with unseen composition. While on one side this is reasonable, as QM7x is exhaustively complete for the molecules up to 7 heavy atoms, this also hints at limitations of this approach for obtaining molecules with a

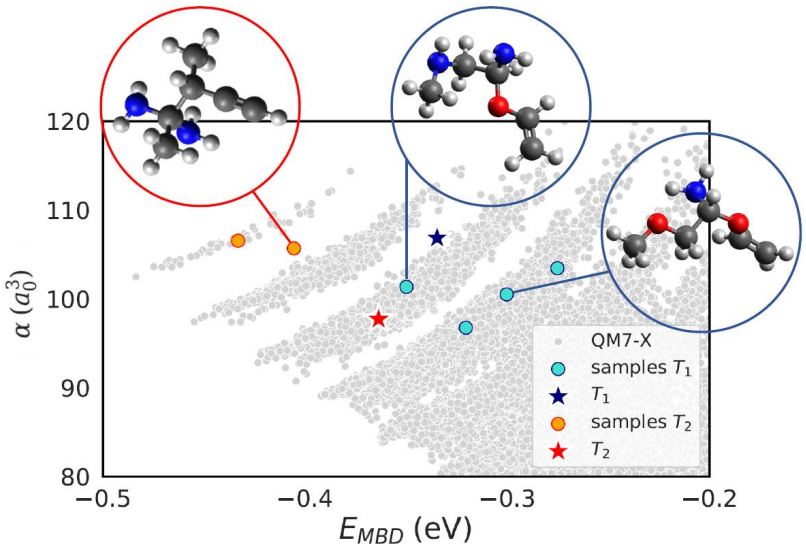


Figure 4.7: Generation of molecules with eight heavy atoms in (α, E_{MBD}) space. Results from modified QIM model trained with fragment-independent masking procedure, showing generated eight-atom molecules (colored circles) for targets T_1 (blue star) and T_2 (red star), with QM7-X reference molecules as gray dots. Representative novel structures shown for each target, with atoms colored as: carbon (gray), oxygen (red), nitrogen (blue), hydrogen (white).

higher number of atoms. This is likely due to using a padded representation which treats the molecule as a whole, and hence lacks the ability to treat molecular fragments independently. In this sense, one can modify the training procedure to reduce bias towards existing molecular fragments by using a masking method that randomly masks atoms during training (with probability p = 0.5) and applies leaky masking to padded entries. While this slightly reduces reconstruction accuracy ($\langle$RMSD$\rangle$ decreased by 0.05 Å), it enables generation of unseen molecular scaffolds, particularly those containing eight heavy atoms. Testing this modified model across the (α, E_{MBD}) space revealed that generation of unseen compositions was primarily confined to regions of low dataset coverage, characterized by high α and large $|E_{MBD}|$ values (see Fig. 4.7). For target T_1, the model achieved errors of approximately 10% for E_{MBD} and 8% for α - comparable to results on seven-atom molecules. Target T_2 showed higher errors of approximately 20% for E_{MBD} and 16% for α. Analysis of the generated structures revealed varying chemical diversity depending on target location. T_2 samples consisted exclusively of (C,N)-based molecules, while T_1 samples showed greater chemical diversity, including (C,N,O)-based compositions. All generated molecules exhibited

physically reasonable property values for eight-atom systems (-0.6 eV $< E_{MBD} < -0.09$ eV and 66 a.u. $< \alpha <$ 160 a.u.) and maintained chemical validity. The increased structural complexity of T_1 molecules naturally corresponded to larger $|E_{MBD}|$ values compared to similar targets for seven-atom systems, aligning with chemical intuition.

These results, while showing the limitations of a proof-of-concept implementation, validate that this methodology successfully adapts the QIM framework for conditional molecular generation while maintaining the advantages of a physically-grounded, differentiable mapping between properties and structures.

4.5 Energy Barrier and Transition Structures Estimation in One Method

Throughout the preceding sections, the versatility of the QIM model in learning a CCS parameterization that supports both molecular reconstruction and targeted generation was displayed. A notable observation from the analysis is that the latent space exhibits clear structure with respect to energetic properties that show high attribution values, even under linear transformations like PCA (see Fig. 4.3). This characteristic suggests that linear interpolation in the latent space could generate structures with smoothly varying energetic properties - a direct consequence of linear transformations preserving convexity properties up to a sign. This insight, naturally leads to exploring whether the learned latent representation could serve as an intrinsic coordinate system for generating transition geometries between conformational isomers in QM7-X. The geodesic interpolation algorithm for VAEs from [198] can indeed be adapted to find curves in property space that are geodesics with respect to the metric induced by the latent space encoding. For this investigation, three pairs of conformational isomers were selected: $C_4H_9NO_2$ (case I), $C_4H_5NO_2$ (case II), and C_5H_5NO (case III). The selection criteria required that the structures be achiral and reconstructed with RMSD $\leq$ 0.2 Å. The interpolated geometries shown in Fig. 4.8(a-c) illustrate the model's ability to generate plausible transition paths between conformers, with the exceptions of some unphysical transitions that are observed in cases II and III between steps 1 and 2. These are

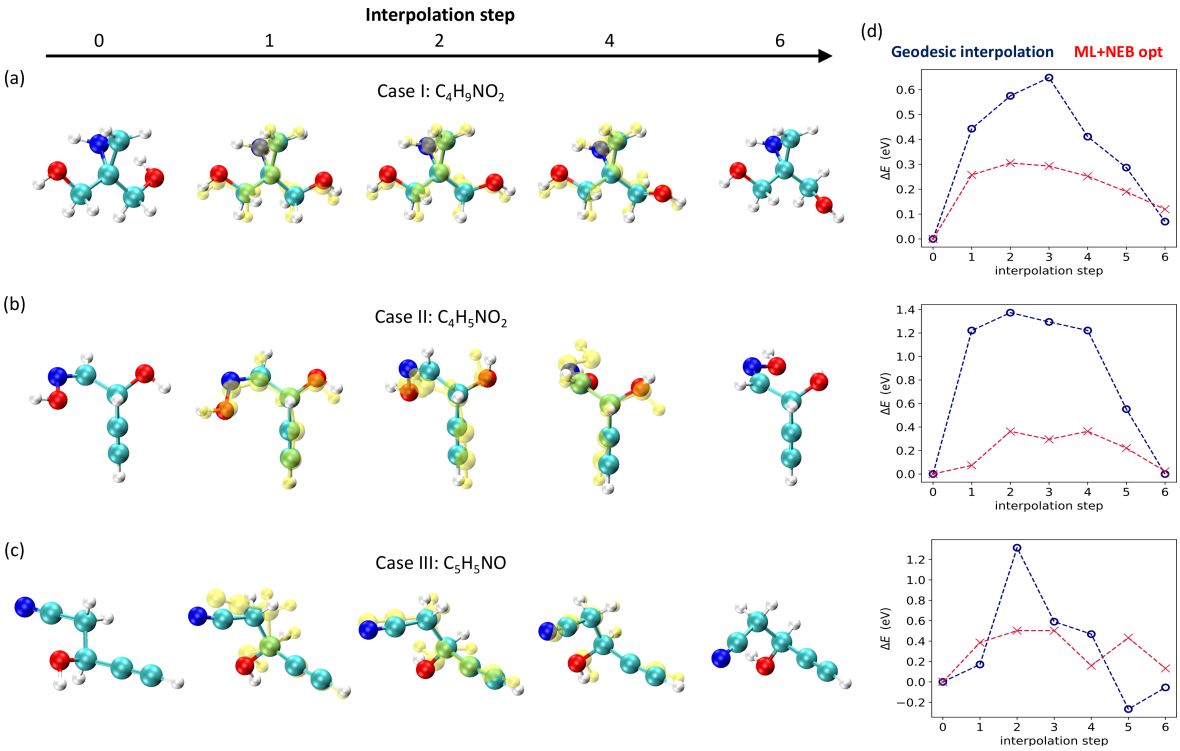


Figure 4.8: Interpolated geometries generated by geodesic interpolation in VAE latent space for three conformational isomer pairs: (a) $C_4H_9NO_2$ (case I), (b) $C_4H_5NO_2$ (case II), and (c) C_5H_5NO (case III). QIM-generated structures shown as yellow balls, with optimized transition structures from ML-NEB calculations shown as solid colored balls (atoms colored as: carbon (cyan), oxygen (red), nitrogen (blue), hydrogen (white)). (d) Relative energy profiles $\Delta E_i = E_i - E_0$ versus interpolation step i for each isomerization, comparing geodesic interpolation in property space (blue) with ML-NEB optimized paths (red). Results demonstrate the effectiveness of the learned latent space as an intrinsic coordinate system for predicting transition paths between conformers in QM7-X.

likely due to two factors: (1) the CM representation's sensitivity to small changes, leading to large mirror-like transformations, and (2) degraded model performance in unexplored latent space regions. Remarkably, without explicit energy optimization, the relative energy profiles ($\Delta E_i = E_i - E_0$) of geodesics exhibit barrier-like behavior, with energy barriers between 0.6 eV and 1.4 eV (Fig. 4.8d). This is noteworthy since the model was trained solely on equilibrium geometries and had no exposure to non-equilibrium structures. To validate these predictions, the generated geometries were used as initial guesses for nudged elastic band (NEB) calculations, following the ODE method [199]. Hydrogen atoms were added via OpenBabel, and endpoint geometries were optimized before performing NEB calculations

with a machine learning force field [200], trained on PBE0+MBD energies and forces from QM7-X. The resulting energy profiles (Fig. 4.8d) reveal that geodesic interpolations consistently overestimate energy barriers but capture the qualitative features of the transitions. The RMSD between interpolated and NEB-optimized heavy-atom structures ranged from 0.14 Å to 0.35 Å, indicating reasonable accuracy. These results highlight a connection between this property-based interpolation and geodesic transition path methods [201]. Moreover, they demonstrate that the model's latent space effectively captures essential physical aspects of molecular conformational changes, despite being trained exclusively on equilibrium structures, suggesting possible applications in rapid reaction pathway estimation.

Chapter 5

Quantum Chemistry Data for Better Drug Discovery

Throughout this thesis, different ways of representing and understanding molecular systems have been explored through the lens of quantum mechanics. The first part of Chapter 3 demonstrated how quantum mechanical calculations can be used to cover portions of interest of the vast CCS and produce both structures and properties at very high level of quality. The second part of Chapter 3 showed how Neural Networks can be used to push the boundaries of quantum chemistry in emerging technologies like quantum computing, while Chapter 4 illustrated how these networks can be used together with quantum chemistry data to gain insights into the structure-property relationship and achieve a better navigation of the CCS. As a natural progression, this chapter shifts the focus of the synergy between quantum chemistry and deep learning to explore how quantum chemistry data can be leveraged in a representation learning capacity to produce better deep learning models for molecular properties that cannot be directly computed from first principles. Particularly, the focus lies

on absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties - crucial endpoints for drug discovery that are only accessible through experimental measurements that are often costly, time-consuming, and prone to variability. Through a systematic study of a Graph Transformer network pretrained on different types of quantum mechanical data, this work investigates how various pretraining strategies influence the quality of learned molecular representations. Importantly, the validation extends beyond public benchmark data, which are known to be small and sometimes unreliable, to include a large internal pharmaceutical dataset of microsomal clearance measurements, demonstrating how these methods translate to real-world drug discovery applications.

5.1 The Current Issues with ADMET Modeling

Absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties play a crucial role in determining a drug candidate's success. These properties govern how a drug moves through and interacts with biological systems, ultimately determining its safety and efficacy [202]. A major challenge in developing effective machine learning models for ADMET properties stems from data availability [203]. Unlike quantum mechanical properties which can be computed systematically for any molecule, ADMET properties must be measured experimentally. These experiments are costly, time-consuming, and prone to variability, leading to relatively small datasets. This is evident in public benchmarks like the Therapeutic Data Commons (TDC) [204], where individual ADMET tasks often contain only hundreds to a few thousand samples. Such limited data makes it difficult to draw reliable conclusions about model performance or establish robust benchmarks. As discussed in Chapter 2, the success of deep learning approaches relies heavily on two key ingredients: sophisticated model architectures and large amounts of high-quality training data. While the field has made significant advances in architectural design - from molecular fingerprints to graph neural networks and transformers - the fundamental limitation of experimental data scarcity remains. This creates an imbalance in the recipe for success, where increasingly powerful models exist but there is insufficient data to fully leverage their capabilities. The limited data issue is

particularly problematic given the complexity of biological systems and the high-dimensional nature of molecular space. As shown in recent studies [205, 206, 207, 208, 209], machine learning models often struggle with generalization when applied to molecules outside their training distribution. This is especially challenging for drug discovery applications, where novel chemical scaffolds are constantly being explored. Traditional approaches to improving ADMET characterization have focused primarily on engineering better molecular descriptors or developing more sophisticated architectures. However, these approaches alone may not be enough to overcome the fundamental constraint of limited experimental data. This suggests the need for alternative strategies that can leverage more abundant sources of molecular information to enhance ADMET modeling capabilities. In this sense, more recent work has explored the possibility to use additional data for representation learning by pretraining [210, 211, 212, 213, 214]. Quantum mechanical calculations, as explored in Chapters 2 and 3, are a promising option in this direction as they are known to be related to fundamental aspects of molecular behavior [215, 216, 217, 218]. Unlike experimental ADMET measurements, quantum mechanical properties can be computed systematically at scale for any molecule. If this approach can demonstrably produce better and more reliable representations of molecules, then the increasing number of public QM datasets such as the ones mentioned in Chapter 3 will find a very strong and useful case of application for drug discovery. In the following sections, the chapter explores how pretraining on quantum mechanical data can help address the data scarcity challenge, but more importantly how and why this can produce better molecular representations for ADMET modeling in real-world applications.

5.2 Pretraining on Quantum Mechanical Data for Better Performance

Starting from the challenges just mentioned in the previous section, this study systematically evaluates different pretraining strategies based on quantum mechanical data. For what concerns the model of choice, the case of Graphormer architecture (see Chapter 2) is con-

sidered as an increasingly popular model for handling molecular data at the intersection between GNNs and Transformers. In particular, this study employs the architecture introduced in [219], where the centrality encoding incorporates both explicit atoms and implicit hydrogens. This allows the model to handle atom hybridization implicitly, avoiding the need for a separate edge encoder component. The output values for atomic properties are then obtained by applying a linear layer to the last layer latent representations of the atom tokens, while the representation of the global token (called CLS token) is used in a similar fashion for global properties estimation. Here the focus is not on optimizing the model's absolute performance, but rather on the evaluation of pretraining effects on deep models. With this in mind, to ensure a fair comparison, the hyperparameters are kept consistent across all models during both pretraining and fine-tuning stages, and a configuration with 20 hidden layers is chosen, which is deeper compared to typical models, while maintaining a comparable parameter count to other well-known Transformer-based approaches. For what concerns the pretraining part, the decision was made to train the model on three distinct types of pretraining data: atom-level quantum mechanical properties, molecular quantum mechanical properties, and self-supervised masking. For atom-level properties, a dataset containing approximately 136,000 organic molecules totaling over 2M heavy atoms is utilized [220]. Each atom is characterized by multiple quantum mechanical properties calculated using DFT methods (B3LYP/def2svp): atomic charges, NMR shielding constants, and electrophilic and nucleophilic Fukui function indices. These properties were chosen as they represent different aspects of electronic structure and chemical reactivity. Models are pretrained both on individual properties and in a multi-task setting combining all properties. For molecular-level pretraining, the PCQM4Mv2 dataset containing HOMO-LUMO gaps calculated for over 2M molecules is used [221]. This provides a comparison point using a quantum mechanical property defined at the whole-molecule level with a number of molecular structures comparable to the number of labeled atoms in the atom-level QM dataset. Finally, the masking pretraining, which serves as the self-supervised baseline, follows standard BERT-like approaches by randomly masking 15% of atom nodes and training the model to restore the correct atom types. After pretraining, all models are fine-tuned on 22 ADMET tasks from the

Therapeutic Data Commons benchmark [204]. These tasks span various molecular properties crucial for drug development, including membrane permeability (Caco2), blood-brain barrier penetration, toxicity, and metabolic stability. The evaluation metrics vary by task, including MAE for regression tasks and ROC-AUC or PR-AUC for classification. Results are summarized in Table 5.1 in terms of mean and standard deviation computed over 5 predefined training/validation splitting seeds. Here, the best results are highlighted based on their mean values and then, for every other model, a t-test paired by seed is performed to test the hypothesis that the best model is significantly better than the others. All the models that have results that are not significantly worse than the best one are highlighted as well.

Table 5.1: Global results obtained from the ADMET group of TDC are presented. Each row corresponds to a specific task, along with the metric used for evaluation. Columns represent different pretrainings considered. Highlighted values denote the best performance achieved among our models, based on the average value and t-tests paired across seeds. Additionally, cases where our results surpass in mean value the top-performing model in the TDC leaderboard are marked with an asterisk (*).

task	metric	scratch	all	charges	nmr	fukui_n	fukui_e	masking	homo-lumo
caco2_wang	MAE↓	0.442 ± 0.041	**0.354 ± 0.015**	**0.404 ± 0.069**	**0.364 ± 0.046**	**0.346 ± 0.034**	0.483 ± 0.036	0.471 ± 0.080	0.381 ± 0.040
hia_hou	ROC-AUC ↑	**0.972 ± 0.015**	**0.982 ± 0.003**	**0.973 ± 0.027**	**0.977 ± 0.011**	0.967 ± 0.011	0.908 ± 0.019	**0.981 ± 0.013**	0.869 ± 0.037
pgp_broccatelli	ROC-AUC ↑	0.892 ± 0.011	**0.913 ± 0.015**	0.902 ± 0.019	**0.917 ± 0.009**	0.896 ± 0.020	**0.911 ± 0.008**	**0.921 ± 0.003**	0.870 ± 0.016
bioavailability_ma	ROC-AUC ↑	0.606 ± 0.040	**0.673 ± 0.028**	**0.662 ± 0.071**	0.640 ± 0.040	0.663 ± 0.025	0.616 ± 0.082	**0.698 ± 0.035**	**0.667 ± 0.031**
lipophilicity_astrazeneca	MAE ↓	0.539 ± 0.036	**0.393 ± 0.005***	**0.425 ± 0.023***	0.424 ± 0.007*	0.457 ± 0.008*	0.463 ± 0.011*	0.462 ± 0.005*	0.451 ± 0.011*
solubility_aqsoldb	MAE ↓	0.878 ± 0.031	**0.720 ± 0.010***	0.726 ± 0.011*	0.728 ± 0.014*	0.756 ± 0.012	0.771 ± 0.015	0.769 ± 0.007	0.772 ± 0.019
bbb_martins	ROC-AUC ↑	0.860 ± 0.016	**0.872 ± 0.021**	**0.874 ± 0.011**	**0.869 ± 0.014**	0.848 ± 0.018	0.845 ± 0.014	**0.861 ± 0.025**	**0.883 ± 0.007**
ppbr_az	MAE ↓	8.477 ± 0.483	**7.589 ± 0.203**	**7.668 ± 0.236**	**7.542 ± 0.215**	**7.530 ± 0.318**	8.026 ± 0.222	**8.056 ± 0.340**	**7.874 ± 0.287**
vdss_lombardo	Spearman ↑	0.554 ± 0.049	0.624 ± 0.020	**0.637 ± 0.022**	0.616 ± 0.034	0.616 ± 0.015	**0.652 ± 0.012**	0.620 ± 0.023	0.580 ± 0.029
cyp2d6_veith	PR-AUC ↑	0.549 ± 0.043	**0.621 ± 0.046**	**0.675 ± 0.014**	**0.643 ± 0.036**	0.660 ± 0.009	0.638 ± 0.011	0.612 ± 0.021	0.612 ± 0.028
cyp3a4_veith	PR-AUC ↑	0.799 ± 0.012	0.797 ± 0.029	**0.847 ± 0.022**	**0.824 ± 0.021**	**0.838 ± 0.016**	**0.828 ± 0.018**	0.817 ± 0.014	0.794 ± 0.018
cyp2c9_veith	PR-AUC ↑	0.706 ± 0.014	0.703 ± 0.022	**0.726 ± 0.024**	**0.739 ± 0.011**	**0.722 ± 0.021**	**0.734 ± 0.014**	**0.736 ± 0.014**	0.708 ± 0.010
cyp2d6_substrate_carbonmangels	PR-AUC ↑	0.546 ± 0.042	**0.648 ± 0.031**	**0.634 ± 0.050**	**0.653 ± 0.023**	**0.619 ± 0.057**	0.578 ± 0.052	**0.677 ± 0.022**	0.582 ± 0.036
cyp3a4_substrate_carbonmangels	ROC-AUC ↑	0.637 ± 0.027	0.630 ± 0.015	0.646 ± 0.020	0.642 ± 0.009	0.645 ± 0.015	0.635 ± 0.031	0.641 ± 0.030	**0.685 ± 0.015**
cyp2c9_substrate_carbonmangels	PR-AUC ↑	0.360 ± 0.022	0.374 ± 0.028	**0.404 ± 0.027**	0.394 ± 0.024	**0.405 ± 0.036**	0.375 ± 0.030	**0.396 ± 0.024**	**0.439 ± 0.043**
half_life_obach	Spearman ↑	0.373 ± 0.076	0.462 ± 0.154	**0.559 ± 0.034**	0.487 ± 0.045	0.486 ± 0.030	0.476 ± 0.015	0.462 ± 0.052	0.426 ± 0.039
clearance_microsome_az	Spearman ↑	0.448 ± 0.038	0.548 ± 0.029	**0.620 ± 0.007**	**0.613 ± 0.014**	0.554 ± 0.019	0.513 ± 0.022	0.555 ± 0.022	0.565 ± 0.032
clearance_hepatocyte_az	Spearman ↑	0.336 ± 0.050	0.382 ± 0.032	**0.456 ± 0.015**	0.460 ± 0.019	0.374 ± 0.021	0.353 ± 0.028	**0.478 ± 0.018**	0.413 ± 0.030
herg	ROC-AUC ↑	0.709 ± 0.080	0.788 ± 0.029	0.824 ± 0.046	0.834 ± 0.030	0.752 ± 0.042	0.758 ± 0.053	**0.880 ± 0.003**	0.790 ± 0.031
ames	ROC-AUC ↑	0.772 ± 0.022	**0.822 ± 0.005**	**0.821 ± 0.010**	**0.833 ± 0.014**	**0.820 ± 0.009**	**0.823 ± 0.012**	0.801 ± 0.008	0.808 ± 0.008
dili	ROC-AUC ↑	0.856 ± 0.037	**0.892 ± 0.033**	**0.859 ± 0.055**	**0.898 ± 0.022**	0.847 ± 0.016	0.812 ± 0.122	**0.906 ± 0.021**	0.854 ± 0.017
ld50_zhu	MAE ↓	0.593 ± 0.038	0.559 ± 0.016	0.571 ± 0.012	**0.538 ± 0.014***	0.592 ± 0.029	0.618 ± 0.014	0.577 ± 0.010	0.582 ± 0.031
Number of best models		1	12	17	13	7	5	11	5

Overall, while all pretraining strategies generally lead to improvements, pretraining on HLG excels distinctly in only one property, though it ranks among the top-performing models in four additional cases. Similarly, masking pretraining outperforms others significantly in just one instance but matches the highest performance across ten other tasks. Models pretrained with atom-level quantum mechanical (QM) properties collectively demonstrate strong performance, with at least one model outperforming both masking and HLG in ten tasks and tying for the best in twenty out of twenty-two cases. Within this group, pretraining on

properties like charges, NMR shifts, and comprehensive atomic QM attributes yields more consistent top results compared to models trained on Fukui functions. Additionally, for solubility, lipophilicity, and toxicity (LD50), the results surpass those of the best models currently listed on the TDC leaderboard. To validate the benefits of different pretraining strategies

Table 5.2: Results of the fine-tuning on internal microsomal clearance dataset. Results are reported for both values of clearance in the dataset and for all pretraining strategies both in terms of R^2 coefficient and in terms of Spearman's rank coefficient.

	METRIC	SCRATCH	ALL	CHARGES	NMR	FUKUI_N	FUKUI_E	MASKING	HOMO-LUMO
CLEARANCE_1	$R^2 \uparrow$	0.505 ± 0.010	**0.640 ± 0.004**	0.629 ± 0.006	**0.635 ± 0.006**	0.599 ± 0.004	0.593 ± 0.004	0.580 ± 0.012	0.602 ± 0.006
	SPEARMAN ↑	0.728 ± 0.008	**0.807 ± 0.003**	0.799 ± 0.004	**0.801 ± 0.003**	0.785 ± 0.003	0.785 ± 0.001	0.774 ± 0.007	0.786 ± 0.004
CLEARANCE_2	$R^2 \uparrow$	0.534 ± 0.006	**0.653 ± 0.004**	0.633 ± 0.003	**0.643 ± 0.005**	0.598 ± 0.007	0.610 ± 0.008	0.597 ± 0.002	0.607 ± 0.005
	SPEARMAN ↑	0.750 ± 0.005	**0.818 ± 0.003**	0.807 ± 0.004	**0.811 ± 0.002**	0.789 ± 0.002	0.795 ± 0.006	0.786 ± 0.002	0.794 ± 0.002

beyond public benchmarks, the models are evaluated on a large internal dataset of human liver microsome intrinsic clearance measurements containing approximately 138,000 compounds. The training is here conducted over 3 different training/validation splittings and results are reported as mean and standard deviation in Table 5.2. This evaluation on a more extensive dataset reveals clearer distinctions between pretraining strategies and also some differences w.r.t. what was found in the public benchmark. The models pretrained on all atomic QM properties significantly outperform all other approaches except NMR pretraining, achieving R^2 values of 0.640 ± 0.004 and 0.653 ± 0.004 for the two clearance assays. This represents a substantial improvement over models trained from scratch (R^2 of 0.505 ± 0.010 and 0.534 ± 0.006) and notably better performance than masking pretraining (R^2 of 0.580 ± 0.012 and 0.597 ± 0.002). Pretraining only on NMR is found to have a lower mean performance (R^2 of 0.635 ± 0.006 and 0.643 ± 0.005), but to not be significantly worse than the best model, while charges also provides close results (R^2 of 0.629 ± 0.006 and 0.633 ± 0.003). Models pretrained only on Fukui functions are found to have worse results that are comparable to the ones from models pretrained on HOMO-LUMO gap (R^2 of 0.602 ± 0.006 and 0.607 ± 0.005). Finally, in contrast with what found on the public benchmark, models pretrained with masking provide the worst results among pretrained models (R^2 of 0.580 ± 0.012 and 0.597 ± 0.002) while still improving over the non-pretrained baseline. Overall, these results demonstrate

that quantum mechanical pretraining can effectively address the data scarcity challenge in ADMET modeling producing models that are better than models trained from scratch. The superior performance of atom-level pretraining over HOMO-LUMO gap, despite the drastic difference in dataset size, suggests that these features provide a much more efficient way for pretraining than its global counterpart. Calculation of atomic properties, in fact, typically requires only a fraction of overall computational resources spent on geometry optimization and electronic structure refinement during QM modeling, hence providing a finer grade physical description of molecular structures with a non-dramatic overhead in the data generation phase. While the comparable performances of masking pretraining on the public benchmark seem to point at a less expensive pretraining strategy (as masking does not require any quantum calculation), on the larger internal HLM dataset it is found to provide the worst results, not confirming its usefulness in this particular experimental setting. The stronger differentiation between pretraining strategies on the larger clearance dataset strongly highlights the limitations of evaluating models solely on smaller public benchmarks.

5.3 The Effects of Atom-Level Pretraining on Graphormer

To understand how pretraining with quantum mechanical data influences model behavior beyond raw performance metrics, a comprehensive analysis of model representations and behavior is conducted. This analysis spans multiple aspects of the model's learned representations and reveals key insights into why atom-level quantum mechanical pretraining proves particularly effective.

5.3.1 Preservation of Pretraining Information

A crucial question when employing pretraining strategies is whether the learned representations retain information about the pretraining task after fine-tuning. To address this, the latent representations obtained in the last layer of the fine-tuned models are analyzed using a simple linear probe approach. For each model, the network is frozen after fine-tuning on ADMET tasks, and a sample of 5000 molecules from the pretraining datasets is encoded.

These latent representations are then split into equal size train/test sets and fit with a regularized linear regressor [222] to assess to what extent the representation still preserves linear correlation with the pretraining labels. The analysis is conducted in an all-to-all fashion, meaning that every differently pretrained model was tested for prediction of its own property but also for the properties used to pretrain the other models (only pretraining strategies with labels were considered, hence excluding masking).

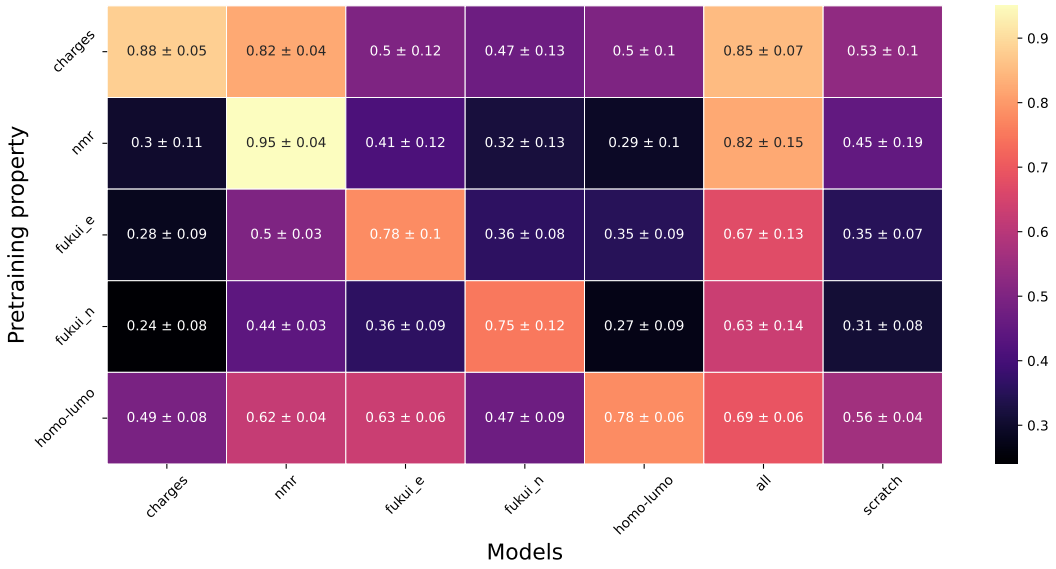


Figure 5.1: The R^2 for the regression tasks using the representations of a sample of the pretraining data obtained with fine-tuned models is reported. The mean and standard deviation are computed over all fine-tuning cases (mean and standard deviation over twenty-two cases).

The results, shown in Fig. 5.1, reveal that fine-tuned models retain substantial linear correlation with their pretraining tasks. Models pretrained on NMR shifts maintain their predictive power quite well, achieving $R^2 \approx 0.95$ for their own property while also maintaining good correlation with atomic charges ($R^2 \approx 0.82$). This suggests that NMR shifts contain richer information about local chemical environments compared to other atomic properties, aligning with their known sensitivity to electronic and steric effects. Models pretrained on Fukui function values and HOMO-LUMO gap exhibited slightly lower correlations with their respective pretraining tasks but still outperformed models trained from scratch, particularly when accounting for the standard deviation of results. Pretraining on all four atomic prop-

erties provided versatile representations, with strong correlations observed especially for charges and NMR shifts. Also, an asymmetry emerged in cross-task correlations: models pretrained on NMR shifts, Fukui indices, or all atomic properties exhibited some degree of alignment with HOMO-LUMO gap labels, whereas models pretrained on HOMO-LUMO gap showed limited correlation with atomic properties. This difference could stem from how the final network layers handle latent representations - atomic-property pretraining influences both atom-level and molecular CLS token representations, while HOMO-LUMO gap pretraining primarily impacts CLS token representations.

5.3.2 Latent Expressivity Across Layers

To gain deeper insights into how different pretraining strategies shape the network's internal representations, an analysis of the expressivity of learned features across network layers is performed. A common phenomenon in deep graph neural networks and transformers alike is the tendency for representations to become increasingly similar or "collapse" as information propagates through the layers. This effect has been extensively studied [223, 224, 225, 226, 227], in particular recent work on transformers [228] studies this in terms of representation rank collapse. Following this line of research, the similarity of the latent token representations at each layer is measured using the metric:

$$\rho_L = \frac{||res(GT_L(X))||1,\infty}{||GT_L(X)||1,\infty} \tag{5.1}$$

where $GT_L(X)$ is the latent representation at layer L, $res(X) = X - \mathbf{1}x^T$ with $x = \arg\min_x ||X - \mathbf{1}x^T||$, and $||\cdot||_{1,\infty} = \sqrt{||\cdot||1||\cdot||\infty}$. This metric measures how close the representation is to having all tokens equal to the same vector, with higher values indicating more distinct token representations. The results, shown in Fig. 5.2, reveal distinct patterns across pretraining strategies. All pretraining strategies mitigate the collapse in latent expressivity seen in models trained from scratch. However, the trend of ρ_L across layers varies significantly between approaches. Models pretrained on all atomic quantum properties exhibit a strong increase in expressivity in the early layers of the network, reaching higher maximum values

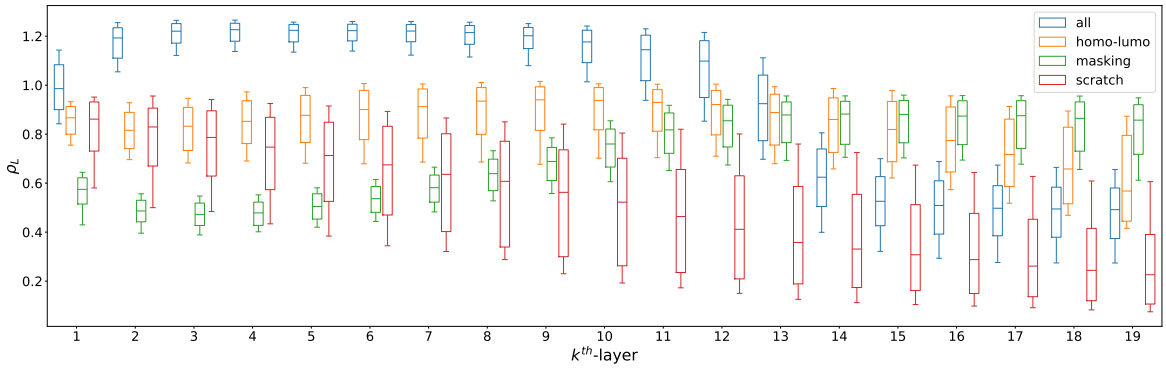


Figure 5.2: The expressivity of the latent representation, measured using the quantity ρ_L, is plotted as a function of layer number. This quantity is computed for a sample of 2200 structures, extracted uniformly from all the fine-tuning test sets (100 structures for each of the 22 tasks). The results are reported as boxplots for each layer. This analysis is conducted for models pretrained on HLG, models pretrained on all atom-level QM properties, models pretrained with masking, and models trained from scratch. The whiskers represent the 15th to the 85th percentiles to better visualize trends, and outliers are excluded for the same reason.

than other approaches, followed by a decrease in the last layers closer to the regression head. In contrast, models pretrained on HOMO-LUMO gap maintain a relatively constant level of expressivity across layers. Models pretrained using masking present a contrasting behavior, showing more similar atomic latent representations in the first few layers and more dissimilar representations in the last ones. This pattern appears opposite to models pretrained on atomic quantum properties, suggesting a fundamental difference in how information is processed through the network depending on the pretraining objective. For models pretrained on individual atomic properties, NMR shifts and charges models show similar behavior reaching higher maximum expressivity, while those pretrained on Fukui indices achieve lower maximum expressivity more similar to models pretrained on HOMO-LUMO gap. The absence of complete expressivity collapse in pretrained models likely stems from the much higher number of examples they were trained on compared to models trained from scratch. The sharp decrease in expressivity in the final layers of models pretrained on atomic properties may be related to the need for atom representations to capture correlations within molecular structures when close to the last layer, as these properties depend on surrounding atoms. This interpretation is supported by the opposing trend found in models

pretrained with masking, which require maximally distinct atomic representations near the network output for atom type classification, as this task is trained using a loss function that rewards higher certainties.

5.3.3 Spectral Analysis of Attention

To analyze how pretrained models mix information from different tokens, a natural approach is to use the attention rollout matrix [229]. This method iteratively stacks attention matrices from various layers and is defined as follows:

$$\tilde{A}(l_i) = \begin{cases} A(l_i)\tilde{A}(l_{i-1}) & \text{if } i > j \\ A(l_i) & \text{if } i = j \end{cases} \tag{5.2}$$

where $A(l_i)$ represents the attention matrix averaged over heads at layer l_i. This approach serves as a useful proxy for analyzing how a model mixes information from different input tokens. In particular, a spectral analysis of the attention rollout matrices from various models reveals an intriguing observation: the attention rollout matrix of a model pretrained on atom-level QM properties closely resembles the low-frequency eigenmodes of the Laplacian of the input molecular graph (see Fig. 5.3). This motivates a detailed examination of how different pretraining strategies influence the model's processing of molecular graph information.

The method used to explore the relationship between the eigenstructure of the attention rollout matrix $\tilde{A}$ and the graph Laplacian eigenmodes of input molecules is novel and outlined here. The attention rollout matrix can be decomposed as:

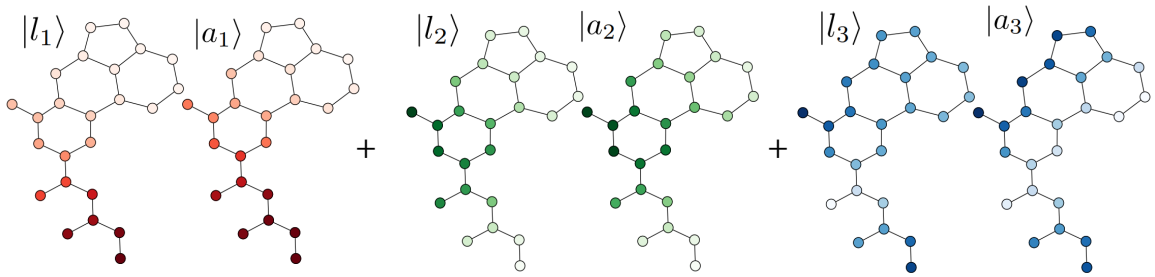
$$\tilde{A} = \sum_{i=0}^{N-1} a_i |a_i\rangle\langle a_i| \tag{5.3}$$

where a_i are eigenvalues sorted by magnitude and $|a_i\rangle$ are the corresponding eigenvectors. Similarly, for the molecular graph Laplacian L, the decomposition is:

$$L = \sum_{i=0}^{N-1} l_i |l_i\rangle\langle l_i| \tag{5.4}$$

$$|a_0| \geq |a_1| \geq |a_2| \geq |a_3| \geq \dots$$

$$\tilde{A} = a_0 \, |a_0\rangle \, \langle a_0| + a_1 \, |a_1\rangle \, \langle a_1| + a_2 \, |a_2\rangle \, \langle a_2| + a_3 \, |a_3\rangle \, \langle a_3| + \dots$$



$$L = l_0 \, |l_0\rangle \, \langle l_0| + l_1 \, |l_1\rangle \, \langle l_1| + l_2 \, |l_2\rangle \, \langle l_2| + l_3 \, |l_3\rangle \, \langle l_3| + \dots$$

$$l_0 \leq l_1 \leq l_2 \leq l_3 \leq \dots$$

Figure 5.3: A visual representation of a molecule from the TDC dataset comparing the most relevant eigenvectors of the attention rollout matrix from a model pretrained on atom-level QM properties with the low-frequency eigenvectors of the graph Laplacian associated with the molecular structure. $|a_i\rangle$ are the eigenvectors of the attention rollout matrix $\tilde{A}$ with eigenvalue a_i, and $|l_i\rangle$ are the eigenvectors of the graph Laplacian L with eigenvalue l_i.

with eigenvalues sorted in ascending order. By analyzing the overlap matrix $C_{ij} = |\langle l_i | a_j \rangle|$, it is possible to assess how the model's attention aligns with the natural modes of the graph. To quantify the degree of graph-spectral perception, the following definition is used:

$$\zeta = \eta \sum_{i=1}^{N-1} \Theta(\max_j C_{i,j} - 0.9) \tag{5.5}$$

where η is the fraction of non-trivial eigenvalue magnitude contained in Laplacian-like modes, Θ is the Heaviside function, and the threshold 0.9 identifies strong overlap. The value of ζ is then evaluated as an average over the test set for each downstream task, with the distribution across tasks reported for each pretraining condition. The results, reported in Fig. 5.4 as a distribution of 22 values over the downstream tasks per each group of studied models. This reveal that models trained from scratch exhibit values of ζ close to 0, indicating minimal perception of graph Laplacian eigenmodes. In contrast, every pretrained model shows nonzero values ranging from ~ 1 to ~ 6, with models pretrained on atom-level quantum prop-

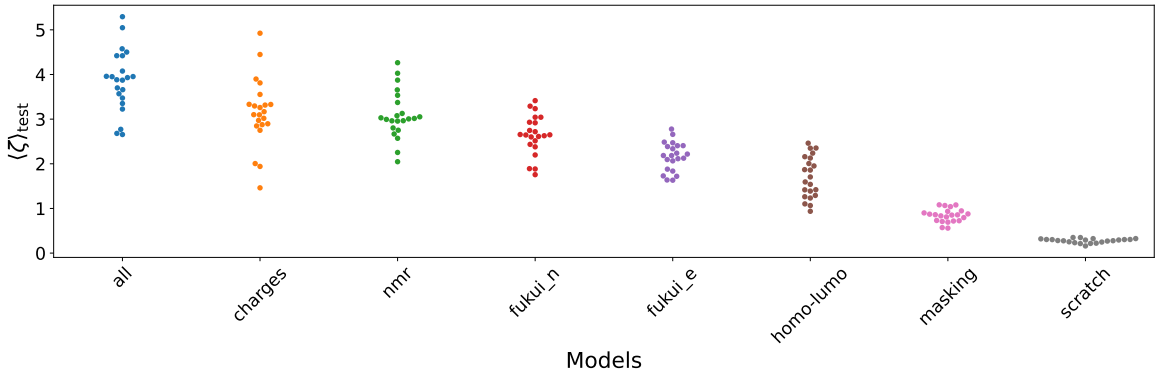


Figure 5.4: The spectral perception of the input graphs for the models fine-tuned on the TDC datasets is reported, grouped by pretraining strategy. This is presented in the form of swarm plots of the values of ζ, averaged across each of the 22 fine-tuning test sets for a fixed pretraining strategy.

erties showing the strongest graph-spectral perception. The model pretrained on all properties in a multi-task fashion achieves the highest ζ values, followed by models pretrained on charges, NMR shifts, and Fukui function indices. Models pretrained on HOMO-LUMO gap show lower but still significant graph-spectral perception, while masking pretraining yields the lowest values among pretrained models. This analysis suggests that atomic quantum mechanical pretraining enhances the model's ability to process molecular graph structure through attention mechanisms that naturally mimics a low-pass Fourier convolution with a wider filter than its non-pretrained counterpart. This improved structural perception may help explain the superior performance of these models on large-scale ADMET prediction tasks, and is also a sign of reduced oversmoothing, a well-known problem in GNN literature [230].

5.3.4 Neighbor Sensitivity Analysis

Building on the spectral analysis that revealed how models process global molecular graph structure, the focus now shifts to examining their ability to capture local chemical environments. This analysis is particularly relevant in light of the known challenge of oversquashing in graph neural networks [73], where bottlenecks in the message passing mechanism hinder proper information propagation between distant nodes. This issue can be interpreted as characterizing the effective receptive field of each atom's latent representation. To ad-

dress this, the Jacobian matrix of the network is used to analyze how sensitive each atomic representation is to changes in the input embedding of its k-th neighbors within the same molecule, analysing a quantity closely related to the one used in [231]:

$$S_k = \left\langle \sqrt{\sum_{\nu=0}^{d}\sum_{\mu=0}^{d}\left(\frac{\delta GT(X)i,\nu}{\delta Xk,\mu}\right)^2}_{k\in\mathcal{K}i} \right\rangle_{i\in\mathcal{M}} \tag{5.6}$$

where $\mathcal{K}_i$ is the set of k-th neighbors of atom i and $\mathcal{M}$ is the set of atoms in the molecule. For each molecule, the vector of sensitivities $[S_0, S_1, \ldots]$ is standardized by subtracting the minimum value and dividing by the maximum (which is usually S_0). The results, shown in

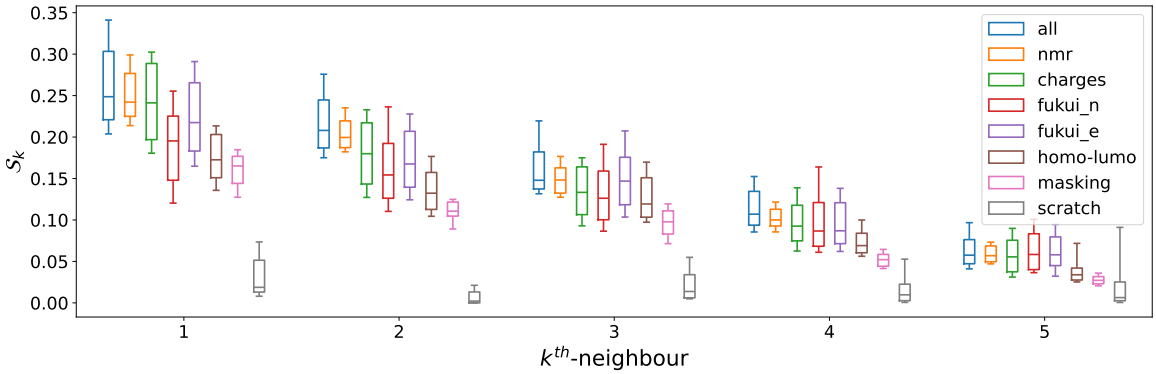


Figure 5.5: Boxplots of the k-th neighbor normalized sensitivities S_k for $k \in [1, \ldots, 5]$ are presented. Each boxplot summarizes a sample of 1100 structures, extracted uniformly from all the fine-tuning test sets (50 structures for each of the 22 tasks). This quantity is reported for all studied pretraining strategies, as well as for the models trained from scratch. The whiskers cover the values from the 15th to the 85th percentile to better visualize trends, and outliers are excluded for the same reason.

Fig. 5.5, reveal distinct patterns across pretraining strategies. Models trained from scratch exhibit constant and low sensitivity to neighboring atoms, suggesting an inability to develop meaningful local chemical environment representations. In contrast, all pretrained models show a clear descending trend with topological distance, indicating the ability to process atomic information with a good perception of the underlying molecular graph. Models pretrained on all atomic QM properties demonstrate the strongest sensitivity pattern, particularly for first and second neighbors, suggesting the development of the most refined representation of local chemical environments. This aligns with previous findings regarding enhanced

graph-spectral perception. Models pretrained on HOMO-LUMO gap show moderate sensitivity patterns, higher than those achieved through masking pretraining but not reaching the level of atom-level QM pretraining. Among models trained on individual atomic QM properties, sensitivity ranges tend to overlap but consistently position themselves between models pretrained on HOMO-LUMO gap and those pretrained on all atomic QM properties. This hierarchy in local environment sensitivity mirrors the one found for the analysis of ζ, and more importantly that of performance trends observed in modeling results on the large-scale clearance dataset, suggesting that the ability to effectively capture local chemical environments is crucial for modeling molecular properties in biological contexts. The analysis indicates that atomic quantum mechanical pretraining helps models develop more sophisticated representations of molecular structure, characterized by pronounced sensitivity to local atomic environments that naturally decays with topological distance. This behavior not only aligns with chemical intuition about the importance of local structural features but also suggests that QM pretraining helps mitigate the oversquashing problem by establishing more effective mechanisms for processing local chemical information.

Chapter 6

Discussion and perspectives

The convergence of multiple decades of computational science in the form of quantum mechanical methods for chemistry and machine learning has emerged as a powerful paradigm for advancing our understanding and design of molecular systems. This thesis has explored this synergy through multiple complementary approaches, which together populate the forefront of the development of a new and interdisciplinary way of approaching chemistry and drug discovery problems which has rapidly progressed in recent years. The contributions puth forth in this thesis include advancements from dataset development and novel computational methods, up to using neural networks for revealing both fundamental insights and practical applications. Here we synthesize our key findings and discuss their broader implications for the field.

Our investigation began with the development and analysis of the Aquamarine dataset. This is aimed at investigating the effects of non-covalent interactions and solvation on the conformational landscape of a set of molecules covering a similar chemical space to the one of a pharmaceutical company in the realm of medicinal chemistry. The analysis revealed the critical importance of many-body dispersion effects and solvation in drug-like molecules, and demonstrated that these collective effects become increasingly significant with molecular size and flexibility. In particular, strong differences were found in the total forces in the analysed structures when comparing molecules optimized with and without implicit solvation models, indicating that the inclusion of these more realistic non-covalent

effects has a strong impact on the forces at play, the form and hence the functionality of the molecules. This finding has immediate implications for drug discovery pipelines, suggesting that accurate modeling of these effects is essential for reliable prediction of molecular behavior in biological environments, as is the case in datasets developed for training neural network potentials. Furthermore, when correlating the many body and pairwise dispersion corrections in conformer ensambles of solvated structures, it was found that the pairwise approximation underestimates the corrections due to van der Waals interactions increasingly with molecular size.

Considering that the two main approximations involved in the treatment of dispersion corrections are the truncation of the many-body expansion to pairwise interactions and the use of the localized dipole approximation for charge fluctuations, we then considered the problem limiting many body approaches like MBD from a full-Coulombic treatment of non-covalent interactions. This led us to explore the possibility to use quantum computing approaches, specifically through photonic quantum simulation, in order to simulate this kind of interaction without dipole approximation. Our proof-of-concept study demonstrated the feasibility of maintaining full Coulomb interactions between quantum Drude oscillators without requiring truncated multipole expansions, and was made possible by the use of a neural network ansatz for the quantum state of the pair of cQDOs. The analysis of the results obtained revealed intriguing quantum mechanical phenomena, among which the presence of a binding curve following a Mueller potential in a specific range of model parameters, and also the emergence of cat states at the transition point of the binding curve. This last result is particularly interesting as it explains very well a peak in the entanglement entropy of the quantum state of the system. While currently limited to this simple case and of difficult implementation on real hardware, this work establishes a foundation for quantum-classical hybrid approaches that could eventually enable more accurate modeling of complex molecular systems, but also shows how the interpretation of results obtained from a neural network can give rise to interesting observations and span new research directions.

Building on the insights into molecular interactions coming from dataset analysis, we then tried to answer the natural question of whether it is possible to reduce the exploration

complexity of the chemical compound space by using quantum mechanical properties as coordinates. To this end we developed the Quantum Inverse Mapping (QIM) framework, establishing a differentiable parameterization of chemical compound space using quantum mechanical properties as intrinsic coordinates. This approach, which leverages on a VAE architecture, provides a novel way to navigate chemical space, where the differentiability is leveraged to reveal fundamental insights about its organization and also finds application in transition state interpolation. This proof-of-concept implementation, in fact, not only provided the answer to the initial scientific question, but was also used to reveal hierarchies in the properties used for navigation via explainability methods applied to the neural networks composing the model. Also, other than the possibility to retrieve the well-known conditional generative paradigm, the framework's ability to predict transition pathways between conformers as well as barrier-like behaviours in the energy coordinates, despite training only on equilibrium structures, hints at some degree of abstraction to the underlying physical mechanisms associating properties to structures.

Finally, the use of quantum mechanical data is put to the test with an application that is closer to the applicability in every day situations in the drug development processes. This is done by systematically analysing the impact of pretraining Graph Transformer models on quantum mechanical data as a mean to obtain better representations for the modeling of ADMET properties, which notoriously suffer from data scarcity and experimental noise. This evaluation, done both on the public TDC benchmark as well as on the internal JNJ HLM dataset, demonstrated that there is practical value of incorporating quantum mechanical information into machine learning pipelines as a mean to obtain more meaningful and robust molecular representation from deep learning models. This hypothesis is then put to the test, revealing through multiple model weight analysis that the superior performance achieved through atom-level quantum mechanical pretraining, indeed comes from molecular representations which have a better perception of the input molecular graph. The results from the multiple analysis conducted, and in particular the Fourier analysis of the attention weights, could potentially provide guidance for future model development.

Overall, looking forward, several promising research directions emerge from our findings.

Beyond the specific results outlined above, this body of work establishes a broader methodology for addressing research problems where mechanistic and physics-grounded understanding exists but proves hardly scalable due to intractable emerging complexities. This is evident across all works presented here, from the combinatorial intractability of chemical space itself, spanning an inconceivable number of structures, to the intractability of many-body electron wavefunctions in electronic structure problems, and finally to the complex interactions governing drug-like molecules in ADMET properties and drug discovery. Our approach addresses these challenges through three main steps: using local mechanistic knowledge to create focused simulations of complexity, delegating the complexity to appropriate machine learning methods (particularly leveraging the universal function approximation theorem of neural networks), and analyzing the resulting tools to gain new scientific insights about both the problem and the successful model's inner mechanisms. While this methodology has proven particularly effective in our proof-of-concept implementations, scaling these approaches to practical applications presents significant challenges. The computational cost of accurate quantum mechanical calculations remains a fundamental constraint, particularly evident in our Aquamarine dataset analysis where the gap between more and less accurate methods for dispersion correction widens with system size. When combined with the needs for conformational sampling and solvation effects, this creates substantial computational barriers. Similarly, while our quantum Drude oscillator simulation revealed interesting quantum phenomena, extending such approaches to practically relevant molecular systems faces both implementation challenges and hardware limitations inherent to quantum computing. In the realm of molecular representation, our QIM framework demonstrated that quantum mechanical properties can provide natural coordinates for understanding molecular behavior, though currently limited to small molecules with restricted chemical diversity. The work on Graphormer suggests that integrating atom-level quantum mechanical information can significantly enhance model performance, even on internal pharmaceutical data, but questions remain about optimal pretraining strategies and chemical space coverage. The complex nature of biological systems suggests that purely quantum mechanical approaches may need to be complemented with other types of domain knowledge and multimodality,

while maintaining rigorous analysis of how additional data affects model behavior beyond simple performance metrics. Nevertheless, this thesis demonstrates the potential of combining mechanistic understanding with machine learning approaches, providing both practical tools for molecular design and deeper understanding of fundamental principles. The future of molecular modeling likely lies in strongly interdisciplinary approaches that leverage both quantum mechanical accuracy, machine learning efficiency and chemical domain knowledge while maintaining some degree of interpretability. As computational capabilities advance and our understanding deepens, this synthesis of physical understanding, chemical intuition and data-driven methods will play an increasingly central role in advancing molecular science and drug discovery, with this thesis providing both concrete tools and a methodological framework for future development in this rapidly evolving field.

Bibliography

[1] Rutherford, E. The Scattering of α and β Particles by Matter and the Structure of the Atom. *Philosophical Magazine* **1911**, *21*, 669–688.

[2] Planck, M. Zur Theorie des Gesetzes der Energieverteilung im Normalspektrum. *Annalen der Physik* **1901**, *4*, 553–563.

[3] Einstein, A. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik* **1905**, *17*, 132–148.

[4] Bohr, N. I. On the constitution of atoms and molecules. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1913**, *26*, 1–25.

[5] De Broglie, Louis Recherches sur la théorie des Quanta. *Ann. Phys.* **1925**, *10*, 22–128.

[6] Schrödinger, E. Quantisierung als Eigenwertproblem. *Annalen der Physik* **1926**, *384*, 361–376.

[7] Breuer, H.-P.; Petruccione, F. *The Theory of Open Quantum Systems*; Oxford University Press, 2007.

[8] Dirac, P. A. M. A new notation for quantum mechanics. *Mathematical Proceedings of the Cambridge Philosophical Society* **1939**, *35*, 416–418.

[9] Gerlach, W.; Stern, O. Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld. *Zeitschrift für Physik* **1922**, *9*, 349–352.

[10] Pauli, W. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Physik* **1925**, *31*, 765–783.

[11] der Wissenschaften zu Berlin., D. A. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin*; Berlin, Deutsche Akademie der Wissenschaften zu Berlin, 1882-1918, 1882; Vol. Jan-Mai 1882; p 656.

[12] Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, *389*, 457–484.

[13] Slater, J. C. The Theory of Complex Spectra. *Phys. Rev.* **1929**, *34*, 1293–1322.

[14] Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Mathematical Proceedings of the Cambridge Philosophical Society* **1928**, *24*, 89–110.

[15] Fock, V. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Zeitschrift für Physik* **1930**, *61*, 126–148.

[16] Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

[17] Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **1992**, *46*, 6671–6687.

[18] Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **1996**, *105*, 9982–9985.

[19] Ostwald, W.; Hoff, J. H. v. Zeitschrift für physikalische Chemie (Leipzig). Abteilung A, Chemische Thermodynamik, Kinetik, Elektrochemie, Eigenschaftslehre. 1943.

[20] Ambrosetti, A.; Ferri, N.; DiStasio, R. A.; Tkatchenko, A. Wavelike charge density fluctuations and van der Waals interactions at the nanoscale. *Science* **2016**, *351*, 1171–1176.

[21] Hermann, J.; DiStasio, R. A. J.; Tkatchenko, A. First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chemical Reviews* **2017**, *117*, 4714–4758, PMID: 28272886.

[22] Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

[23] Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

[24] Reilly, A. M.; Tkatchenko, A. van der Waals dispersion interactions in molecular materials: beyond pairwise additivity. *Chem. Sci.* **2015**, *6*, 3289–3301.

[25] Yang, S.; Jiang, Y.; Li, S.; Liu, W. Many-body dispersion effects on the binding of TCNQ and F4-TCNQ with graphene. *Carbon* **2017**, *111*, 513–518.

[26] Stöhr, M.; Tkatchenko, A. Quantum mechanics of proteins in explicit water: The role of plasmon-like solute-solvent interactions. *Science Advances* **2019**, *5*, eaax0024.

[27] Aizerman, M. A.; Braverman, E. M.; Rozonoer, L. I. Theoretical foundation of potential functions method in pattern recognition. 2019.

[28] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **1989**, *2*, 303–314.

[29] Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **1989**, *2*, 359–366.

[30] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.

[31] Robbins, H. E. A Stochastic Approximation Method. *Annals of Mathematical Statistics* **1951**, *22*, 400–407.

[32] Hoeffding, W. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **1963**, *58*, 13–30.

[33] Geman, S.; Bienenstock, E.; Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **1992**, *4*, 1–58.

[34] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems. 2012.

[35] Zhao, X.; Wang, L.; Zhang, Y.; Han, X.; Deveci, M.; Parmar, M. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review* **2024**, *57*, 99.

[36] Mienye, I. D.; Swart, T. G.; Obaido, G. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information* **2024**, *15*.

[37] Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132.

[38] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; Polosukhin, I. Attention is All you Need. Advances in Neural Information Processing Systems. 2017.

[39] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics. 2019.

[40] Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. 2018.

[41] Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science* **2022**, *12*, e1608.

[42] Alakhdar, A.; Poczos, B.; Washburn, N. Diffusion Models in De Novo Drug Design. *Journal of Chemical Information and Modeling* **2024**, *64*, 7238–7256, PMID: 39322943.

[43] Raghunathan, S.; Priyakumar, U. D. Molecular representations for machine learning applications in chemistry. *International Journal of Quantum Chemistry* **2022**, *122*, e26870.

[44] Boldini, D.; Ballabio, D.; Consonni, V.; Todeschini, R.; Grisoni, F.; Sieber, S. A. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *Journal of Cheminformatics* **2024**, *16*, 35.

[45] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, PMID: 20426451.

[46] Škuta, C.; Cortés-Ciriano, I.; Dehaen, W.; Kříž, P.; van Westen, G. J. P.; Tetko, I. V.; Bender, A.; Svozil, D. QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *Journal of Cheminformatics* **2020**, *12*, 39.

[47] Cortés-Ciriano, I.; Škuta, C.; Bender, A.; Svozil, D. QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. *Journal of Cheminformatics* **2020**, *12*, 41.

[48] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

[49] Goh, G.; Hodas, N.; Siegel, C.; Vishnu, A. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. **2017**,

[50] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.;

Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **2018**, *4*, 268–276, PMID: 29532027.

[51] Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics* **2017**, *9*, 48.

[52] Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *Journal of Chemical Information and Modeling* **2020**, *60*, 1175–1183, PMID: 31904964.

[53] Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph neural networks for materials science and chemistry. *Communications Materials* **2022**, *3*, 93.

[54] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for Quantum chemistry. Proceedings of the 34th International Conference on Machine Learning - Volume 70. 2017; p 1263–1272.

[55] Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations. 2017.

[56] Bruna, J.; Zaremba, W.; Szlam, A.; Lecun, Y. Spectral networks and locally connected networks on graphs. International Conference on Learning Representations (ICLR2014), CBLS, April 2014. 2014.

[57] Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling* **2024**, *64*, 9–17, PMID: 38147829.

[58] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

[59] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters* **2015**, *6*, 2326–2331, PMID: 26113956.

[60] Huang, B.; Symonds, N. O.; Lilienfeld, O. A. v. *Handbook of Materials Modeling*; Springer International Publishing, 2018; p 1–27.

[61] Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* **2011**, *134*, 074106.

[62] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

[63] Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in Neural Information Processing Systems. 2017.

[64] Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications* **2022**, *13*, 2453.

[65] Frank, T.; Unke, O.; Müller, K.-R. So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems. Advances in Neural Information Processing Systems. 2022; pp 29400–29413.

[66] Batatia, I.; Kovács, D. P.; Simm, G. N. C.; Ortner, C.; Csányi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2024.

[67] Thölke, P.; Fabritiis, G. D. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. *ArXiv* **2022**, *arXiv:2202.02541*.

[68] Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. Proceedings of the 38th International Conference on Machine Learning. 2021; pp 9377–9388.

[69] Notwell, J. H.; Wood, M. W. ADMET property prediction through combinations of molecular fingerprints. *ArXiv* **2023**, *arXiv:2310.00174*.

[70] Correia, J.; Capela, J.; Rocha, M. DeepMol: An Automated Machine and Deep Learning Framework for Computational Chemistry. *bioRxiv* **2024**,

[71] Guo, Z.; Guo, K.; Nan, B.; Tian, Y.; Iyer, R.; Ma, Y.; Wiest, O.; Zhang, X.; Wang, W.; Zhang, C.; Chawla, N. Graph-based Molecular Representation Learning. 2023.

[72] Harnik, Y.; Milo, A. A focus on molecular representation learning for the prediction of chemical properties. *Chem. Sci.* **2024**, *15*, 5052–5055.

[73] Rusch, T. K.; Bronstein, M. M.; Mishra, S. A Survey on Oversmoothing in Graph Neural Networks. *ArXiv* **2023**, *arXiv:2303.10993*.

[74] Topping, J.; Giovanni, F. D.; Chamberlain, B. P.; Dong, X.; Bronstein, M. M. Understanding over-squashing and bottlenecks on graphs via curvature. International Conference on Learning Representations. 2022.

[75] Yun, S.; Jeong, M.; Kim, R.; Kang, J.; Kim, H. J. Graph Transformer Networks. Advances in Neural Information Processing Systems. 2019.

[76] Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do Transformers Really Perform Badly for Graph Representation? Advances in Neural Information Processing Systems. 2021.

[77] Müller, L.; Galkin, M.; Morris, C.; Rampášek, L. Attending to Graph Transformers. *Transactions on Machine Learning Research* **2024**,

[78] Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.

[79] Honda, S.; Shi, S.; Ueda, H. R. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. *ArXiv* **2019**, *arXiv:1911.04738*.

[80] Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019**, *5*, 1572–1583, PMID: 31572784.

[81] Li, J.; Jiang, X. Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction. *Wireless Communications and Mobile Computing* **2021**, *2021*, 1–7.

[82] Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M. H. S.; Meyers, J.; Fiscato, M.; Ahmed, M. Molecular representation learning with language models and domain-relevant auxiliary tasks. *CoRR* **2020**, *arXiv:2011.13230*.

[83] Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **2021**, *3*, 144–152.

[84] Satorras, V. G.; Hoogeboom, E.; Welling, M. E(n) Equivariant Graph Neural Networks. Proceedings of the 38th International Conference on Machine Learning. 2021; pp 9323–9332.

[85] Zheng, X.; Tomiura, Y. A BERT-based pretraining model for extracting molecular structural information from a SMILES sequence. *Journal of Cheminformatics* **2024**, *16*, 71.

[86] Chilingaryan, G.; Tamoyan, H.; Tevosyan, A.; Babayan, N.; Hambardzumyan, K.; Navoyan, Z.; Aghajanyan, A.; Khachatrian, H.; Khondkaryan, L. BartSmiles: Generative Masked Language Models for Molecular Representations. *Journal of Chemical Information and Modeling* **2024**, *64*, 5832–5843, PMID: 39054761.

[87] Li, S.; Zhang, L.; Wang, Z.; Wu, D.; Wu, L.; Liu, Z.; Xia, J.; Tan, C.; Liu, Y.; Sun, B.; Li, S. Z. Masked Modeling for Self-supervised Representation Learning on Vision and Beyond. *arXiv* **2024**,

[88] Shoghi, N.; Kolluru, A.; Kitchin, J. R.; Ulissi, Z. W.; Zitnick, C. L.; Wood, B. M. From Molecules to Materials: Pre-training Large Generalizable Models for Atomic Property Prediction. The Twelfth International Conference on Learning Representations. 2024.

[89] Batatia, I. et al. A foundation model for atomistic materials chemistry. *arXiv* **2024**, *arXiv:2401.00096*.

[90] Sanchez-Fernandez, A.; Rumetshofer, E.; Hochreiter, S.; Klambauer, G. CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications* **2023**, *14*, 7339.

[91] Liu, Y.; Zhang, R.; yuan, Y.; Ma, J.; Li, T.; Yu, Z. A Multi-view Molecular Pre-training with Generative Contrastive Learning. *Interdisciplinary Sciences: Computational Life Sciences* **2024**, *16*, 741–754.

[92] Xie, J.; Wang, Y.; Rao, J.; Zheng, S.; Yang, Y. Self-Supervised Contrastive Molecular Representation Learning with a Chemical Synthesis Knowledge Graph. *Journal of Chemical Information and Modeling* **2024**, *64*, 1945–1954, PMID: 38484468.

[93] Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; Chen, H. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence* **2023**, *5*, 542–553.

[94] Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692–1701.

[95] von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry* **2020**, *4*, 347–358.

[96] Perdew, J. P.; Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings* **2001**, *577*, 1–20.

[97] Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Research* **2015**, *44*, D1202–D1213.

[98] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

[99] Tingle, B. I.; Tang, K. G.; Castanon, M.; Gutierrez, J. J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y. S.; Irwin, J. J. ZINC-22 A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *Journal of Chemical Information and Modeling* **2023**, *63*, 1166–1176, PMID: 36790087.

[100] Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *Journal of Chemical Information and Modeling* **2007**, *47*, 342–353, PMID: 17260980.

[101] Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society* **2009**, *131*, 8732–8733, PMID: 19505099.

[102] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875, PMID: 23088335.

[103] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics* **2013**, *15*, 095003.

[104] Hoja, J.; Medrano Sandonas, L.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio Jr., R. A.; Tkatchenko, A. QM7-X, a comprehensive dataset of quantum-

mechanical properties spanning the chemical space of small organic molecules. *Scientific Data* **2021**, *8*, 43.

[105] Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **2009**, *180*, 2175–2196.

[106] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, 140022.

[107] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

[108] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

[109] Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1972**, *56*, 2257–2261.

[110] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2011**, *40*, D1100–D1107.

[111] Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671, PMID: 30741547.

[112] Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006**, *27*, 1787–1799.

[113] Hellweg, A.; Rappoport, D. Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD,

def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations. *Phys. Chem. Chem. Phys.* **2015**, *17*, 1010–1017.

[114] Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data* **2017**, *4*, 170193.

[115] Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data* **2020**, *7*, 134.

[116] Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation* **2020**, *16*, 4192–4202, PMID: 32543858.

[117] Axelrod, S.; Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data* **2022**, *9*, 185.

[118] Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *Journal of Chemical Information and Modeling* **2016**, *56*, 1936–1949, PMID: 27689393.

[119] Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules. *The Journal of Physical Chemistry A* **2021**, *125*, 4039–4054, PMID: 33688730.

[120] Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chemical Reviews* **2021**, *121*, 9816–9872, PMID: 34232033.

[121] Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.;

Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chemical Reviews* **2021**, *121*, 10142–10186, PMID: 33705118.

[122] Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. S. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. The Eleventh International Conference on Learning Representations. 2023.

[123] Ashtawy, H. M.; Mahapatra, N. R. Molecular Docking for Drug Discovery: Machine-Learning Approaches for Native Pose Prediction of Protein-Ligand Complexes. Computational Intelligence Methods for Bioinformatics and Biostatistics. Cham, 2014; pp 15–32.

[124] Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. International Conference on Learning Representations. 2022.

[125] Kim, J.; Chang, W.; Ji, H.; Joung, I. Quantum-Informed Molecular Representation Learning Enhancing ADMET Property Prediction. *Journal of Chemical Information and Modeling* **2024**, *64*, 5028–5040, PMID: 38916580.

[126] Galante, M.; Tkatchenko, A. Anisotropic van der Waals dispersion forces in polymers: Structural symmetry breaking leads to enhanced conformational search. *Phys. Rev. Res.* **2023**, *5*, L012028.

[127] Decherchi, S.; Cavalli, A. Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chemical Reviews* **2020**, *120*, 12788–12833, PMID: 33006893.

[128] Oko, O.; Antai, E.; Adindu, E.; Godfrey, O.; Bassey, I.; Nwaokolrie, F.; Owolabi, A.; Nkang, A.; Gber, T.; Edim, M.; Louis, H. Unraveling the Impact of Polar Solvation on the Molecular geometry Spectroscopy (FT-IR, UV, NMR), Reactivity (ELF, NBO, HOMO-LUMO) and Antiviral Inhibitory Potential of Cissampeline by Molecular Docking approach. *Chemical Physics Impact* **2023**, *7*, 100346.

[129] Matczak, P.; Domagała, M. Heteroatom and solvent effects on molecular properties of formaldehyde and thioformaldehyde symmetrically disubstituted with heterocyclic groups C4H3Y (where Y = O–Po). *Journal of Molecular Modeling* **2017**, *23*, 268.

[130] Ensing, B.; Meijer, E. J.; Blöchl, P. E.; Baerends, E. J. Solvation Effects on the SN2 Reaction between CH3Cl and Cl- in Water. *The Journal of Physical Chemistry A* **2001**, *105*, 3300–3310.

[131] Gorges, J.; Grimme, S.; Hansen, A.; Pracht, P. Towards understanding solvation effects on the conformational entropy of non-rigid molecules. *Phys. Chem. Chem. Phys.* **2022**, *24*, 12249–12259.

[132] Eastman, P.; Behara, P. K.; Dotson, D. L.; Galvelis, R.; Herr, J. E.; Horton, J. T.; Mao, Y.; Chodera, J. D.; Pritchard, B. P.; Wang, Y.; De Fabritiis, G.; Markland, T. E. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Scientific Data* **2023**, *10*, 11.

[133] Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry* **1996**, *17*, 520–552.

[134] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 490–519.

[135] Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 553–586.

[136] Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. conformational energies and geometries for MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 587–615.

[137] Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *Journal of Computational Chemistry* **1996**, *17*, 616–641.

[138] Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

[139] Xie, L.; Liu, H. The treatment of solvation by a generalized Born model and a self-consistent charge-density functional theory-based tight-binding method. *Journal of Computational Chemistry* **2002**, *23*, 1404–1415.

[140] Seifert, G.; Porezag, D.; Frauenheim, T. Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme. *International Journal of Quantum Chemistry* **1996**, *58*, 185–192.

[141] Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260–7268.

[142] Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *Journal of Chemical Theory and Computation* **2011**, *7*, 931–948.

[143] Mortazavi, M.; Brandenburg, J. G.; Maurer, R. J.; Tkatchenko, A. Structure and Stability of Molecular Crystals with Many-Body Dispersion-Inclusive Density Functional Tight Binding. *The Journal of Physical Chemistry Letters* **2018**, *9*, 399–405, PMID: 29298075.

[144] Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2010**, *50*, 572–584, PMID: 20235588.

[145] Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *Journal of Chemical Information and Modeling* **2020**, *60*, 2044–2058, PMID: 32155061.

[146] Larsen, A. et al. The Atomic Simulation Environment — A Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, *29*.

[147] Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *The Journal of Physical Chemistry A* **2007**, *111*, 5678–5684, PMID: 17567110.

[148] Hoja, J.; Ko, H.-Y.; Neumann, M. A.; Car, R.; DiStasio, R. A.; Tkatchenko, A. Reliable and practical computational description of molecular crystal polymorphs. *Science Advances* **2019**, *5*, eaau3338.

[149] Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional. *The Journal of Chemical Physics* **1999**, *110*, 5029–5036.

[150] Reilly, A. M.; Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *The Journal of Chemical Physics* **2013**, *139*, 024705.

[151] Ringe, S.; Oberhofer, H.; Hille, C.; Matera, S.; Reuter, K. Function-Space-Based Solution Scheme for the Size-Modified Poisson–Boltzmann Equation in Full-Potential DFT. *Journal of Chemical Theory and Computation* **2016**, *12*, 4052–4066, PMID: 27323006.

[152] Abidi, N.; Lim, K. R. G.; Seh, Z. W.; Steinmann, S. N. Atomistic modeling of electrocatalysis: Are we there yet? *WIREs Computational Molecular Science* **2021**, *11*, e1499.

[153] Gauthier, J. A.; Dickens, C. F.; Heenen, H. H.; Vijay, S.; Ringe, S.; Chan, K. Unified Approach to Implicit and Explicit Solvent Simulations of Electrochemical Reaction Energetics. *Journal of Chemical Theory and Computation* **2019**, *15*, 6895–6906, PMID: 31689089.

[154] Ringe, S.; Hörmann, N. G.; Oberhofer, H.; Reuter, K. Implicit Solvation Methods for Catalysis at Electrified Interfaces. *Chemical Reviews* **2022**, *122*, 10777–10820, PMID: 34928131.

[155] Ringe, S.; Clark, E. L.; Resasco, J.; Walton, A.; Seger, B.; Bell, A. T.; Chan, K. Understanding cation effects in electrochemical CO2 reduction. *Energy Environ. Sci.* **2019**, *12*, 3001–3014.

[156] Ambrosetti, A.; Reilly, A. M.; DiStasio, J., Robert A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *The Journal of Chemical Physics* **2014**, *140*, 18A508.

[157] Góger, S.; Sandonas, L. M.; Müller, C.; Tkatchenko, A. Data-driven tailoring of molecular dipole polarizability and frontier orbital energies in chemical compound space. *Phys. Chem. Chem. Phys.* **2023**, *25*, 22211–22222.

[158] Medrano Sandonas, L.; Hoja, J.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio, R. A.; Tkatchenko, A. "Freedom of design" in chemical compound space: towards rational in silico design of molecules with targeted quantum-mechanical properties. *Chem. Sci.* **2023**, *14*, 10702–10717.

[159] Whitfield, T. W.; Martyna, G. J. A unified formalism for many-body polarization and dispersion: The quantum Drude model applied to fluid xenon. *Chemical Physics Letters* **2006**, *424*, 409–413.

[160] Cao, Y.; Romero, J.; Olson, J. P.; Degroote, M.; Johnson, P. D.; Kieferová, M.; Kivlichan, I. D.; Menke, T.; Peropadre, B.; Sawaya, N. P. D.; Sim, S.; Veis, L.; Aspuru-Guzik, A. Quantum Chemistry in the Age of Quantum Computing. *Chemical Reviews* **2019**, *119*, 10856–10915, PMID: 31469277.

[161] Peruzzo, A.; McClean, J.; Shadbolt, P.; Yung, M.-H.; Zhou, X.-Q.; Love, P. J.; Aspuru-Guzik, A.; O'Brien, J. L. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications* **2014**, *5*, 4213.

[162] Kandala, A.; Mezzacapo, A.; Temme, K.; Takita, M.; Brink, M.; Chow, J. M.; Gambetta, J. M. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature* **2017**, *549*, 242–246.

[163] Nam, Y. et al. Ground-state energy estimation of the water molecule on a trapped-ion quantum computer. *npj Quantum Information* **2020**, *6*, 33.

[164] Anderson, L. W.; Kiffner, M.; Barkoutsos, P. K.; Tavernelli, I.; Crain, J.; Jaksch, D. Coarse-grained intermolecular interactions on quantum processors. *Phys. Rev. A* **2022**, *105*, 062409.

[165] Nielsen, M. A.; Chuang, I. L. *Quantum Computation and Quantum Information*; Cambridge University Press, 2000.

[166] Schrödinger, E. Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften* **1935**, *23*, 807–812.

[167] Yurke, B.; Stoler, D. Generating quantum mechanical superpositions of macroscopically distinguishable states via amplitude dispersion. *Phys. Rev. Lett.* **1986**, *57*, 13–16.

[168] Tilly, J.; Chen, H.; Cao, S.; Picozzi, D.; Setia, K.; Li, Y.; Grant, E.; Wossnig, L.; Rungger, I.; Booth, G. H.; Tennyson, J. The Variational Quantum Eigensolver: A review of methods and best practices. *Physics Reports* **2022**, *986*, 1–128, The Variational Quantum Eigensolver: a review of methods and best practices.

[169] Killoran, N.; Bromley, T. R.; Arrazola, J. M.; Schuld, M.; Quesada, N.; Lloyd, S. Continuous-variable quantum neural networks. *Phys. Rev. Res.* **2019**, *1*, 033063.

[170] Arrazola, J. M.; Bromley, T. R.; Izaac, J.; Myers, C. R.; Brádler, K.; Killoran, N. Machine learning method for state preparation and gate synthesis on photonic quantum computers. *Quantum Science and Technology* **2019**, *4*, 024004.

[171] Okamoto, K. In *Fundamentals of Optical Waveguides (Second Edition)*, second edition ed.; Okamoto, K., Ed.; Academic Press: Burlington, 2006; pp 209–259.

[172] Serafini, A. *Quantum Continuous Variables: A Primer of Theoretical Methods*; CRC Press, 2017; Chapter 5, pp xx–yy.

[173] Sadhukhan, M.; Manby, F. R. Quantum mechanics of Drude oscillators with full Coulomb interaction. *Phys. Rev. B* **2016**, *94*, 115106.

[174] Herzberg, G. *Molecular spectra and molecular structure*; D. van Nostrand: New York, 1945.

[175] Apanavicius, J.; Feng, Y.; Flores, Y.; Hassan, M.; McGuigan, M. Morse Potential on a Quantum Computer for Molecules and Supersymmetric Quantum Mechanics. *arXiv* **2021**, *arXiv:2102.05102*.

[176] Le Roy, R. J.; Huang, Y.; Jary, C. An accurate analytic potential function for ground-state N2 from a direct-potential-fit analysis of spectroscopic data. *The Journal of Chemical Physics* **2006**, *125*, 164310.

[177] Roy, R. J. L.; Henderson, R. D. E. A new potential function form incorporating extended long-range behaviour: application to ground-state Ca2. *Molecular Physics* **2007**, *105*, 663–677.

[178] Chiu, S.-W.; Scott, H. L.; Jakobsson, E. A Coarse-Grained Model Based on Morse Potential for Water and n-Alkanes. *Journal of Chemical Theory and Computation* **2010**, *6*, 851–863, PMID: 26613312.

[179] Serafini, A. *Quantum Continuous Variables: A Primer of Theoretical Methods*; CRC Press, 2017; Chapter 4.

[180] Chabaud, U.; Emeriau, P.-E.; Grosshans, F. Witnessing Wigner Negativity. *Quantum* **2021**, *5*, 471.

[181] Arkhipov, I. I.; Barasiński, A.; Svozilík, J. Negativity volume of the generalized Wigner function as an entanglement witness for hybrid bipartite states. *Scientific Reports* **2018**, *8*, 16955.

[182] Grimm, A.; Frattini, N. E.; Puri, S.; Mundhada, S. O.; Touzard, S.; Mirrahimi, M.; Girvin, S. M.; Shankar, S.; Devoret, M. H. Stabilization and operation of a Kerr-cat qubit. *Nature* **2020**, *584*, 205–209.

[183] Puri, S.; Boutin, S.; Blais, A. Engineering the quantum states of light in a Kerr-nonlinear resonator by two-photon driving. *npj Quantum Information* **2017**, *3*, 18.

[184] Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *Journal of the American Chemical Society* **2023**, *145*, 8736–8750, PMID: 37052978.

[185] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. Advances in Neural Information Processing Systems. 2014.

[186] De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models* **2018**,

[187] Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems. 2020; pp 6840–6851.

[188] Hoogeboom, E.; Satorras, V. G.; Vignac, C.; Welling, M. Equivariant Diffusion for Molecule Generation in 3D. Proceedings of the 39th International Conference on Machine Learning. 2022; pp 8867–8887.

[189] Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2022**, *arXiv:1312.6114*.

[190] Ochiai, T.; Inukai, T.; Akiyama, M.; Furui, K.; Ohue, M.; Matsumori, N.; Inuki, S.; Uesugi, M.; Sunazuka, T.; Kikuchi, K.; Kakeya, H.; Sakakibara, Y. Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. *Communications Chemistry* **2023**, *6*, 249.

[191] Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics* **2018**, *10*, 31.

[192] Dollar, O.; Joshi, N.; Pfaendtner, J.; Beck, D. A. C. Efficient 3D Molecular Design with an E(3) Invariant Transformer VAE. *The Journal of Physical Chemistry A* **2023**, *127*, 7844–7852, PMID: 37670244.

[193] Dokmanic, I.; Parhizkar, R.; Ranieri, J.; Vetterli, M. Euclidean Distance Matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine* **2015**, *32*, 12–30.

[194] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

[195] Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* **2014**, *arXiv:1312.6034*.

[196] Collins, T. C.; Euwema, R. N.; Stukel, D. J.; Wepfer, G. G. Valence electron density of states of ZnSe obtained from an energy dependent exchange approximation. *International Journal of Quantum Chemistry* **1970**, *5*, 77–85.

[197] Rincón, L.; Alvarellos, J. E.; Almeida, R. Electron density, exchange-correlation density, and bond characterization from the perspective of the valence-bond theory. II. Numerical results. *J. Chem. Phys.* **2005**, *122*.

[198] Shao, H.; Kumar, A.; Fletcher, P. T. The Riemannian Geometry of Deep Generative Models. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018; pp 428–4288.

[199] Makri, S.; Ortner, C.; Kermode, J. R. A preconditioning scheme for minimum energy path finding methods. *J. Chem. Phys.* **2019**, *150*, 094109.

[200] Unke, O.; Chmiela, S.; Gastegger, M.; Schütt, K.; Sauceda, H. E.; Müller, K.-R. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature Communications* **2021**, *12*, 7273.

[201] Zhu, X.; Thompson, K.; Martinez, T. Geodesic interpolation for reaction pathways. *J. Chem. Phys.* **2019**, *150*, 164103.

[202] Daoud, N. E.-H.; Borah, P.; Deb, P. K.; Venugopala, K. N.; Hourani, W.; Alzweiri, M.; Bardaweel, S. K.; Tiwari, V. ADMET Profiling in Drug Discovery and Development: Perspectives of In Silico, In Vitro and Integrated Approaches. *Current Drug Metabolism* **2021**, *22*, 503–522.

[203] Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chemical Reviews* **2023**, *123*, 8736–8780, PMID: 37384816.

[204] Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology* **2022**, *18*, 1033–1036.

[205] Born, J.; Markert, G.; Janakarajan, N.; Kimber, T. B.; Volkamer, A.; Martínez, M. R.; Manica, M. Chemical representation learning for toxicity prediction. *Digital Discovery* **2023**, *2*, 674–691.

[206] Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *Journal of Cheminformatics* **2019**, *11*, 69.

[207] Ektefaie, Y.; Shen, A.; Bykova, D.; Marin, M.; Zitnik, M.; Farhat, M. Evaluating generalizability of artificial intelligence models for molecular datasets. *bioRxiv* **2024**, *doi:10.1101/2024.02.25.581982*.

[208] Broccatelli, F.; Trager, R.; Reutlinger, M.; Karypis, G.; Li, M. Benchmarking Accu-

racy and Generalizability of Four Graph Neural Networks Using Large In Vitro ADME Datasets from Different Chemical Spaces. *Molecular Informatics* **2022**, *41*, 2100321.

[209] David Z Huang, J. C. B.; Bahmanyar, S. S. The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction. *Expert Opinion on Drug Discovery* **2021**, *16*, 1045–1056, PMID: 33739897.

[210] Kaufman, B.; Williams, E. C.; Underkoffler, C.; Pederson, R.; Mardirossian, N.; Watson, I.; Parkhill, J. COATI: Multimodal Contrastive Pretraining for Representing and Traversing Chemical Space. *Journal of Chemical Information and Modeling* **2024**, *64*, 1145–1157, PMID: 38316665.

[211] Wang, Y.; Xu, C.; Li, Z.; Barati Farimani, A. Denoise Pretraining on Nonequilibrium Molecules for Accurate and Transferable Neural Potentials. *Journal of Chemical Theory and Computation* **2023**, *19*, 5077–5087, PMID: 37390120.

[212] Xia, J.; Zhao, C.; Hu, B.; Gao, Z.; Tan, C.; Liu, Y.; Li, S.; Li, S. Z. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. The Eleventh International Conference on Learning Representations. 2023.

[213] Xia, J.; Zhu, Y.; Du, Y.; Li, S. Z. A Systematic Survey of Chemical Pre-trained Models. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23. 2023; pp 6787–6795, Survey Track.

[214] Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. International Conference on Learning Representations. 2020.

[215] Beck, M. E. Do Fukui Function Maxima Relate to Sites of Metabolism? A Critical Case Study. *Journal of Chemical Information and Modeling* **2005**, *45*, 273–282, PMID: 15807488.

[216] Wang, X.; Wang, L.; Wang, S.; Ren, Y.; Chen, W.; Li, X.; Han, P.; Song, T. Quan-

tumTox: Utilizing quantum chemistry with ensemble learning for molecular toxicity prediction. *Computers in Biology and Medicine* **2023**, *157*, 106744.

[217] Beck, M. E.; Schindler, M. *Pesticide Chemistry*; John Wiley & Sons, Ltd, 2007; Chapter 24, pp 227–238.

[218] Göller, A. H. The art of atom descriptor design. *Drug Discovery Today: Technologies* **2019**, *32-33*, 37–43, Artificial Intelligence.

[219] Nugmanov, R.; Dyubankova, N.; Gedich, A.; Wegner, J. K. Bidirectional Graphormer for Reactivity Understanding: Neural Network Trained to Reaction Atom-to-Atom Mapping Task. *Journal of Chemical Information and Modeling* **2022**, *62*, 3307–3315, PMID: 35792579.

[220] Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chemical Science* **2021**, *12*, 2198–2208.

[221] Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling* **2017**, *57*, 1300–1308.

[222] Zou, H.; Hastie, T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2005**, *67*, 301–320.

[223] Shi, H.; GAO, J.; Xu, H.; Liang, X.; Li, Z.; Kong, L.; Lee, S. M. S.; Kwok, J. Revisiting Over-smoothing in BERT from the Perspective of Graph. International Conference on Learning Representations. 2022.

[224] Noci, L.; Li, C.; Li, M.; He, B.; Hofmann, T.; Maddison, C. J.; Roy, D. The Shaped Transformer: Attention Models in the Infinite Depth-and-Width Limit. Advances in Neural Information Processing Systems. 2023; pp 54250–54281.

[225] Noci, L.; Anagnostidis, S.; Biggio, L.; Orvieto, A.; Singh, S. P.; Lucchi, A. Signal Propagation in Transformers: Theoretical Perspectives and the Role of Rank Collapse. Advances in Neural Information Processing Systems. 2022; pp 27198–27211.

[226] Roth, A.; Bause, F.; Kriege, N. M.; Liebig, T. Preventing Representational Rank Collapse in MPNNs by Splitting the Computational Graph. The Third Learning on Graphs Conference. 2024.

[227] Roth, A.; Liebig, T. Rank Collapse Causes Over-Smoothing and Over-Correlation in Graph Neural Networks. Proceedings of the Second Learning on Graphs Conference. 2024; pp 35:1–35:23.

[228] Dong, Y.; Cordonnier, J.-B.; Loukas, A. Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth. *arXiv* **2023**, *arXiv:2103.03404*.

[229] Abnar, S.; Zuidema, W. H. Quantifying Attention Flow in Transformers. *arXiv* **2020**, *arXiv:2005.00928*.

[230] Rusch, T. K.; Bronstein, M. M.; Mishra, S. A Survey on Oversmoothing in Graph Neural Networks. *arXiv* **2023**, *arXiv:2303.10993*.

[231] Topping, J.; Giovanni, F. D.; Chamberlain, B. P.; Dong, X.; Bronstein, M. M. Understanding over-squashing and bottlenecks on graphs via curvature. International Conference on Learning Representations. 2022.