SCHOOL OF NATURAL SCIENCES

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Quantum Science and Technology

# Towards Geometric Neural Wave-Functions

**Benjamin Classen**

SCHOOL OF NATURAL SCIENCES

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Quantum Science and Technology

# Towards Geometric Neural Wave-Functions

# Ansatz zu Geometrischen Neuronalen Wellenfunktionen

| | |
|---|---|
| Author: | Benjamin Classen |
| Supervisor: | Professor Christian Mendl |
| Submission Date: | 19. Juli 2024 |

I confirm that this Master's thesis in Quantum Science and Technology is my own work and I have documented all sources and material used.

München, 19. Juli 2024 Benjamin Classen

# Acknowledgments

To Lucy, who always kept up my curiosity, and who taught me to stop searching for answers and start looking for questions.

# Abstract

Deep neural networks provide a novel and highly successful avenue to calculate notoriously challenging electronic ground state wave-functions for molecular Hamiltonians. Whereas initial architectures were trained separately for any given molecular Hamiltonian, neural networks have recently been proposed which simultaneously calculate the ground state wave-functions for multiple geometrical arrangements of a given molecule. This work constitutes a further foray in that vein, proposing a neural architecture which explicitly recognizes rotations of a molecule's atomic positions. Such explicit recognition of spatial symmetries has a celebrated history in related fields of research as a means to reduce training effort and increase expressiveness of the ansatz. While our proposed neural wave-function is indeed able to recognise spatial symmetries as desired, it nevertheless fails the tests of empiricism, being far from competitive with the state-of-the-art, which we corroborate with a series of numerical experiments.

# Kurzfassung

Neuronale Netze stellen einen neuartigen und äußerst erfolgreichen Ansatz zur Berechnung der notorisch schwierigen elektronischen Grundzustandswellenfunktionen für molekulare Hamiltonians dar. Während initiale Architekturen separat für jeden gegebenen molekularen Hamiltonian trainiert wurden, wurden kürzlich neuronale Netze vorgeschlagen, die simultan die Grundzustandswellenfunktionen für mehrere geometrische Anordnungen eines gegebenen Moleküls berechnen. Die vorliegende Arbeit schlägt in dieselbe Kerbe, indem sie eine neuronale Architektur vorschlägt, die explizit Rotationen der atomaren Positionen eines Moleküls erkennt. Eine solche explizite Erkennung räumlicher Symmetrien hat sich in verwandten Forschungsbereichen als sehr gewinnbringend dafür erwiesen, Trainingszeit zu reduzieren und die Expressivität des Ansatzes zu steigern. Die von uns vorgeschlagene neuronale Wellenfunktion ist zwar in der Lage, räumliche Symmetrien wie gewünscht zu erkennen, ist aber nicht konkurrenzfähig mit dem aktuellen Stand der Technik, was wir mit einer Reihe von numerischen Experimenten untermauern.

# Contents

# 1. Introduction

Artificial intelligence in its various flavors has proved to be truly paradigm-shifting across a wide span of domains. The advent of deep neural networks demonstrating image-recognition of unprecedented accuracy [1] heralded an era of seemingly ubiquitous breakthroughs in diverse areas like mastering chess and go [2], folding proteins [3], the most recent prodigy ChatGPT and countless more.

While a satisfying understanding of the actual mechanism underlying AI's seeming omnipotence is sill mostly elusive [4], the ability of deep neural networks to model arbitrary functions to any degree desired has been known for decades [5]. Nevertheless, only relatively recently have researchers begun to exploit neural networks as approximations for a particularly notorious class of functions, namely the wave-functions corresponding to quantum mechanical Hamiltonians. The initial foray of Carleo et al. in 2017 [6] modelling a many-body quantum state of $N$ spins with a deep neural net spurred lots of investigations how to model the wave-functions of other systems via neural nets. One class of such systems are quantum chemical Hamiltonians governing the behaviour of molecules. The two concurrently published models FermiNet [7] and PauliNet [8] constituted the first attempts to calculate the ground state wave-function of molecular Hamiltonians. Both reported impressive levels of accuracy, and, interestingly, doing so by calculating the wave-function in first quantization. This starkly contrasts conventional quantum chemistry methods which mostly work in the framework of second quantization. More in the spirit of conventional quantum chemistry methods, Choo et al. [9] examined with considerable success the use of neural networks to parametrize wave-functions of second-quantized molecular Hamiltonians.

Since then, the two lines of research of parametrizing molecular wave-functions in first and second quantization respectively have been pursued somewhat in parallel. In both, much effort has been invested on methodological fine-tuning resulting in reduced computational cost and increased accuracy. Especially however for first-quantized neural network wave-functions - or 1Q-NN-WFs, as we will refer to them - the sights have been set on more ambitious goals. The vanilla versions of FermiNet and PauliNet described the wave-function for any particular molecular geometry only. In practice however, often relative energies between different molecular geometries are of interest [10], prompting the search for 1Q-NN-WFs with the ability to model the wave-functions corresponding to different molecular geometries with one set of parameters at once. This is not only desirable on practical grounds, alleviating the massive computational burden resulting from retraining a model for every molecular configuration

[11], but furthermore incentivises the neural network to learn features of electronic correlation in general, rather than (over-)fitting to specific geometries [12]. One fruitful approach to slim down training efforts is to minimize the amount of unphysical information the neural network has to learn. For example, initial 1Q-NN-WFs were not able to recognise rotated molecular geometries as such and thus had to learn geometrical symmetries from vast amount of data. Incorporating such kinds of symmetries explicitly in the neural architecture has proven to reduce computational costs significantly in closely related fields of research [13]. However, a neural wave-function that fully recognizes rotational symmetries of the molecule has not yet been proposed and is an active field of research [14].

In this thesis, we aim to construct a neural wave-function that recognizes rotated molecular geometries and is thus guaranteed to treat them on equal footing. To that order, we will harness concepts from 1Q-NN-WFs as well as neural wave-functions working in second quantization, or 2Q-NN-WFs, as we will refer to them.

# 2. A primer on conventional quantum chemistry

> I think I can safely say that nobody understands quantum mechanics.
>
> *Richard Feynman*

We will in the following provide a synopsis of the quintessential concepts and jargon of conventional quantum chemistry methods that calculate a molecule's ground state wavefunction as a means to determine the ground state energy. Whereas 1Q-NN-WFs deviate signicifantly from these conventional approaches, 2Q-NN-WFs do not, making it indispensable to be familiar with the basic notions.

## 2.1. The molecular Hamiltonian

The non-relativistic time-independent Hamiltonian describing molecules in a quantum mechanical framework encompasses single-particle energies as well as pairwise electromagnetic interaction between electrons and atoms respectively. It is given by

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{U}_{en} + \hat{U}_{ee} + \hat{U}_{nn} \tag{2.1}$$

where

$$\hat{T}_n = -\sum_a \frac{\hbar}{2M_a} \nabla^2_{\mathbf{R}_a} \tag{2.2}$$

captures the kinetic energy of each atomic nucleus,

$$\hat{T}_e = -\sum_i \frac{\hbar}{2m_e} \nabla^2_{\mathbf{r}_i} \tag{2.3}$$

captures the kinetic energy of each electron

$$\hat{U}_{en} = -\sum_{a,i} \frac{Z_a e^2}{4\pi\varepsilon_0 |\mathbf{R}_a - \mathbf{r}_i|} \tag{2.4}$$

captures the pairwise electromagnetic interaction between each atom and electron,

$$\hat{U}_{ee} = \sum_{i<j} \frac{e^2}{4\pi\varepsilon_0 |\mathbf{r}_i - \mathbf{r}_j|} \tag{2.5}$$

captures the pairwise electromagnetic interactions among the electrons, and

$$\hat{U}_{nn} = \sum_{a_1<a_2} \frac{Z_{a_1} Z_{a_2}}{4\pi\varepsilon_0 |\mathbf{R}_{a_1} - \mathbf{R}_{a_2}|} \tag{2.6}$$

captures the pairwise electromagnetic interactions among the atoms[15]. The many-body interactions involved necessitate the introduction of approximations to be made in order for any plausible solution to be computationally attainable. A particularly popular approximation is the classic approximation of Born and Oppenheimer which is motivated by the starkly different velocity of the atoms and electrons[16]. Since electrons move on a timescale that is orders of magnitude smaller than the one of the much heavier nuclei, the atoms can be considered static for any time frame of observed electronic motion. This not only annihilates the atomic kinetic energies in the original Hamiltonian, 2.2, but furthermore reduces the atomic interactions energies to mere constants, allowing to remove them from the Hamiltonian. Employing atomic units (i.e. $\hbar = m_e = 4\pi\varepsilon_0 = e = 1$) then yields

$$\hat{H}_{el} = -\sum_i \nabla^2_{\mathbf{r}_i} - \sum_{a,i} \frac{Z_a}{|\mathbf{R}_a - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.7}$$

**Potential energy surface:** The stationary nature of the atomic position within the electronic structure problem implies that any spatial configuration of a molecule, i.e. a specimen of all $3M$ spatial coordinates $\{\mathbf{R}_a\}_{a=1}^{M}$ for a molecule consisting of $M$ atoms, fully defines the electronic Hamiltonian $\hat{H}_{el}$. Associating for any the given molecular spatial configuration the lowest eigenvalue of its corresponding electronic Hamiltonian, i.e. the ground state energy, defines a map $\mathbf{E} : \mathbb{R}^{3M} \to \mathbb{R}$. This map is known as the potential energy surface.

## 2.2. A mathematical toolkit

We now turn to the essential quantum chemical methods developed over the last decades aiming to solve the molecular Hamiltonian. The description of these techniques detailed here draws rests in large part on the seminal introductory book on the topic by Szabo and Ostlund [15]. After introducing some basic notation and terminology, we present a synopsis of common quantum chemical procedures using the low-level Hartree-Fock theory as stepstone for more accurate post-Hartree-Fock methods.

**Atomic orbitals as basic building blocks:** The electronic wave-function of a molecule's ground state is the eigenfunction of the electronic Hamiltonian 2.2 with minimal corresponding eigenvalue. Since solving the Hamiltonian analytically is intractable due to the many-body interactions, it is commonplace to reduce the search space to a vector space constructed by a finite set of basis functions. The size of this set is determined by how one positions oneself on the tradeoff of accuracy vs. computational cost, as a bigger set of basis functions enables higher accuracy at the price of increased computational cost. The basis functions itself are referred to as atomic orbitals. An atomic orbital $\phi_\mu$ is a function

$$\phi_\mu \colon \mathbb{R}^3 \to \mathbb{R} \tag{2.8}$$

$$\mathbf{r} \mapsto \phi_\mu(\mathbf{r}) \tag{2.9}$$

The set of atom orbitals are in general not orthonormal. The set of pairwise scalar products define the overlap matrix

$$S_{\mu\nu} := \langle \phi_\mu, \phi_\nu \rangle = \int d^3\mathbf{r}\, \phi_\mu(\mathbf{r})\phi_\nu(\mathbf{r}) \tag{2.10}$$

**From atomic orbitals to molecular wave-functions:** The guiding principle underlying the composition of an N-electron wave-function $\Psi(r_1, ..., r_N) \equiv \Psi(\vec{r})$ with $\vec{r} = (r_1, ..., r_N)$ is to construct it as a product of N atomic orbitals. To account for fermionic antisymmetry, only properly antisymmetrized linear combinations of such atomic orbitals constitute a legitimate wave-function. A particularly convenient and almost exclusively utilized manner of antisymmetrization is to construct Slater-determinants of atomic orbitals as

$$\Psi(\vec{r}) = \langle \vec{r} | \Psi \rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{r}_1) & \dots & \phi_N(\mathbf{r}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{r}_N) & \dots & \phi_N(\mathbf{r}_N) \end{vmatrix} = \frac{1}{\sqrt{N!}} \det[\phi_{\mu i}] \equiv \frac{1}{\sqrt{N!}} \det[\boldsymbol{\phi}] \tag{2.11}$$

$$\phi_{\mu i} \equiv \phi_\mu(\mathbf{r}_i) \tag{2.12}$$

Equivalently, we can substitute the atomic orbitals in the Slater-determinant by linear combinations of atomic orbitals. We refer to these linear combinations of atomic orbitals as molecular orbitals $\chi_i$ defined by

$$\chi_i = \sum_\nu C_{\nu i} \phi_\nu \tag{2.13}$$

with a matrix of molecular orbital coefficients $\boldsymbol{C}$. It is usually desirable to choose $\boldsymbol{C}$ in a way that the resulting molecular orbitals $\chi_i$ are orthonormal, i.e.

$$\langle \chi_i, \chi_j \rangle = \delta_{ij} \tag{2.14}$$

We now turn to integrating electronic spin into our formalism. We introduce spin orbitals as the products of spatial molecular orbitals and the spin function $w$, defined by

$$\chi_i \rightarrow \chi_i^{\uparrow}, \chi_i^{\downarrow} \tag{2.15}$$
$$\chi_i^{\alpha} \rightarrow \chi_i w(\alpha) \tag{2.16}$$

modifying the orthonormality relations to

$$\langle \chi_i^{\alpha}, \chi_j^{\beta} \rangle = \delta_{ij}\delta_{\alpha\beta} \tag{2.17}$$

Each spatial orbital can accommodate two electrons with spins up and down respectively. We will in the following refer to $\chi_i$ as spin orbitals and assume the spin-related double counting to be absorbed into the enumerating index $i$.

## 2.3. Hartree-Fock Theory

Hartree-Fock theory assumes the wave-function to take the form of a single determinant constituted of orthonormal molecular orbitals as in eq. 2.11. The wave-function corresponding to the physical ground state, $|\Psi_0\rangle$ is the one minimizing the energy, i.e.

$$|\Psi_0\rangle = \arg\min_{\Psi} \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle} = \arg\min_{\Psi} E[\Psi] \tag{2.18}$$

Assuming the trial wave-function $\Psi$ to be normalized, $E[\Psi]$ can be expressed as

$$E[\Psi] = \langle \Psi | \hat{H} | \Psi \rangle = \sum_i \langle \chi_i | \mathrm{h}_i | \chi_i \rangle + \frac{1}{2} \sum_{mn} [\chi_m \chi_m | \chi_n \chi_n] - [\chi_m \chi_n | \chi_n \chi_m] \tag{2.19}$$

where the one-particle term $\mathrm{h}_i$

$$\mathrm{h}_i \equiv \mathrm{h}(\mathbf{r}_i) = -\nabla_{\mathbf{r}_i} - \sum_a \frac{1}{|\mathbf{R}_a - \mathbf{r}_i|} \tag{2.20}$$

captures each of the single-electron energy contributions, and terms of the form

$$[\chi_i\chi_j|\chi_k\chi_l] = \iint d^3\mathbf{r}_1 d^3\mathbf{r}_2 \frac{\chi_i(\mathbf{r}_1)\chi_j(\mathbf{r}_1)\chi_k(\mathbf{r}_2)\chi_l(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{2.21}$$

capture the electron-electron Coulomb interaction. Minimizing the functional $E[\Psi]$ via variational methods yields a set of descriptive integro-differential equations for the molecular orbitals $\chi_\mu$ that form $|\Psi_0\rangle$:

$$\mathrm{h}_i\chi_\mu(\mathbf{r}_i) + \sum_{\nu\neq\mu}\left[\int d^3\mathbf{r}_2\frac{\chi_\nu(\mathbf{r}_2)^2}{|\mathbf{r}_1-\mathbf{r}_2|}\right]\chi_\mu(\mathbf{r}_i) - \sum_{\nu\neq\mu}\left[\int d^3\mathbf{r}_2\frac{\chi_\nu(\mathbf{r}_2)\chi_\mu(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}\right]\chi_\nu(\mathbf{r}_1) = \varepsilon_\mu\chi_\mu(\mathbf{r}_1) \tag{2.22}$$

with associated orbital-energies $\varepsilon_\mu$. Introducing the Coulomb operator

$$\mathbf{J}_\mu(\mathbf{r}_i) = \int d^3\mathbf{r}_2\frac{\chi_\mu^2(\mathbf{r}_i)}{|\mathbf{r}_i-\mathbf{r}_2|} \tag{2.23}$$

as well as the exchange operator

$$\mathbf{K}_\mu(\mathbf{r}_i)\chi_\mu(\mathbf{r}_i) = \left[\int d^3\mathbf{r}_2\frac{\chi_\mu(\mathbf{r}_2)\chi_\nu(\mathbf{r}_2)}{|\mathbf{r}_i-\mathbf{r}_2|}\right]\chi_\mu(\mathbf{r}_i) \tag{2.24}$$

which, contrasting the Coulomb operator, can only be expressed via its action on the molecular orbitals, and the Fock operator

$$\mathbf{f}(\mathbf{r}_i) = \mathrm{h}(\mathbf{r}_i) + \sum_{\nu\neq\mu}\mathbf{J}_\nu(\mathbf{r}_i) - \mathbf{K}_\nu(\mathbf{r}_i) = \mathrm{h}(\mathbf{r}_i) + \sum_\nu \mathbf{J}_\nu(\mathbf{r}_i) - \mathbf{K}_\nu(\mathbf{r}_i) \tag{2.25}$$

where the second equality follows from

$$\left[\mathbf{J}_\mu(\mathbf{r}_i) - \mathbf{K}_\mu(\mathbf{r}_i)\right]\chi_\mu(\mathbf{r}_i) = 0 \tag{2.26}$$

allows to compact eq. 2.22 as

$$\mathbf{f}(\mathbf{r}_i)\chi_\mu(\mathbf{r}_i) = \varepsilon_\mu\chi_\mu(\mathbf{r}_i) \tag{2.27}$$

Expressing the spin orbitals $\chi_j$ in the set of atomic orbitals $phi_\mu$ as

$$\chi_j = \sum_\mu C_{\mu j}\phi_\mu \tag{2.28}$$

and introducing the Fock matrix

$$F_{\mu\nu} = \int d^3\mathbf{r}\, \phi_\mu(\mathbf{r}_i)\mathbf{f}(\mathbf{r}_i)\phi_\nu(\mathbf{r}_i) \tag{2.29}$$

allows to express eq. 2.27 in the form of the Roothaan equations

$$\mathbf{FC} = \mathbf{SC}\varepsilon \tag{2.30}$$

with the overlap integral **S** defined as in 2.10 and the diagonal orbital energy matrix $\boldsymbol{\varepsilon} = \mathrm{diag}(\varepsilon_1, ..., \varepsilon_N)$.

The iterative procedure employed to solve 2.30, commonly referred to as the self-consistent field method or SCF in short, yields molecular coefficients ***C*** which fully specify the Hartree-Fock wave-function. As is obvious from eq. 2.27, the set of resulting orbitals $\chi_j$ form the set of eigenvectors of the Fock matrix. While the hermicity of the Fock matrix enforces the orbitals to be orthogonal, they are further normalized to be pairwise orthonormal, i.e.

$$\int d^3\mathbf{r}\chi_i(\mathbf{r})\chi_j(\mathbf{r}) = \delta_{ij} \tag{2.31}$$

This orthonormality relation implies together with the definitions of **C** and **S**

$$\mathbf{C}^\dagger\mathbf{SC} = \mathbb{1} \tag{2.32}$$

In a molecule containing $N$ electrons, the $N$ orbitals $\chi_1, ..., \chi_N$ with the lowest associated orbital energy $\varepsilon_1, ..., \varepsilon_N$ form the Hartree-Fock wave-function. The occupied and non-occupied orbitals are referred to as active and virtual orbitals respectively.

**Accuracy of Hartree-Fock theory:** Hartree-Fock theory often yields qualitatively correct within a computational runtime that is deemed acceptable, being in the order of $\mathcal{O}(N^4)$, with $N$ being the number of the molecule's electrons, assuming a naive implementation [17]. While the energy of the Hartree-Fock wave-function recovers around 99% of the actual energy, this is far from sufficient for quantum chemical applications. Furthermore, a host of rather banal chemical phenomena are not predicted to an even qualitatively correct degree by the HF method, such as the behaviour of strongly correlated systems [18] or the bond breaking of a dissociating molecule [19]. Hartree-Fock theory is howeverinsdispensable as stepstone towards more accurate quantum chemistry, which are tellingly referred to as post-Hartree-Fock methods.

An insight on the limitations of Hartree-Fock theory can be gleaned by scrutinizing the Fock operator

$$\mathbf{f}(\mathbf{r}_i) = \mathrm{h}(\mathbf{r}_i) + \sum_{\nu} \mathbf{J}_\nu(\mathbf{r}_i) - \mathbf{K}_\nu(\mathbf{r}_i) \tag{2.33}$$

where pairwise electron-electron interactions are encoded in the Coulomb and exchange operators. They are formed as sums over all of the orbitals, hence representing a mean-field potential for the orbital $\chi_i$ and thus taking account electronic correlation only in a crude fashion.

## 2.4. On the nature of electronic correlation

Electronic correlation arises due to the antisymmetry principle as well as the pairwise Coulomb interaction between the electrons. The correlation resulting from the antisymmetry principle is referred to as Fermi correlation and is recovered by Slater-determinantal wave-functions such as the Hartree-Fock wave-function. The remaining correlation, which is not recovered by a single-Slater-determinantal wave-function, can be divided into static and dynamic correlation. Static correlation arises from near-degeneracy of the ground-state, while dynamic correlation results from electronic movement [20]. Resolution of both types of correlation demands to consider linear combinations of Slater-determinants rather than just a single Slater-determinant. This idea forms the backbone of Configuration Interaction (CI) and Coupled Cluster (CC) methods, to which we will return in full swing at a later stage.

**Electronic antisymmetry and cusp conditions:** Electronic-electronic interactions imprint some prominent features on the structure of $\Psi(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$, an especially prominent one being the antisymmetry principle. Another feature are sharp peaks of $\Psi(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$ at certain locations. One set of such locations are those where two electrons coalesce. The existence of such sharp peaks can be understood from inspecting the electronic Hamiltonian eq. 2.7 which contains terms of the form $\frac{1}{\|\boldsymbol{r}_i - \boldsymbol{r}_j\|}$. These terms diverge for $\boldsymbol{r}_i \to \boldsymbol{r}_i$, necessitating counterbalancing diverging contributions from the kinetic terms corresponding to $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. Diverging kinetic terms imply diverging gradients of $\Psi$ with the respect to $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, resulting in infinitely sharp cusps of $\Psi$. While folklore has it that these electronic cusps pose a significant obstacle in correct modelling of the wave-function, some counterarguments have been raised, see [21]. Regardless, attempts have been undertaken to introduce modifying factors to a Slater-determinantal wave-function which explicitly incorporate the electronic-electronic cusp conditions. A particularly celebrated example is the introduction of the called Jastrow-factor, first introduced by Robert Jastrow in 1955 [22]. Given a fully antisymmetric wave-function $\Psi(\mathbf{r}_1, ..., \mathbf{r}_N)_S$, e.g. a single Slater-determinant or a sum thereof, the Slater-Jastrow wave-function is given by

$$\Psi(\mathbf{r}_1, ..., \mathbf{r}_N)_{SJ} = \exp[J(\mathbf{r}_1, ..., \mathbf{r}_N; \mathbf{R}_1, ..., \mathbf{R}_M)]\Psi(\mathbf{r}_1, ..., \mathbf{r}_N)_S \tag{2.34}$$

with the Jastrow-factor $J$ being a fully symmetric function with respect to the electronic coordinates $\mathbf{r}_i$. The functional form of $J$ is chosen in a way such that it enforces the electronic cusps[23]. The inclusion of such a deceivingly simple Jastrow-factor allows to retrieve a surprising amount of the electronic correlation lacking in the Hartree-Fock wave-function [24]

A second set of locations introducing cusps into the wave-function are those where the position of one electron and nuclues coalesce. The reasoning is identical to the reasoning for electronic cusps, however with the culprit now being terms of the form $\frac{1}{\|\boldsymbol{r}_i - \boldsymbol{R}_m\|}$ with some atomic position $\boldsymbol{R}_m$. Such terms demand cusps for the wave-function whenever $\boldsymbol{r}_i = \boldsymbol{R}_m$ for some electron $i$ and some atom $m$. These nuclear cusps are arranged for by choosing the atomic orbitals to implement the atomic cusps automatically. We will address how this is achieved in the next section alongside more general considerations regarding the nature of the atomic orbitals.

## 2.5. Choosing an optimal basis set

No specifications were introduced so far regarding the exact form of the atomic and molecular spin orbitals $\phi(\mathbf{r})$ and $\chi(\mathbf{r})$ respectively. The fundamental dilemma lying at the heart of committing to any concrete set of atomic orbitals consists of leveraging the trade-off between expressiveness and computational complexity, both straightforwardly depending on the number of basis functions. The standard route is to choose chemically-informed orbital shapes and to choose the number of included basis functions depending on the required accuracy of the task at hand [25]. A panoply of different bases have been developed, many of them suiting subtly different purposes [26]. We review here the fundamental building blocks of these basis sets.

The most common atomic orbital basis functions are of the form

$$\phi_{nlm,j}(\mathbf{r}; \boldsymbol{R}_j) = \mathrm{R}_{nl}(\|\boldsymbol{R}_j - \boldsymbol{r}\|)\mathrm{Y}_{lm}(\theta, \phi) \tag{2.35}$$

being comprised of the radial component $\mathrm{R}_{nl}$ and an angular component $\mathrm{Y}_{lm}$, usually chosen to be a spherical harmonic. The triple index $nlm$ enumerates the atomic orbitals per atom by the quantum numbers $n = s, p, d, ..., l = 0, ..., n-1$ and $m = -l, ..., l-1, l$. The angles $\theta$ and $\phi$ are defined such that

$$\boldsymbol{R}_j - \boldsymbol{r} = \begin{pmatrix} \sin(\theta)\cos(\phi)\|\boldsymbol{R}_j - \boldsymbol{r}\| \\ \sin(\theta)\sin(\phi)\|\boldsymbol{R}_j - \boldsymbol{r}\| \\ \cos(\theta)(\|\boldsymbol{R}_j - \boldsymbol{r}\|) \end{pmatrix} \tag{2.36}$$

Slater-type orbitals (STOs) as first introduced by John Slater in 1930 [27] determine the radial part up to normalization to be of the form

$$R_{l;\mathrm{STO}}(r) \propto r^l e^{-\alpha r} \tag{2.37}$$

with $r = \|\mathbf{r}\|$ and $\alpha$ being some exponent of choice. STOs became fashionable due to their innate correct modelling of the nuclear cusp conditions[28]. Nonetheless, STOs are usually not the function of choice due to the high cost of evaluating one- and two electron integrals [25]. Instead, Gaussian functions usually constitute the radial part

$$R_{l;\mathrm{GTO}}(r) \propto r^l e^{-\alpha r^2} \tag{2.38}$$

thus forming so-called Gaussian-type orbitals (GTOs). While this necessitates a larger set of basis functions compared to the case of employing STOs, this is far outweighed by the vastly reduced computational effort of evaluating integrals of products of GTOs [25]. In order to still at least approximately fulfill the cusp condition, basis functions are built from linear combinations from GTOs as

$$\phi_{nlm,j}(\mathbf{r}) = r^l \left( \sum_s c_s e^{-\alpha_s r^2} \right) Y_{lm}(\theta, \phi) \tag{2.39}$$

with as sharp a peak at the nucleus's position as possible.

The basis set of a molecule is the union of the atomic orbital sets for all constituting atoms. In order for any basis set to at least qualitatively be able to describe the wave-function's behaviour correctly, it must contain at least as many orbitals as electrons that are present. Basis sets including one orbital per electron are termed minimal basis sets, such as the STO-nG basis sets, where n indicates the number of GTOs per orbital. Empirically it hols true though that, in order to capture electronic correlation adequately, one needs to include way more basis functions at least per valence electron, yielding basis sets such as the cc-pVDZ basis with 14 basis functions for each atom with more than two electrons [26].

Before we now turn to covering CI and CC methods, a short primer on the terminology of second quantization is in order, as both methods are most conveniently formulated in the framework of second quantization.

## 2.6. Second quantization in quantum chemistry

Because of the ubiquitous nature of the formalism of second quantization, we will restrict ourselves to shortly introduce the common notation and conventions as employed in this work. The conventions are adopted from Szabo and Ostlund's classic [15]. An in-depth description of the use of second quantization in quantum chemistry, which we in part also lean on, can be found in [29].

Given a set of orthonormal spin orbitals $\{\chi_\mu\}_\mu$ we assign a creation operator $\hat{c}^\dagger_\mu$ and annihilation operator $\hat{c}_\mu$ to each spin orbital, equipped with the usual fermionic anticommutation relations

$$\begin{aligned} \left\{\hat{c}^\dagger_\mu, \hat{c}^\dagger_\nu\right\} &= \{\hat{c}_\mu, \hat{c}_\nu\} = 0 \\ \left\{\hat{c}^\dagger_\mu, \hat{c}_\nu\right\} &= \delta_{\mu\nu} \end{aligned} \tag{2.40}$$

For a Slater-determinant $|\Psi\rangle$ consisting of $N$ spin-orbitals we introduce the notation

$$|\Psi\rangle = |\chi_{\mu_1}\chi_{\mu_2}...\chi_{\mu_N}\rangle \tag{2.41}$$

and define the action of the creation and annihilation operators on such states as

$$\begin{aligned} \hat{c}^\dagger_\nu |\chi_\alpha\chi_\beta...\chi_\omega\rangle &= |\chi_\nu\chi_\alpha\chi_\beta...\chi_\omega\rangle \\ \hat{c}_\nu |\chi_\nu\chi_\alpha\chi_\beta...\chi_\omega\rangle &= |\chi_\alpha\chi_\beta...\chi_\omega\rangle \\ \hat{c}^\dagger_\nu |0\rangle &= |\chi_\nu\rangle \\ \hat{c}_\nu |0\rangle &= 0 \\ (\hat{c}^\dagger_\nu)^\dagger &= \hat{c}_\nu \end{aligned} \tag{2.42}$$

where we furthermore introduced the vacuum state $|0\rangle$. In a slight overload of notation, we will refer to kets such as those on the right side of the above equation as Slater-determinants. The ordering of the orbitals within any Slater-determinants therefore enshrines, via the ordering of the creation operators corresponding to that Slater-determinant, the wave-functions' antisymmetry.

**Second-quantized wave-functions and operators:** We can now connect the dots of first and second quantization by defining a mapping between the spin orbitals employed so far of the form $\chi_\mu(\mathbf{r})$ and the respective kets $|\chi_\mu\rangle$ by

$$\chi_\mu(\mathbf{r}) = \langle\mathbf{r}|\chi_\mu\rangle \tag{2.43}$$

We identify a generic ket $|\chi_1\chi_2...\chi_N\rangle$ with a N-electron Slater-determinant consisting of spin orbitals $\chi_1, \chi_2, ..., \chi_N$, i.e.

$$|\chi_1\chi_2...\chi_N\rangle = \frac{1}{\sqrt{N!}}\det[\chi_{\mu i}] \tag{2.44}$$

as was hinted at in the terminology above and in eq. 2.11.

Establishing an analogous mapping between operators expressed in first and second quantization respectively enables to conduct calculations in the handy framework of second quantization. Such a mapping is achieved by operators of the form

$$\hat{H}^{(n)} = \sum_{\mu_{\vec{a}},\mu_{\vec{b}}} h^{(n)}_{\vec{a},\vec{b}} \, \hat{c}^{\dagger}_{a_n} \hat{c}^{\dagger}_{a_{n-1}} ... \hat{c}^{\dagger}_{a_1} \hat{c}_{b_1} ... \hat{c}_{b_{n-1}} \hat{c}_{b_n} \tag{2.45}$$

Since the electronic Hamiltonian exclusively contains terms including up to two electrons interacting, it suffices to consider operators of the above form for $n = 1$ and $n = 2$. The electronic Hamiltonian can thus be expressed as

$$\hat{H} = \hat{H}^{(1)} + \hat{H}^{(2)} \tag{2.46}$$

with the one-body and two-body operators being defined via

$$\begin{aligned} h^{(1)}_{\mu\nu} &= \int d^3\mathbf{r} \, \chi^*_\mu(\mathbf{r}) \left( -\frac{1}{2}\nabla_\mathbf{r} - \sum_m \frac{Z_m}{|\mathbf{r} - \mathbf{R}_m|} \right) \chi_\nu(\mathbf{r}) \\ h^{(2)}_{\mu\nu\alpha\beta} &= \iint d^3\mathbf{r}_1 d^3\mathbf{r}_2 \, \chi^*_\mu(\mathbf{r}_1)\chi_\nu(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \chi^*_\alpha(\mathbf{r}_2)\chi_\beta(\mathbf{r}_2) \end{aligned} \tag{2.47}$$

The energy's expectation value of any (normalized) wave-function $|\Psi\rangle$ is thus given by

$$\begin{aligned} \langle\Psi|\hat{H}|\Psi\rangle &= \sum_{\mu\nu} h^{(1)}_{\mu\nu} \langle\Psi|\hat{c}^\dagger_\mu \hat{c}_\nu|\Psi\rangle + \sum_{\mu\nu\alpha\beta} h^{(2)}_{\mu\nu\alpha\beta} \langle\Psi|\hat{c}^\dagger_\mu \hat{c}^\dagger_\nu \hat{c}_\beta \hat{c}_\alpha|\Psi\rangle \\ &= \sum_{\mu\nu} h^{(1)}_{\mu\nu}\Gamma^{(1)}_{\mu\nu} + \frac{1}{2}\sum_{\mu\nu\alpha\beta} h^{(2)}_{\mu\nu\alpha\beta}\Gamma^{(2)}_{\mu\nu\alpha\beta} \end{aligned} \tag{2.48}$$

where we introduced the 1- and 2-reduced density matrices (1- and 2-RDMs) respectively as

$$\begin{aligned} \Gamma^{(1)}_{\mu\nu} &= \langle\Psi| \, \hat{c}^\dagger_\mu \hat{c}_\nu \, |\Psi\rangle \\ \Gamma^{(2)}_{\mu\nu\alpha\beta} &= \langle\Psi| \, \hat{c}^\dagger_\mu \hat{c}^\dagger_\nu \hat{c}_\beta \hat{c}_\alpha \, |\Psi\rangle \end{aligned} \tag{2.49}$$

## 2.7. Configuration Interaction theory

The main philosophy behind Configuration interaction (CI) theory is to construct the N-electron-wave-function not as a single Slater determinant, but as a linear combination of Slater-determinants, each comprised of a different set of $N$ spin orbitals. CI theory chooses the single-determinantal HF wave-function

$$|\Psi_{\mathrm{HF}}\rangle = |\chi_1\chi_2...\chi_N\rangle = \hat{c}_N^\dagger...\hat{c}_2^\dagger\hat{c}_1^\dagger\,|0\rangle \tag{2.50}$$

as starting point and postulates a more accurate wave-function as

$$\begin{aligned} |\Psi_{\mathrm{CI}}\rangle &= \lambda_0\,|\Psi_{\mathrm{HF}}\rangle + \sum_{\substack{\mathrm{a}\in\mathcal{A}\\ \mathrm{r}\in\mathcal{V}}} \lambda_a^r\,|\Psi_a^r\rangle + \sum_{\substack{\mathrm{a,\,b}\in\mathcal{A}\\ \mathrm{r,\,s}\in\mathcal{V}}} \lambda_{ab}^{rs}\,|\Psi_{ab}^{rs}\rangle + ... \\ &\equiv |\Psi_{CI}(\boldsymbol{\lambda})\rangle \end{aligned} \tag{2.51}$$

where we defined $\mathcal{A}$ and $\mathcal{V}$ as the active and virtual space of spin orbitals respectively. In this case, the active space is defined as the set of spin orbitals comprising the Hartree-Fock wave-function. Kets of the form $|\Psi_a^r\rangle$ resemble singly-excited determinants, terms of the form $|\Psi_{ab}^{rs}\rangle$ resemble doubly-excited determinants and so forth. Put precisely,

$$\begin{aligned} |\Psi_a^r\rangle &= |\chi_1...\chi_{a-1}\chi_r\chi_{a+1}...\chi_N\rangle = \hat{c}_N^\dagger...\hat{c}_{a+1}^\dagger\hat{c}_r^\dagger\hat{c}_{a-1}^\dagger...\hat{c}_1^\dagger\,|0\rangle \\ &= \hat{c}_r^\dagger\hat{c}_a\,|\Psi_{HF}\rangle \\ |\Psi_{ab}^{rs}\rangle &= \hat{c}_r^\dagger\hat{c}_s^\dagger\hat{c}_b\hat{c}_a\,|\Psi_{HF}\rangle \end{aligned} \tag{2.52}$$

The set of parameters $\boldsymbol{\lambda}^*$ corresponding the wave-function with minimal energy can thus be determined via the standard Rayleigh-Ritz functional

$$\boldsymbol{\lambda}^* := \arg\min_{\lambda} \frac{\langle\Psi_{CI}(\boldsymbol{\lambda})|\,\hat{H}\,|\Psi_{CI}(\boldsymbol{\lambda})\rangle}{\langle\Psi_{CI}(\boldsymbol{\lambda})|\Psi_{CI}(\boldsymbol{\lambda})\rangle} \tag{2.53}$$

**CI wave-functions in binary notation:** We will additionally introduce a second notation for CI wave-functions that will come in handy in later stages of this work. For a ket of the form $|\chi_{i_1}\chi_{i_2}...\chi_{i_N}\rangle$ we define a canonical ordering of the spin via their respective orbital energies as defined by the Fock matrix in eq. 2.27. Assuming now $|\chi_{i_1}\chi_{i_2}...\chi_{i_N}\rangle$ to be such a canonically ordered ket with index set $I = \{i_1, i_2, ..., i_N\}$, we then identify

$$|\boldsymbol{x}\rangle := |\chi_{i_1}\chi_{i_2}...\chi_{i_N}\rangle \tag{2.54}$$

$$\boldsymbol{x} = (x_1, x_2, ..., x_{N_{\mathrm{orb}}}) \text{ with } x_i = \mathbb{1}_I(x_i) \tag{2.55}$$

with the boolean indicator function $\mathbb{1}_A(x)$ indicating the presence of $x$ in $A$. For a molecule with ten spin orbitals and four electrons, the Hartree-Fock wave-function would correspond to $\boldsymbol{x}_{HF} = (1,1,1,1,0,0,0,0,0,0)$ whereas a single excited state would e.g. correspond to $\boldsymbol{x}_{SE} = (1,1,1,0,1,0,0,0,0,0)$. We can therefore compactly denote the full CI wave-function as

$$|\Psi\rangle = \sum_{\boldsymbol{x}} \psi(\boldsymbol{x})\,|\boldsymbol{x}\rangle = \sum_{\boldsymbol{x}} \psi_{\boldsymbol{x}}\,|\boldsymbol{x}\rangle \tag{2.56}$$

Note that the coefficients $\psi(\boldsymbol{x})$ do not inhabit any antisymmetry relations, as the antisymmetry of $|\Psi\rangle$ is fully captured in the basis states $|\boldsymbol{x}\rangle$.

**Computation of CI wave-functions**: The full CI-optimized wave-function will yield the closest possible approximation to the actual ground state wave-function within the space spanned by the orbital basis set, rendering CI-calculations invaluable for benchmarking other approximative methods [15]. Unfortunately, as the number of determinants constituting the full CI wave-functions grows factorially with the number of orbitals emplyoed, conducting full CI calculations is prohibitive for most systems of interest and only possible for small molecules and small basis sets [30]. Recently, the largest-ever full CI calculation was carried out, calculating the ground state CI wave-function of propane in the STO-3G basis encompassing 1.3 trillion determinants, distributed over 256 servers [31]. The computational burden can be alleviated significantly by, instead of treating all determinants on equal footing, restricting to a subset of presumably predominant determinants, an approach commonly referred to as selected CI. Studies employing such techniques have blossomed in the recent past, see e.g. [32, 33]. In a similar vein, Monte Carlo procedures, where relevant determinants are sampled stochastically, are applied to an increasing extent, yielding highly encouraging results, see e.g. [34, 35]. Furthermore, machine-learning tools have been employed in aiding to select the relevant determinants, either in a supervised fashion [36] or in an unsupervised manner via reinforcement learning [37].

## 2.8. Coupled Cluster theory

Coupled cluster theory does not introduce a separate parameter for each excited Slater-determinant, but rather parametrizes the excitation process itself. Concretely, it postulates the wave-function to be of the form

$$|\Psi\rangle_{CC} = \exp(\hat{T})\,|\Psi\rangle_{HF} \tag{2.57}$$

with

$$\hat{T} = \sum_i \hat{T}_i \tag{2.58}$$

$$\hat{T}_1 = \sum_{\substack{\mathrm{a}\in\mathcal{A}\\ \mathrm{r}\in\mathcal{V}}} t_{ar}\hat{c}_r^\dagger \hat{c}_a \tag{2.59}$$

$$\hat{T}_2 = \sum_{\substack{\mathrm{a,b}\in\mathcal{A}\\ \mathrm{r,s}\in\mathcal{V}}} t_{abrs}\hat{c}_r^\dagger \hat{c}_s^\dagger \hat{c}_a \hat{c}_b \tag{2.60}$$

and $\hat{T}_3, \hat{T}_4, ...$ defined equivalently. The parameters $t_{ar}, t_{abrs}, ...$ are referred to as the CC-amplitudes. Note the drastic difference between the CC wave-function and its CI counterpart, which we could identically denote as $|\Psi\rangle_{CI} = \hat{T}\,|\Psi\rangle_{HF}$. Whereas we could set all $\hat{T}_i$ for $i > 2$ to zero in the CC case and still obtain a wave-function containing excitations of up to N electrons due to the exponential coupling excitations of all orders, the CI wave-function would be restricted to determinants with up to two excitations. Such a restriction scheme is in fact often done in practice, and its compatibility with CC is one of the main reasons that CC is often referred to as the gold-standard of quantum chemistry [38]. It has furthermore been refined to a variety of flavours and for a host of use cases [39]. However, the great strength of CC theory is in fact Janus-faced. As the CC wave-function contains excitations of order up to N, evaluating the energy of the CC wave-function in the same fashion as the CI wave-function in eq. 2.53 is not possible in polynomial time. Thus, one has to resort to other computational schemes in order to determine the optimal set of CC-amplitudes. These schemes do not however guarantee that the obtained CC wave-function corresponds to one actually attainable by the molecule and might yield an energy that is lower than the ground state energy. In short, CC is not variational. CC shares this shortcoming with other post-Hartree-Fock methods such as Møller–Plesset perturbation theory [40] we will therefore not cover in much detail.

## 2.9. Dividing and conquering the electronic correlation: local correlation

Even when introducing significant simplifications such as trimming the CC approach to only contain singly and doubly excited determinants, expensive scaling still hinders application for large molecules or large basis sets. A particularly promising facet of further improvement is to exploit the local nature of electronic correlation. Despite this character being recognized for many decades [41], investigations on how to best capitalize on that locality have been rather scarce and are an active field of research [42].

**Localized orbitals**: One key primary insight on how to operationalize locality is that spin orbitals, being as a linear combination of atomic orbitals in general delocalized, can be transformed into locally concentrated and equally expressive spin orbitals. In such a localized scheme, a preponderance of pairwise orbital interactions is negligible, implying favourable scaling [43]. Localized spin orbital can be generated by making use of a certain degree of freedom that Hartree-Fock theory presents us with. Namely, applying a unitary rotation $O$ within the spin orbital space of the form

$$\chi'_{nu} = \sum_{\nu\mu} \boldsymbol{O}_{\nu\mu}\chi_{\mu} \tag{2.61}$$

with $|\det(\boldsymbol{O})| = 1$ leaves a determinantal wave-function up to an unphysical phase factor invariant, as

$$\boldsymbol{O}\,|\Psi\rangle = \frac{1}{\sqrt{N!}}\mathrm{det}[\mathbf{O}\chi] = \frac{1}{\sqrt{N!}}\mathrm{det}[\mathbf{O}]\mathrm{det}[\chi] = \frac{(-1)^k}{\sqrt{N!}}\mathrm{det}[\chi] \tag{2.62}$$

We can now employ different strategies to obtain orbitals with as little spatial span as possible [44]. These localized orbitals allow best to capture essentially local phenomena and in many instances decrease computational efforts significantly [45]. A series of different localization schemes have been developed. Notable examples are the Foster-Boys [46] and the Pipek-Mezey [47] algorithm, both of which define loss functions quantifying the amount of spatial extension. A different category of localized orbitals consists of diagonalizing density matrices (which will be introduced in a later section) and yield Natural Bond Orbitals, which are especially well suited to capture electronic correlation [48].

## 2.10. Quo vadis, quantum chemistry?

Herculean efforts to develop approaches of tackling the electronic structure problem notwithstanding, the list of unsolved problems remains long. At the heart of them all resides the unfavorable scaling up to large systems and to large basis sets enabling sufficient accuracy. It is the essence of quantum mechanics, which, returning to Richard Feynman, who opened this section, are probably understood by nobody. But if nobody, or no body, understands quantum mechanics, a machine may do just fine?

# 3. Quantum chemistry in the age of artificial intelligence

> With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.
>
> *John von Neumann*

With astonishing breakthroughs and achievements being heralded almost in a daily manner, the advent of AI evermore proves to be truly paradigm-shifting to the world of sciences and society at large. The seeming omnipotence of AI rests not in the least on the guarantee that certain mathematical architectures such as multilayer feedforward network can be twisted in ways to approximate arbitrary functions on Euclidean spaces to any desired level of accuracy [5].

The empirical success earning AI its laurels has motivated many forays into the use of machine learning techniques in the context of quantum chemistry. A majority of applications attempted to predict molecular properties based on training data sets obtained via CC, DFT or similar methods with the goal of achieving comparable success with significantly decreased computational effort. Notable examples include the identification of possible novel molecules [49], molecular dynamics [50], electron densities [51] and many more. The archetypal undertaking in this regard however is the prediction of a molecule's energy based on its Hamiltonian alone, or equivalently on the atomic coordinates alone as they fully specify the Hamiltonian. An accurate such model would in effect bypass any laborious ab-initio calculations and simply output the ground state energy, thus allowing to straightforwardly model a molecule's potential energy surface. This can in principle be achieved via one of two routes: either one trains the neural net in a supervised manner on a pre-computed set of reference data, which itself is obtained via costly but accurate techniques, such as CCSD(T) or advanced DFT methods. Alternatively, one instead parameterized the electronic wave-function via the neural net and minimizes its energy expectation value directly without reliance on any external data. The former approaches we will subsume under the category neural-network-potential-energy-surfaces (NN-PESs) and the latter one we will subsume under the category neural-network-wave-functions. We will provide short synopses on the pros and cons of both approaches in the following. Based on the understanding we gain of advantages and disadvantages of certain NN-WFs we will afterwards propose the architecture of NN-WF ourselves.

## 3.1. AI for PES: the wiggling elephant in the room?

NN-PESs calculate the energy of a molecular geometry via a black-box functional, receiving, in most cases, only the atomic positions as input. Paradigmatic specimens of neural nets belonging to this class, such as GemNet [52] or NequIP [53], consist of sophisticated neural architectures with parameter amounts in the millions, which almost exclusively rely on the geometric information contained in the spatial configuration of the atoms of a molecule. Hence, only a modicum of physical domain knowledge is implemented. Such neural nets are trained on representative datasets such as the MD-17 dataset [54]. The MD-17 dataset contains the energies and forces of samples of molecular dynamics simulations of eight different organic molecules, with the reference values being obtained via DFT. State-of-the-art NN-PESs easily reach an accuracy well within the regime of the so-called chemical accuracy of less than 1 kcal/mol [55], or 1.6 Millihartree, utilizing on the order of magnitude of one million parameters [52].

On the flip side, the appealing simplicity of NN-PESs is in fact Janus-faced: while easing implementation and leaving much leeway for the neural net to learn optimal representations and functionals itself, the very reductionist model appears to hinder an effective learning of the underlying physics. Perhaps unsurprisingly, extrapolation of NN-PESs to inputs outside the training domain generally yields poor results [56], possibly suggesting a high-resolution interpolation within the training domain rather than supreme modelling of the relevant physical laws. Or, in the words of John von Neumann, the elephant is taught to wiggle his trunk. Much of the promise of NN-PESs rests however on a faithful generalization to unseen molecules, as their accuracy is bottlenecked by the reference DFT calculations, and hence replacing DFT techniques by NN-PESs could only be warranted by a gain in computational complexity.

Various endeavors have aimed at increasing the generalizability of NN-PESs by furthermore including more domain knowledge. Examples for such endeavors include SpookyNet [57], including information on the molecule's total charge and spin state, UNiTE [58], which predicts the energy based upon cheap low-level electronic structure calculations similar to Hartree-Fock or DeepHF [59], which predicts the energy difference between a Hartree-Fock calculation and an expensive post-Hartree-Fock method based upon the Hartree-Fock orbitals. Other approaches tweak the neural architecture with encouraging results, such as the Allegro model [60].

The aforementioned non-exhaustive list of state-of-the-art NN-PESs is characterised by the common paradigm of utilizing an essentially black-box functional determining the energy.

Since this field of inquiry is still in its infancy, it is to be expected that many of the downsides of current approaches will be mitigated in due time and NN-PESs will contend with conventional quantum chemistry techniques for relevant applications. The immanent bottleneck of high-quality reference data can in principle however not be overcome. We thus turn to a somewhat orthogonal line of research focusing on parametrizing molecular wave-functions via neural networks as intermediaries for obtaining the molecular energy. It is such an approach we will pursue ourselves in the later stages of this work.

## 3.2. AI for wave-functions: of artificial and human intelligence

Directly modelling the wave-function relieves the necessity for external data. It is thus an avenue where the accuracy of what the artificial intelligence is able to achieve is not bottlenecked by what human intelligence was able to achieve beforehand. Approaches aiming at parametrizing physically plausible wave-functions necessitate a much more explicit and deliberate construction of the neural architecture to ensure that physical constraints such as the antisymmetry with respect to electron exchange are fulfilled. We start this section by outlining some key models put forward over the last years. Based on the state-of-the art, we will then derive some desiderata for a NN-WF explicitly tailored to maximize the capacity of the NN-WF for "insight" into the underlying physics. We do this in the hopes of leveraging the enormous potential of pattern recognition that neural nets have showcased in many application areas.

We can divide the totality of neural wave-functions into those operating in the framework of first quantization, 1Q-NN-WFs, and those working within second quantization, 2Q-NN-WFs.

## 3.3. Neural wave-functions in first quantization

Neural wave-functions in first quantization construct wave-functions in their spatial representation $\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = \Psi(\vec{\mathbf{r}})$. The particular wave-function of interest is the ground state wave-function, the eigenfunction associated with the minimal eigenvalue of the electronic Hamiltonian's

$$\hat{H} = -\sum_i \nabla^2_{\mathbf{r}_i} - \sum_{m,i} \frac{Z_m}{|\mathbf{R}_m - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{3.1}$$

First-quantized neural wave-functions rests on a series of pillars we will illuminate now.

**Jastrow-Slater wave-function:** Pioneering neural wave-functions such as FermiNet [7] and PauliNet [8] represent the wave-function as

$$\Psi(\vec{\mathbf{r}};\vec{\mathbf{R}},\boldsymbol{\theta}) = e^{J(\vec{\mathbf{r}};\vec{\mathbf{R}},\boldsymbol{\theta})} \sum_{k=1}^{N_{\text{det}}} c_k \begin{vmatrix} \phi_1^k(\mathbf{r}_1;\vec{\mathbf{r}},\vec{\mathbf{R}},\boldsymbol{\theta}) & \dots & \phi_N^k(\mathbf{r}_1;\vec{\mathbf{r}},\vec{\mathbf{R}},\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \phi_1^k(\mathbf{r}_N;\vec{\mathbf{r}},\vec{\mathbf{R}},\boldsymbol{\theta}) & \dots & \phi_N^k(\mathbf{r}_N;\vec{\mathbf{r}},\vec{\mathbf{R}},\boldsymbol{\theta}) \end{vmatrix} = e^{J(\vec{\mathbf{r}};\vec{\mathbf{R}},\boldsymbol{\theta})} \sum_{k=1}^{N_{\text{det}}} c_k \det[\boldsymbol{\Phi}_k] \tag{3.2}$$

with $\vec{\mathbf{r}} = (\mathbf{r}_1, ..., \mathbf{r}_N)$ denoting the N electrons' positions, $\vec{\mathbf{R}} = (\mathbf{R}_1, ..., \mathbf{R}_N)$ denoting the set of atomic coordinates, which are parameters of the function rather than variables, $\boldsymbol{\theta}$ a set of trainable parameters, a fixed number of determinants $N_{\text{det}}$ and $\phi_n^k(\mathbf{r}_m;\vec{\mathbf{r}},\vec{\mathbf{R}},\boldsymbol{\theta})$ being backflow-generalized molecular orbitals, which we elaborate upon in the following section. For the sake of completeness, we point out the spin-explicit wave-function of FermiNet and PauliNet to be of the form

$$\Psi(\vec{\mathbf{r}}^{\uparrow},\vec{\mathbf{r}}^{\downarrow};\vec{\mathbf{R}},\boldsymbol{\theta}) = e^{J(\vec{\mathbf{r}}^{\uparrow},\vec{\mathbf{r}}^{\downarrow};\vec{\mathbf{R}},\boldsymbol{\theta})} \sum_{k=1}^{N_{\text{det}}} c_k \det[\boldsymbol{\Phi}_k^{\uparrow}] \det[\boldsymbol{\Phi}_k^{\downarrow}] \tag{3.3}$$

with $\boldsymbol{\Phi}_k^{\uparrow}, \boldsymbol{\Phi}_k^{\downarrow}$ being determinants akin to the one employed in eq. 3.2, but depending solely on the positions of spin-up electrons and spin-down electrons respectively. The resulting wave-function is thus only antisymmetric with respect to exchange of electrons of differing spin. Nevertheless, the such constructed wave-function yields correct expectation values for spin-independent observables [7]. Utilizing the determinant's properties for block diagonal matrices, we can immediately reconcile eq. 3.3 with the generic wave-function eq. 3.2 by defining

$$\boldsymbol{\Phi}_k = \begin{pmatrix} \boldsymbol{\Phi}^{\uparrow} & 0 \\ 0 & \boldsymbol{\Phi}^{\downarrow} \end{pmatrix} \tag{3.4}$$

and hence $\det[\Phi] = \det[\Phi_k^{\uparrow}] \det[\Phi_k^{\downarrow}]$. The spin-explicit wave-function in eq. 3.3 this constitutes the special case of the generic wave-function eq. 3.2 with the off-diagonal orbitals being identically equal to zero. We will in the following only employ spin-explicit notation if it is necessary for understanding and omit it otherwise.

**Neural backflow orbitals:** The concept of backflow was originally introduced by Feynman [61] in order to describe the purely quantum mechanical effect of diametrically opposed probability current and momentum of a wave-function. This effect can be modelled by postulating fictitious electronic positions [62]

$$\tilde{\boldsymbol{r}}_m = \boldsymbol{r}_m + \sum_{i \neq m} \eta_{im}(\boldsymbol{r}_i - \boldsymbol{r}_m) \tag{3.5}$$

with free parameters $\eta_{im}$ and by evaluating the orbitals accordingly as $\phi_n(\tilde{\mathbf{r}}_m) \rightarrow \phi_n(\boldsymbol{r}_m^b)$. The slight variation of that theme consisting of backflow-adapted orbitals instead

$$\tilde{\phi}_n(\mathbf{r}_m; \vec{\mathbf{r}}) = \phi_n(\mathbf{r}_m) + \sum_{i \neq m} \eta_{im} \phi_n(\mathbf{r}_i) \tag{3.6}$$

has been exploited extensively as basic building block for the electronic wave-function [63] and can thus accurately be seen as antecedents to 1Q-NN-WFs of the form eq. 3.2. Luo and Clark [64] recognized the concomitant optimization of an overall wave-function in the form of eq. 3.2 and the backflow-optimization of the basis orbitals $\phi_n$ to be reminiscent of the layered structure of neural nets and introduced generalized backflow-orbitals

$$\tilde{\phi}_n(\mathbf{r}_m; \vec{\mathbf{r}}) = \phi_n(\mathbf{r}_m) + f_n(\mathbf{r}_m; \vec{\mathbf{r}}; \boldsymbol{\theta}) \tag{3.7}$$

with $f_n$ being a neural net with trainable parameters $\boldsymbol{\theta}$. Furthermore introducing multiplicative corrective terms yields the most generic form of backflow-orbitals as

$$\tilde{\phi}_n(\mathbf{r}_m; \vec{\mathbf{r}}) = \phi_n(\mathbf{r}_m) f_n^{\otimes}(\mathbf{r}_m; \vec{\mathbf{r}}; \boldsymbol{\theta}) + f_n^{\oplus}(\mathbf{r}_m; \vec{\mathbf{r}}; \boldsymbol{\theta}) \tag{3.8}$$

where we drop the explicit parameter-like dependence of the orbitals on the atomic positions. Note that for reasons of readability we omitted the tildes in the determinant in eq. 3.2. Such generalized molecular orbitals do not compromise the wave-function's antisymmetry due to the employment of determinants in eq. 3.2. Furthermore, a single determinant consisting of such modified orbitals is vastly more expressive compared to conventional one-electron-orbital Slater-determinants due to the freedom granted by the corrective terms [7]. In fact, any antisymmetric N-electron wave-function can in principle be represented by a single Slater-determinant comprised of such backflow orbitals [65]. A popular choice for backflow orbitals is exemplified by FermiNet's orbitals

$$\phi_i^k(\mathbf{r}_m; \vec{\mathbf{r}}) = (\mathbf{w}_i^k \cdot \mathbf{h}_m^L(\vec{\mathbf{r}}) + g_i^k) \sum_j \pi_{ij}^k \exp(-\|\boldsymbol{\Sigma}_{ij}^k(\mathbf{r}_m - \mathbf{R}_j)\|) \tag{3.9}$$

with trainable parameters $\mathbf{w}_i^k \in \mathbb{R}^F, g_i^k \in \mathbb{R}$ and $\boldsymbol{\Sigma}_{ij}^k \in \mathbb{R}^{3\times 3}$ and a learnable function $\mathbf{h}_j^L(\vec{\mathbf{r}}) \in \mathbb{R}^F$. PauliNet on the other hand relies on pre-computed multi-reference-HF one-electron orbitals which are subjected to backflow modification in the training process.

**Parameter optimization:** The parameters are optimized in a variational manner via refinement of gradient descent methods [66] such as Kronecker Factorized Approximate Curvature (KFAC) [67] in FermiNet or stochastic gradient descent in PauliNet. The loss function is the Hamiltonian itself [68]

$$E[\Psi_\theta] = \frac{\int d^3\vec{\mathbf{r}}\,\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})\hat{H}\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})}{|\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})|^2} \tag{3.10}$$

Costly integral evaluations can be sufficiently substituted by stochastic variational Monte Carlo (VMC) methods, as heavily used in the study of correlated systems [69]. VMC reformulates the above energy expectation value as

$$E[\Psi_\theta] = \int d^3\vec{\mathbf{r}}\,\frac{|\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})|^2}{\int d^3\vec{\mathbf{r}}|\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})|^2}\frac{\hat{H}\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})}{\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})} = \mathbb{E}_{\vec{\mathbf{r}}\sim|\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})|^2}\left[E_{\text{loc}}(\vec{\mathbf{r}})\right] \tag{3.11}$$

where $E_{\text{loc}}(\vec{\mathbf{r}}) = \frac{\hat{H}\Psi(\vec{\mathbf{r}};\vec{\boldsymbol{R}},\boldsymbol{\theta})}{\Psi(\vec{\mathbf{r}};\vec{\boldsymbol{R}},,\boldsymbol{\theta}})$. Drawing samples from the distribution $p(\vec{\mathbf{r}}) \propto |\Psi(\vec{\mathbf{r}};\vec{\boldsymbol{R}},\boldsymbol{\theta})|^2$ allows to stochastically evaluate the expectation value to a satisfactory degree of accuracy. A common strategy of obtaining samples is making use of Monte Carlo Markov chains such as the Metropolis-Hastings algorithm [68]. The gradient guiding the parameter's update is then given by

$$\nabla_{\boldsymbol{\theta}}E[\Psi_\theta] = 2\mathbb{E}_{\vec{\mathbf{r}}\sim|\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})|^2}\left[\left(E_{\text{loc}}(\vec{\mathbf{r}}) - \mathbb{E}_{\vec{\mathbf{r}}\sim|\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})|^2}[E_{\text{loc}}(\vec{\mathbf{r}})]\right)\nabla_{\boldsymbol{\theta}}\Psi(\vec{\mathbf{r}};\boldsymbol{\theta})\right] \tag{3.12}$$

**Computational complexity:** The vanilla versions of FermiNet and PauliNet, while quite similar in spirit, offer differing trade-offs between accuracy and complexity. While FermiNet is able to reach somewhat lower energies - details about accuracy and results will be discussed in the following - it does so at the expense of utilizing many more training parameters and thus necessitating many more parameters. The theoretical scaling of $\mathcal{O}(N^4)$ with $N$ being the number of electrons, which is very moderate comparing to costly conventional quantum chemistry approaches such as CCSD(T) with a scaling of $\mathcal{O}(N^7)$, is in practice overshadowed by an enormous prefactor [56]. An architecturally slimmed down version of FermiNet still demanded five times the computational cost of vanilla PauliNet [70]. Whereas PauliNet, albeit confined to simple molecular systems such as $H_2$, LiH, B, Be, $H_{10}$, could be trained within time spans ranging from tens of minutes to a couple of hours, training FermiNet took much longer, up to a staggering 1104 hours of GPU for the admittedly much complexer molecule cyclobutadiene [70]. Digesting how PauliNet achieved highly accurate results within a much narrower parameter space spurred much subsequent, and in fact still ongoing, research. It was hypothesized that the heavily physics-inspired orbital construction of PauliNet consisting of additively corrected Hartree-Fock orbitals, or additively and multiplicatively corrected Hartree-Fock orbitals in a generalization of PauliNet [71], constitutes a very significant inductive bias contrasting the orbitals of FermiNet, which are to a large degree relearned from scratch. On the other hand, some experimental data contradicts this hypothesis, as will be discussed in the following section.

**Ground state energy accuracy:** Vanilla PauliNet and FermiNet, while being limited to small molecules, were able to showcase an impressive degree of accuracy when calculating the ground state energy. Various attempts have been undertaken to improve upon the architecture of FermiNet and PauliNet to boost the accuracy even further. Gerard et al. [12] synthesized the architectures of PauliNet and FermiNet yielding more accurate results, in some instances providing the most accurate results ever recorded.

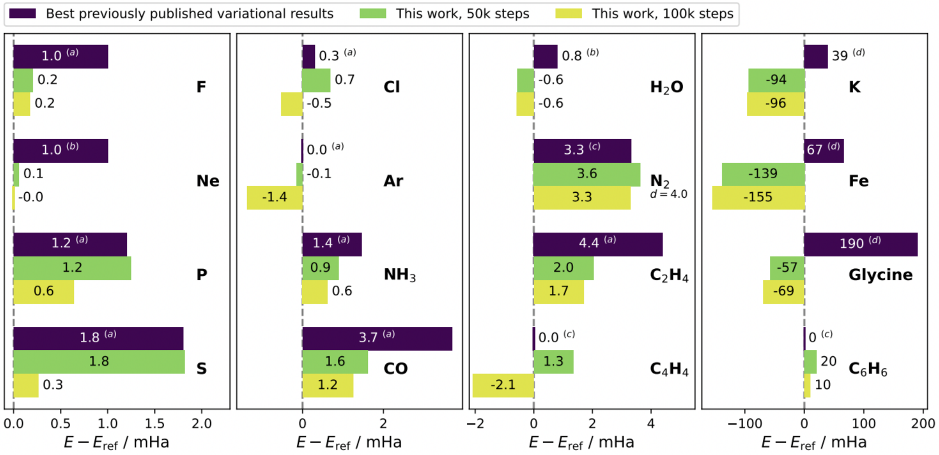

Figure 3.1.: The results from [12]. Reference energies were calculated with a non-variational conventional quantum chemistry approach, thereby potentially underestimating the true ground state energy.

Lots of effort has been invested to fine-tune the Monte Carlo simulation component. Variants such as Diffusion Monte Carlo [72] have been put to use with FermiNet-type wave-functions improving upon initial results [73, 74], though not outperforming Gerard et al. [12].

Other approaches tweaked the neural architecture to a more extensive degree. von Glehn et al. [75] introduced the seemingly omnipotent transformer architecture [76] to replace convolutional layers from the FermiNet architecture, enabling accuracies on par or better than the one displayed in fig. 3.1.

One common feature of variational Monte Carlo methods such as the 1Q-NN-WFs is to allow the treatment of certain systems, where conventional quantum chemistry approaches fail to even yield qualitatively correct results. Systems out of equilibrium geometries serve as prime examples, since CC methods often yield highly dubious results [7]. A paradigmatic model system small enough to be tractable for a FCI calculation is given by a four-hydrogen rectangle parametrized by the hydrogen atoms' distance to the center of the mass as well as the spanned angle. Vanilla FermiNet was able to in essence match the FCI result for the hydrogen rectangle.

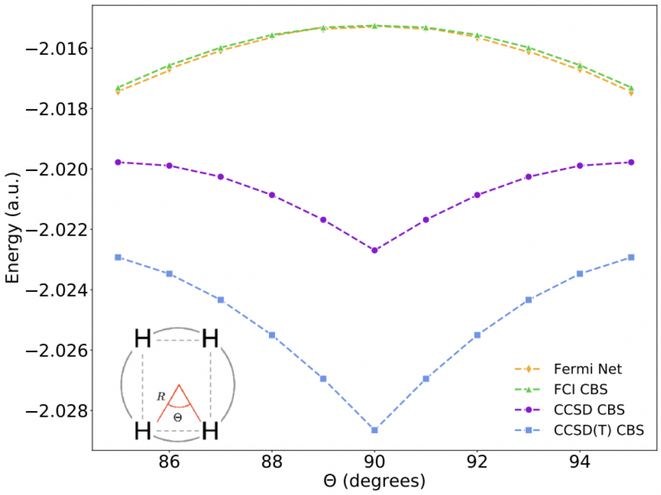

Figure 3.2.: The results from FermiNet for the hydrogen rectangle as taken from [7]. FermiNet achieves results on par with a FCI calculation in the complete basis set limit, which consists of FCI calculations in some bases extrapolated via some extrapolation formula.

**Potential energy surfaces with 1Q-NN-WFs:** The neural wave-functions discussed above share the drawback that they have to be retrained for every molecular geometry anew, and a fortiori for every different molecule anew. This not only prohibits a large-scale implementation of 1Q-NN-WFs to model potential energy surfaces, but also raises questions how much of the accuracy of the 1Q-NN-WFs are due to fitting to a given atomic geometry rather than awareness of universal features of electronic correlation [17]. Attempts to generalize the architecture of FermiNet and PauliNet to be able to model arbitrary electronic wave-functions have been put forward.

Scherbela et al. [10] developed a trial wave-function very similar in spirit to PauliNet, where within the training process it is enforced that, while training the model for different spatial geometries of the same molecule, a preponderance of learnable weights are shared within the models for different geometries. This empirically not only showed to speed up the training process for different spatial arrangements of the same molecule by an order of magnitude while retaining or improving accuracy, but furthermore suggests that general characteristics of electronic correlation, which are physically highly similar between different geometries, are being captured by the model.

A different technique to realize the joint training of multiple atomic geometries was developed by Gao et al. [56] by partnering a neural wave-function such as FermiNet for a given spatial configuration with another neural net which reparametrizes the wave-function model to the spatial geometry at hand.

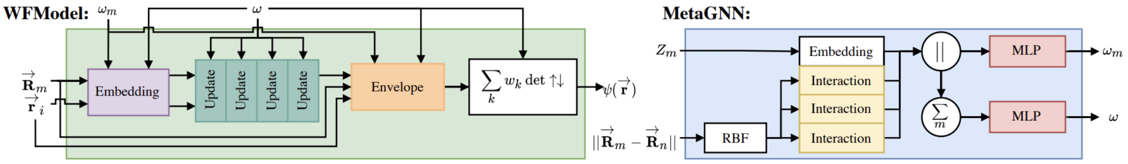

Figure 3.3.: The neural architecture in [56]. The network MetaGNN on the right calculates parameters of the neural wave-function from the atomic positions only.

In a subsequent publication, the authors extended this technique to the case of treating different molecules within the same model [11]. As the way this is realized depends crucially on the neural architecture chosen, which we introduce shortly, we will elaborate upon the philosophy behind [11] at at later stage. Recently, Scherbela at el. [14] proposed a modularized version of the FermiNet architecture, which calculates atom-wise parameters characterising the atomic orbitals and thereby allows to compose bigger molecules by combining pre-learned atomic features and conduct some additional learning on the whole structure for the purpose of fine-tuning.

## 3.4. Neural wave-functions in second quantization

Neural wave-functions in second quantization (2Q-NN-WFs) consider a second-quantized Hamiltonian of the form

$$\hat{H} = \sum_{ij} h_{ij}\hat{c}_i^\dagger \hat{c}_j + \sum_{ijkl} h_{ijkl}\hat{c}_i^\dagger \hat{c}_j^\dagger \hat{c}_l \hat{c}_k \tag{3.13}$$

2Q-NN-WFs thus model wave-functions in the same Hilbert space of Slater-determinants as e.g. CI techniques do. Current 2Q-NN-WFs are restricted to the case of $S = 0$-wave-functions, i.e. wave-funcitons with an equal number of spin up and spin down electrons. The quintessential inspiration for 2Q-NN-WFs is the mathematically substantiated hope that highly accurate CI-wave-functions can be determined in polynomial time despite the exponential size of the parameter space [77]. The aim with neural nets is then to model the exponentially sized parameter space accurately with a limited amount of parameters.

Any N-electron wave-function can be expressed via a basis set composed of $N_{orb}$ spin-orbitals as

$$|\Psi_{\boldsymbol{\theta}}\rangle = \sum_{\boldsymbol{x}} \psi(\boldsymbol{x};\boldsymbol{\theta})\,|x_1, ..., x_{N_{\text{orb}}}\rangle \tag{3.14}$$

$$= \sum_{\boldsymbol{x}} \psi(\mathbf{x};\boldsymbol{\theta})\,|\boldsymbol{x}\rangle \tag{3.15}$$

with $\mathbf{x} = (x_1, ..., x_{N_{orb}})$, $x_j \in \{0, 1\}$, $\sum_i^{N_{orb}} x_i = N < N_{orb}$ and coefficients $\psi(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{x}|\psi_{\boldsymbol{\theta}}\rangle$. As the antisymmetry is contained in the kets $|x\rangle$ th coefficients $\psi(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{x}|\psi_{\boldsymbol{\theta}}\rangle$ do no need to fulfill according constraints. 2Q-NN-WFs thus naturally embed the electronic antisymmetry and eliminate the necessity to explicitly incorporate cusp conditions via a Jastrow-factor by utilizing conventional basis sets. This drops much of the constraints on the neural architecture that 1Q-NN-WFs have to grapple with and allows to achieve high degrees of accuracy with simple architectures, yet they do so at the expense of giving up the flexibility of learnable basis functions [17]. Second-quantized neural wave-functions thus present with a unique portfolio of strengths and weaknesses and therefore, compared to their first-quantization counterparts, rest on fundamentally different building blocks, which we will spotlight now.

**Wave-function parametrization:** The canonical strategy in 2Q-NN-WFs is to calculate the amplitudes $\psi(\boldsymbol{x};\boldsymbol{\theta})$ from eq. 3.14 as the output of the neural net as done in multiple 2Q-NN-WFs [78, 79, 80, 81]. The state-of-the art investigation in [80] utilizes simple multilayer-perceptrons calculating matrices $\phi_{ij}^k$ from occupation strings $\boldsymbol{x}$ as input, as showcased in fig. 3.4

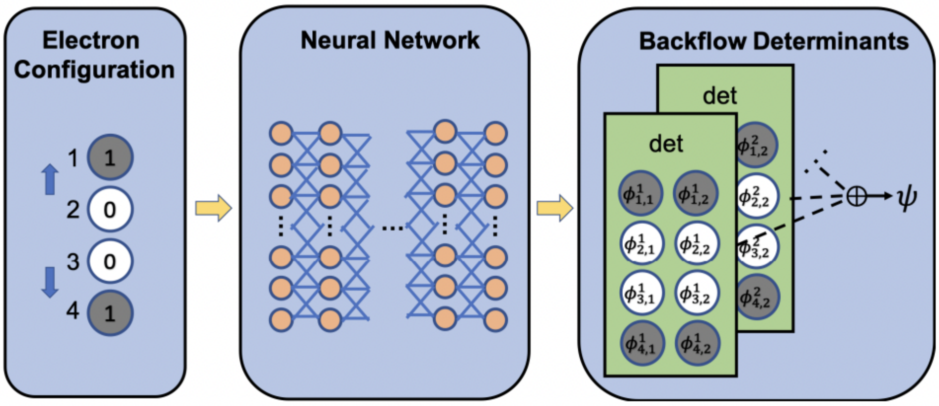

Figure 3.4.: The neural architecture in [80]. The amplitudes are calculated straightforwardly via MLPs which work on the occupation strings as input. The term backflow is used differently in this context comparing to ours and can thus be ignored.

from which the amplitudes are derived as

$$\psi(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{D} \det[\phi^{k}_{\{i=\{l:x_l=1\},j\}}] \tag{3.16}$$

As 2Q-NN-WFs cannot optimize the employed basis set functions, they rely on utilizing a large number of Slater-determinants to capture electronic correlation, as conventional quantum chemistry approaches do [78]. It is a common empirical finding however that the CI-expansion in eq. 3.14 is dominated heavily by a very limited amount of Slater-determinants. This fuels the hope that efficient reparametrizations of the amplitudes $\psi(\boldsymbol{x};\boldsymbol{\theta})$ via $\boldsymbol{\theta}$ exist [77], contrasting a convential FCI-expansion with all the factorially many amplitudes $\psi(\boldsymbol{x}) = \psi_{\boldsymbol{x}}$ being free parameters.

**Mapping orbitals onto spin systems:** Alternatively, as the binary structure of the wave-function in eq. 3.14 insinuates, the electronic problem can alternatively be mapped onto a virtual set of interacting qubits. This is equally achieved by the Jordan-Wigner [82] or Bravyi-Kitaev mapping [83], which can jointly be expressed as

$$\begin{aligned}
\hat{c}_j &\rightarrow \frac{1}{2} \prod_{i\in U(j)} \hat{\sigma}^x_i \times \left( \hat{\sigma}^x_j \prod_{i\in P(j)} \hat{\sigma}^z_i - i\hat{\sigma}^y_j \prod_{i\in R(j)} \hat{\sigma}^z_i \right) \\
\hat{c}^\dagger_j &\rightarrow \frac{1}{2} \prod_{i\in U(j)} \hat{\sigma}^x_i \times \left( \hat{\sigma}^x_j \prod_{i\in P(j)} \hat{\sigma}^z_i + i\hat{\sigma}^y_j \prod_{i\in R(j)} \hat{\sigma}^z_i \right)
\end{aligned} \tag{3.17}$$

with Pauli-matrices $\hat{\sigma}$, an update set of qubits $U(j)$, a parity set of qubits $P(j)$ and a rest set of qubits $R(j)$, which depend on the particular mapping of choice. Choo et al. [9] settle for the popular Jordan-Wigner transformation, where $U(j) = j, P(j) = \{0, 1, ..., j-1\}, R(j) = P(j)$, implicating a mapping of the form

$$\begin{aligned}
\hat{c}_j &\rightarrow \left( \prod_{i=0}^{j-1} \hat{\sigma}^z_i \right) \hat{\sigma}^-_j \\
\hat{c}^\dagger_j &\rightarrow \left( \prod_{i=0}^{j-1} \hat{\sigma}^z_i \right) \hat{\sigma}^+_j
\end{aligned} \tag{3.18}$$

with $\hat{\sigma}^{\mp}_j = \hat{\sigma}^x_j \pm \hat{\sigma}^y_j$. Following the groundbreaking initial foray into neural wave-functions by Carleo et al. [6], Choo et al. parametrize the ensuing wave-function as

$$\psi(\boldsymbol{\theta}) \equiv \Psi_{\boldsymbol{\theta}} = e^{\sum_i a_i \hat{\sigma}^z_i} \prod_{i=1}^{M} 2\cosh\left[ b_i + \sum_j W_{ij} \hat{\sigma}^z_j \right] \tag{3.19}$$

with trainable weights $\boldsymbol{\theta} = \{a_{ij}, b_{ij}, W_{ij}\}$.

**Parameter optimization:** The weights are analogously to neural networks working in first quantization optimized via stochastic gradient-descent-like techniques with the wavefunction's expectation value

$$E_{\boldsymbol{\theta}} = \frac{\langle \Psi_{\boldsymbol{\theta}} | \hat{H} | \Psi_{\boldsymbol{\theta}} \rangle}{\langle \Psi_{\boldsymbol{\theta}} | \Psi_{\boldsymbol{\theta}} \rangle} \tag{3.20}$$

which can analogously to its real-space equivalent eq. 3.10 be evaluated via Monte-Carlo techniques. To that end, Choo et al. [9] rewrite the above expectation value as

$$E_{\boldsymbol{\theta}} = \sum_{\boldsymbol{\sigma}} p_{\boldsymbol{\theta}}(\boldsymbol{\sigma}) E_{\text{loc}}(\boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{\sigma} \sim p_{\boldsymbol{\theta}}(\boldsymbol{\sigma})} [E_{\text{loc}}(\boldsymbol{\sigma})] \tag{3.21}$$

with a vector of qubit configurations $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_M)$, the local energy

$$E_{\text{loc}}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} \frac{\Psi_{\boldsymbol{\theta}(\boldsymbol{\sigma}')}}{\Psi^*_{\boldsymbol{\theta}(\boldsymbol{\sigma})}} \langle \boldsymbol{\sigma}' | \hat{H} | \boldsymbol{\sigma} \rangle \tag{3.22}$$

and a probability distribution

$$p_{\boldsymbol{\theta}}(\boldsymbol{\sigma}) = \frac{|\Psi_{\boldsymbol{\theta}}(\boldsymbol{\sigma})|^2}{\sum_{\boldsymbol{\sigma}} |\Psi_{\boldsymbol{\theta}}(\boldsymbol{\sigma})|^2} \tag{3.23}$$

Note that we employ $\boldsymbol{\sigma}$ for configurations in the qubit space and $\boldsymbol{x}$ for configurations in the original orbital space. Resorting once again to sampling strategies like the Metropolis-Hastings algorithm allows to evaluate the expectation value in eq. 3.21 stochastically. Concretely, sampling from $p_{\boldsymbol{\theta}}(\boldsymbol{\sigma}) \sim |\Psi_{\boldsymbol{\theta}}(\boldsymbol{\sigma})|^2$ is performed via Monte Carlo chains of occupation configurations $\boldsymbol{\sigma}_0 \rightarrow \boldsymbol{\sigma}_1 \rightarrow \boldsymbol{\sigma}_2 \rightarrow ...$ where at each iteration $k$ the transition to a proposed configuration $\boldsymbol{\sigma}_{\text{prop}}$ is accepted with probability

$$P(\boldsymbol{\sigma}_{\text{prop}} = \boldsymbol{\sigma}_{\boldsymbol{k}+1}) = \min\left(1, \left| \frac{\Psi_{\boldsymbol{\theta}}(\boldsymbol{\sigma}_{\text{prop}})}{\Psi_{\boldsymbol{\theta}}(\boldsymbol{\sigma}_{\boldsymbol{k}})} \right|^2 \right) \tag{3.24}$$

The resulting Markov chain is downsampled according to a rate $K$ to obtain the final set of configurations $\{\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_K, \boldsymbol{\sigma}_{2K}, ..., \boldsymbol{\sigma}_{MK}\}$. These samples are used to calculate the expectation value from eq. 3.21.

**Alternatives to MCMC sampling**: As this MCMC-based optimization scheme proves to be inefficient for molecular CI expansions which are dominated by the Hartree-Fock determinant as well as a couple of excited determinants [80], a series of investigations have aimed at improving this sampling procedure. One line of research has been centered around constructing the amplitudes $\psi(\boldsymbol{x};\boldsymbol{\theta})$ in an autoregressive fashion[78, 84] since the distribution is heavily skewed towards a few configurations, thus resulting for any sampling procedure in the same configurations being drawn quite often. We will however not delve into these autoregressive neural architectures, as we instead follow the concurrent approach of drawing inspiration from selected-CI methods [80, 79]. These approaches circumvent the stochastic MCMC-sampling by introducing, for each step $n$ in the gradient descent procedure, a core space $\mathcal{V}^n$ of orbital configurations $\boldsymbol{x}_i$. The gradient is calculated as $\nabla \tilde{E}_{\boldsymbol{\theta}}$ with

$$\tilde{E}_{\boldsymbol{\theta}} = \sum_{k \in \mathcal{V}} \frac{|\psi(\boldsymbol{x}_k)|^2}{\sum_{i \in \mathcal{V}} |\psi(\boldsymbol{x}_i)|^2} E_{\text{loc}}(\boldsymbol{x}_k) \tag{3.25}$$

being a surrogate for the actual energy $E$. As the set $\mathcal{V}$ is quite limited, this evaluation can be done exactly without the need to invoke any stochastic methods. Ensuring that the gradient calculated via the surrogate quantity $\tilde{E}$ resembles the hypothetical gradient calculated with respect to the actual energy faithfully is achieved via slightly varying techniques in [80] and [79] respectively. In [79], the authors ensure that all relevant configurations are gradually included in the core space $\mathcal{V}$. This is achieved by defining the connected space

$$\mathcal{C}^n = \{\boldsymbol{x} : \exists \tilde{\boldsymbol{x}} \in \mathcal{V} \text{ s.t. } |\langle \boldsymbol{x}|\hat{H}|\tilde{\boldsymbol{x}}\rangle| \geq \varepsilon\} \tag{3.26}$$

with the hyperparameter $\varepsilon$ and governing the core space via the update rule

$$\mathcal{V}^{n+1} = \mathcal{V} \cup \mathcal{C}^n \tag{3.27}$$

as illustrated in figure 3.5

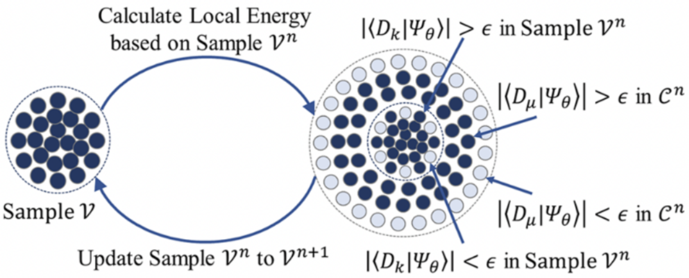
Calculate Local Energy based on Sample $\mathcal{V}^n$
$|\langle D_k|\Psi_\theta\rangle| > \epsilon$ in Sample $\mathcal{V}^n$
$|\langle D_\mu|\Psi_\theta\rangle| > \epsilon$ in $\mathcal{C}^n$
Sample $\mathcal{V}$
$|\langle D_\mu|\Psi_\theta\rangle| < \epsilon$ in $\mathcal{C}^n$
Update Sample $\mathcal{V}^n$ to $\mathcal{V}^{n+1}$
$|\langle D_k|\Psi_\theta\rangle| < \epsilon$ in Sample $\mathcal{V}^n$

Figure 3.5.: The update rule of the core space as taken from [79].

To avoid an overly bloated core space $\mathcal{V}$, the authors in [80] instead use a core space with fixed size $N_u$, where the update rule consists of defining the space $\mathcal{V}^{n+1}$ as the $N_u$ configurations $\boldsymbol{x}_1, ..., \boldsymbol{x}_{N_u}$ with largest amplitudes $|\psi(\boldsymbol{x}_i)|$. Whereas the continually expanded core space $\mathcal{V}$ in [79] eventually encompasses enough configurations in order for $\tilde{E}$ to be a reasonably close approximation to $E$, the same cannot be necessarily said about the fixed-size core space in [80]. In order to calculate the overall energy with a trained model, the authors in [80] thus fall back on a standard MCMC-calculation scheme. This incurs negligible computational cost, as the MCMC-scheme is invoked only once as the capstone of the forward pass, whereas it is invoked for every single step of gradient descent during the training in previous approaches.

Since [80] will serve as a benchmark later, we hereby explicitly introduce it by name as Neural Network Backflow (referred to as NNBF).

**Computational complexity:** 2Q-NN-WFs usually employ orders of magnitude fewer parameters than their first-quantized counterparts and hence naturally offer better computation times. The main computational bottlenecks for a broad amount of 2Q-NN-WFs from the MCMC sampling of the wave-function, which is highly time-consuming due to the stark dominance by the Hartree-Fock determinant within most occurring wave-functions [9]. It is for that precise reason that the class of autoregressive neural wave-functions was developed, leading to vastly reduced sampling efforts [78], reaching CC-benchmarks within a few minutes of training [84]. The few recent selected-CI based methods [80, 84] also are able to circumvent the significant computational cost associated with the wave-function sampling, rendering the computation of the local energy eq. 3.22 with a scaling of $\mathcal{O}(N_O)$, with $N_O$ the total number of a molecule's orbitals, the most burdensome computational task.

**Ground state energy accuracy:** We limit ourselves to a much shorter discussion of the accuracy reported for 1Q-NN-WFs, since 2Q-NN-WFs so far were restricted to minimal basis sets, hence being far from competitive with overall state-of-the-art methods. For these basis sets however, chemical accuracy was achieved with orders of magnitude fewer parameters than necessary for a FCI calculation. As current and future research concentrates on scaling 2Q-NN-WFs to larger bases [80], better comparability of 1Q-NN-WFs and 2Q-NN-WFs can be expected.

## 3.5. Desiderata for a neural wave-function

Equipped with a rough understanding of the current state-of-the-art of neural wave-functions, we now turn to the goal of designing a neural wave-function with maximal capacity for gaining insight into the underlying physics. To that end, we postulate a handful of core tenets that such a neural wave-function architecture needs to possess, and we address to what degree these tenets are fulfilled in current approaches.

**Fulfilling physical constraints:** Obvious constraints that any neural wave-function needs to adhere to are the wave-function's normalizability - i.e. being a square-integrable function - as well as correct behaviour under exchange of two electrons and under nearing of two electrons

- i.a. antisymmetry and cusps. Conventional quantum chemistry approaches employ basis functions to construct molecular orbitals and subsequent wave-functions which adhere to these constraints constraints as discussed in section 2.4. As 2Q-NN-WFs inherit the wave-functions from conventional quantum chemistry approaches, there is no need for further accommodation of physical constraints. The approaches working in first quantization on the other hand need to implement the physicality constraints manually, which they achieve by using suitable Gaussian envelope functions for the molecular orbitals, Slater-determinants in the construction of the molecular wave-function and Jastrow factors for cusps, as described in section 2.4.

**Locality and size-consistency:** As the Coulomb interaction is a local phenomenon, any neural wave-function needs to properly account for this locality. Thus, a proper wave-function of a molecule consisting of two far-apart composites will have to factorize, a property referred to as size-consistency [85]. Recent 1Q-NN-WFs [11] claimed to have successfully constructed size-consistent wave-functions of the form

$$\Psi(\vec{\mathbf{r}};\boldsymbol{\theta}) = e^{J(\vec{\mathbf{r}};\boldsymbol{\theta})} \sum_{k=1}^{N_{\text{det}}} c_k \begin{vmatrix} \phi_1^k(\mathbf{r}_1;\vec{\mathbf{r}};\boldsymbol{\theta}) & \dots & \phi_N^k(\mathbf{r}_1;\vec{\mathbf{r}};\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \phi_1^k(\mathbf{r}_N;\vec{\mathbf{r}};\boldsymbol{\theta}) & \dots & \phi_N^k(\mathbf{r}_N;\vec{\mathbf{r}};\boldsymbol{\theta}) \end{vmatrix} = e^{J(\vec{\mathbf{r}};\boldsymbol{\theta})} \sum_{k=1}^{N_{\text{det}}} c_k \det[\phi_{mn}^k] \tag{3.28}$$

by using localized orbitals $\phi_j^k$, where we have dropped the explicit dependence on the $\vec{\boldsymbol{R}}$ for reasons of readability. They are localized in the sense that, considering two molecules A and B far apart, orbitals localized at A will evaluate to zero when evaluated at positions of electrons of B. It may however be possible that the size-consistency as reported in [11] is rather an artefact of the methodology used in asserting the size-consistency. We will sketch in appendix A.1 why we deem wave-functions of the form of 3.28 not necessarily suited for achieving size-consistency.

The limited amount of NN-WFs working with generic electronic Hamiltonians in second quantization [9, 78, 84, 79, 80] parametrize the wave-function in its occupation number representation

$$|\Psi\rangle = \sum_{x} \psi(\mathbf{x}) |\mathbf{x}\rangle \tag{3.29}$$

While the aforementioned neural nets do not explicitly enforce size-consistency, their ansatz can easily be accommodated to be size-consistent. Consider again the thought experiment of a molecule consisting of two disjoint composites $A_1$ and $A_2$ described by $N_{orb}$ spin orbitals each. We reorder the orbital string $\boldsymbol{x_i}$ such that the entries in position 1 to $N_{orb}$ correspond to the spin orbitals of $A_1$ and the entries in position $N_{orb}+1$ to $2N_{orb}$ correspond to spin orbitals of $A_2$. We compactly denote this as $\boldsymbol{x} = \boldsymbol{x_1 x_2}$. If now the coefficient function $\psi$ is constructed in a manner such that it factorizes as

$$\psi(\boldsymbol{x}) = \psi_1(\boldsymbol{x_1})\psi_2(\boldsymbol{x_2}) \tag{3.30}$$

the overall wave-function of the molecule is given by

$$\begin{aligned} |\Psi\rangle_{A_1A_2} &= \sum_{\boldsymbol{x}} \psi(\mathbf{x})\,|\mathbf{x}\rangle = \sum_{x_1}\sum_{x_2} \psi_1(\boldsymbol{x_2})\psi_2(\boldsymbol{x_2})\,|\boldsymbol{x_1 x_2}\rangle & (3.31)\\ &= \left(\sum_{\boldsymbol{x_1}} \psi(\boldsymbol{x_1})\,|\boldsymbol{x_1}\rangle\right)\left(\sum_{\boldsymbol{x_2}} \psi(\boldsymbol{x_2})\,|\boldsymbol{x_2}\rangle\right) = |\Psi\rangle_{A_1}|\Psi\rangle_{A_2} & (3.32) \end{aligned}$$

and hence is size-consistent [85].

**Recognition of system symmetries:** In lines of research similar to NN-WFs it crystallized early on omitting to explicitly encode symmetries of the input domain induces significant training difficulties and leads to ballooning parameter spaces [13]. This empirical finding was recently substantiated with a theoretical foundation [86, 87, 88]. The nature of the symmetries in the input domain can be easily grasped for the case of NN-PESs, where substantial effort has been invested to encode such symmetries, see [52, 53, 89]. For a NN-PES $\mathcal{E}(\vec{\boldsymbol{R}};\boldsymbol{\theta})$ predicting the energy of an $M$-atom molecule based on the geometric configuration $\vec{\boldsymbol{R}} = (\boldsymbol{R}_1, ..., \boldsymbol{R}_M) \in \mathbb{R}^{3\times M}$ of the $M$ atoms it is obvious that the energy should be invariant with respect to rotations, translations and inversions of the molecular geometry, i.e.

$$\mathcal{E}(\boldsymbol{O}\vec{\boldsymbol{R}} + \boldsymbol{T};\boldsymbol{\theta}) = \mathcal{E}(\vec{\boldsymbol{R}};\boldsymbol{\theta}) \text{ for } \boldsymbol{O} \in O(3),\, \boldsymbol{T} \in \mathbb{R}^{3\times M} \tag{3.33}$$

This does not however translate trivially to wave-functions, as wave-functions are, unlike scalar energies, of geometric character. We clearly do not want a wave-function that is invariant to spatial rotations, as this would in effect yield a spherical wave-function. We thus need a more fanciful notion than simple invariance. We first try to roughly capture our intuition as to what behaviour we would expect. Consider two dislocated copies of the same molecule, one with molecular geometry $\vec{\boldsymbol{R}}_1$ and the other with molecular geometry $\vec{\boldsymbol{R}}_1$. We assume the two composites to be exact replicas of each other, implying that there is some map $\boldsymbol{O} \in \mathbb{R}^{3\times 3}$ combining a translation and a rotation such that $\boldsymbol{O}\vec{\boldsymbol{R}}_1 = \vec{\boldsymbol{R}}_2$. We refer to the respective wave-functions as $\Psi_1(\vec{\boldsymbol{r}}_1;\vec{\boldsymbol{R}}_1)$ and $\Psi_2(\vec{\boldsymbol{r}}_2,;\vec{\boldsymbol{R}}_2)$. Intuition now dictates that evaluating $\Psi_1$ at the electronic positions $\vec{\boldsymbol{r}}_1$ should yield the same values as evaluating $\Psi_2$ at $\vec{\boldsymbol{r}}_2 = \boldsymbol{O}\vec{\boldsymbol{r}}_1$. Put alternatively, were we to realign the two geometries, the two wave-functions should coincide exactly.

While we postpone a more rigorous mathematical treatment to a later stage, suffice it say here that a preponderance of hitherto proposed approaches for NN-WFs in first quantization does not explicitly encode correct transformational behaviour, i.e. were we to calculate the wave-functions of two rotated molecules. Globe [11] and the related approach PESNet [56] circumvent this by defining for each molecular configuration a principal-component-based coordinate frame which itself rotates alongside the molecule and expressing all coordinates in that frame, in effect hiding rotations of the molecule from the neural net. In the latest publication on the matter as of the writing of this manuscript, Scherbela et al. [14] state the incorporation of correct behaviour for rotated inputs as open problem in the field.

NN-WFs working in second quantization are unbothered by considerations regarding spatial symmetry, as they use the second-quantized Hamiltonian as starting point, the scalar elements of which are invariant to rotations of the molecule, cf. section 2.6.

**Transferability to unseen molecules:** Recall our guiding doctrine of aiming to construct a NN-WF with maximal capacity for grasping underlying physical patterns. Such a NN-WF would ideally, once being trained on a small set of representative molecules, be able to predict the wave-function of arbitrary molecules, as they are all governed by the same laws of quantum mechanics. While seeming to be a dauntingly ambitious goal, some NN-WFs in first quantization have indeed achieved some success in employing a model that is transferable to arbitrary and therefore also unseen molecules [56, 14]. The key paradigm employed is to parametrize local constituents of the wave-function and compose the overall wave-function from these local components. We will elucidate the underlying mathematical finesse in the next chapter.

The matter of transferability of wave-functions for NN-WFs in second quantization is dealt with quickly, as this simply was not a concern in hitherto studies and therefore also not addressed. All the studies focused on individual Hamiltonians without the ambition to generalize across different Hamiltonians resulting from different geometries. For example, the recent state-of-the-art 2Q-NN-WF proposed in [80] calculates the amplitudes $\psi(\boldsymbol{x};\boldsymbol{\theta})$ without incorporating features of the Hamiltonian or the molecular geometry in any way. This clearly necessitates the retraining for every novel molecular geometry.

We will now proceed to introduce the mathematical armamentarium necessary to realize the above desiderata.

## 3.6. Enabling locality and transferability: Message-passing graph neural networks

Graph neural networks are neural networks that model functions acting on a graph $G = (\mathcal{V}, \mathcal{W})$ with nodes $V$ and edges $W$ which have enjoyed tremendous success in a variety of domains [90]. A particular subclass of graph neural networks that has proved to be suitable for many problems within quantum chemistry are message-passing neural networks (MPNNs). In their seminal work, Gilmer et al. [91] developed the first use case of message-passing graph neural networks (MMPNs) within the context of quantum chemistry. MPNNs consist of so-called node feature vectors $\{\boldsymbol{f}_v \in \mathbb{R}^F : v \in \mathcal{V}\}$ - equivalently referred to as hidden or latent states - as well as edge feature vectors $\{\boldsymbol{e}_{vw} \in \mathbb{R}^F : \langle v, w \rangle \in \mathcal{W}\}$ are being attributed to each node and edge respectively. Both the feature vectors as well as the edge vectors are learnable. The forward pass of MPNNs consists of two phases, a message passing phase and a readout phase. The messaging phase consists of a finite number $T$ updates of the nodes' and edges' feature vectors, i.e. $\boldsymbol{f}_v \equiv {}^t\boldsymbol{f}_v$ with $t \in \{0, ..., T\}$. The update is governed by the update rule

$$\begin{aligned} {}^{t+1}\boldsymbol{f}_v &= U_t\left({}^t\boldsymbol{f}_v, {}^t\boldsymbol{m}_v\right) \\ {}^t\boldsymbol{m}_v &= \sum_{w \in \mathcal{N}(v)} M_t\left({}^t\boldsymbol{f}_v, {}^t\boldsymbol{f}_w, \boldsymbol{e}_{vw}\right) \end{aligned} \tag{3.34}$$

with node update functions $U_t$, message functions $M_t$ and a node's neighbourhood $\mathcal{N}(v)$. The readout phase calculates the desired output quantity $\hat{y}$ from the feature vectors as

$$\hat{y} = R(\{{}^T\boldsymbol{f}_v : v \in G\}) \tag{3.35}$$

with some readout function $R$.

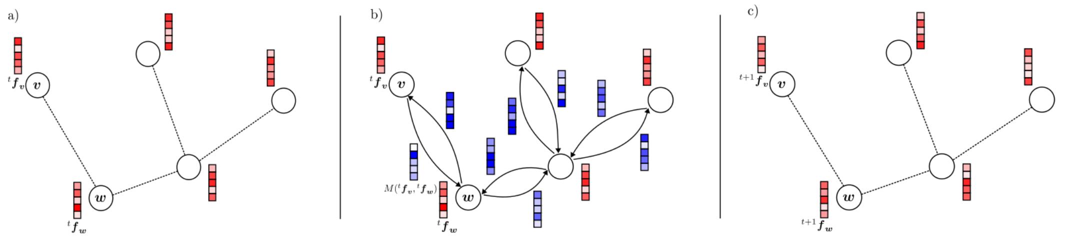

Figure 3.6.: An illustration a) the graph structure, b) message passing and c) the graph with updated feature vectors.

This framework is of particular use for quantum chemical applications, as not in the least indicated by the number of citation of [91], reaching the whooping mark of 8000. The graph structure of MPNNs allows to neatly describe molecules by e.g. identifying the graph's nodes with the molecule's atoms. Still, the architecture leaves enough leeway regarding the exact form of the message, update and readout functions to encompass domain-knowledge or use-case specific constraints. In fact, as Gilmer et al. point out and as still holds true, many of the paradigmatic studies employ neural architectures that can be formulated as MPNNs across the whole of AI-for-quantum-chemistry applications. That includes a preponderance of the past and state-of-the-art NN-PESs, among them SchNet [92], PhysNet [93], GemNet [52] and its predecessor DimeNet(++) [13], UNiTE [58] and NequIP [53], as well as some very recent state-of-the-art NN-WFs [11, 56, 14]. For example, Gao et al. [11] imbue the atoms of a molecule with feature vectors and construct the orbital-wise outputs of their neural architecture by instantiating additional nodes for the orbitals which receive messages from the atoms without sending messages on their own.

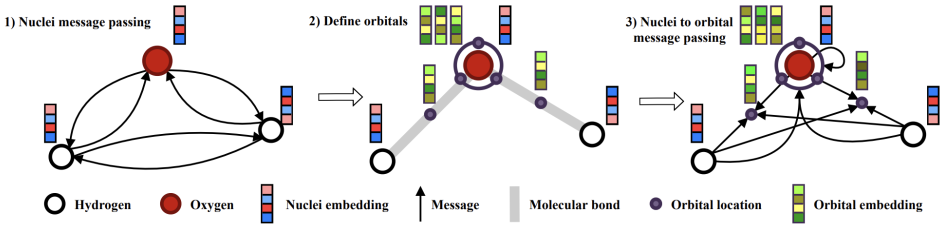

Figure 3.7.: The graph architecture in [11]. Illustrated are the inter-atomic message passing in a), the instantiation of orbital-wise additional nodes in b) and the construction of orbital-wise feature vectors via the surrounding atoms in c).

On the other hand, certain features of the MPNN design have received some scrutiny as well, cf. [60], as the interdependence between atomic features hinders parallel computation and hence scaling to large molecules.

We would like to note that we are aware that to some degree, attention has - literally - shifted towards attention-based transformer architectures, see e.g. [75] or upcoming work from Günnemann et al. (Nicholas Gao, personal communication). We nevertheless settle for a conventional MPNN architecture, as such an architecture could be relatively effortlessly transformed into an attention-based architecture [94] and can thus reasonably be interpreted as a first progenitor towards attention-based neural networks.

## 3.7. Recognizing symmetries: Equivariant neural networks

We now pick up the thread of how to incorporate system symmetries into the architecture of the neural net architecture. Before going abstract, we first want to concretely provide an even more native example for the relevance of encoding symmetries correctly.

Consider the calculation of atomic forces from a given NN-PES $\mathcal{F}_{\boldsymbol{\theta}}$. This use case is in fact far from hypothetical and has been investigated extensively, see e.g. [95]. We can obtain these forces by differentiating the neural net with respect to the atomic positions [96], such that the force acting on atom $a$ is given as

$$\mathbf{F}_a = -\nabla_{\mathbf{R}_a}\mathcal{F}_{\boldsymbol{\theta}}(R) := \mathcal{Q}_{\boldsymbol{\theta}}(R) \tag{3.36}$$

where we introduced the neural net $\mathcal{Q}_{\boldsymbol{\theta}}$ as the gradient of $\mathcal{F}_{\boldsymbol{\theta}}$ with respect to $\mathbf{R}_a$. It is intuitively clear that, if one were to rotate the original molecule and thus the atomic positions by a rotation matrix $O \in SO(3)$, the force should rotate accordingly, as the rotation does not bear any meaningful physical effect and merely constitutes a re-orientation of the coordinate frame. For rotated coordinates $\boldsymbol{R}' = \boldsymbol{O}\mathbf{R}$ we thus expect the final force to be given by $\mathbf{F}'_a = \boldsymbol{O}\mathbf{F}_a$. Formulated in terms of the neural net this amounts to

$$\mathcal{Q}_{\boldsymbol{\theta}}(\boldsymbol{O}\boldsymbol{R}) = \mathcal{Q}_{\boldsymbol{\theta}}(R') = \mathbf{F}'_a = \boldsymbol{O}\mathbf{F}_a = \boldsymbol{O}\mathcal{Q}_{\boldsymbol{\theta}}(\boldsymbol{R}) \tag{3.37}$$

This example also highlights the fact that the correct behaviour is not trivially guaranteed at all. Here, the map $\mathcal{Q}_{\boldsymbol{\theta}}$ needs to commute with any rotation matrix $\boldsymbol{O}$, which of course imposes significant constraints on the anatomy of $\mathcal{Q}_{\boldsymbol{\theta}}$.

Comparing the first and last term of the above chain of equalities allows us then to introduce the idea of so-called equivariance as a generalization of invariance. The neural net $\mathcal{Q}_{\boldsymbol{\theta}}$ is not *invariant* with respect to rotations, but *equivariant* - it changes "in the same manner".

**A mathematical primer on equivariance**: Group theory provides us with the vocabulary to formalise the concept of equivariance [97, 98]. A function $f : \mathcal{X} \to \mathcal{Y}$ with vector spaces $\mathcal{X}$ and $\mathcal{Y}$ is called equivariant with respect to a group $G$ and representations $\mathcal{R}^{\mathcal{X}}$ and $\mathcal{R}^{\mathcal{Y}}$ if for all $g \in G, x \in \mathcal{X}$

$$f(\mathcal{R}^{\mathcal{X}}(g)x) = \mathcal{R}^{\mathcal{Y}}(g)f(x) \tag{3.38}$$

A representation $\mathcal{R}_G$ of a group $G$ is a homomorphism $\mathcal{R}_G : G \to \mathbb{C}^{d\times d}$ for some $d$, thus satisfying

$$\mathcal{R}_G(g)\mathcal{R}_G(h) = \mathcal{R}_G(gh)\forall g,h \in G \tag{3.39}$$

We will restrict ourselves the case $G = SO(3)$ for now. We abuse notation by referring to $SO(3)$-equivariance when we just state equivariance. We make the group explicit in case it is not $SO(3)$. From the above definition it also follows immediately that compositions of equivariant functions are again equivariant.

Representations of the group $SO(3)$ exist for all $d = 2l + 1, l \in \mathbb{N}_0$. For $l = 1$, any rotation matrix $\boldsymbol{O} \in SO(3)$ is straightforwardly represented by itself, as seen in the introductory example above. For a general $l$, it is known that any representation can up to a similarity transformation be expressed as the direct sum of irreducible representations, or irreps. For $SO(3)$, the irreps $\mathcal{R}_{SO(3)}(\boldsymbol{O}) = D^l(\boldsymbol{O}) \in \mathbb{C}^{(2l+1)\times(2l+1)}$ are referred to as the Wigner-D matrices [99]. Any Wigner-D matrix $D^l$ is spanned by the $2l + 1$-dimensional basis set

$\{|lm\rangle : -l \leq m \leq l\}$ as $D^l(\boldsymbol{O}) = \sum_{m,m'} D^l_{m,m'}(\boldsymbol{O}) \,|lm\rangle\langle lm'|$. The spatial expression of the basis states is given by the spherical harmonics

$$\langle \mathbf{r}|lm\rangle = Y_{lm}(\mathbf{r}) \tag{3.40}$$

As said before, every representation of a given rotation matrix $\boldsymbol{O}$ can up to a similarity transformation be expressed as the direct sum of irreducibles

$$S\mathcal{R}_{SO(3)}(\boldsymbol{O})S^{-1} = \bigoplus_{l} \bigoplus_{i=1}^{\tau_l} D^l_i(\boldsymbol{O}) \tag{3.41}$$

with $\mathcal{R}_{SO(3)}(\boldsymbol{O}) \in \mathbb{C}^{(2D+1)\times(2D+1)}$ and multiplicities $\tau_l$ such that $\sum_l \tau_l(2l+1) = 2D+1$. Representations that can be expressed as a non-trivial direct sum of irreducibles are referred to as reducibles. A subclass of reducibles that is of particular relevance to our endeavor are those generated by tensor products of representations. A tensor product of representations $\mathcal{R}_1(\boldsymbol{O}) \otimes \mathcal{R}_2(\boldsymbol{O}) = \mathcal{R}(\boldsymbol{O})$ is again a representation. The transformation between the bases $\{|l_1m_1\rangle \otimes |l_2m_2\rangle = |l_1m_1;l_2m_2\rangle : -l_1 \leq m_1 \leq l_1; -l_2 \leq m_2 \leq l_2\}$ of the representations $\mathcal{R}_1$ and $\mathcal{R}_2$ and the basis $\{|LM\rangle\}$ of $\mathcal{R}$ is determined by

$$|LM\rangle = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} |l_1m_1;l_2m_2\rangle \,\langle l_1m_1;l_2m_2|LM\rangle = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} C^{Ll_2l_1}_{Mm_2m_1} \,|l_1m_1;l_2m_2\rangle \tag{3.42}$$

with $|l_1m_1;l_2m_2\rangle = |l_1m_1\rangle \otimes |l_2m_2\rangle$ and the Clebsch-Gordan coefficients [100]

$$C^{Ll_2l_1}_{Mm_2m_1} = \langle l_1m_1;l_2m_2|LM\rangle \tag{3.43}$$

Only terms where $|l_1 - l_2| \leq L \leq l_1 + l_2$ and $M = m_1 + m_2$ are non-zero.

**Equivariance in practice:** The concept of group equivariance has for the most part attained much higher status in the neural network community than simply being mathematical acrobatics, being implemented over a wide spectrum of use cases [101], even within the transformer architecture [102]. Within the machine-learning-for-quantum-chemistry community there has however been debate however on the benefit-to-cost ratio of equivariant versus more banal invariant neural nets. A significant amount of attempts, see e.g. [60, 53, 103, 57], have been undertaken to construct MPNNs with equivariant feature vectors due to their presumed heightened expressiveness [103]. Yet empirically the results have not unequivocally favoured the more complex $SO(3)$-equivariant network designs over $SO(3)$-invariant counterparts [13, 52], which only rely on $SO(3)$-invariant inputs such as distances and angles, spurring some debate over the usefulness of equivariant network designs. For NN-WFs however, where the output quantity is not scalar anymore, an equivariant model is evermore desirable, as in order for the wave-function's energy to be SO(3)-invariant, the wave-function needs to be equivariant [56]. Some attempts have been made to achieve equivariance, though rather in a circumventing fashion [11] with the search for a fully equivariant neural wave-function still an open question [14].

For instructive purposes, we will shortly showcase cornerstones of prototypical implementation for a $SO(3)$-equivariant NN-PES inspired by [96]. Whereas the state-of-the-art of equivariant NN-PESs may be more accurately be represented by NequIP [53], its architectural finetunings are not as relevant to our approach.

As neural networks are usually a composition of multiple layers, equivariance of the full neural net corresponds to equivariance of each layer on its own in order to propagate the rotation of the input through the neural net. The scheme in [96] represents a MPNN utilizing feature vectors - or, more accurately, matricized feature vectors - of the form

$$\boldsymbol{f}_v = \bigoplus_{l=0}^{l_{max}} \boldsymbol{f}_v^{(l)} \in \mathbb{R}^{L\times F} \tag{3.44}$$

with $L = \sum_{l=0}^{l_{max}} 2l+1$. The basis vectors are the spherical harmonics. Thus, the $j$-th column of $\boldsymbol{f}$ ought to be understood as

$$(\boldsymbol{f})_{:,j} = \begin{pmatrix} f_{0,0;j} \\ f_{1,-1;j} \\ f_{1,0;j} \\ f_{1,0;j} \\ ... \\ f_{l_{max},l_{max};j} \end{pmatrix} = f_{0,0;j}Y_{0,0} + f_{1,-1;j}Y_{1,-1} + ... + f_{l_{max},l_{max};j}Y_{l_{max},l_{max}} \tag{3.45}$$

where we make use of NumPy-ian indexing, i.e. $(\boldsymbol{f})_{:,j}$ denoting the j-th column of the matrix $\boldsymbol{f}$. We can recognize these matrix feature vectors to be equivalent to the vectorial feature vectors described in section 3.6 by regarding a matrix as unstacked vector and vice versa.

The feature vectors ${}^0\boldsymbol{f} \equiv {}^0\boldsymbol{f}(E(\vec{\boldsymbol{R}}))$ as constructed from the molecular geometry $\vec{\boldsymbol{R}}$ via some $SO(3)$-equivariant function $E$. We thus know them to transform as

$$\vec{\boldsymbol{R}} \mapsto \boldsymbol{O}\vec{\boldsymbol{R}} \tag{3.46}$$

$${}^0\boldsymbol{f}_v^{(l)} \mapsto D^l(\boldsymbol{O})^0\boldsymbol{f}_v^{(l)} \tag{3.47}$$

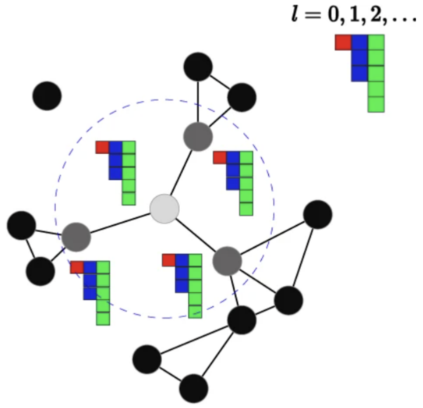

Figure 3.8.: An illustration of a graph with SO(3)-equivariant feature vectors, taken from [53]. For all practical purposes, we concatenate the differently-colored components into one vector.

thus rendering the whole feature vector ${}^{0}\boldsymbol{f}$ equivariant. The key challenge for MMPNs to overcome is to preserve the equivariance throughout the message-passing, that is, to couple feauture vectors non-trivially while preserving $SO(3)$-equivariance. This coupling is achieved by considering the tensor product of two neighbouring feature vectors $\boldsymbol{f} \otimes \boldsymbol{g}$. We have seen above that tensor products of spherical harmonics $|l_1 m_1\rangle \otimes |l_2 m_2\rangle$ yields another set of spherical harmonics $|LM\rangle$ with the transformation between the two sets governed by the Clebsch-Gordan coefficients. Therefore, the resulting tensor product is still SO(3)-equivariant. All that remains is to express the coefficients of $|l_1 m_1\rangle \otimes |l_2 m_2\rangle$ in the new basis $|LM\rangle$. Each pairs of components $l_1$ and $l_2$ generate a set of feature vectors of degree $l_3$ with $l_3 \in \{|l_2 - l_1|, ..., l_2 + l_1\}$ coupling of feature vectors $\boldsymbol{f}$ and $\boldsymbol{g}$ is given element-wise by

$$\left(\boldsymbol{f}^{(l_1)} \otimes_{CG} \boldsymbol{g}^{(l_2)}\right)_{l_3, m_3} = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} C^{l_3, l_2, l_1}_{m_3, m_2, m_1} \boldsymbol{f}_{l_1, m_1} \odot \boldsymbol{g}_{l_2, m_2} \tag{3.48}$$

with the element-wise product $\odot$ (note that we are dealing with $F$-dimensional vectors, hence element-wise product). For example, the fragments $\boldsymbol{f}^{(1)}$ and $\boldsymbol{f}^{(1)}$ will yield equivariant fragments with values $l_3 = 0, 1, 2$. In general, we will have a series of resulting equivariant fragments. For example, the pairs $l_1 = l_2 = 0$, $l_1 = l_2 = 1$ and $l_1 = l_2 = 2$ all yield a fragment with $l_3 = 0$. The simplest way to recombine all obtained equivariant fragments into a single feature vector is by summing over all fragments with a given value of $l_3$, which we implicitly mean whenever we speak about this type of tensor product. We refer to this equivariance-preserving technique as Clebsch-Gordan contraction. By truncating the resulting to a certain $l_{max}$ it is possible to work with tensors of constant shape.

For later purposes we also define a vector $\boldsymbol{Y}(\boldsymbol{R}) \in \mathbb{R}^F$ of spherical harmonics as

$$\boldsymbol{Y}(\boldsymbol{R}) = \begin{pmatrix} Y_{0,0}(\boldsymbol{R}) \\ Y_{1,-1}(\boldsymbol{R}) \\ Y_{1,0}(\boldsymbol{R}) \\ Y_{1,0}(\boldsymbol{R}) \\ ... \\ Y_{l_{max},l_{max}}(\boldsymbol{R}) \end{pmatrix} \tag{3.49}$$

We can recognize eq. 3.45 to essentially describe a molecular orbital. The coefficient vector of a molecular orbital thus constitutes a naturally equivariant quantity. As molecular orbitals are equivariant, so are products of molecular orbitals and linear combinations thereof, implying that CI wave-functions are equivariant. This implies that the amplitudes of a CI wave-function $|\Psi\rangle = \sum_{\boldsymbol{x}} \psi(\boldsymbol{x}; \boldsymbol{\theta}) |\boldsymbol{x}\rangle$ calculated by a neural wave-function need to be *SO*(3)*-invariant*, as the equivariance of the wave-function is fully captured by the *SO*(3)-equivariance of $|\boldsymbol{x}\rangle$.

**HF-$\mathbb{Z}_2$-equivariance:** This insight translates to another kind of equivariance which was highlighted in a recent study on NN-WFs [14] and will thus be shortly addressed by us. As the molecular orbitals are obtained as being the eigenvectors of the Fock matrix, they are only defined up to a phase factor of ±1. We can treat this gauge freedom adequately in the equivariance paradigm. We assume now that we receive for spin orbital $j$ the vector of molecular coefficients $-\boldsymbol{c}_j$ instead of $\boldsymbol{c}_j$. We refer to this flipped sign as the HF-sign-flip. For a given CI wave-function $|\Psi\rangle = \sum_{\boldsymbol{x}} \psi(\boldsymbol{x}) |\boldsymbol{x}\rangle$, the basis kets will change as

$$|\boldsymbol{x}\rangle \mapsto |\tilde{\boldsymbol{x}}\rangle := (-1)^{x_j} |\boldsymbol{x}\rangle \tag{3.50}$$

The kets are thus equivariant to the HF-sign-flip. We dub this equivariance the HF-$\mathbb{Z}_2$-equivariance. We define a unitary $\boldsymbol{U}_j$ via $\boldsymbol{U}_j : |\boldsymbol{x}\rangle \mapsto |\tilde{\boldsymbol{x}}\rangle$. Occurrence of the HF-sign-flip thus corresponds to a basis change $\{|\boldsymbol{x}\rangle\}_{\boldsymbol{x}} \mapsto \{|\tilde{\boldsymbol{x}}\rangle\}_{\tilde{\boldsymbol{x}}}$. The Hamiltonian $\hat{H} = \sum_{\boldsymbol{x},\boldsymbol{y}} H_{\boldsymbol{x},\boldsymbol{y}} |\boldsymbol{x}\rangle\langle\boldsymbol{y}|$ will under this basis change transform as $\hat{H} \mapsto \boldsymbol{U}_j \hat{H} \boldsymbol{U}^\dagger$. In order for the energy $\langle\Psi|\hat{H}|\Psi\rangle$ to be HF-$\mathbb{Z}_2$-invariant, $|\Psi\rangle$ has to transform as $|\Psi\rangle \mapsto \boldsymbol{U}_j |\Psi\rangle$, i.e. it has to be HF-$\mathbb{Z}_2$-equivariant. As the kets $|\boldsymbol{x}\rangle$ themselves are HF-$\mathbb{Z}_2$-equivariant, it is necessary for the amplitudes $\psi(\boldsymbol{x})$ to be HF-$\mathbb{Z}_2$-invariant. The same line of reasoning can be applied to the necessity of *SO*(3)-equivariance of the wave-function for *SO*(3)-invariance of the energy [11].

# 4. MFCNet: towards geometric neural wave-functions

We now turn towards crafting a neural architecture which adheres to all the desiderata laid out in section 3.5. We will employ a MPNN to that end, prompting us to structure this section along the lines of the core notions of a MPNN: we need to define what constitutes the nodes, how we construct their corresponding feature vectors, how we update feature vectors based upon neighbouring feature vectors, and how we read out the neural wave-function from the final graph. We present a high-level description of the core tenets before elaborating on the mathematical machinery.

**Calculation of $SO(3)$-invariant and HF-$\mathbb{Z}_2$-invariant CI amplitudes:** Our model represents a 2Q-NN-WF and thus calculates the amplitudes $\psi(\boldsymbol{x};\boldsymbol{\theta})$ of a CI wave-function $|\Psi\rangle = \sum_{\boldsymbol{x}} \psi(\boldsymbol{x};\boldsymbol{\theta})\,|x\rangle$ based on the molecular geometry. The amplitudes are SO(3)-invariant and HF-$\mathbb{Z}_2$-invariant, thus yielding a fully equivariant wave-function. As current 2Q-NN-WFs do, see section 3.4, we work exclusively with spinless systems, $S = 0$, implying the Hilbert space to be spanned only by determinants with equal number of spin up and down electrons.

**Spin orbital as nodes:** Our model proposes a two-layered graph. The first layer corresponds to the molecular geometry with atoms as nodes. We extract a second layer from the first layer consisting of one node per spin orbital. While graphs based on atomic nodes are posed with the problem how to construct a possibly varying number of spin orbitals from atomic nodes [11], we can straightforwardly map nodes to orbitals one-to-one. Message-passing takes place among the nodes of the second layer.

$SO(3)$**-equivariant feature vectors:** We imbue the spin orbital nodes with $SO(3)$-equivariant feature vectors, where we import the geometric structure from a Hartree-Fock precalculation. While we output a scalar amplitude rendering the employment of geometric feature vectors not strictly necessary, believe in the enhanced expressiveness of geometric feature vectors for our use case due to the molecular orbital coefficients being naturally equivariant.

**Constant parameter count, locality and transferability:** As our model is completely modular and all learnable parameters describe either local atomic feature vectors or their local interaction, there are no a priori limitations on training the model for arbitrary molecules and molecule geometries at once.

After a thorough description of the key components of the architecture, we will shortly reflect upon which further finetunings are necessary, before we put our proposed architecture to the numerical test. We dub our model **M**olecular-orbital-**F**eature-**C**entric-**N**et, or MFCNet in short.

## 4.1. The embedding phase: constructing equivariant molecular orbital features

MFCNet obtains as input the molecular geometry $\vec{\boldsymbol{R}} \in \mathbb{R}^{3\times M}$. As is commonly done in 1Q-NN-WFs, see e.g. [8, 14, 71, 56], we first conduct a Hartree-Fock (HF) precalculation. We assume the HF precalculation to be conducted in the STO-3G basis [104] which equips atoms with atomic number $Z > 2$ with orbitals $1s, 2s$ and $2p$ and the $1s$-orbital only if $Z \leq 2$, i.e. for the atoms Hydrogen and Helium. We construct the first layer of the graph by attributing learnable feature vectors $\boldsymbol{x}_{Z,nl} \in \mathbb{R}^F$ to each orbital of each atom. For example, in the STO-3G the nitrogen atom with seven electrons basis gets equipped with three feature vectors, namely $\boldsymbol{x}_{7,s_1} \in \mathbb{R}^F$, $\boldsymbol{x}_{7,s_2} \in \mathbb{R}^F$ and $\boldsymbol{x}_{7,p} \in \mathbb{R}^F$. We thus deviate slightly from exemplary MPNN-based neural wave-functions such as [11] and [14], which utilize a single $F$-dimensional feature vector per atom. This also prohibits to reuse the same model when basis sets other than STO-3G are employed. This drawback is however not exclusive to MFCNet, but rather implicitly present for all 2Q-NN-WFs.

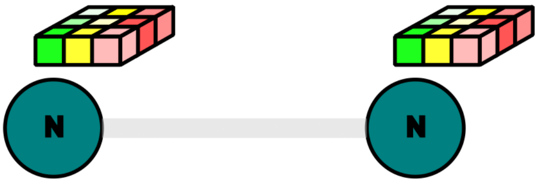

Figure 4.1.: Exemplary atomic feature vectors $\boldsymbol{x}_{7,s_1}$ (green), $\boldsymbol{x}_{7,s_2}$ (yellow) and $\boldsymbol{x}_{7,p}$ (red) for $F = 3$. Since we work with multidimensional feature vectors, we choose the $F$ channels to be depicted along the inward-axis.

From the atomic graph we then proceed to construct the second layer of the graph, where the atomic feature vectors are recombined and synthesized with the HF coefficients $\boldsymbol{C} \in \mathbb{R}^{N_{orb}\times N_{orb}}$, with $N_{orb}$ denoting the number of molecular orbitals, to yield feature vectors $\boldsymbol{f}_v$ for all spin orbitals $v$ separately. This explicit introduction of spin-orbital-wise feature vectors differentiates MFCNet from MPNN-based NN-WFs [11, 56, 14], which purely utilize atomic nodes and construct the electronic wave-functions from atomic feature vectors. We design these feature vectors $\boldsymbol{f}_v$ to be SO(3)-equivariant as well as HF-$\mathbb{Z}_2$-invariant. We achieve the former by imposing the geometrical character of the HF coefficients $\boldsymbol{C}$ onto the feature vectors $\boldsymbol{f}_v$. Note that utilizing $\boldsymbol{C}$ to realize SO(3)-equivariant NN-WFs was lined out in [14] as a future avenue of research.

Recall a molecular orbital to be defined as a linear combination of atomic orbitals,

$$\chi_v(\mathbf{r}) = \sum_j c_{jv}\phi_j(\mathbf{r}) \tag{4.1}$$

with the index $j$ running over all orbitals of all atoms of a molecule and with $\boldsymbol{C}_{jv} = c_{jv}$. As alluded to above, the coefficients $c_{jv}$ will prove to be our gateway to constructing equivariant feature vectors from the scalar atomic feature fectors $\mathbf{x}_{ZO}$. To elucidate where the geometric structure originates, we first consider the toy example of a molecular orbital containing only atomic orbitals from a single atom $A$ with $Z_A > 2$. Explicitly spelling this scenario out and assuming the atom $A$ to contain two s-orbitals and one p-orbital, the molecular orbital $\chi_v$ is defined as

$$\chi_v(\mathbf{r}) = c_{1s}\phi_{1s,A}(\mathbf{r}) + c_{2s}\phi_{2s,A}(\mathbf{r}) + c_{2p_x}\phi_{2p_x,A}(\mathbf{r}) + c_{2p_y}\phi_{2p_y,A}(\mathbf{r}) + c_{2p_z}\phi_{2p_z,A}(\mathbf{r}) \tag{4.2}$$

Recall the atomic orbitals to be of the form

$$\phi_{nlm}(\mathbf{r}) = R_n(\|\mathbf{r}\|)Y_{lm}(\mathbf{r}) \tag{4.3}$$

with $Y_{lm}$ denoting the real spherical harmonics [105]. Plugging this definition into the above eq. 4.2 and absorbing $R_n(\|\mathbf{r}\|)$ into the coefficients $\tilde{c}_{nlm} = R_n(\|\mathbf{r}\|)c_{nlm}$ yields

$$\chi_v(\mathbf{r}) = \tilde{c}_{1s}Y_{0,0}(\mathbf{r}) + \tilde{c}_{2s}Y_{0,0}(\mathbf{r}) + \tilde{c}_{2p_x}Y_{1,-1}(\mathbf{r}) + \tilde{c}_{2p_y}Y_{1,0}(\mathbf{r}) + \tilde{c}_{2p_z}Y_{1,1}(\mathbf{r}) \tag{4.4}$$

Considering section 3.7 and defining a concatenation operation $[\cdot\|\cdot]$, we can now immediately recognize the vector

$$\boldsymbol{g}_{\chi_v} := \left[\tilde{c}_{1s}\|\tilde{c}_{2s}\|\tilde{c}_{2p_x}\|\tilde{c}_{2p_y}\|\tilde{c}_{2p_z}\right] \equiv \begin{pmatrix} \tilde{c}_{1s} \\ \tilde{c}_{2s} \\ \tilde{c}_{2p_x} \\ \tilde{c}_{2p_y} \\ \tilde{c}_{2p_z} \end{pmatrix} \in \mathbb{R}^5 \tag{4.5}$$

to be $SO(3)$-equivariant. Formally, the transformational behaviour of $\boldsymbol{g}_{\chi_v}$ is given by

$$\vec{\boldsymbol{R}} \mapsto \boldsymbol{O}\vec{\boldsymbol{R}} \text{ with } \boldsymbol{O} \in SO(3) \implies \boldsymbol{g}_{\chi_v} \mapsto \boldsymbol{D_O} \tag{4.6}$$

with

$$\boldsymbol{D_O} = \mathbb{1}_1 \oplus \mathbb{1}_1 \oplus \boldsymbol{O} \in \mathbb{R}^{5\times 5} \tag{4.7}$$

and hence rendering $\boldsymbol{g}_{\chi_v}$ SO(3)-equivariant.

While in practice molecular orbitals will rarely contain contributions from a single atom only, we can easily accommodate for the general case. Consider the molecular orbital $\chi_v$ to now contain contributions from $M$ atoms. The matrix $\boldsymbol{G}_{\chi_v}$ obtained from stacking vectors of the form of $\boldsymbol{g}_{\chi_v}$

$$\boldsymbol{G}_{\chi_v} := \left[\tilde{c}_{1s}\|\tilde{c}_{2s}\|\tilde{c}_{2p_x}\|\tilde{c}_{2p_y}\|\tilde{c}_{2p_z}\right] \equiv \begin{pmatrix} \tilde{c}_{1s,1} & \tilde{c}_{1s,2} & & \tilde{c}_{1s,M} \\ \tilde{c}_{2s,1} & \tilde{c}_{2s,2} & & \tilde{c}_{2s,M} \\ \tilde{c}_{2p_x,1} & \tilde{c}_{2p_x,2} & \cdots & \tilde{c}_{2p_x,M} \\ \tilde{c}_{2p_y,1} & \tilde{c}_{2p_y,2} & & \tilde{c}_{2p_y,M} \\ \tilde{c}_{2p_z,1} & \tilde{c}_{2p_z,2} & & \tilde{c}_{2p_z,M} \end{pmatrix} \in \mathbb{R}^{5\times M} \tag{4.8}$$

is again equivariant. As we will utilize $\boldsymbol{G}_{\chi_v}$ as a means to construct equivariant and learnable feature vectors, we want to standardize its shape across different basis sets, as in its current form its shape explicitly depends upon how many s- and p-orbitals a basis set provides any given atom with. We note that we can aggregate all blocks with identical angular momentum along the rows of $\boldsymbol{G}_{\chi_v}$ without torpedoing the SO(3)-equivariance of $\boldsymbol{G}_{\chi_v}$. In the case of working in the STO-3G basis, which contains more than one orbital only for $l = 0$, this amounts to

$$\tilde{\boldsymbol{G}}_{\chi_v} = \begin{pmatrix} \lambda\tilde{c}_{2s,1} + \mu\tilde{c}_{1s,1} & \lambda\tilde{c}_{2s,2} + \mu\tilde{c}_{1s,2} & & \lambda\tilde{c}_{2s,M} + \mu\tilde{c}_{1s,M} \\ \tilde{c}_{2p_x,1} & \tilde{c}_{2p_x,2} & \cdots & \tilde{c}_{2p_x,M} \\ \tilde{c}_{2p_y,1} & \tilde{c}_{2p_y,2} & & \tilde{c}_{2p_y,M} \\ \tilde{c}_{2p_z,1} & \tilde{c}_{2p_z,2} & & \tilde{c}_{2p_z,M} \end{pmatrix} \in \mathbb{R}^{4\times M} \tag{4.9}$$

with learnable parameters $\lambda, \mu$, which we can equivalently define by $\tilde{\boldsymbol{G}}_{\chi_v} = \boldsymbol{W}\boldsymbol{G}_{\chi_v}$ with a suiting partly learnable matrix $\boldsymbol{W} \in \mathbb{R}^{4\times 5}$. We merge $\tilde{\boldsymbol{G}}_{\chi_v} \in \mathbb{R}^{4\times M}$ now with the atom-and-orbital-wise feature vectors $\boldsymbol{x}_{Z,nl} \in \mathbb{R}^F$ by introducing the quantity $\boldsymbol{t}_v \in \mathbb{R}^{4\times F\times M}$ defined elementwise by

$$(\boldsymbol{t}_v)_{nlm,i,j} = \tilde{c}_{nlm,j}(\boldsymbol{x}_{Z_j,nl})_i \tag{4.10}$$

which ought to be understood as a "matrix of vectors"

$$\boldsymbol{t}_v = \begin{pmatrix} (\lambda\tilde{c}_{2s,1}\boldsymbol{x}_{Z_1,s_1} + \mu\tilde{c}_{1s,1}\boldsymbol{x}_{Z_1,s_2}) & \lambda\tilde{c}_{2s,2}\boldsymbol{x}_{Z_2,s_1} + \mu\tilde{c}_{1s,2}\boldsymbol{x}_{Z_2,s_2} & \\ \tilde{c}_{2p_x,1}\boldsymbol{x}_{Z_1,p} & \tilde{c}_{2p_x,2}\boldsymbol{x}_{Z_2,p} & \cdots \\ \tilde{c}_{2p_y,1}\boldsymbol{x}_{Z_1,p} & \tilde{c}_{2p_y,2}\boldsymbol{x}_{Z_2,p} & \\ \tilde{c}_{2p_z,1}\boldsymbol{x}_{Z_1,p} & \tilde{c}_{2p_z,2}\boldsymbol{x}_{Z_2,p} & \end{pmatrix} \tag{4.11}$$

We furthermore want to augment $\boldsymbol{t}_v$ with spatial information regarding the relative positions of the atoms 1 to M. To that end, we first define a weighted center for the molecular orbital as

$$\boldsymbol{R}_{\chi_v} := \frac{\sum_{j=1}^{M} C_j^2 \boldsymbol{R}_j}{\sum_{j=1}^{M} C_j^2} \text{ with } C_j^2 = \sum_{n,l,m} c_{nlm,j}^2 \tag{4.12}$$

with $\boldsymbol{R}_j$ denoting the position of the j-th atom contributing some atomic orbital to $\chi_v$. Note that $C_j$ is $SO(3)$-invariant, being the squared norm of the $j$-th column of the SO(3)-equivariant quantity $\boldsymbol{G}_{\chi_v}$. Hence, $\boldsymbol{R}_{\chi_v}$ is $SO(3)$-equivariant. We import information regarding the relative atomic positions via the L-dimensional vectors $\boldsymbol{Y}(\boldsymbol{R}_k - \boldsymbol{R}_{\chi_v})$, utilizing the CG-contraction $\otimes_{CG}$ introduced in section 3.7 to define a new tensor $\tilde{\boldsymbol{f}}_v \in \mathbb{R}^{L\times F\times M}$ column-wise as

$$\left(\tilde{\boldsymbol{f}}_v\right)_{:,j,k} = b_j(\|\boldsymbol{R}_k - \boldsymbol{R}_{\chi_v}\|) \left[\left[\boldsymbol{t}_v \| \boldsymbol{0}_{(L-4)\times M\times F}\right]_{:,j,k} \otimes_{CG} \boldsymbol{Y}(\boldsymbol{R}_k - \boldsymbol{R}_{\chi_v})\right] \tag{4.13}$$

with $\boldsymbol{0}_{(L-4)\times F\times M}$ being a tensor of zeros of shape $(L-4)\times F\times M$ and $b_j$ with $j = 1, ..., F$ being the $j$-th spherical Bessel polynomial [100]

$$b_j(r) = (-r)^j \left(\frac{1}{r}\frac{d}{dr}\right)^j \frac{\sin(r)}{r} \tag{4.14}$$

Employing Bessel polynomials as a means to diffuse radial information across the different channels was popularized by Gasteiger et al. in DimeNet [13] and henceforth constitutes a common design element in numerous PES-NNs and WF-NNs.

Since $\tilde{\boldsymbol{f}}_v$ is linear in all molecular orbital coefficients, it is equivariant to the HF-sign-flip, since this sign flip equates to $c_{nlm,j} \mapsto -c_{nlm,j}$. Since we aim for HF-$\mathbb{Z}_2$-invariance instead of equivariance, we Clebsch-Gordan-contract $\tilde{\boldsymbol{f}}_v$ with itself,

$$\tilde{\tilde{\boldsymbol{f}}}_v = \tilde{\boldsymbol{f}}_v \otimes_{CG} \tilde{\boldsymbol{f}}_v \tag{4.15}$$

to render it quadratic with regards to the molecular orbital coefficients.

One last drawback preventing the tensor $\tilde{\tilde{\boldsymbol{f}}}_v$ from being used as feature vector for the molecular orbital $\chi_v$ is the explicit dependence of its shape on the number of atoms $M$ with non-vanishing contributing atomic orbitals. This is undesirable would necessitate cumbersome mathematical interventions in order to render feature vectors of different molecular orbitals with possibly different values of $M$ comparable. We alleviate this by tracing out the atom-enumerating last dimension and use as feature vector

$$\boldsymbol{f}_{\chi_v} = \mathrm{Tr}_3[\tilde{\tilde{\boldsymbol{f}}}_{\chi_v}] \equiv {}^0\boldsymbol{f}_{\chi_v} \equiv {}^0\boldsymbol{f}_{\chi_v}(\boldsymbol{c}_{\chi_v}) \tag{4.16}$$

with $\boldsymbol{c}_{\chi_M}$ the column of HF-coefficients belonging to $\chi_M$ and the superscript zero indicating that these are the initial feature vectors after zero steps of message passing.

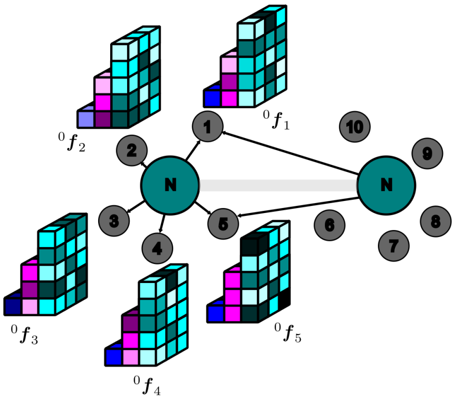

Figure 4.2.: Exemplary feature vectors for five of the ten molecular orbitals for $N_2$. Arrows indicate which atoms contribute to which molecular orbitals. In practice, for small molecules all atoms will contribute. This visualization serves only illustrative purposes.

We furthermore visualize the computational graph in fig. 4.3.

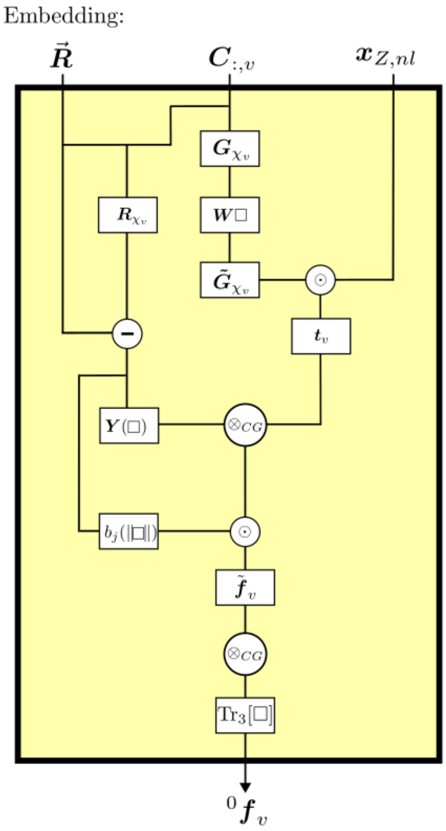

Figure 4.3.: Computational graph of how the initial feature vectors are calculated. Squares denote input into the computational operation.

## 4.2. The message-passing phase: preserving equivariance

We now turn to concocting a message-phasing scheme that allows the nodes of different molecular orbitals to communicate and incorporate information of neighbouring nodes while preserving equivariance of each node's feature vector. We first have to formalise the concept of neighbourhood. Contrasting atomistic MPNNs, where distance in euclidean space presents a canonical choice to define neighbourhood between nodes in the graph [13], we have to redefine the concept of neighbourhood when considering molecular orbitals as nodes. We settle on defining the neighbourhood $N(v)$ of a molecular orbital $v$ as

$$N(v) := \left\{ w \neq v \in G : h^{(2)}_{vvww} = \frac{\iint d^3\mathbf{r}_1 d^3\mathbf{r}_2 |\chi_v(\mathbf{r}_1)|^2 |\chi_w(\mathbf{r}_1)|^2}{\|\mathbf{r}_1 - \mathbf{r}_2\|} > \epsilon \right\} \tag{4.17}$$

with $\epsilon$ being a tunable hyperparameter. Note that this integral, referred to as the Schwartz integral, has been employed as useful heuristic in other use cases in quantum chemistry [106]. Favourably, it preserves locality for localized MOs [107]. On the other hand, introducing the cutoff $\epsilon$ poses the threat of introducing discontinuities within the neural net if two molecules exhibit a value for the Schwartz integral that hovers around the cutoff value. We mitigate this by curating the input data in a way such that these potential discontinuities are smoothened out. We elaborate on this procedure in section 4.4.

Remember the updates of the feature vectors in the $t$-the step of message passing to to be governed by

$$
\begin{aligned}
{}^{t+1}\boldsymbol{f}_v &= {}^{t}U\left({}^{t}\boldsymbol{f}_v, {}^{t+1}\boldsymbol{m}_v\right) && (4.18)\\
{}^{t+1}\boldsymbol{m}_v &= \sum_{w \in \mathcal{N}(v)} {}^{t}M\left({}^{t}\boldsymbol{f}_v, {}^{t}\boldsymbol{f}_w\right) && (4.19)
\end{aligned}
$$

with update functions $U^{(t)}$ and mixing operations $M^{(t)}$. Note that contrasting the original definition in section 3.6, we are not employing any edge weights in our model. A commplace method of non-linear $SO(3)$-equivariance-preserving mixing of two feature vectors $\boldsymbol{f}_v$ and $\boldsymbol{f}_w$ is the use of the CG-contraction, as described in section 3.7. While this CG-contraction allows mixing of different channels along the same degree of angular momentum as well as $SO(3)$-equivariance-preserving mixing across different levels of angular momentum, it has been criticized for its large computational cost [108]. Though we are aware of the ingenious techniques developed to slim down computational efforts, we can not straightforwardly implement them for our use case [109]. We thus aim for a more minimalist technique allowing to mix equivariant feature vectors of different nodes, relying on theoretical underpinnings exploring the expressiveness for certain types of mixing operations [110]. Proposition 5 of [110] circumscribes the amount of design freedom we have in creating a mixing operation by stating that, given any $SO(d)$-equivariant function $h : \mathbb{R}^{d\times F} \to \mathbb{R}^d$, we can find $F$ $SO(d)$-invariant differentiable functions $k_i : \mathbb{R}^{d\times F} \to \mathbb{R}$; $i = 1, ..., F$ such that

$$
h(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_F) = \sum_{i=1}^{F} k_i(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_F)\boldsymbol{v}_i \tag{4.20}
$$

with $\boldsymbol{v}_i \in \mathbb{R}^d$. We note that this can be trivially generalized to an $SO(d)$-equivariant function $h : \mathbb{R}^{d\times F} \to \mathbb{R}^{d\times F}$ by stacking $F$ functions

$$
h_j(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_F) = \sum_{i=1}^{N} k_{ij}(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_F)\boldsymbol{v}_i \text{ for } j = 1, ..., F \tag{4.21}
$$

as $h = [h_1 \| h_2 \| ... \| h_F]^T$. We now remember our feature vectors to be of the form $\boldsymbol{f} = \bigoplus_{l=0}^{l_{max}} \boldsymbol{f}^{(l)}$. We compose an SO(3)-equivariant mixing operation $M(\boldsymbol{f}, \boldsymbol{g})$ via

$$M(\boldsymbol{f},\boldsymbol{g}) = \bigoplus_{l=0}^{l_{max}} M^{(l)}(\boldsymbol{f}^{(l)},\boldsymbol{g}^{(l)}) \tag{4.22}$$

where each of the $M^{(l)}$ are $SO(2l+1)$-equivariant functions. We quickly show that $M$ is indeed $SO(3)$-equivariant - that is, for $\boldsymbol{R} \in SO(3)$, and $\boldsymbol{D} = \bigoplus_{l=0}^{l_{max}} \boldsymbol{D}^l(\boldsymbol{R})$ with the Wigner-D-matrices $\boldsymbol{D}^{\boldsymbol{l}}$, it holds that

$$M(\boldsymbol{Df},\boldsymbol{Dg}) = \boldsymbol{D}M(\boldsymbol{f},\boldsymbol{g}) \tag{4.23}$$

This however follows immediately from the Wigner-D-matrices obeying the set of orthonormality relations [111]

$$\sum_k D^l_{km'}(\boldsymbol{R})^* D^l_{km}(\boldsymbol{R}) = \delta_{mm'} \tag{4.24}$$

implying $\boldsymbol{D}^l(\boldsymbol{R})^\dagger \boldsymbol{D}^l(\boldsymbol{R}) = \mathbb{1}_{2l+1}$ and hence $\boldsymbol{D}^l(\boldsymbol{R}) \in SO(2l+1)$. The $SO(2l+1)$-equivariance of $M^{(l)}$ now yields

$$\begin{aligned} M(\boldsymbol{Df},\boldsymbol{Dg}) &= \bigoplus_{l=0}^{l_{max}} M^{(l)}(\boldsymbol{D}^l(\boldsymbol{R})\boldsymbol{f}^{(l)},\boldsymbol{D}^l(\boldsymbol{R})\boldsymbol{g}^{(l)}) = \bigoplus_{l=0}^{l_{max}} \boldsymbol{D}^l(\boldsymbol{R})M^{(l)}(\boldsymbol{f}^{(l)},\boldsymbol{g}^{(l)}) && (4.25)\\ &= \boldsymbol{D}\bigoplus_{l=0}^{l_{max}} M^{(l)}(\boldsymbol{f}^{(l)},\boldsymbol{g}^{(l)}) = \boldsymbol{D}M(\boldsymbol{f},\boldsymbol{g}) && (4.26) \end{aligned}$$

where the second-to-last equality follows from the block-diagonality of $\boldsymbol{D}$.

We now define $\boldsymbol{f}_j := \boldsymbol{f}_{:,j}$ and $\boldsymbol{h}_j := \boldsymbol{h}_{:,j}$. By identifying the set $\{\boldsymbol{f}_j^{(l)}\}_j \cup \{\boldsymbol{g}_j^{(l)}\}_j$ with the set $\{\boldsymbol{v}_n\}_n$ from eq. 4.21 we recognize the functions $M^{(l)}$ to be of the form of $h$ from eq. 4.20, implying the decomposition $M^{(l)} = [M_1^{(l)} \| M_2^{(l)} \| ... \| M_F^{(l)}]$ with

$$M_i^{(l)} = \sum_{j=1}^{F} {}_1M_{ij}^{(l)}(\boldsymbol{f}_1^{(l)},...,\boldsymbol{f}_F^{(l)},\boldsymbol{g}_1^{(l)},...,\boldsymbol{g}_F^{(l)})\boldsymbol{f}_j^{(l)} + \sum_{j=1}^{F} {}_2M_{ij}^{(l)}(\boldsymbol{f}_1^{(l)},...,\boldsymbol{f}_F^{(l)},\boldsymbol{g}_1^{(l)},...,\boldsymbol{g}_F^{(l)})\boldsymbol{g}_j^{(l)} \tag{4.27}$$

with ${}_nM_{ij}$ being scalar functions. We can enforce $SO(2l+1)$-invariance of ${}_nM_{ij}$, $n = 1,2$ by

$$\begin{aligned} {}_nM_{ij}^{(l)}(\boldsymbol{f}_1^{(l)},...,\boldsymbol{f}_F^{(l)},\boldsymbol{g}_1^{(l)},...,\boldsymbol{g}_F^{(l)}) &\equiv {}_nM_{ij}^{(l)}(\{\langle \boldsymbol{x}_p^{(l)},\boldsymbol{y}_q^{(l)}\rangle : p,q = 1,...F;\ \boldsymbol{x},\boldsymbol{y} \in \{\boldsymbol{f},\boldsymbol{g}\}\}) && (4.28)\\ &= {}_nM_{ij}^{(l)}(\mathcal{A}^{(l)}) && (4.29) \end{aligned}$$

with $\mathcal{A}^{(l)} = \{\langle \boldsymbol{x}_p^{(l)}, \boldsymbol{y}_q^{(l)} \rangle : p, q = 1, ...F; \boldsymbol{x}, \boldsymbol{y} \in \{\boldsymbol{f}, \boldsymbol{g}\}\}$, i.e. we only hand in $SO(2l+1)$-invariant scalars to start with. This current setup does not however mix feature fragments $\boldsymbol{f}^{(l_1)}, \boldsymbol{f}^{(l_2)}$ if $l_1 \neq l_2$. We remedy this by augmenting

$$ {}_nM_{ij}^{(l)}(\mathcal{A}^{(l)}) \mapsto {}_nM_{ij}^{(l)} \left( \bigcup_{l=0}^{l_{max}} \mathcal{A}^{(l)} \right) \tag{4.30} $$

which is innocuous from an equivariance-preserving standpoint since we only input $SO(3)$-invariant scalar products. We implement the functions ${}_nM_{ij}^{(l)}$ by multi-layer perceptrons (MLPs), which are stacks of layers $\mathcal{L}$ of the form

$$ \mathcal{L}(\boldsymbol{X}) = \sigma(\boldsymbol{W}\boldsymbol{X} + \boldsymbol{b}) \tag{4.31} $$

with an element-wise nonlinear activation function $\sigma$ and learnable parameters $\boldsymbol{W}$ and $\boldsymbol{b}$.

The resulting mixing-operation $M$ is very much in the spirit of [108], where the authors presented evidence that such mixing-operations are as expressive as CG-contraction with a fraction of the amount of trainable parameters.

Lastly, we want to define a coupling strength $\boldsymbol{J}_{v,w} \in \mathbb{R}^{L \times F}$ between different feature vectors $v$ and $w$. We do so by defining

$$ \left( \tilde{\boldsymbol{J}}_{v,w} \right)_{lf} = b_f(h^{(2)}_{vvww} - \epsilon)) \tag{4.32} $$

with the Bessel functions $b_f$ as above and $\epsilon$ being a tunable hyperparameter. The final coupling strength is defined as

$$ \boldsymbol{J}_{\boldsymbol{v},\boldsymbol{w}} = \mathrm{MLP}\tilde{\boldsymbol{H}}_{\boldsymbol{v},\boldsymbol{w}} \tag{4.33} $$

with MLP a column-wise applied MLP. We now succinctly state the message passing scheme as

$$ {}^{t+1}\boldsymbol{f}_v = \tilde{M}(\boldsymbol{f}_v, \boldsymbol{f}_v) + \sum_{w \in (v)} \boldsymbol{J}_{v,w} \odot M(\tilde{M}(\boldsymbol{f}_v, \boldsymbol{f}_v), \boldsymbol{f}_w) \tag{4.34} $$

with the element-wise product $\odot$.

We visualize the computations in a graph in fig. 4.4

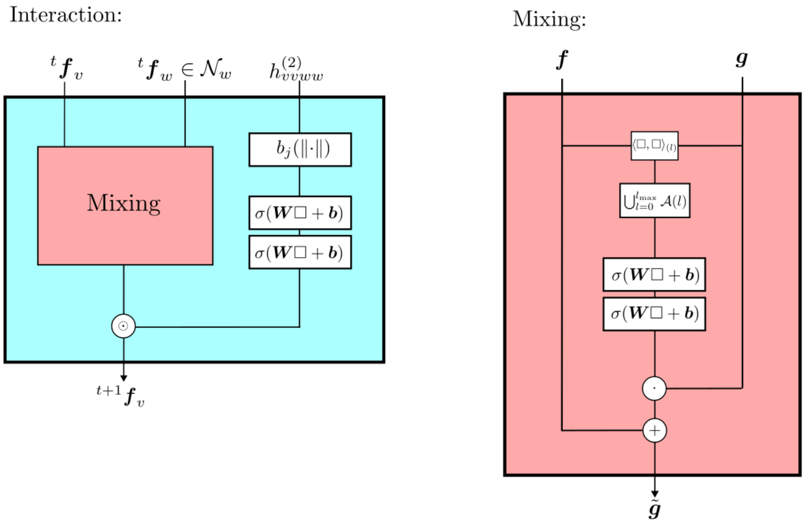
Interaction:
$^{t}\boldsymbol{f}_v$
$^{t}\boldsymbol{f}_w \in \mathcal{N}_w$
$h^{(2)}_{vvww}$
Mixing
$b_j(\|\cdot\|)$
$\sigma(\boldsymbol{W}\Box + \boldsymbol{b})$
$\sigma(\boldsymbol{W}\Box + \boldsymbol{b})$
$\odot$
$^{t+1}\boldsymbol{f}_v$
Mixing:
$\boldsymbol{f}$
$\boldsymbol{g}$
$\langle \Box, \Box \rangle_{(l)}$
$\bigcup_{l=0}^{l_{\max}} \mathcal{A}(l)$
$\sigma(\boldsymbol{W}\Box + \boldsymbol{b})$
$\sigma(\boldsymbol{W}\Box + \boldsymbol{b})$
$\cdot$
$+$
$\tilde{\boldsymbol{g}}$

Figure 4.4.: The computations within the interaction layer.

## 4.3. The read-out phase I: assembling the wave-function

The final piece of the forward pass consists of the readout function $R$ calculating the amplitudes $\psi(\boldsymbol{x};\boldsymbol{\theta})$ as

$$R(\boldsymbol{x}; \{^{T}\boldsymbol{f}_v\}_v) = \psi(\boldsymbol{x};\boldsymbol{\theta}) = \text{MFCNet}(\boldsymbol{x_1x_2}; \vec{\boldsymbol{R}}_A, \boldsymbol{C}) \tag{4.35}$$

while preserving locality, i.e. for a molecule $A$ consisting of two non-interacting composites $A_1$ and $A_2$, we want the wave-function to factorise, i.e.

$$\text{MFCNet}(\boldsymbol{x_1x_2}; \vec{\boldsymbol{R}}_A, \boldsymbol{C}) = \text{MFCNet}(\boldsymbol{x_1}; \vec{\boldsymbol{R}}_{A_1}, \boldsymbol{C})\text{MFCNet}(\boldsymbol{x_2}; \vec{\boldsymbol{R}}_{A_2}, \boldsymbol{c}) \tag{4.36}$$

We reintroduce explicit spin notation by defining separate indicator variables $\boldsymbol{x}^{\uparrow} = (x_1^{\uparrow}, ..., x_{N_{orb}/2}^{\uparrow})$ and $\boldsymbol{x}^{\downarrow} = (x_1^{\downarrow}, ..., x_{N_{orb}/2}^{\downarrow})$ for spin up and spin down orbitals. Since we treated the two spin orbitals for any given spatial orbital on equal footing, we know the feature vectors to be of the form

$$^{T}\boldsymbol{f}_v \equiv \boldsymbol{f}_v = \boldsymbol{f}_v^{\uparrow} = \boldsymbol{f}_v^{\downarrow} \tag{4.37}$$

where we drop $T$ for readability and where $v$ now enumerates the $N_{orb}/2$ spatial orbitals. We furthermore define for a given determinant $\boldsymbol{x}$ the sets of feature vectors corresponding to occupied spin-$\alpha$ orbitals $\mathcal{F}_{\boldsymbol{x}}^{\alpha} = \{\boldsymbol{f}_v^{\alpha} : x_v^{\alpha} = 1\}$ for $\alpha \in \{\uparrow, \downarrow\}$. As we only consider wave-functions with an equal number of spin-up and spin-down electrons, we know these sets to be of equal size. For both sets we define an average feature vector

$$\bar{\boldsymbol{f}}^{\alpha}_{\boldsymbol{x}^{\alpha}} = \frac{1}{|\mathcal{F}^{\alpha}_{v}|} \sum_{v=1}^{N_{\mathrm{orb}}/2} x^{\alpha}_{v} \boldsymbol{f}^{\alpha}_{v} \tag{4.38}$$

We now introduce helper variables $\boldsymbol{I}^{\alpha,\beta} \in \mathbb{R}^{N_{orb}/2 \times N_{orb}/2 \times F \times F}$ as

$$\boldsymbol{I}^{\alpha,\beta}_{vw} = h^{(2)}_{vvww} \langle (M^{\alpha\beta}_{4}(\boldsymbol{f}^{\alpha}_{v}, \bar{\boldsymbol{f}}^{\alpha}_{\boldsymbol{x}^{\alpha}}), M^{\alpha\beta}_{3}(M^{\alpha\beta}_{2}(\boldsymbol{f}^{\beta}_{v}, \bar{\boldsymbol{f}}^{\beta}_{\boldsymbol{x}^{\beta}}), M^{\alpha\beta}_{1}(\boldsymbol{f}^{\alpha}_{v}, \boldsymbol{f}^{\beta}_{w})) \rangle_{L} \tag{4.39}$$

with mixing functions $M^{\alpha\beta}_{i}$ as defined in section 4.2. Via weight-sharing we ensure that $M^{\uparrow\uparrow}_{i} = M^{\downarrow\downarrow}_{i}$ and $M^{\uparrow\downarrow}_{i} = M^{\downarrow\uparrow}_{i}$. By considering the matrices $\boldsymbol{I}^{\alpha,\beta}_{vw} \in \mathbb{R}^{F \times F}$ as vectors of length $F^2$ we define new helper variables $\tilde{\boldsymbol{I}}^{\alpha,\beta} \in \mathbb{R}^{N_{orb}/2 \times N_{orb}/2}$

$$\boldsymbol{I}^{\alpha,\beta}_{vw} = \mathrm{MLP}_{\mathrm{o}}(\tilde{\boldsymbol{I}}^{\alpha,\beta}_{vw}) \tag{4.40}$$

with $\mathrm{MLP}_{\mathrm{o}}$ being a MLP mapping vectors of length $F^2$ to scalars. We use MLPs with one hidden layer with $k$ units, with $k$ being a hyperparamter. Finally, we define $R(\boldsymbol{x}; \{\boldsymbol{f}_v\})$ as

$$R(\boldsymbol{x}; \{\boldsymbol{f}_v\}_v) := \begin{vmatrix} \tilde{\boldsymbol{I}}^{\uparrow,\uparrow} & \tilde{\boldsymbol{I}}^{\uparrow,\downarrow} \\ \tilde{\boldsymbol{I}}^{\downarrow,\uparrow} & \tilde{\boldsymbol{I}}^{\downarrow,\downarrow} \end{vmatrix} \tag{4.41}$$

Note that, reassuringly, $R$ is invariant with regards to the physically irrelevant ordering of the orbitals $v$. Additionally, $R$ is invariant to a global spin flip, consistent with the lack of dependence of the system's Hamiltonian on any spin variables. Both properties follow immediately from the determinant's invariance under identical permutations of rows and columns.

It is now straightforward to validate adherence of $R$ to the demands of $SO(3)$-invariance and locality. $SO(3)$-invariance of $R$ follows since in eq. 4.39 equivariant feature vectors are via scalar products downprojected to SO(3)-invariant scalars. Locality follows from the fact that, for $v$ and $w$ being spin orbitals belonging to different and non-interacting molecule composites $\mathrm{A}_1$ and $\mathrm{A}_2$, $h^{(2)}_{vvww} = 0$, therefore leading to block-diagonal matrices $\tilde{\boldsymbol{I}}^{\alpha,\beta}$. By reordering the rows and columns we can achieve $R(\boldsymbol{x}; \{\boldsymbol{f}_v\}_v)$ to be block-diagonal, where each of the non-zero blocks corresponds to one of $\mathrm{A_i}$. As the determinant factorizes for block-diagonal matrices, we can thus conclude that indeed

$$R(\boldsymbol{x}_1 \boldsymbol{x}_2; \{\boldsymbol{f}_v\}_v) = R(\boldsymbol{x}_1; \{\boldsymbol{f}_v\}_v) R(\boldsymbol{x}_2; \{\boldsymbol{f}_v\}_v) \tag{4.42}$$

We visualize the computations in a graph in fig. 4.5.

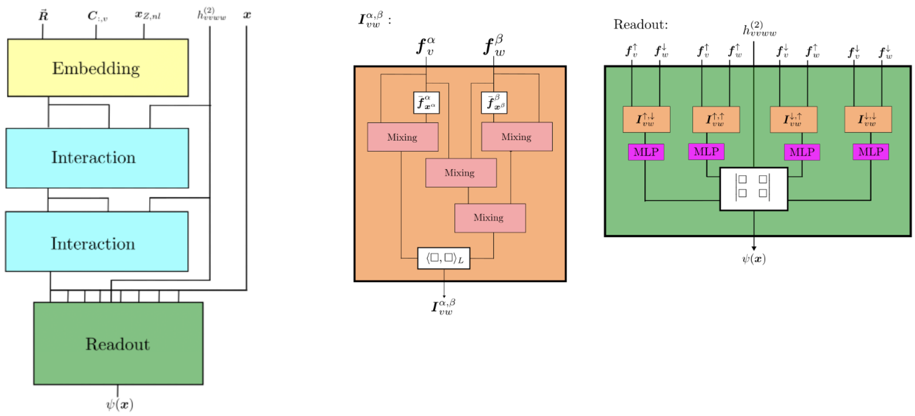
$\vec{R}$
$C_{:,v}$
$x_{Z,nl}$
$h^{(2)}_{vvww}$
$x$
Embedding
Interaction
Interaction
Readout
$\psi(x)$
$I^{\alpha,\beta}_{vw}$ :
$f^{\alpha}_{v}$
$f^{\beta}_{w}$
$\bar{f}^{\alpha}_{x^{\alpha}}$
$\bar{f}^{\beta}_{x^{\beta}}$
Mixing
Mixing
Mixing
Mixing
$\langle \square, \square \rangle_L$
$I^{\alpha,\beta}_{vw}$
Readout:
$h^{(2)}_{vvww}$
$f^{\uparrow}_{v}$ $f^{\downarrow}_{w}$ $f^{\uparrow}_{v}$ $f^{\uparrow}_{w}$ $f^{\downarrow}_{v}$ $f^{\uparrow}_{w}$ $f^{\downarrow}_{v}$ $f^{\downarrow}_{w}$
$I^{\uparrow,\downarrow}_{vw}$
$I^{\uparrow,\uparrow}_{vw}$
$I^{\downarrow,\uparrow}_{vw}$
$I^{\downarrow,\downarrow}_{vw}$
MLP
MLP
MLP
MLP
$\begin{vmatrix} \square & \square \\ \square & \square \end{vmatrix}$
$\psi(x)$

Figure 4.5.: The overall computation of $\psi(x)$ via MFCNet denoted as a computational graph.

## 4.4. The read-out phase II: energy calculation and optimization procedure

We optimize the free parameters using the standard ADAM optimizer [112]. While in the context of solid-state neural wave-functions more advanced optimization mechanisms have been developed [113] which are tailored to the rugged energy landscapes typically seen for neural wave-functions [114], we nevertheless settle for the vanilla ADAM optimizer as the recent state-of-the-art architecture in [80] also utilized the ADAM optimizer and thus serves as more than respectable benchmark already. We furthermore follow [79, 80] by segregating the way we calculate the energy during the optimization scheme and after the optimization with the fully trained model. Identically in spirit to [80] we define a core space $\mathcal{V}^n$ containing $N_u$ configurations, and a connected space $\mathcal{C}^n$ consisting of all configuration differing by any configuration in $\mathcal{V}^n$ by one excitation. However, we replace the not necessarily variational loss function from [80]

$$\tilde{E}_{\boldsymbol{\theta}} = \sum_{x \in \mathcal{V}^n} \frac{|\psi(x;\boldsymbol{\theta})|^2}{\sum_{y \in \mathcal{V}^n} |\psi(y;\boldsymbol{\theta})|^2} \mathrm{E}_{\mathrm{loc}}(x) = \frac{\sum_{x \in \mathcal{V}^n} \langle \psi_{\boldsymbol{\theta}} | x \rangle \, \langle x | \hat{H} | \psi_{\boldsymbol{\theta}} \rangle}{\sum_{y \in \mathcal{V}^n} |\langle y | \psi_{\boldsymbol{\theta}} \rangle|^2} \tag{4.43}$$

which is not necessarily variational, with the variational adaptation

$$\tilde{E}^{\mathcal{V}^n}_{\boldsymbol{\theta}} = \frac{\sum_{x,z \in \mathcal{V}^n} \langle \psi_{\boldsymbol{\theta}} | x \rangle \, \langle x | \hat{H} | z \rangle \, \langle z | \psi_{\boldsymbol{\theta}} \rangle}{\sum_{y \in \mathcal{V}^n} | \, \langle y | \psi_{\boldsymbol{\theta}} \rangle \, |^2} \tag{4.44}$$

which corresponds to the energy of the wave-function downprojected to the core space $\mathcal{V}^n$. This was motivated by our empirical observation that in practice eq. 4.43 proved to be highly non-variational, contrasting the high correspondence between eq. 4.43 and eq. 4.44 observed in [79]. This was not the only empirical anomaly arising as we will touch upon in section 6.

Once the model is trained, we either employ eq. 4.44 or a MCMC-sampling as described earlier to obtain the wave-function's energy, based on the use case at hand.

## 4.5. Finishing touches I: a note on further geometric invariances

Whereas our discussion so far was centered around the proper incorporation of $SO(3)$-equivariance, there are further geometric symmetries we want to respect. The trivial symmetry is the symmetry with respect to translations of the molecule's geometry of the form $\vec{R} \mapsto \vec{R} + \vec{T}$, which obviously should not influence the molecule's energy prediction. We regard this symmetry as trivial, since spatial information only enters MFCNet in the form of the translationally invariant HF coefficients and in the form of translationally invariant relative positions, rendering MFCNet trivially translationally invariant.

$O(3)$**-invariance**: We furthermore want to pick up the thread of considering generic rotations from $O(3)$ rather than only those preserving spatial orientation from $SO(3)$. Since $O(3) = SO(3) \times \{\pm \mathbb{1}_3\}$, the only additional cases we need to consider are rotations that are a combination of a $SO(3)$-rotation and an inversion with respect to the origin. Physical intuition again demands that the energy of a molecule has to be invariant to an inversion of the molecule's geometry of the form $\vec{R} \mapsto -\vec{R}$. Indeed, $O(3)$-invariance has been incorporated into some PES-NNs [53] by employing $O(3)$-equivariant feature vectors. This demands to introduce two separate feature vectors per molecular orbital $f_v^o$ and $f_v^e$ with odd and even parity respectively, defined by according respective behaviour under inversion as

$$\vec{R} \mapsto -\vec{R} \implies f_v^o \mapsto -f_v^o;\, f_v^e \mapsto f_v^e \tag{4.45}$$

As this would significantly bloat the neural architecture, we circumvent this explicit coverage of inversion-equivariance. We rather settle on curating the input data in a way such that we neither have to explicitly incorporate inversion-equivariance nor have to ensure that the neural net learns to recognize inversion symmetry from raw data. We achieve this by first defining the handedness $h_M \in \{+1, -1\}$ of the molecular geometry $\vec{R}_M$ of a molecule $M$ in an algorithmic manner. Due to the translational invariance described above, we can w.l.o.g. assume the molecule to be positioned in space in a way that the averaged position of the $N$ atoms coincides with the origin.

The case of planar molecules is already covered by the $SO(3)$-equivariance. Given a planar molecule which is w.l.o.g. positioned in the xy-plane. For such a molecule, inversion with respect to the origin is achieved by the transformation $e_x \mapsto -e_x$, $e_y \mapsto -e_y$. Such a transformation is achieved by a $\pi$-rotation around the z-axis, which is a $SO(3)$-rotation itself, implying that the action of any $O(3)$-rotation is equivalent to the action of some $SO(3)$-rotation.

As we can for all practical purposes assume the input molecular configurations to be samples drawn from a random distribution [115], the subset of molecular geometries with two atomic positions having identical norm or two atomic positions being linearly dependent is of measure zero and will thus occur with probability zero. Choosing then the unique three atomic positions $\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{R}_3$ with largest norm and such that $\|\boldsymbol{R}_1\| > \|\boldsymbol{R}_2\| > \|\boldsymbol{R}_3\| > 0$ allows us to define the handedness as

$$h_M := \frac{\langle \boldsymbol{R}_1 \times \boldsymbol{R}_2, \boldsymbol{R}_3 \rangle}{|\langle \boldsymbol{R}_1 \times \boldsymbol{R}_2, \boldsymbol{R}_3 \rangle|} \tag{4.46}$$

Since $\boldsymbol{R}_1 \times \boldsymbol{R}_2 \perp \boldsymbol{R}_3$ with probability zero, $h$ is well-defined. Furthermore, $h$ is equivariant with respect to inversion. By convention, we only train and evaluate MFCNet on geometric configurations with handedness $+1$. This does not reduce MFCNet's expressiveness, as we can transform any molecular geometry to one with $h = +1$ by applying an inversion.

**Preprocessing to achieve HF-$\mathbb{Z}_2$-invariance?** One might at this stage reasonably wonder why a similar kind of preprocessing is not conducted on the molecular orbital coefficients $\boldsymbol{C}$ in order to avoid having to explicitly accommodate for HF-$\mathbb{Z}_2$-invariance in the neural architecture. However, due to the Berry phase there are molecular geometries where it is not possible to realign the sign of the molecular orbital coefficients [116].

## 4.6. Finishing touches II: on locality

While locality is inherent in a message-passing scheme and, as we laid out in section 4.3 also preserved in our readout function, we did not explicitly address the issue of locality when describing the composition of the spin orbital feature vectors in section Ȯne may legitimate wonder, whether the across-atom aggregation conducted in eq. 4.8 tacitly undermines locality, as vanilla molecular coefficients are in general delocalized [45]. However, as briefly explored in section 2.4, unitary localization schemes exist. They modify the molecular coefficients such that they are overwhelmingly dominated by the contributions of a couple of neighboured atoms [45]. Even so, the contributions of far-distant atoms will not be strictly zero, though vanishingly small. In order to prohibit those coefficients from making an entrance in eq. 4.8, we apply the same logic that is underlying graph neural networks, defining some cutoff value beyond which nodes do not interact anymore. We define some threshold value $\epsilon_c$ for coefficient values to be recognized and manually enforce that coefficients with lower value are irrelevant. We cannot use a standard step-function, as this would induce unwanted discontinuities. We rather define a differentiable step-function function $f : \mathbb{R} \to \mathbb{R}$ by

$$f_{\epsilon_c,\delta_c}(x) = \begin{cases} 0 & x \leq 0 \\ \left(1 - \cos\left(\pi\left(\frac{x-\epsilon_c}{\delta_c}\right)\right)\right) & \epsilon_c \leq x \leq \epsilon_c + \delta_c \\ 1 & \epsilon_c + \delta_c \leq x \end{cases} \tag{4.47}$$

with tunable hyperparameters $\epsilon_c, \delta_c$. Consider again the $SO(3)$-equivariant quantity

$$\boldsymbol{G}_{\chi_M} := \left[\tilde{c}_{1s} \| \tilde{c}_{2s} \| \tilde{c}_{2p_x} \| \tilde{c}_{2p_y} \| \tilde{c}_{2p_z}\right] \equiv \begin{pmatrix} \tilde{c}_{1s,1} & \tilde{c}_{1s,2} & & \tilde{c}_{1s,M} \\ \tilde{c}_{2s,1} & \tilde{c}_{2s,2} & & \tilde{c}_{2s,M} \\ \tilde{c}_{2p_x,1} & \tilde{c}_{2p_x,2} & \dots & \tilde{c}_{2p_x,M} \\ \tilde{c}_{2p_y,1} & \tilde{c}_{2p_y,2} & & \tilde{c}_{2p_y,M} \\ \tilde{c}_{2p_z,1} & \tilde{c}_{2p_z,2} & & \tilde{c}_{2p_z,M} \end{pmatrix} \in \mathbb{R}^{5 \times M} \tag{4.48}$$

as in eq. 4.8. Right-multiplying $\boldsymbol{g}_{\chi_v}$ with any $M \times M$-matrix does not interfere with $SO(3)$-equivariance. We thus preprocess $\boldsymbol{g}_{\chi_M}$ as

$$\boldsymbol{g}_{\chi_v} \mapsto \boldsymbol{g}_{\chi_v} \mathrm{diag}(f_{\epsilon_c,\delta_c}(C_1), ..., f_{\epsilon_c,\delta_c}(C_m)) \tag{4.49}$$

with $C_j = \sum_{nlm} c^2_{nlm,j}$ denoting the column-wise norm of $\boldsymbol{g}_{\chi_v}$ as in eq. 4.12. This preprocess filters out negligibly small atomic contributions and thus ensures locality in our initial embedding aggregation scheme as described in section 4.1.

# 5. Numerical experiments

We provide a series of numerical experiments to firstly validate the adherence MFCNet to its proclaimed mathematical desiderata and secondly to probe the accuracy of MFCNet. Concretely, we demonstrate the Hartree-Fock coefficients to showcase $SO(3)$-equivariace. We furthermore underscore the relevance of HF-$\mathbb{Z}_2$-invariance by illustrating that Hartree-Fock precalculations apparently exhibit some ambiguity with respect to the HF-$\mathbb{Z}_2$-phase, thus necessitating MFCNet to be designed in an explicitly HF-$\mathbb{Z}_2$-invariant fashion to eradicate that ambiguity. We then proceed to benchmark the accuracy of MFCNet for two molecules and discuss potential shortcomings. For the generation of all training data we utilize the Python-based quantum chemistry framework PySCF [117]. Furthermore, unless otherwise stated, we settle on using the minimal STO-3G basis for all calculations for reasons of computational efficiency.

## 5.1. Characteristics and behaviour of the Hartree-Fock precalculation

As the molecular orbital coefficients obtained by a Hartree-Fock calculation form one of the centerpieces of MFCNet, we investigate the characteristics of these coefficients for the exemplary molecule ammonia. We conduct a Hartree-Fock calculation close to ammonia's typical bond length of 1.008 Å [118] to obtain some exemplary physically plausible molecular orbitals. In the STO-3G basis, each of the three hydrogen atoms gets equipped with a single 1s-orbital only, whereas the nitrogen atom gets equipped with a 1s-orbital, a 2s-orbital and a 1p-orbital, yielding in total the atomic orbitals 1s, 2s, $1p_x$, $1p_y$, $1p_z$.

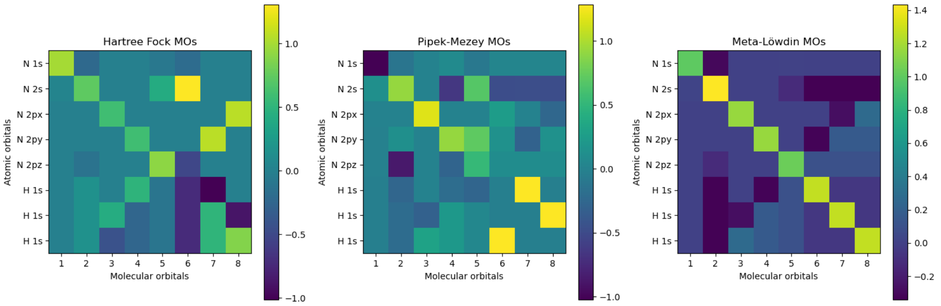

Figure 5.1.: Heatmaps of the molecular orbital coefficients for the vanilla Hartree-Fock calculation as well as the Pipek-Mezey localization scheme [47] and the Meta-Löwdin localization scheme [119].

The Pipek-Mezey and Meta-Löwdin localization respectively differ significantly in their degree of their localization. This comes as no surprise, as the Meta-Löwdin localization produces a set of molecular orbitals that has maximal overlap with the non-orthogonal set of atomic orbitals, yielding the markedly diagonal coefficient matrix. The Pipek-Mezey localization does not alter the coefficients as fundamentally as the Meta-Löwdin localization does, which we attribute to the small size of the molecule.

We furthermore investigate the behaviour of the molecular orbital coefficients $\boldsymbol{C}$ upon rotation of the molecule's coordinates $\vec{\boldsymbol{R}}$ by some rotation matrix $\boldsymbol{O}$. We hypothesize the molecular orbital coefficients to transform equivariantly to the molecular coordinates

$$\vec{\boldsymbol{R}} \to \boldsymbol{O}\vec{\boldsymbol{R}} \implies \boldsymbol{C} \to \boldsymbol{D}(\boldsymbol{O})\boldsymbol{C} \text{ with } \boldsymbol{D}(\boldsymbol{O}) = \boldsymbol{D}^{(0)}(\boldsymbol{O}) \oplus \boldsymbol{D}^{(1)}(\boldsymbol{O}) \oplus \bigoplus_{i=1}^{3} \boldsymbol{D}^{(0)}(\boldsymbol{O}) \tag{5.1}$$

with the Wigner-D matrices $\boldsymbol{D}^{(l)}(\boldsymbol{O})$. The structure of the direct sum of $\boldsymbol{D}(\boldsymbol{O})$ follows from the ordering of the atomic orbitals as can be seen in fig. 5.7. We thus benchmark the molecular coefficients calculated for a rotated molecule against the molecular coefficients manually rotated via $\boldsymbol{D}^{(l)}(\boldsymbol{O})$.

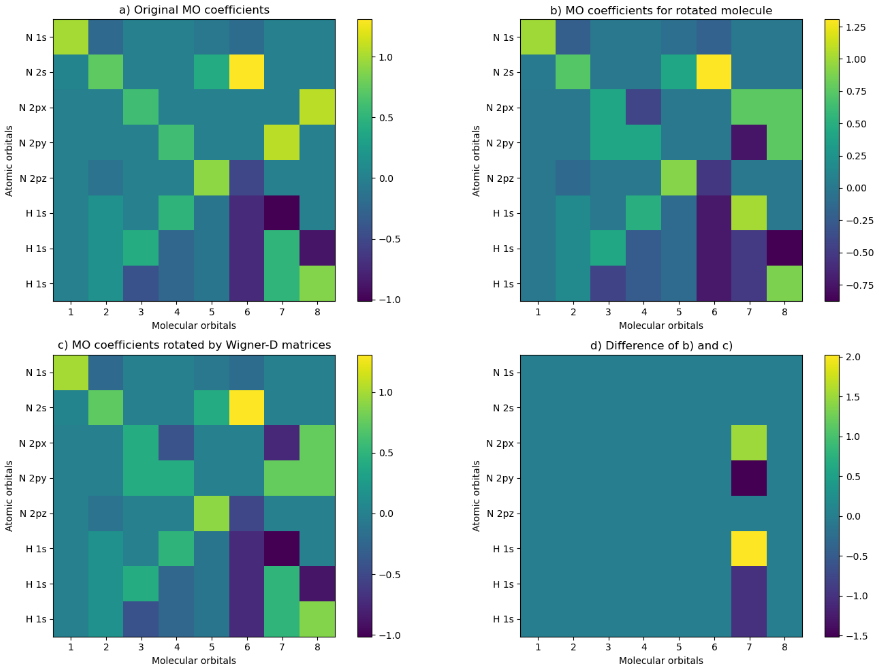

Figure 5.2.: Molecular coefficients $\boldsymbol{C}$ for a) the original geometry, b) the geometry rotated by $\boldsymbol{O}$ and c) as obtained by $\boldsymbol{D}^{(l)}(\boldsymbol{O})\boldsymbol{C}$.

For almost all molecular orbitals we recognize the coefficients from b) and d) to be in perfect agreement with the exception of molecular orbital 7. The divergence for molecular orbital 7 however only results from a global phase difference of -1 between the seventh column in b) and c), thus underscoring that molecular coefficients as calculated via the Hartree-Fock method indeed vary by phase factors even for identical physical systems.

## 5.2. Accuracy of MFCNet

We now proceed to evaluate the accuracy of MFCNet for sufficiently interesting benchmark molecules. We first showcase the correctness of our implementation as well as the adherence to one of our main tenants, namely $SO(3)$-invariance, for the water molecule. Afterwards, we choose the nitrogen dimer $N_2$ as the main testbed, which has been the focal point of many investigations of NN-WFs due to its notorious recalcitrance to conventional quantum chemistry methods [120], and where previous NN-WFs have shined, achieving the most accurate results ever recorded [7]. Foreshadowing results, which are far from competitive with the state-of-the-art, we do not test MFCNet on a battery of other molecules but rather try to elucidate which components of MFCNet seem to be bottlenecks by testing various truncated versions of MFCNet.

### 5.2.1. Training specifics

In the spirit of all existing NN-WFs [75, 8, 80] we conduct a pretraining to ensure proper convergence of the wave-function and to avoid getting stuck in local minima. As in [80] we choose a configuration-interaction-singles-doubles (CISD) calculation resembling a selected configuration interaction scheme with only singly and doubly excited determinants,

$$|\Psi\rangle_{\text{CISD}} = \sum_{\boldsymbol{x} \in \mathcal{X}_1 \cup \mathcal{X}_2} \psi(\boldsymbol{x}) \, |\boldsymbol{x}\rangle \tag{5.2}$$

with $\mathcal{X}_i$ the set of $i$-fold excited determinants. Again following [80] we train MFCNet as a pretraining routine using as loss function the negative logarithm of the fidelity between the wave-function calculated by MFCNet

$$|\Psi\rangle_{\text{MFC}} = \sum_{\boldsymbol{x}} \text{MFCNet}(\boldsymbol{x}; \vec{\boldsymbol{R}}_M, \boldsymbol{c}) \, |\boldsymbol{x}\rangle \tag{5.3}$$

and the CISD wave-function, i.e. the loss $\mathcal{L}_{\text{pre}}$ is given by

$$\mathcal{L}_{\text{pre}}(\boldsymbol{\theta}) = -\ln \frac{| \langle \Psi_{\boldsymbol{\theta}} | \Psi_T \rangle |^2}{\langle \Psi_{\boldsymbol{\theta}} | \Psi_{\boldsymbol{\theta}} \rangle \, \langle \Psi_T | \Psi_T \rangle} \tag{5.4}$$

We implemented all components in TensorFlow [121]. All molecule-wise operations are implemented in a parallelized way, enabling to parallelize batch calculations over multiple

| | **Hyperparameter** | **Value** |
|---|---|---|
| Pretraining | Steps | 500 |
| | Basis | STO-3G |
| | Method | CISD |
| MFCNet | F | 10 |
| | $l_{max}$ | 2 |
| | $n_{int}$ | 2 |
| | k | 60 |
| | Total parameters | 183.630 |
| Optimization | Steps | 10000 |
| | $N_u$ | 4096 |
| | Optimizer | Adam |
| | Learning rate | 10e-4 |
| | Max grad norm | 10 |

Table 5.1.: The defaul choice of hyperparameters for MFCNet.

processing units. During gradient descent we employ gradient clipping [122] with a maximal gradient norm of 10. We use the ADAM optimizer [112] with a learning rate of 10e-4. All experiments are run on AMD Epyc 7402 CPU with 24 cores and a frequency of 2.80GHz. While training on GPUs was planned, this could not be realized in practice due to technical difficulties. This did not prove problematic for the small investigated chemical systems, with modest training times in the order of magnitude of one hour.

All the default values for hyperparameters for the training process as well as all those characterising MFCNet are displayed in table 5.1.

To test the expressiveness of MFCNet we benchmark it against a self-implemented version of NNBF [80] which consists of a plain MLPs predicting $\psi(\boldsymbol{x};\boldsymbol{\theta})$ based on the binary input string $\boldsymbol{x}$ with a parameter count of 143.128, hence roughly equalling MFCNet in parameter volume. We did not have access to the original implementation of NNBF as we were unfortunately unable to contact the authors successfully. We hope to elucidate whether the significant amount of constraints we impose on the neural architecture compromises its expressiveness compared to the very general neural architecture of NNBF which does not encode any domain constraints. The hyperparameters chosen for NNBF are identical to those listed in the publication [80].

### 5.2.2. Results for $H_2O$

**Characteristics of the data set:** We generate a set of $H_2O$ geometries using the equilibrium geometry of water [9] as a starting point, with an angle of $\sphericalangle H_1OH_2 = 101.76^\circ$ and with bond lengths $b = \overline{OH_1} = \overline{OH_1} = 0.96$ Å. We create an ensemble of geometries with different bond lengths in the range $0.5b$ to $2.8b$ all with the same angle. In the STO-3G basis, the full

Slater-determinant Hilbert space of $H_2O$ consists of 441 determinants. As the molecule is tiny, we assume all feature vectors to be neighbours to each other.

**Results for pointwise training:** We first present the results when retraining the model for every molecular geometry as state-of-the-art 2Q-NN-WFs do. This serves rather as a sanity check than an actual benchmark, considering that the FCI result is obtained with 441 parameters only. For such a small system, we can furthermore use the full Hilbert space as core space $\mathcal{V}^n$ during the optimization process, cf. section 3.4. Hence, the minimal value obtained during the training process is the best approximation to the ground-state MFCNet is able to achieve with the specific choice of hyperparameters. We additionally plot the FCI energies, as they constitute the best result attainable with any given basis set and thus serve as ground truth.

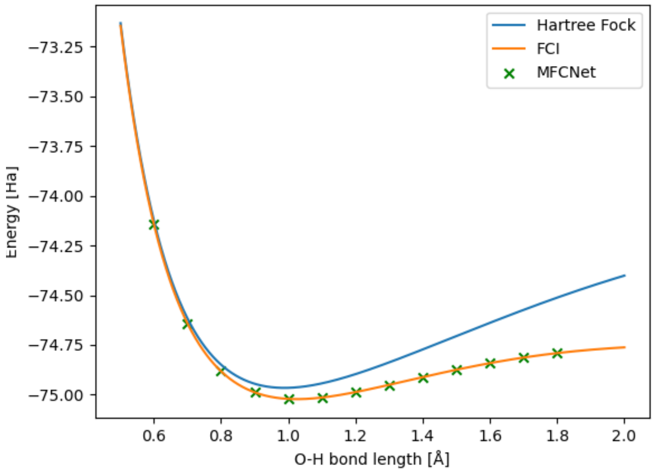

Figure 5.3.: Energies calculated by MFCNet in Hartrees for a series of different bond lengths. For each point a separate model was trained.

While the results of MFCNet seem to pass the test of visual inspection, they dishearteningly already fail to achieve chemical accuracy even for the primitive system of $H_2O$ with 441 Slater-determinants with a mean difference of $\Delta(E_{\mathrm{MFC}} - E_{\mathrm{FCI}}) = 2.8\mathrm{mHa}$, which is significantly higher than the threshold of 1.6 mHa for chemical accuracy. We will discuss empirical shortcomings at a later stage.

**Results for training on multiple geometries simultaneously:** We additionally train MFCNet on multiple geometries simultaneously, i.e. one model with one set of parameters is asked to predict the wave-functions of multiple geometries, hence of multiple Hamiltonians simultaneously. To the best of our knowledge, no counterpart in conventional quantum chemistry exists for such a model [56]. The plethora of existing 2Q-NN-WFs have also hitherto not attempted to construct a model with the ability to calculate wave-functions across different geometries.

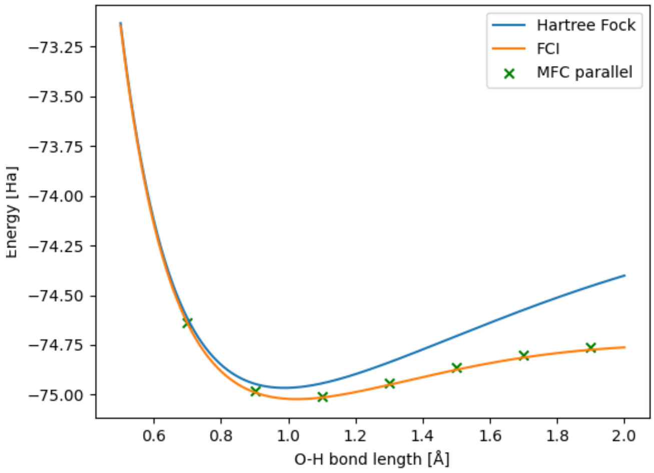

Figure 5.4.: Energies calculated for various bond lengths when using a single MFCNet model.

While the results still coincide roughly with the FCI energies, numerical inspection reveals a markedly worsened accuracy, with an average deviation of the FCI energies of $\Delta(E_{\text{MFC}} - E_{\text{FCI}}) = 8.9\text{mHa}$, or more than five times the threshold for chemical accuracy. In this setup, alle the bond lengths marked in 5.4 were employed in the training procedure. Ideally we would also like our model to perform well on similar but different geometries compared to those in the training set, something we - slighlty inaccurately - refer to as out-of-distribution-training. While our model is in principle able to predict the wave-functions of unseen geometries, due to unforeseen technical difficulties we were not able to extract and analyse the results of out-of-distribution training.

**Results for rotated geometries:** We train MFCNet on four rotated geometries simultaneously, expecting identical results for each geometry.

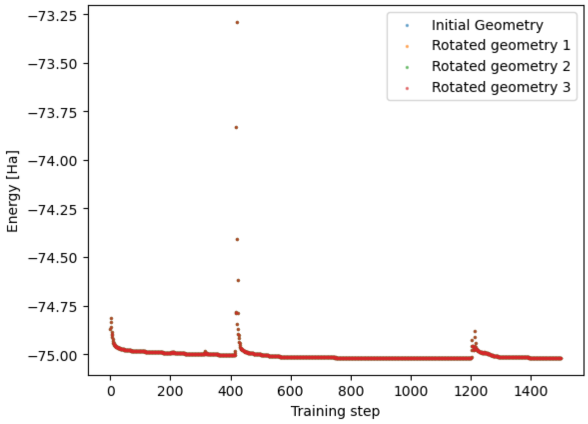

Figure 5.5.: Energies calculated by MFCNet for four rotated geometries in parallel. As they align and overlay essentially perfectly, mostly the color corresponding to geometry 3 is visible.

As we can see, the energies coincide almost exactly, serving as strong numerical evidence that indeed MFCNet calculates rotationally invariant amplitudes. To drive this point home, we calculate for each step the maximal energetic deviation between two of the four geometries.

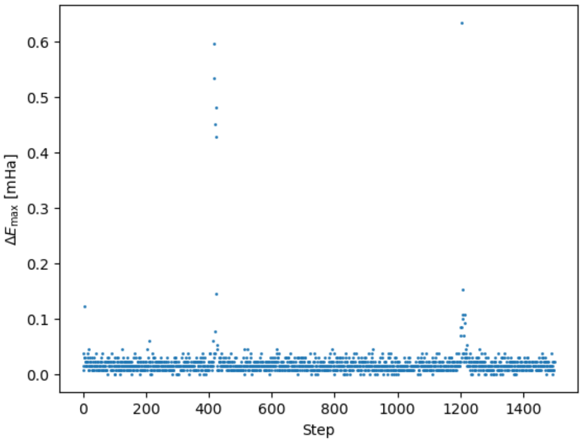

Figure 5.6.: Maximal difference between two of the four energies for each training step.

Even for outliers this maximal difference lies well within the range of chemical accuracy. We thus conclude once more that MFCNet indeed calculates $SO(3)$-invariant amplitudes.

### 5.2.3. Results for $N_2$

**Characteristics of the data set:** We generate a set of geometries with bonds lengths $b \in [0.5, 2.5]$ of the nitrogen dimer sampled from its dissociation curve [123], each of which is characterised by the distance between the two nitrogen atoms. The Hilbert space spanned by all Slater-determinants has dimension 14400 for $N_2$ in the STO-3G basis. As for $H_2O$, we assume all feature vectors to be neighbours to each other.

**Results for pointwise training:** We again train separate MFCNet models for a series of geometries with differing bond lengths. As further benchmark, we calculate the same set of energies with our implementation of NNBF [80].

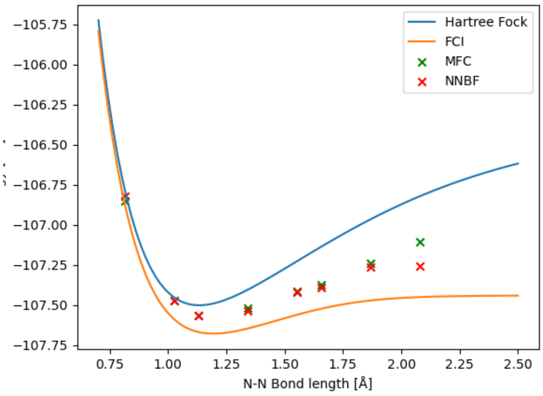

Figure 5.7.: Energies calculated for a series of differing bond lengths as calculated by MFCNet and NNBF.

We observe MFCNet to achieve quite poorly, capturing roughly half of the correlation energy or even less. Even more concerningly, our NNBF implementation also performs dismally, even though we know NNBF to usually achieve accuracies well in the range of chemical accuracy [80]. We will return to this observation in the discussion.

**Results for training on multiple geometries simultaneously:** As for $H_2O$ we train MFCNet for multiple geometries simultaneously.

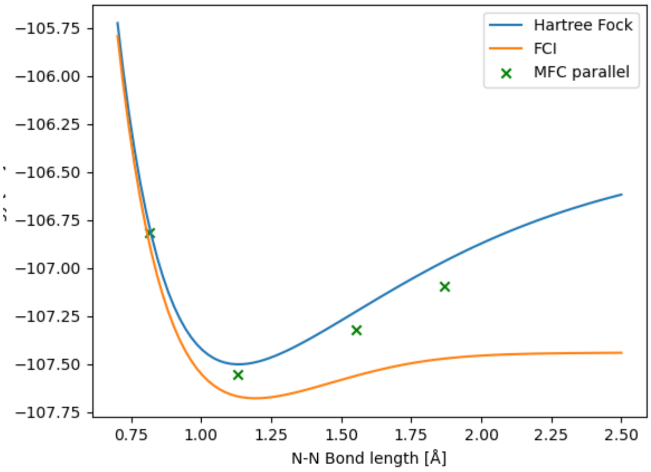

Figure 5.8.: Energies calculated for various bond lengths when using a single MFCNet model.

We again observe MFCNet to perform quite poorly and, consistent with the results of $H_2O$, showcasing a drop in performance when comparing joint training to pointwise training.

### 5.2.4. Hyperparameter tuning

We tested various choices of hyperparameters to finetune the architecture of MFCNet. As even training on a single point only turned out to be in no way competitive with state-of-the-art methods, we decided to some bond length - in our case $b = 1.35$Å - as a single training point and try to boost performance for that task.

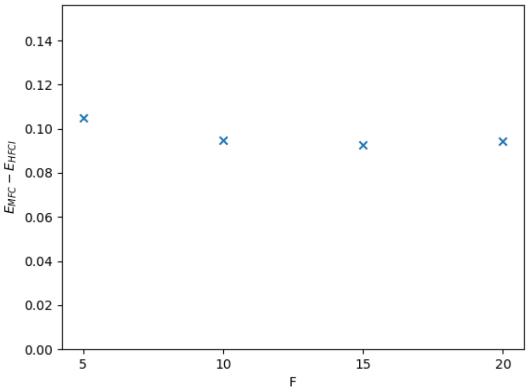

Figure 5.9.: Energies calculated by MFCNet for various choices of $F$ benchmarked against the default valued of $F = 10$. The y-axis is chosen in a way such that $y_{\max} = E_{\mathrm{HF}} - E_{\mathrm{FCI}}$.

For the number of channels employed, $F$, we do not observe significant effects on training success except a small drop at $F = 5$.

We additionally test varying the number of units $k$ of the MLPs which occur at a variety of locations in our architecture, cf. section 4.3.

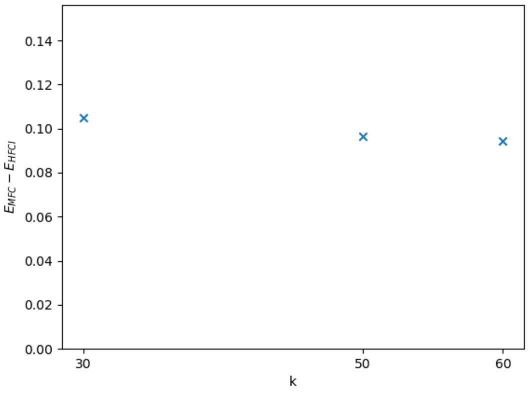

Figure 5.10.: Energies calculated by MFCNet for various choices of $k$ benchmarked against the default valued of $k = 60$. The y-axis is chosen in a way such that $y_{\text{max}} = E_{\text{HF}} - E_{\text{FCI}}$.

Again, except a slight decrease in performance for $k = 30$ we again do not observe significant effects on training outcome.

Lastly, we explore any possible beneficial effects of utilizing more interaction blocks.

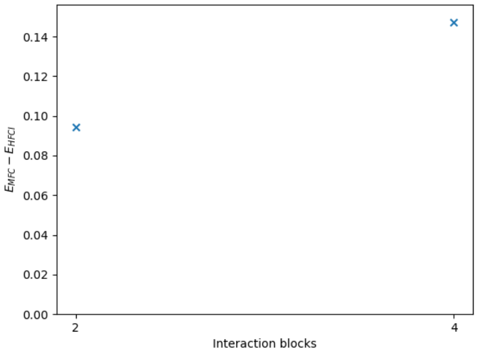

Figure 5.11.: Energies calculated by MFCNet for two and four interactions blocks. The y-axis is chosen in a way such that $y_{\text{max}} = E_{\text{HF}} - E_{\text{FCI}}$.

In fact however, the opposite seems to be case: utilizing four instead of two interaction blocks significantly hinders the optimization process while offering many more parameters.

### 5.2.5. Ablation studies

As our architecture is rather plain, we do not conduct ablation studies in the form of using a modified architecture. Instead, we conduct ablation studies in the form of ablating inputs that are thought to be physically informative.

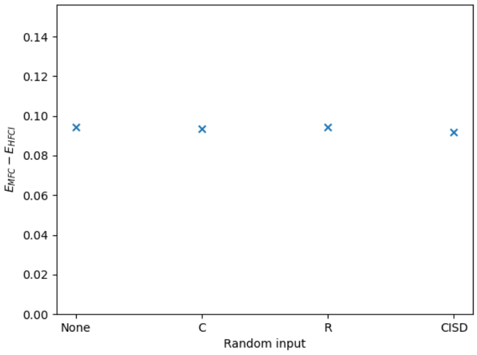

Figure 5.12.: Energies calculated by MFCNet when substituting the respective input quantity with random values. The CISD variable refers to the CISD-amplitudes employed for the pretraining. The y-axis is chosen in a way such that $y_{\text{max}} = E_{\text{HF}} - E_{\text{FCI}}$.

Disappointingly, there does not seem to be any deleterious effect on MFCNet's performance when comparing actual physical inputs to randomly generated noise.

All in all, little empirical support could be garnered for MFCNet in its current setup. With no significant dependence of hyperparameters or inputs discernible, we did not saw a very clear avenue to further pursue. We thus refrained from further testing, assuming major overhauls to the architecture or implementation details necessary to escape the rut.

# 6. Discussion and outlook

In this thesis we proposed the neural architecture for a ground state neural wave-function for quantum chemical Hamiltonians based on the state-of-the-art and ongoing lines of research. We aimed to design a neural architecture that recognizes rotated molecular geometries and treats them indifferently, as physical intuition dictates. We furthermore aimed for a design that capitalizes on the predominantly local nature of electronic correlation, learning essential features of electronic correlation rather than fitting to specific geometries. It was hoped that such a neural net would be able to transfer that knowledge to similar but different geometries, enabling to calculate the ground state wave-functions for different Hamiltonians simultaneously by one trained neural network.

We tested our proposed neural wave-function MFCNet on simple systems, firstly to validate its adherence to the desiderata we defined and secondly to evaluate its performance for relevant use cases. While indeed MFCNet incorporates geometric information in a way that enables it to recognize rotational symmetries and to treat rotated geometries on equal footing, its abysmal accuracy renders it far from competitive with state-of-the-art neural wave-functions. We tried to elucidate which hyperparameters matter most as determinants of MFCNet's accuracy, however to no avail, as we could not identify significant influences on training accuracy for our choices of hyperparameters. Even more disconcertingly, we observed MFCNet to perform equally well, regardless whether actual physical parameters or random values were used as input.

**Is there a baby in the bathwater?** All the concerns raised in the last paragraph seriously beg the question whether MFCNet is in fact a doomed neural architecture. Especially the seemingly complete independence on the quality of the input data strongly insinuates that the current architecture of MFCNet is not able to develop any grasp on the underlying physics. It is however possible of course that it was simply the type of physical information that was included, with the Hartree-Fock coefficients as their centerpiece, that proved to be not informative enough.

A third option also cannot be ruled out, an option which implies that fatal verdicts on MFCNet's core design features may be premature. Along the way there were a series of numerical anomalies showing up, for many of which we could not get to the bottom of. For example, significant numerical problems arose during training from plain TensorFlow functions such as its determinant function and norm function, prompting us to manually implement numerically stable versions of said functions. Still, divergent behaviour and exploding gradients of MFCNet and especially our implementation of NNBF was relatively frequent. While we cannot rule out similar troubles being existent but undocumented in published neural wave-functions, we doubt this to be the case. Even more concerningly, we observed strong deviations from published literature in a couple of regards. Most prominently, our implementation of NNBF was - by the standards of a quantum chemist - orders of magnitude more imprecise than its official progenitor [80]. Though we tried to meticulously follow the publication for our implementation, lacking access to the original code, we assume unnoticed bugs to be the culprit for the strongly deviant behaviour of our code. Similarly, whereas the authors in [79] use an empirically quite faithful approximation to the exact energy during the optimisation procedure, the same approximation yields wildly implausible results in our implementation. This again to points to unnoticed bugs in our code.

All in all, issues with the current version of MFCNet seem to be manifold. We are hopeful that some of the more promising design features, such as the successful implementation of equivariance as well as the formulation via a graph neural network, enabling at least in theory the model to be transferable across geometries, are simply overshadowed by the flaws in the more mechanical aspects inherent in the optimisation procedure. It does not seem too far-fetched to imagine a polished version of MFCNet perform on par with state-of-the-art neural wave-functions while still exhibiting our core design tenets of equivariance and locality. In fact, on paper MFCNet constitutes mostly a collage of best-practices from different lines of research as they were perceived by us - equivariant message-passing graph neural networks as proper architecture to exploit spatial symmetries, selected-configuration optimisation schemes as conducted in state-of-the-art investigations. No overly experimental features with unknown effects on performance were implemented.

We thus maintain a modest optimism regarding the potential of MFCNet and believe that, coupled with strong expertise in the training of deep neural networks, the baby that actually is in the bathwater might one day be nurtured to live out its full potential.

# A. Appendix

## A.1. Size-consistency of neural wave-functions in first quantization

We now skim why we deem wave-functions of the form

$$\Psi(\vec{\mathbf{r}};\boldsymbol{\theta}) = e^{J(\vec{\mathbf{r}};\boldsymbol{\theta})} \sum_{k=1}^{N_{\text{det}}} c_k \begin{vmatrix} \phi_1^k(\mathbf{r}_1;\vec{\mathbf{r}};\boldsymbol{\theta}) & \dots & \phi_N^k(\mathbf{r}_1;\vec{\mathbf{r}};\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \phi_1^k(\mathbf{r}_N;\vec{\mathbf{r}};\boldsymbol{\theta}) & \dots & \phi_N^k(\mathbf{r}_N;\vec{\mathbf{r}};\boldsymbol{\theta}) \end{vmatrix} = e^{J(\vec{\mathbf{r}};\boldsymbol{\theta})} \sum_{k=1}^{N_{\text{det}}} c_k \det[\phi_{mn}^k] \tag{A.1}$$

not suited for size-consistent wave-functions. Consider two molecules $A_1$ and $A_2$ far apart. The authors in [11] construct the orbitals in a way such the matrices $\boldsymbol{\phi}^k$ are diagonal as

$$\boldsymbol{\phi}^k = \begin{pmatrix} {}^1\boldsymbol{\phi}^k & 0 \\ 0 & {}^2\boldsymbol{\phi}^k \end{pmatrix} \tag{A.2}$$

where ${}^1\boldsymbol{\phi}^k$ denotes the block of orbitals corresponding to $A_1$ and ${}^2\boldsymbol{\phi}^k$ denotes the block of orbitals belonging to $A_2$. We denote the positions of the electrons belonging to A as $\vec{\boldsymbol{r}}_A$, the positions of the electrons belonging to B as $\vec{\boldsymbol{r}}_B$ and the total wave-function $\Psi(\vec{\boldsymbol{r}}_A, \vec{\boldsymbol{r}}_B)_{AB}$. For the case of non-interacting molecules we expect the electronic density $\rho(\vec{\boldsymbol{r}}_A, \vec{\boldsymbol{r}}_B)_{AB} = |\Psi(\vec{\boldsymbol{r}}_A, \vec{\boldsymbol{r}}_B)_{AB}|^2$ to factorize as $\rho(\vec{\boldsymbol{r}}_A, \vec{\boldsymbol{r}}_B)_{AB} = \rho(\vec{\boldsymbol{r}}_A)_A \rho(\vec{\boldsymbol{r}}_B)_B$ for some $\rho(\vec{\boldsymbol{r}}_A)_A$ and $\rho(\vec{\boldsymbol{r}}_B)_B$. However, a wave-function $\Psi(\vec{\boldsymbol{r}}_A, \vec{\boldsymbol{r}}_B)_{AB}$ of the form eq. A.1 yields a density

$$|\Psi(\vec{\boldsymbol{r}}_A, \vec{\boldsymbol{r}}_B)_{AB}|^2 = e^{2J(\vec{\mathbf{r}};\boldsymbol{\theta})} \sum_{k_1,k_2=1}^{N_{\text{det}}} \rho_A^{(k_1)}(\vec{\boldsymbol{r}}_A) \rho_B^{(k_2)}(\vec{\boldsymbol{r}}_B) \tag{A.3}$$

which in general for $N_{\text{det}} > 1$ does not factorize, though it should. Hence $\Psi(\vec{\boldsymbol{r}}_A, \vec{\boldsymbol{r}}_B)_{AB}$ was not a proper size-consistent wave-function to begin with.

# List of Tables

# Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60 (2012), pp. 84–90. URL: https://api.semanticscholar.org/CorpusID:195908774.

[2] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”. In: *Science* 362.6419 (Dec. 2018), pp. 1140–1144. ISSN: 1095-9203. DOI: 10.1126/science.aar6404. URL: http://dx.doi.org/10.1126/science.aar6404.

[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (July 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: http://dx.doi.org/10.1038/s41586-021-03819-2.

[4] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu. “Explaining deep neural networks: A survey on the global interpretation methods”. en. In: *Neurocomputing* 513 (Nov. 2022), pp. 165–180.

[5] K. Hornik, M. Stinchcombe, and H. White. “Multilayer feedforward networks are universal approximators”. In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366. ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8. URL: http://dx.doi.org/10.1016/0893-6080(89)90020-8.

[6] G. Carleo and M. Troyer. “Solving the quantum many-body problem with artificial neural networks”. In: *Science* 355.6325 (Feb. 2017), pp. 602–606. ISSN: 1095-9203. DOI: 10.1126/science.aag2302. URL: http://dx.doi.org/10.1126/science.aag2302.

[7] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes. “Ab initio solution of the many-electron Schrödinger equation with deep neural networks”. In: *Physical Review Research* 2.3 (Sept. 2020). ISSN: 2643-1564. DOI: 10.1103/physrevresearch.2.033429. URL: http://dx.doi.org/10.1103/PhysRevResearch.2.033429.

[8] J. Hermann, Z. Schätzle, and F. Noé. “Deep-neural-network solution of the electronic Schrödinger equation”. In: *Nature Chemistry* 12.10 (Sept. 2020), pp. 891–897. ISSN: 1755-4349. DOI: 10.1038/s41557-020-0544-y. URL: http://dx.doi.org/10.1038/s41557-020-0544-y.

[9] K. Choo, A. Mezzacapo, and G. Carleo. "Fermionic neural-network states for ab-initio electronic structure". en. In: *Nat. Commun.* 11.1 (May 2020), p. 2368.

[10] M. Scherbela, R. Reisenhofer, L. Gerard, P. Marquetand, and P. Grohs. *Solving the electronic Schrödinger equation for multiple nuclear geometries with weight-sharing deep neural networks*. 2021. DOI: `10.48550/ARXIV.2105.08351`. URL: `https://arxiv.org/abs/2105.08351`.

[11] N. Gao and S. Günnemann. *Generalizing Neural Wave Functions*. 2023. DOI: `10.48550/ARXIV.2302.04168`. URL: `https://arxiv.org/abs/2302.04168`.

[12] L. Gerard, M. Scherbela, P. Marquetand, and P. Grohs. "Gold-standard solutions to the Schrödinger equation using deep learning: How much physics do we need?" In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho. 2022. URL: `https://openreview.net/forum?id=nX-gReQ0OT`.

[13] J. Gasteiger, J. Groß, and S. Günnemann. *Directional Message Passing for Molecular Graphs*. 2020. DOI: `10.48550/ARXIV.2003.03123`. URL: `https://arxiv.org/abs/2003.03123`.

[14] M. Scherbela, L. Gerard, and P. Grohs. "Towards a transferable fermionic neural wavefunction for molecules". In: *Nature Communications* 15.1 (Jan. 2024). ISSN: 2041-1723. DOI: `10.1038/s41467-023-44216-9`. URL: `http://dx.doi.org/10.1038/s41467-023-44216-9`.

[15] A. Szabo and N. S. Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[16] M. Born and R. Oppenheimer. "Zur Quantentheorie der Molekeln". In: *Annalen der Physik* 389.20 (Jan. 1927), pp. 457–484. ISSN: 1521-3889. DOI: `10.1002/andp.19273892002`. URL: `http://dx.doi.org/10.1002/andp.19273892002`.

[17] J. Hermann, J. Spencer, K. Choo, A. Mezzacapo, W. M. C. Foulkes, D. Pfau, G. Carleo, and F. Noé. "Ab initio quantum chemistry with neural-network wavefunctions". In: *Nature Reviews Chemistry* 7.10 (Aug. 2023), pp. 692–709. ISSN: 2397-3358. DOI: `10.1038/s41570-023-00516-8`. URL: `http://dx.doi.org/10.1038/s41570-023-00516-8`.

[18] C. Zhou, M. R. Hermes, D. Wu, J. J. Bao, R. Pandharkar, D. S. King, D. Zhang, T. R. Scott, A. O. Lykhin, L. Gagliardi, and D. G. Truhlar. "Electronic structure of strongly correlated systems: recent developments in multiconfiguration pair-density functional theory and multiconfiguration nonclassical-energy functional theory". In: *Chemical Science* 13.26 (2022), pp. 7685–7706. ISSN: 2041-6539. DOI: `10.1039/d2sc01022d`. URL: `http://dx.doi.org/10.1039/d2sc01022d`.

[19] C. Sherrill. "Chapter 4 Bond Breaking in Quantum Chemistry". In: *Annual Reports in Computational Chemistry*. Elsevier, 2005, pp. 45–56. DOI: `10.1016/s1574-1400(05)01004-2`. URL: `http://dx.doi.org/10.1016/S1574-1400(05)01004-2`.

[20] C. L. Benavides-Riveros, N. N. Lathiotakis, and M. A. L. Marques. "Towards a formal definition of static and dynamic electronic correlations". In: *Physical Chemistry Chemical Physics* 19.20 (2017), pp. 12655–12664. ISSN: 1463-9084. DOI: `10.1039/c7cp01137g`. URL: `http://dx.doi.org/10.1039/C7CP01137G`.

[21] D. Prendergast, M. Nolan, C. Filippi, S. Fahy, and J. C. Greer. "Impact of electron–electron cusp on configuration interaction energies". In: *The Journal of Chemical*

*Physics* 115.4 (July 2001), pp. 1626–1634. ISSN: 1089-7690. DOI: 10.1063/1.1383585. URL: http://dx.doi.org/10.1063/1.1383585.

[22] R. Jastrow. "Many-Body Problem with Strong Forces". In: *Physical Review* 98.5 (June 1955), pp. 1479–1484. ISSN: 0031-899X. DOI: 10.1103/physrev.98.1479. URL: http://dx.doi.org/10.1103/PhysRev.98.1479.

[23] N. D. Drummond, M. D. Towler, and R. J. Needs. "Jastrow correlation factor for atoms, molecules, and solids". In: *Physical Review B* 70.23 (Dec. 2004). ISSN: 1550-235X. DOI: 10.1103/physrevb.70.235119. URL: http://dx.doi.org/10.1103/PhysRevB.70.235119.

[24] N. R. Walet and R. F. Bishop. "The Unreasonable accuracy of the Jastrow approach in many body physics". In: (July 2003). arXiv: physics/0307069.

[25] B. Nagy and F. Jensen. *Basis Sets in Quantum Chemistry*. Apr. 2017. DOI: 10.1002/9781119356059.ch3. URL: http://dx.doi.org/10.1002/9781119356059.ch3.

[26] J. G. Hill. "Gaussian basis sets for molecular applications". In: *International Journal of Quantum Chemistry* 113.1 (Oct. 2012), pp. 21–34. ISSN: 1097-461X. DOI: 10.1002/qua.24355. URL: http://dx.doi.org/10.1002/qua.24355.

[27] J. C. Slater. "Atomic Shielding Constants". In: *Physical Review* 36.1 (July 1930), pp. 57–64. ISSN: 0031-899X. DOI: 10.1103/physrev.36.57. URL: http://dx.doi.org/10.1103/PhysRev.36.57.

[28] T. Kato. "On the eigenfunctions of many-particle systems in quantum mechanics". In: *Communications on Pure and Applied Mathematics* 10.2 (Jan. 1957), pp. 151–177. ISSN: 1097-0312. DOI: 10.1002/cpa.3160100201. URL: http://dx.doi.org/10.1002/cpa.3160100201.

[29] P. Surján. *Second Quantized Approach to Quantum Chemistry: An Elementary Introduction*. Springer-Verlag, 1989. ISBN: 9783540511373. URL: https://books.google.de/books?id=EvGMQgAACAAJ.

[30] O. Christiansen, H. Koch, P. Jørgensen, and J. Olsen. "Excitation energies of H2O, N2 and C2 in full configuration interaction and coupled cluster theory". In: *Chemical Physics Letters* 256.1–2 (June 1996), pp. 185–194. ISSN: 0009-2614. DOI: 10.1016/0009-2614(96)00394-6. URL: http://dx.doi.org/10.1016/0009-2614(96)00394-6.

[31] H. Gao, S. Imamura, A. Kasagi, and E. Yoshida. "Distributed Implementation of Full Configuration Interaction for One Trillion Determinants". In: *Journal of Chemical Theory and Computation* 20.3 (Feb. 2024), pp. 1185–1192. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.3c01190. URL: http://dx.doi.org/10.1021/acs.jctc.3c01190.

[32] Z. Rolik, Á. Szabados, and P. R. Surján. "A sparse matrix based full-configuration interaction algorithm". In: *The Journal of Chemical Physics* 128.14 (Apr. 2008). ISSN: 1089-7690. DOI: 10.1063/1.2839304. URL: http://dx.doi.org/10.1063/1.2839304.

[33] J. B. Schriber and F. A. Evangelista. "Communication: An adaptive configuration interaction approach for strongly correlated electrons with tunable accuracy". In: *The Journal of Chemical Physics* 144.16 (Apr. 2016). ISSN: 1089-7690. DOI: 10.1063/1.4948308. URL: http://dx.doi.org/10.1063/1.4948308.

[34] N. Liebermann, K. Ghanem, and A. Alavi. "Importance-sampling FCIQMC: Solving weak sign-problem systems". In: *The Journal of Chemical Physics* 157.12 (Sept. 2022). ISSN: 1089-7690. DOI: 10.1063/5.0107317. URL: http://dx.doi.org/10.1063/5.0107317.

[35] W. Dobrautz. *Development of full configuration interaction quantum Monte Carlo methods for strongly correlated electron systems*. en. 2019. DOI: 10.18419/OPUS-10593. URL: http://elib.uni-stuttgart.de/handle/11682/10610.

[36] A. Glielmo, Y. Rath, G. Csányi, A. De Vita, and G. H. Booth. "Gaussian Process States: A Data-Driven Representation of Quantum Many-Body Physics". In: *Physical Review X* 10.4 (Nov. 2020). ISSN: 2160-3308. DOI: 10.1103/physrevx.10.041026. URL: http://dx.doi.org/10.1103/PhysRevX.10.041026.

[37] J. J. Goings, H. Hu, C. Yang, and X. Li. "Reinforcement Learning Configuration Interaction". In: *Journal of Chemical Theory and Computation* 17.9 (Aug. 2021), pp. 5482–5491. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.1c00010. URL: http://dx.doi.org/10.1021/acs.jctc.1c00010.

[38] R. O. Ramabhadran and K. Raghavachari. "Extrapolation to the Gold-Standard in Quantum Chemistry: Computationally Efficient and Accurate CCSD(T) Energies for Large Molecules Using an Automated Thermochemical Hierarchy". In: *Journal of Chemical Theory and Computation* 9.9 (Aug. 2013), pp. 3986–3994. ISSN: 1549-9626. DOI: 10.1021/ct400465q. URL: http://dx.doi.org/10.1021/ct400465q.

[39] T. D. Crawford and H. F. Schaefer. *An Introduction to Coupled Cluster Theory for Computational Chemists*. Jan. 2000. DOI: 10.1002/9780470125915.ch2. URL: http://dx.doi.org/10.1002/9780470125915.ch2.

[40] D. Cremer. "Møller–Plesset perturbation theory: from small molecule methods to methods for thousands of atoms". In: *WIREs Computational Molecular Science* 1.4 (May 2011), pp. 509–530. ISSN: 1759-0884. DOI: 10.1002/wcms.58. URL: http://dx.doi.org/10.1002/wcms.58.

[41] R. K. Nesbet. *Electronic Correlation in Atoms and Molecules*. Jan. 1965. DOI: 10.1002/9780470143551.ch4. URL: http://dx.doi.org/10.1002/9780470143551.ch4.

[42] Z. Ni, Y. Guo, F. Neese, W. Li, and S. Li. "Cluster-in-Molecule Local Correlation Method with an Accurate Distant Pair Correction for Large Systems". In: *Journal of Chemical Theory and Computation* 17.2 (Jan. 2021), pp. 756–766. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.0c00831. URL: http://dx.doi.org/10.1021/acs.jctc.0c00831.

[43] M. Schwilk, D. Usvyat, and H.-J. Werner. "Communication: Improved pair approximations in local coupled-cluster methods". In: *The Journal of Chemical Physics* 142.12 (Mar. 2015). ISSN: 1089-7690. DOI: 10.1063/1.4916316. URL: http://dx.doi.org/10.1063/1.4916316.

[44] N. Ben Amor, S. Evangelisti, T. Leininger, and D. Andrae. "Local Orbitals in Quantum Chemistry". In: *Basis Sets in Computational Chemistry*. Springer International Publishing, 2021, pp. 41–101. ISBN: 9783030672621. DOI: 10.1007/978-3-030-67262-1_3. URL: http://dx.doi.org/10.1007/978-3-030-67262-1_3.

[45] J. J. P. Stewart. "An examination of the nature of localized molecular orbitals and their value in understanding various phenomena that occur in organic chemistry". In:

*Journal of Molecular Modeling* 25.1 (Dec. 2018). ISSN: 0948-5023. DOI: 10.1007/s00894-018-3880-8. URL: http://dx.doi.org/10.1007/s00894-018-3880-8.

[46] J. M. Foster and S. F. Boys. "Canonical Configurational Interaction Procedure". In: *Rev. Mod. Phys.* 32 (2 Apr. 1960), pp. 300–302. DOI: 10.1103/RevModPhys.32.300. URL: https://link.aps.org/doi/10.1103/RevModPhys.32.300.

[47] J. Pipek and P. G. Mezey. "A fast intrinsic localization procedure applicable for ab-initio and semiempirical linear combination of atomic orbital wave functions". In: *The Journal of Chemical Physics* 90.9 (May 1989), pp. 4916–4926. ISSN: 1089-7690. DOI: 10.1063/1.456588. URL: http://dx.doi.org/10.1063/1.456588.

[48] F. Weinhold. "Natural bond orbital analysis: A critical overview of relationships to alternative bonding perspectives". In: *Journal of Computational Chemistry* 33.30 (July 2012), pp. 2363–2379. ISSN: 1096-987X. DOI: 10.1002/jcc.23060. URL: http://dx.doi.org/10.1002/jcc.23060.

[49] O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. "Exploring chemical compound space with quantum-based machine learning". In: *Nature Reviews Chemistry* 4.7 (June 2020), pp. 347–358. ISSN: 2397-3358. DOI: 10.1038/s41570-020-0189-9. URL: http://dx.doi.org/10.1038/s41570-020-0189-9.

[50] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi. "Machine Learning for Molecular Simulation". In: *Annual Review of Physical Chemistry* 71.1 (Apr. 2020), pp. 361–390. ISSN: 1545-1593. DOI: 10.1146/annurev-physchem-042018-052331. URL: http://dx.doi.org/10.1146/annurev-physchem-042018-052331.

[51] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf. "Electron density learning of non-covalent systems". In: *Chemical Science* 10.41 (2019), pp. 9424–9432. ISSN: 2041-6539. DOI: 10.1039/c9sc02696g. URL: http://dx.doi.org/10.1039/c9sc02696g.

[52] J. Gasteiger, F. Becker, and S. Günnemann. *GemNet: Universal Directional Graph Neural Networks for Molecules*. 2021. DOI: 10.48550/ARXIV.2106.08903. URL: https://arxiv.org/abs/2106.08903.

[53] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky. "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials". In: *Nature Communications* 13.1 (May 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-022-29939-5. URL: http://dx.doi.org/10.1038/s41467-022-29939-5.

[54] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. "Machine learning of accurate energy-conserving molecular force fields". In: *Science Advances* 3.5 (May 2017). ISSN: 2375-2548. DOI: 10.1126/sciadv.1603015. URL: http://dx.doi.org/10.1126/sciadv.1603015.

[55] P. Lolur, M. Skogh, W. Dobrautz, C. Warren, J. Biznárová, A. Osman, G. Tancredi, G. Wendin, J. Bylander, and M. Rahm. "Reference-State Error Mitigation: A Strategy for High Accuracy Quantum Computation of Chemistry". In: *Journal of Chemical Theory and Computation* 19.3 (Jan. 2023), pp. 783–789. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.2c00807. URL: http://dx.doi.org/10.1021/acs.jctc.2c00807.

[56] N. Gao and S. Günnemann. *Ab-Initio Potential Energy Surfaces by Pairing GNNs with Neural Wave Functions*. 2021. DOI: 10.48550/ARXIV.2110.05064. URL: https://arxiv.org/abs/2110.05064.

[57] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K.-R. Müller. "SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects". In: *Nature Communications* 12.1 (Dec. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-27504-0. URL: http://dx.doi.org/10.1038/s41467-021-27504-0.

[58] Z. Qiao, A. S. Christensen, M. Welborn, F. R. Manby, A. Anandkumar, and T. F. Miller. "Informing geometric deep learning with electronic interactions to accelerate quantum chemistry". In: *Proceedings of the National Academy of Sciences* 119.31 (July 2022). ISSN: 1091-6490. DOI: 10.1073/pnas.2205221119. URL: http://dx.doi.org/10.1073/pnas.2205221119.

[59] Y. Chen, L. Zhang, H. Wang, and W. E. "Ground State Energy Functional with Hartree–Fock Efficiency and Chemical Accuracy". In: *The Journal of Physical Chemistry A* 124.35 (Aug. 2020), pp. 7155–7165. ISSN: 1520-5215. DOI: 10.1021/acs.jpca.0c03886. URL: http://dx.doi.org/10.1021/acs.jpca.0c03886.

[60] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky. "Learning local equivariant representations for large-scale atomistic dynamics". In: *Nature Communications* 14.1 (Feb. 2023). ISSN: 2041-1723. DOI: 10.1038/s41467-023-36329-y. URL: http://dx.doi.org/10.1038/s41467-023-36329-y.

[61] R. P. Feynman and M. Cohen. "Energy Spectrum of the Excitations in Liquid Helium". In: *Physical Review* 102.5 (June 1956), pp. 1189–1204. ISSN: 0031-899X. DOI: 10.1103/physrev.102.1189. URL: http://dx.doi.org/10.1103/PhysRev.102.1189.

[62] L. F. Tocchio, F. Becca, A. Parola, and S. Sorella. "Role of backflow correlations for the nonmagnetic phase of the Hubbard model". In: *Physical Review B* 78.4 (July 2008). ISSN: 1550-235X. DOI: 10.1103/physrevb.78.041101. URL: http://dx.doi.org/10.1103/PhysRevB.78.041101.

[63] Y. Kwon, D. M. Ceperley, and R. M. Martin. "Effects of three-body and backflow correlations in the two-dimensional electron gas". In: *Physical Review B* 48.16 (Oct. 1993), pp. 12037–12046. ISSN: 1095-3795. DOI: 10.1103/physrevb.48.12037. URL: http://dx.doi.org/10.1103/PhysRevB.48.12037.

[64] D. Luo and B. K. Clark. "Backflow Transformations via Neural Networks for Quantum Many-Body Wave Functions". In: *Physical Review Letters* 122.22 (June 2019). ISSN: 1079-7114. DOI: 10.1103/physrevlett.122.226401. URL: http://dx.doi.org/10.1103/PhysRevLett.122.226401.

[65] M. Hutter. *On Representing (Anti)Symmetric Functions*. 2020. DOI: 10.48550/ARXIV.2007.15298. URL: https://arxiv.org/abs/2007.15298.

[66] S.-i. Amari. "Natural Gradient Works Efficiently in Learning". In: *Neural Computation* 10.2 (Feb. 1998), pp. 251–276. ISSN: 1530-888X. DOI: 10.1162/089976698300017746. URL: http://dx.doi.org/10.1162/089976698300017746.

[67] J. Martens and R. Grosse. *Optimizing Neural Networks with Kronecker-factored Approximate Curvature*. 2015. DOI: 10.48550/ARXIV.1503.05671. URL: https://arxiv.org/abs/1503.05671.

[68] D. Ceperley, G. V. Chester, and M. H. Kalos. "Monte Carlo simulation of a many-fermion study". In: *Physical Review B* 16.7 (Oct. 1977), pp. 3081–3099. ISSN: 0556-2805. DOI: 10.1103/physrevb.16.3081. URL: http://dx.doi.org/10.1103/PhysRevB.16.3081.

[69] F. Becca and S. Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, Nov. 2017. ISBN: 9781316417041. DOI: 10.1017/9781316417041. URL: http://dx.doi.org/10.1017/9781316417041.

[70] J. S. Spencer, D. Pfau, A. Botev, and W. M. C. Foulkes. *Better, Faster Fermionic Neural Networks*. 2020. DOI: 10.48550/ARXIV.2011.07125. URL: https://arxiv.org/abs/2011.07125.

[71] Z. Schätzle, J. Hermann, and F. Noé. "Convergence to the fixed-node limit in deep variational Monte Carlo". In: *The Journal of Chemical Physics* 154.12 (Mar. 2021). ISSN: 1089-7690. DOI: 10.1063/5.0032836. URL: http://dx.doi.org/10.1063/5.0032836.

[72] R. J. Needs, M. D. Towler, N. D. Drummond, P. López Ríos, and J. R. Trail. "Variational and diffusion quantum Monte Carlo calculations with the CASINO code". In: *The Journal of Chemical Physics* 152.15 (Apr. 2020). ISSN: 1089-7690. DOI: 10.1063/1.5144288. URL: http://dx.doi.org/10.1063/1.5144288.

[73] M. Wilson, N. Gao, F. Wudarski, E. Rieffel, and N. M. Tubman. *Simulations of state-of-the-art fermionic neural network wave functions with diffusion Monte Carlo*. 2021. DOI: 10.48550/ARXIV.2103.12570. URL: https://arxiv.org/abs/2103.12570.

[74] W. Ren, W. Fu, X. Wu, and J. Chen. "Towards the ground state of molecules via diffusion Monte Carlo on neural networks". In: *Nature Communications* 14.1 (Apr. 2023). ISSN: 2041-1723. DOI: 10.1038/s41467-023-37609-3. URL: http://dx.doi.org/10.1038/s41467-023-37609-3.

[75] I. von Glehn, J. S. Spencer, and D. Pfau. *A Self-Attention Ansatz for Ab-initio Quantum Chemistry*. 2022. DOI: 10.48550/ARXIV.2211.13672. URL: https://arxiv.org/abs/2211.13672.

[76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[77] J. S. Anderson, F. Heidar-Zadeh, and P. W. Ayers. "Breaking the curse of dimension for the electronic Schrödinger equation with functional analysis". In: *Computational and Theoretical Chemistry* 1142 (Oct. 2018), pp. 66–77. ISSN: 2210-271X. DOI: 10.1016/j.comptc.2018.08.017. URL: http://dx.doi.org/10.1016/j.comptc.2018.08.017.

[78] T. D. Barrett, A. Malyshev, and A. I. Lvovsky. *Autoregressive neural-network wavefunctions for ab initio quantum chemistry*. 2021. DOI: 10.48550/ARXIV.2109.12606. URL: https://arxiv.org/abs/2109.12606.

[79] X. Li, J.-C. Huang, G.-Z. Zhang, H.-E. Li, C.-S. Cao, D. Lv, and H.-S. Hu. "A Non-stochastic Optimization Algorithm for Neural-Network Quantum States". In: *Journal of*

*Chemical Theory and Computation* 19.22 (Nov. 2023), pp. 8156–8165. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.3c00831. URL: http://dx.doi.org/10.1021/acs.jctc.3c00831.

[80] A.-J. Liu and B. K. Clark. *Neural network backflow for ab-initio quantum chemistry*. 2024. DOI: 10.48550/ARXIV.2403.03286. URL: https://arxiv.org/abs/2403.03286.

[81] H. Shang, C. Guo, Y. Wu, Z. Li, and J. Yang. *Solving Schrödinger Equation with a Language Model*. 2023. DOI: 10.48550/ARXIV.2307.09343. URL: https://arxiv.org/abs/2307.09343.

[82] P. Jordan and E. Wigner. "Über das Paulische Äquivalenzverbot". de. In: *Eur. Phys. J. A* 47.9-10 (Sept. 1928), pp. 631–651.

[83] S. B. Bravyi and A. Y. Kitaev. "Fermionic quantum computation". en. In: *Ann. Phys. (N. Y.)* 298.1 (May 2002), pp. 210–226.

[84] T. Zhao, J. Stokes, and S. Veerapaneni. "Scalable neural quantum states architecture for quantum chemistry". In: *Machine Learning: Science and Technology* 4.2 (June 2023), p. 025034. ISSN: 2632-2153. DOI: 10.1088/2632-2153/acdb2f. URL: http://dx.doi.org/10.1088/2632-2153/acdb2f.

[85] T. Helgaker, P. Jørgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. Wiley, Aug. 2000. ISBN: 9781119019572. DOI: 10.1002/9781119019572. URL: http://dx.doi.org/10.1002/9781119019572.

[86] N. Dym and H. Maron. *On the Universality of Rotation Equivariant Point Cloud Networks*. 2020. DOI: 10.48550/ARXIV.2010.02449. URL: https://arxiv.org/abs/2010.02449.

[87] J. Brandstetter, R. Hesselink, E. van der Pol, E. J. Bekkers, and M. Welling. *Geometric and Physical Quantities Improve E(3) Equivariant Message Passing*. 2021. DOI: 10.48550/ARXIV.2110.02905. URL: https://arxiv.org/abs/2110.02905.

[88] P. Müller, V. Golkov, V. Tomassini, and D. Cremers. *Rotation-Equivariant Deep Learning for Diffusion MRI*. 2021. DOI: 10.48550/ARXIV.2102.06942. URL: https://arxiv.org/abs/2102.06942.

[89] N. Dym and H. Maron. *On the Universality of Rotation Equivariant Point Cloud Networks*. 2020. DOI: 10.48550/ARXIV.2010.02449. URL: https://arxiv.org/abs/2010.02449.

[90] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. "Graph neural networks: A review of methods and applications". en. In: *AI Open* 1 (2020), pp. 57–81.

[91] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. *Neural Message Passing for Quantum Chemistry*. 2017. DOI: 10.48550/ARXIV.1704.01212. URL: https://arxiv.org/abs/1704.01212.

[92] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. "SchNet – A deep learning architecture for molecules and materials". In: *The Journal of Chemical Physics* 148.24 (Mar. 2018). ISSN: 1089-7690. DOI: 10.1063/1.5019779. URL: http://dx.doi.org/10.1063/1.5019779.

[93] O. T. Unke and M. Meuwly. "PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges". In: *Journal of Chemical Theory and Computation* 15.6 (May 2019), pp. 3678–3693. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.9b00181. URL: http://dx.doi.org/10.1021/acs.jctc.9b00181.

[94] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim. *Graph Transformer Networks*. 2019. DOI: 10.48550/ARXIV.1911.06455. URL: https://arxiv.org/abs/1911.06455.

[95] M. Haghighatlari, J. Li, X. Guan, O. Zhang, A. Das, C. J. Stein, F. Heidar-Zadeh, M. Liu, M. Head-Gordon, L. Bertels, H. Hao, I. Leven, and T. Head-Gordon. *NewtonNet: A Newtonian message passing network for deep learning of interatomic potentials and forces*. 2021. DOI: 10.48550/ARXIV.2108.02913. URL: https://arxiv.org/abs/2108.02913.

[96] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. "Machine Learning Force Fields". In: *Chemical Reviews* 121.16 (Mar. 2021), pp. 10142–10186. ISSN: 1520-6890. DOI: 10.1021/acs.chemrev.0c01111. URL: http://dx.doi.org/10.1021/acs.chemrev.0c01111.

[97] T. S. Cohen and M. Welling. "Group Equivariant Convolutional Networks". In: (2016). DOI: 10.48550/ARXIV.1602.07576. URL: https://arxiv.org/abs/1602.07576.

[98] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. *Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds*. 2018. DOI: 10.48550/ARXIV.1802.08219. URL: https://arxiv.org/abs/1802.08219.

[99] M. A Morrison and G. A Parker. "A Guide to Rotations in Quantum Mechanics". In: *Australian Journal of Physics* 40.4 (1987), p. 465. ISSN: 0004-9506. DOI: 10.1071/ph870465. URL: http://dx.doi.org/10.1071/PH870465.

[100] D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, Aug. 2018. ISBN: 9781107189638. DOI: 10.1017/9781316995433. URL: http://dx.doi.org/10.1017/9781316995433.

[101] J. E. Gerken, J. Aronsson, O. Carlsson, H. Linander, F. Ohlsson, C. Petersson, and D. Persson. "Geometric deep learning and equivariant neural networks". In: *Artificial Intelligence Review* 56.12 (June 2023), pp. 14605–14662. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10502-7. URL: http://dx.doi.org/10.1007/s10462-023-10502-7.

[102] Y.-L. Liao and T. Smidt. *Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs*. 2022. DOI: 10.48550/ARXIV.2206.11990. URL: https://arxiv.org/abs/2206.11990.

[103] B. Anderson, T.-S. Hy, and R. Kondor. *Cormorant: Covariant Molecular Neural Networks*. 2019. DOI: 10.48550/ARXIV.1906.04015. URL: https://arxiv.org/abs/1906.04015.

[104] W. J. Hehre, R. F. Stewart, and J. A. Pople. "Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals". In: *The Journal of Chemical Physics* 51.6 (Sept. 1969), pp. 2657–2664. ISSN: 1089-7690. DOI: 10.1063/1.1672392. URL: http://dx.doi.org/10.1063/1.1672392.

[105] M. A. Blanco, M. Flórez, and M. Bermejo. "Evaluation of the rotation matrices in the basis of real spherical harmonics". In: *Journal of Molecular Structure: THEOCHEM* 419.1–3 (Dec. 1997), pp. 19–27. ISSN: 0166-1280. DOI: 10.1016/s0166-1280(97)00185-1. URL: http://dx.doi.org/10.1016/s0166-1280(97)00185-1.

[106] M. Häser and R. Ahlrichs. "Improvements on the direct SCF method". In: *Journal of Computational Chemistry* 10.1 (Jan. 1989), pp. 104–111. ISSN: 1096-987X. DOI: 10.1002/jcc.540100111. URL: http://dx.doi.org/10.1002/jcc.540100111.

[107] P. Pinski, C. Riplinger, E. F. Valeev, and F. Neese. "Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. I. An efficient and simple linear scaling local MP2 method that uses an intermediate basis of pair natural orbitals". In: *The Journal of Chemical Physics* 143.3 (July 2015). ISSN: 1089-7690. DOI: 10.1063/1.4926879. URL: http://dx.doi.org/10.1063/1.4926879.

[108] N. Wang, C. Lin, M. Bronstein, and P. Torr. "Towards Flexible, Efficient, and Effective Tensor Product Networks". In: *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*. 2023. URL: https://openreview.net/forum?id=947KhgKKGG.

[109] S. Passaro and C. L. Zitnick. *Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs*. 2023. DOI: 10.48550/ARXIV.2302.03655. URL: https://arxiv.org/abs/2302.03655.

[110] S. Villar, D. W. Hogg, K. Storey-Fisher, W. Yao, and B. Blum-Smith. "Scalars are universal: Equivariant machine learning, structured like classical physics". In: (2021). DOI: 10.48550/ARXIV.2106.06610. URL: https://arxiv.org/abs/2106.06610.

[111] M. E. Rose and B. T. Feld. "Elementary Theory of Angular Momentum". In: *Physics Today* 10.11 (Nov. 1957), pp. 30–30. ISSN: 1945-0699. DOI: 10.1063/1.3060162. URL: http://dx.doi.org/10.1063/1.3060162.

[112] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014). URL: https://api.semanticscholar.org/CorpusID:6628106.

[113] A. Chen and M. Heyl. *Efficient optimization of deep neural quantum states toward machine precision*. 2023. DOI: 10.48550/ARXIV.2302.01941. URL: https://arxiv.org/abs/2302.01941.

[114] M. Bukov, M. Schmitt, and M. Dupont. "Learning the ground state of a non-stoquastic quantum Hamiltonian in a rugged neural network landscape". In: (2020). DOI: 10.48550/ARXIV.2011.11214. URL: https://arxiv.org/abs/2011.11214.

[115] Springer New York, 2006. DOI: 10.1007/978-0-387-45528-0. URL: http://dx.doi.org/10.1007/978-0-387-45528-0.

[116] J. Westermayr and P. Marquetand. "Machine Learning for Electronically Excited States of Molecules". In: *Chemical Reviews* 121.16 (Nov. 2020), pp. 9873–9926. ISSN: 1520-6890. DOI: 10.1021/acs.chemrev.0c00749. URL: http://dx.doi.org/10.1021/acs.chemrev.0c00749.

[117] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K.-L. Chan. *The Python-based Simulations of Chemistry Framework (PySCF)*. 2017. DOI: 10.48550/ARXIV.1701.08223. URL: https://arxiv.org/abs/1701.08223.

[118] K. Nilsson. "Coordination chemistry in liquid ammonia and phosphorous donor solvents". In: ().

[119] P.-O. Löwdin. "On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals". In: *The Journal of Chemical Physics* 18.3 (Mar. 1950), pp. 365–375. ISSN: 1089-7690. DOI: 10.1063/1.1747632. URL: http://dx.doi.org/10.1063/1.1747632.

[120] D. I. Lyakh, M. Musiał, V. F. Lotrich, and R. J. Bartlett. “Multireference nature of chemistry: the coupled-cluster view”. en. In: *Chem. Rev.* 112.1 (Jan. 2012), pp. 182–243.

[121] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `https://www.tensorflow.org/`.

[122] A. Brock, S. De, S. L. Smith, and K. Simonyan. *High-Performance Large-Scale Image Recognition Without Normalization*. 2021. DOI: `10.48550/ARXIV.2102.06171`. URL: `https://arxiv.org/abs/2102.06171`.

[123] S. Das, M. Kállay, and D. Mukherjee. “Inclusion of selected higher excitations involving active orbitals in the state-specific multireference coupled-cluster theory”. In: *The Journal of chemical physics* 133 (Dec. 2010), p. 234110. DOI: `10.1063/1.3515478`.