Accelerating Dynamic Correlation Algorithms for Quantum Chemistry

Mikael Alexander Kovtun

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Xiaosong Li, Chair

Ting Cao

Matthew Yankowitz

Program Authorized to Offer Degree:
Physics

University of Washington

Abstract

Accelerating Dynamic Correlation Algorithms for Quantum Chemistry

Mikael Alexander Kovtun

Chair of the Supervisory Committee:
Prof. Xiaosong Li
Chemistry

Relativistic quantum mechanics provides a fundamental description of virtually all chemical phenomena, though in practice quantum chemistry is formulated through approximate solutions of the Dirac equation. Among systematic approaches, configuration interaction (CI) yields the exact solution in the complete-basis-set limit but suffers from factorial growth in computer memory requirements, restricting its application to the smallest systems. The recently developed small tensor product distributed active space (STP-DAS) CI framework significantly shrinks these resource requirements, yet large correlated calculations remain out of reach. In contrast, density functional theory (DFT) offers an approximate solution with much more favorable scaling. While not systematically improvable, DFT is widely used for large molecular systems where CI is not tractable. The landscape of scientific computing today is moving away from traditional processors in favor of graphics processing units (GPUs). GPUs offer the potential of tremendous speedup, but their unique computing paradigm requires careful algorithmic design to maximize their effectiveness. In this dissertation, we present two GPU-accelerated algorithms that accelerate critical rate-limiting steps in DFT and STP-DAS CI calculations, enabling faster correlated methods across length scales. We describe the design considerations of these algorithms, contextualize their role in chemical applications, and demonstrate their performance on large systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

CASCI: complete active space configuration interaction

CI: configuration interaction

CPU: central processing unit

CSF: configuration space function

DAS: distributed active space

DFT: density functional theory

FCI: full configuration interaction

GHF: generalized Hartree–Fock

GPU: graphics processing unit

GTO: gaussian–type orbital

HK: Hohenberg–Kohn

KS: Kohn–Sham

MO: molecular orbital

MRCI: multi–reference configuration interaction

NEO: nuclear electronic orbital

RHF: restricted Hartree–Fock

STP–DAS: small tensor product distributed active space

UHF: unrestricted Hartree–Fock

X2C: exact two–component

ACKNOWLEDGMENTS TO PREVIOUS WORK

Chapter 2 is adapted with permission from M. Kovtun, E. Lambros, A. Liu, D. Tang, D. B. Williams-Young, and X. Li. Accelerating Relativistic Exact-Two-Component Density Functional Theory Calculations with Graphical Processing Units. *Journal of Chemical Theory and Computation*, 20(18):7694, 2024. DOI: 10.1021/acs.jctc.4c00843. Copyright American Chemical Society.

Chapter 3 is adapted with permission from A. Shayit, M. Kovtun, S. Upadhyay, H. Hu, and X. Li. Accelerating Small Tensor Product Distributed Active Space (STP-DAS) Configuration Interaction using Graphics Processing Units (GPUs). Manuscript in preparation, 2026.

DEDICATION

This work is dedicated, in no particular order, to all the teachers, professors, and mentors who guided my studies, to my parents for supporting me all my life, and to Nikolas, Melina, Kevin, Barb, Luke, Mikey, Alex, Marco, Jonny, Richard (Euicheon), Will, Garan, Eric, Aidan, Kevin, Emily, Andrea, to my fellow group members and whoever else happened to make a positive impression on me.

Chapter 1

INTRODUCTION

The seventeenth century was a tumultuous time for natural philosophy, the Aristotlean school of thought dating back to ancient Greece. Precision measurements of planetary motion by Galileo Galilei challenged the geocentric model of the universe, then purported by natural philosophers relying on deductive reasoning (and authority) to explain natural phenomena. The success of Sir Isaac Newton's *Principia Mathematica*[148] in explaining Galileo's observations sent shockwaves through the collective mind, solidifying inductive reasoning and mathematical modeling as the replacement of the old ways. To this day, scientific understanding of the natural world revolves around mathematical models, sets of universal relationships between quantities that allows one to describe and *predict* physical phenomena. While physics enjoyed its revolution with classical mechanics comparatively early, chemistry had yet to have a Newton of its own. It took countless thinkers to evolve chemistry, with giants like Boyle to shake the mystique off of alchemy and advocate for empirical study[20], Lavoisier to develop precision analytical techniques, and Mendeleev to tabulate the elements into the first hints of quantum theory, into a state ready for it's own revolution. Physicists in the golden age of the early twentieth century finally produced an answer in the form of quantum theory. At last it became possible, in principle, to describe all chemical phenomena quantitatively in terms of a universal set of simple mathematical relationships; a microscopic mathematical model.

As is quoted in nearly every text on the historical development of physical chemistry, Paul Adrien Maurice Dirac wrote in his 1929 paper[38]

> *The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the*

> *difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.*

The reductionist view of Dirac appealed to the physicists of the time as quantum mechanics seemed to be the promised microscopic theory of chemistry. While the physicists' search for their own microscopic theory of physics continued through the twentieth century (and to this day), the meeting point of chemistry and quantum mechanics was developing into a brand new field: quantum chemistry.

In the early days, quantum chemistry was a setting for physicists to hash out what's known today as the "old quantum theory", born out of discoveries about the nature of chemical bonds and predicted spectra of atoms and molecules. However, by the early 1960s, the introduction of a new and powerful instrument — the computer — thrust the field into a totally new era. Quantum chemistry suddenly became a *predictive* tool, one where chemically interesting questions can be asked and answered *ab initio*, without using any data as inputs. A flurry of development followed, with both algorithmic and hardware improvements bringing more problems into tractability throughout the rest of the twentieth century.

Today, quantum chemistry is a mature field at the intersection of physics, chemistry, and computer science — as any one of these fields evolves, so does quantum chemistry. It's development benefits fundamental science, from providing highly scalable quantum algorithms to understand the quantum mechanical origin of molecular–biological phenomena[143], to high–precision quantum electrodynamics calculations[189]. Very recently, quantum chemistry finds itself at center stage with the physical realization of quantum computer, which employ a completely new computing paradigm that emerges from quantum mechanics itself[25]. Perhaps this will bring the second quantum chemistry revolution — and be the second time

quantum chemistry changes forever.

Despite it's central role in the computational study of fundamental science, quantum chemistry is far from an academic pursuit. For decades, it has been at the forefront of developing technologies involving anything molecular or electronic, from drugs[212] to catalysts[1] to batteries[219] to solar cells[35] and more[36]. Fast algorithms are the lynchpin of any industrial application of quantum chemistry. Without the ability to calculate ground and excited state properties of the large molecules in complex environments commonly found in industrial applications, scientists would be limited to experimentally-accessible data only, severely hindering their ability to obtain answers to their research questions. Therefore, the development and integration of fast quantum chemistry algorithms, fit for todays ever–evolving high–performance computing landscape, is central to enabling rapid technological development.

This dissertation outlines my work developing quantum chemistry algorithms for (classical) modern computing clusters. These algorithms are specialized to *relativistic* quantum chemistry, which is an extension to quantum chemistry that takes into account the finite speed of light in the universe. In Sec. 1.1, we will outline the basics of nonrelativistic quantum mechanics, including quantum states and how they change with time. Section 1.2 discusses electronic structure theory, which makes up the core of quantum chemistry. Both the theory and some of the methods used in practice will be discussed in some detail. Relativity and its application to quantum mechanics in general are outlined in Sec. 1.3 and Sec. 1.4. Then, Sec. 1.6 will discuss some vital background for the main results of this work, namely relativistic electronic structure theory. Finally, the main results are presented in Chapter 2 and Chapter 3, and we conclude in Chapter 4.

1.1 Quantum Mechanics

The central object of quantum mechanics is the Hilbert space $\mathcal{H}$, which is a complex–valued vector space that contains all the possible configurations of the physical system. An observable $\hat{O}$ is a Hermitian operator on the elements of $\mathcal{H}$; it represents a physical

measurement on the system, such as position $\boldsymbol{x}$, momentum $\boldsymbol{p}$, or spin along the z-axis S_z. An element of the Hilbert space is called a ket, denoted by $|\alpha\rangle$, where α labels the ket in this instance. Kets contain all the information we are allowed to ask about the state, via multiplication with observables:

$$\hat{O} \cdot (|\alpha\rangle) = \hat{O} \, |\alpha\rangle \tag{1.1}$$

The result of multiplying a ket with an observable is, in general, not that same ket times a constant. However, every observable has a set of special kets that do have this property:

$$\hat{A} \, |a_1\rangle = a_1 \, |a_1\rangle \, , \; \hat{A} \, |a_2\rangle = a_2 \, |a_2\rangle \, , \; \ldots \tag{1.2}$$

where $a_1, a_2, \ldots$ are just numbers. Those special kets $\hat{a_1}, \hat{a_2}, \ldots$ and their corresponding numbers $a_1, a_2, \ldots$ are called the eigenkets and eigenvalues of operator $\hat{A}$. For every ket, there exists a bra, denoted by $\langle\alpha| = |\alpha\rangle^\dagger$ where $\dagger$ is the conjugate transpose operation. From this definition follows two fundamental properties of quantum states:

$$\langle\alpha|\beta\rangle = \langle\beta|\alpha\rangle^* \quad , \quad \langle\alpha|\alpha\rangle \geq 0 \tag{1.3}$$

where * denotes the complex conjugate. These imply the existence of orthogonal states, or those with $\langle\alpha|\beta\rangle = 0$, and that we can always normalize states (divide by the inner product with itself) so that $\langle\alpha|\alpha\rangle = 1$.

The quantum theory of measurement is probably the most subtle and difficult aspect of quantum mechanics to understand and is the subject on ongoing debate, so I will not cover it in depth here. To borrow words from Dirac[37] "A measurement always causes the system to jump into an eigenstate of the dynamical variable that is being measured." In other words, before an observable $\hat{A}$ is measured, the system is assumed to be in some linear combination of eigenstates of $\hat{A}$:

$$|\alpha\rangle = \sum_n c_n \, |a_n\rangle = \sum_n |a_n\rangle \, \langle a_n|\alpha\rangle \tag{1.4}$$

When the measurement is performed, the system suddenly shifts into one of the eigenstates of the observable:

$$|\alpha\rangle \longrightarrow |a_k\rangle \tag{1.5}$$

where k indicates the particular eigenstate the system jumped into. The state that existed before the measurement is not the same state after the measurement; it is destroyed by the process of measurement. The only exception is if the state is already in one of the eigenstates of the observable,

$$|\alpha\rangle = |a_k\rangle \longrightarrow |a_k\rangle \tag{1.6}$$

Which eigenstate does the system jump into? One of the central postulates of quantum mechanics is that the probability for a normalized state $|\alpha\rangle$ to jump into some particular $|a_k\rangle$ is

$$\text{Probability of } a_k = |\langle a_k|\alpha\rangle|^2 \tag{1.7}$$

If an experimenter prepares the state $|\alpha\rangle = \sum_c c_n |a_n\rangle$ and performs the measurement many times, each outcome may be different, and the laws of quantum mechanics does not permit us to exactly know what the outcome of a particular measurement is. However, it does permit us to calculate the expectation value of the observable, which is the average expected outcome of a measurement repeated many times on an ensemble of identical states:

$$\langle A\rangle = \langle\alpha|\hat{A}|\alpha\rangle = \sum_n a_n |\langle a_n|\alpha\rangle|^2 \tag{1.8}$$

The important information about an observable is characterized by its statistical properties, such as the dispersion $\langle(\Delta A)^2\rangle$ (also called the variance),

$$\langle(\Delta A)^2\rangle = \langle A^2\rangle - \langle A\rangle^2 \tag{1.9}$$

When the state is an eigenstate of $\hat{A}$, the dispersion becomes zero. Thus the dispersion is a measure of how broad the distribution of measurement outcomes is for an operator.

The probabilistic nature of quantum mechanics implies that we can only gather limited information about the state of a system. The *uncertainty principle* codifies this in an equation:

$$\langle(\Delta A)^2\rangle \langle(\Delta B)^2\rangle \geq \frac{1}{4}|\langle[\hat{A},\hat{B}]\rangle|^2 \tag{1.10}$$

where the brackets represent the commutator $[\hat{A},\hat{B}] = AB - BA$. Observables that commute ($[\hat{A},\hat{B}] = 0$) have no uncertainty principle between them i.e. a state can be in an eigenstate of $\hat{A}$ and $\hat{B}$ simultaneously.

Commuting observables (also called compatible observables) provide a useful way to label quantum states. If there are two (or more) eigenkets of an observable $\hat{A}$ that have the same eigenvalue a_k, those two eigenvalues are called *degenerate*. Suppose we have a set of observables $\hat{A}, \hat{B}, \ldots$ in our system, and that $\hat{A}$ has a doubly degenerate eigenvalue a_k. If $\hat{A}$ and $\hat{B}$ commute, then $|a_k\rangle$ is a simultaneous eigenket of $\hat{B}$ e.g. $|a_k, b_l\rangle$. If we choose the maximal set of commuting observables, that is, the set of all observables that both commute with $\hat{A}$ and amongst each other, their collective eigenvalues specifies the state with maximal specificity. In real quantum systems, if a degeneracy is present in a system there's often an observable that can be introduced to break the degeneracy. One classic example is of the squared angular momentum operator $\hat{L}^2$ and the z-projected angular momentum operator $\hat{L}_z$: if measuring $\hat{L}^2$ gives $\langle L^2\rangle = \hbar^2 l(l+1)$, there are $2l+1$ degenerate eigenkets of $\hat{L}^2$ the state could be in. It takes another measurement from a commuting observable, that of $\hat{L}_z$, to determine the final state $|l, m_l\rangle$.

1.1.1 Quantum dynamics

In quantum mechanics, there are two main philosophies to handling the time evolution of a state. In the "Schrödinger picture", the state kets change with time, while in the "Heisenberg picture", the state kets are stationary while the observables evolve in time. Both are equivalent and can even be mixed together, but for simplicity we will use the Schrödinger picture throughout this work.

The time evolution of a quantum state is given by an operator $U(t, t_0)$ that evolves a state from time t_0 to time t. It has a few intuitive properties, namely the composition property,

$$U(t_2, t_0) = U(t_2, t_1)U(t_1, t_0) \qquad t_2 > t_1 > t_0 \tag{1.11}$$

and unitarity,

$$U^\dagger(t, t_0)U(t, t_0) = 1 \tag{1.12}$$

The composition property reflects the fact that something evolving from $t = 0$ to $t = 2$ can be equivalently said to evolve from $t = 0$ to $t = 1$ and then $t = 1$ to $t = 2$. Unitarity states that sum of probabilities of all measurement outcomes should always sum to 1. These two properties force U to be of the form of an exponential matrix:

$$\hat{U}(t, t_0) = e^{i\hat{\Omega}t} \tag{1.13}$$

where $\hat{\Omega}$ is a Hermitian operator, $\Omega^\dagger = \Omega$. In classical mechanics, the Hamiltonian is the generator of time evolution[74], so we may relate $\hat{\Omega}$ to the Hamiltonian operator $\hat{H}$[186]:

$$\hat{U}(t, t_0) = e^{i\hat{H}t/\hbar} \tag{1.14}$$

The time evolution operator is formed from the Hamiltonian $\hat{H}$, which is a special operator that gives the total energy of the system.

The $t \to t_0$ limit of equation Eq. (1.14) can be taken to write it as a differential equation, and after multiplying on the right by a $t = t_0$ state ket $|\alpha, t_0\rangle$, we obtain the time-dependent Schrödinger equation for a state ket:

$$i\hbar\frac{\partial}{\partial t}|\alpha, t_0; t\rangle = \hat{H}|\alpha, t_0; t\rangle \tag{1.15}$$

This first-order differential equation gives the time evolution of a state ket $|\alpha\rangle$ from t_0 to t, provided that the Hamiltonian operator $\hat{H}$ does not itself depend on time. If a state ket

$|\alpha = a_k, t_0\rangle$ is an eigenstate of $\hat{A}$ and $[\hat{A}, \hat{H}] = 0$, then the time dependence of the state is simply

$$|\alpha, t_0; t\rangle = \exp(iE_{a_k} t/\hbar) \, |\alpha, t_0\rangle \tag{1.16}$$

where E_{a_k} is the eigenvalue of the Hamiltonian associated with the simultaneous eigenvalue of $\hat{A}$, a_k.

Operators in the Schrödinger picture do not change with time, but the time evolution of the states do. For a state $|\alpha, t_0\rangle$ that is an energy eigenstate with energy E,

$$\langle \hat{A}^n \rangle (t) = \langle \alpha, t_0; t| \, \hat{A}^n \, |\alpha, t_0; t\rangle = \\ \langle \alpha, t_0| \exp(iE_\alpha t/\hbar) \hat{A}^n \exp(-iE_\alpha t/\hbar) \, |\alpha, t_0\rangle = \langle \alpha, t_0| \, \hat{A}^n \, |\alpha, t_0\rangle \tag{1.17}$$

Thus for energy eigenstates, all expectation values of any observable are time-independent. As a bonus, since every power n is stationary, all dispersions and higher-order moments are stationary too.

If the Hamiltonian $\hat{H}$ is time-independent, then plugging the evolution of the state from Eq. (1.16) into Eq. (1.15) yields the time-independent Schrödinger equation,

$$\hat{H} \, |\alpha\rangle = E \, |\alpha\rangle \tag{1.18}$$

The solution of this equation can be obtained by diagonalizing the Hamiltonian $\hat{H}$, thus obtaining the spectrum e.g. the energy eigenstates and their eigenvalues. Then, any initial state $|\alpha\rangle$ can be written in terms of the energy eigenstates and the dynamics follow trivially.

1.1.2 Identical particles

To introduce the machinery of dealing with identical particles, consider a bunch of baseballs in a washing machine. When we turn on the washing machine, the drum begins to spin, and the baseballs are jostled and thrown around. However chaotic their motion, it is in principle

possible to keep track of and predict the motion of each baseball by carefully watching each one and keeping track of each individual collision. This is the picture in classical mechanics: each baseball can be tracked individually and measured even if they are identical. If we collect all of the possible configurations of the baseballs (particles) — every possible position and momentum — into an N-particle Hilbert space[1] $\mathcal{H}^{\otimes N}$, constructed as a *tensor product* of single-baseball Hilbert spaces $\mathcal{H}^1$:

$$\mathcal{H}^{\otimes N} = \mathcal{H}^1(1) \otimes \mathcal{H}^1(2) \otimes \cdots \otimes \mathcal{H}^1(N) \tag{1.19}$$

The N-particle wavefunction is then an arbitrary element of this space

$$\Psi_{MB}(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) \in \mathcal{H}^{\otimes N} \tag{1.20}$$

When Hilbert spaces combine in a tensor product $A \otimes B = C$, the elements of $c \in C$ are generated by sets of elements in A and B, i.e.

$$c = \sum_i a_i \otimes b_i \qquad \{a_i\} \in A, \{b_i\} \in B \tag{1.21}$$

Intuitively, wavefunctions describing all of the baseballs can be written as arbitrary combinations of wavefunctions describing each baseball individually. Crucially, each single particle space carries a particle label, which can be used to distinguish the particles even if they are identical. The subscript MB isn't a typo of Major League Baseball, but stands for *Maxwell-Boltzmann* statistics, reflecting the absence of any imposed permutational symmetry constraint on the wavefunction.

The difference between baseballs and electrons is that electrons have spin[2]. Thus, a

---

[1]Technically, the classical analog of a Hilbert space is a phase space, but for the sake of the argument let's pretend we're playing in the quantum league.

[2]There are a few other differences as well, but that's the important one.

single electronic wavefunction is denoted as a *spinor* which carries a spin label τ,

$$\psi_\tau(\boldsymbol{r}) \tag{1.22}$$

The simplest N-particle wavefunction ansatz is the product of single electron spinors,

$$\theta_T(\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N) = \psi_{\tau_1}(\boldsymbol{r}_1)\psi_{\tau_2}(\boldsymbol{r}_2)\cdots\psi_{\tau_N}(\boldsymbol{r}_N) \tag{1.23}$$

where τ_i is either α or β (indicating up and down spin) and $T = \{\tau_1, \cdots, \tau_N\}$ is shorthand for the particular realization of spins $\tau_i = \{\alpha, \beta\}$. It won't be proven here, but if we take a linear combination of all 2^N permutations of the spin string in θ_T, the resulting wavefunction spans the entire Hilbert space in Eq. (1.19)

$$\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \sum_T C_T \theta_T(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) \tag{1.24}$$

where C_T are the expansion coefficients. However, the spin-statistics theorem[201] restricts the symmetry under permutations of particle labels depending on the spin of the particle. In the case of electrons, which are spin-$\frac{1}{2}$, the wavefunction must pick up a minus sign when any two particle labels ($\{\boldsymbol{r}_i, \tau_i\}$) are interchanged, a property called *antisymmetry*. Integer spin particles like bosons are symmetric, meaning there's no sign change.

The antisymmetry requirement of θ_T can be enforced through an operator called the antisymmetrizer, which is defined in terms of permutations π in the permutation group S_N for N coordinates,

$$\mathcal{A} = \frac{1}{N!}\sum_{P \in S_N} (-1)^\pi \hat{P} \tag{1.25}$$

where $(-1)^\pi$ is the parity of the permutation. The effect of this operator is to leave totally antisymmetric functions unchanged, but sets functions to zero that have at least one set of coordinates that can be swapped without changing the overall function. This is exactly the requirement of the Pauli exclusion principle – the antisymmetrizer enforces that no two

Particle type	Fermions	Bosons	Classical
Statistics	Fermi-Dirac	Bose-Einstein	Maxwell-Boltzmann
Spin	half-integer	integer	none
P_{ij} **symmetry**	anti-symmetric	symmetric	none
2 particle states	$\frac{1}{\sqrt{2}}(\|a\rangle \|b\rangle - \|b\rangle \|a\rangle)$	$\|a\rangle \|a\rangle$, $\|b\rangle \|b\rangle$, $\frac{1}{\sqrt{2}}(\|a\rangle \|b\rangle + \|b\rangle \|a\rangle)$	$\|a\rangle \|a\rangle$, $\|a\rangle \|b\rangle$, $\|b\rangle \|a\rangle$, $\|b\rangle \|b\rangle$

Table 1.1. Breakdown of multi-particle states and their statistics.

states can occupy the same spin-orbital. Applying $\mathcal{A}$ to the product ansatz θ_T yields a *Slater determinant*[78,186],

$$\mathcal{A}\theta_T = \Theta_T = \Theta_{\tau_1\tau_2\ldots\tau_N}(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_{\tau_1}(\boldsymbol{r}_1) & \cdots & \psi_{\tau_1}(\boldsymbol{r}_N) \\ \vdots & \ddots & \vdots \\ \psi_{\tau_N}(\boldsymbol{r}_1) & \cdots & \psi_{\tau_N}(\boldsymbol{r}_N) \end{vmatrix} \tag{1.26}$$

Slater determinants do not span the entire Hilbert space like the product ansatz in Eq. (1.24), but they do span the antisymmetric part of it, $\mathcal{A}\mathcal{H}^{\otimes N}$. Just like with θ_T, we can build a linear combination of Θ_T that spans the entire antisymmetric space,

$$\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \sum_T C_T \Theta_T(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) \tag{1.27}$$

Because of the antisymmetry requirement, electronic wavefunctions only occupy a subset of the overall Hilbert space $\mathcal{H}^{\otimes N}$. Since this ansatz enumerates that subset exactly, it's impossible to produce an invalid wavefunction by varying the expansion coefficients C_T.

1.2 Electronic Structure Theory

Electronic structure theory is concerned with the application of quantum mechanics to atoms, molecules, and all condensed matter, with the goal of explaining electron-driven phenomena ranging from chemical bonding to superconductivity. Electrons are spin-1/2

particles and, being able to travel through space, have a position and momentum as well. Position and momentum (famously) do not commute, so their overall state ket is determined by two quantum numbers: spin and position, or spin and momentum.

The (non-relativistic) Hamiltonian is the sum of kinetic energy and potential energy,

$$\hat{H} = \hat{T} + \hat{V} \tag{1.28}$$

For our purposes here, let's consider the case where an electron is attracted to one or more stationary, positively charged nuclei. The potential term couples only to electron position, and the Hamiltonian commutes with the time evolution operator, so in the position basis,

$$\langle \boldsymbol{x} | \hat{V} | \psi \rangle = V(\boldsymbol{x}) \psi(\boldsymbol{x}) \tag{1.29}$$

$$\langle \boldsymbol{x} | \hat{T} | \psi \rangle = \langle \boldsymbol{x} | \frac{\boldsymbol{p}^2}{2m} | \psi \rangle = -\left(\frac{\hbar^2}{2m} \right) \nabla^2 \psi(\boldsymbol{x}) \tag{1.30}$$

$$i\hbar \frac{\partial}{\partial t} \psi(\boldsymbol{x}, t) = -\left(\frac{\hbar^2}{2m} \right) \nabla^2 \psi(\boldsymbol{x}, t) + V(\boldsymbol{x}, t) \psi(\boldsymbol{x}, t) \tag{1.31}$$

Since time evolution commutes, we can define $u_E(\boldsymbol{x})$ to be the energy eigenfunctions, and we arrive at the time-independent Schrödinger wave equation

$$-\left(\frac{\hbar^2}{2m} \right) \nabla^2 u_E(\boldsymbol{x}) + V(x) u_E(\boldsymbol{x}) = E u_E(\boldsymbol{x}) \tag{1.32}$$

Solving this differential equation gives you the spectrum of the Hamiltonian, or the entire family of eigenenergies. These include the ground state (the lowest energy eigenvalue) together with the excited states (higher energy eigenvalues). This equation was first applied to derive the energy spectrum of the hydrogen atom in 1926[192].

Modeling the nuclei as classical point charges, also known as the Born–Oppenheimer approximation[19], is motivated by the fact that electrons are much lighter than nuclei (the mass of a proton is $\approx$ 1836 times larger than the mass of an electron[146]). Thus, the

timescale for the dynamics of nuclei is much longer than that of the electrons, so they can be reasonably treated as classical degrees of freedom, and in our case since we're interested in ground state properties we'll also treat nuclei as stationary. It should be noted that the Born–Oppenheimer approximation can fail for some systems where nuclear dynamics is strongly coupled to electron dynamics, particularly those with light nuclei. In these cases, treating all or some of the light nuclei as a quantum system themselves in the Nuclear Electronic Orbital (NEO) method[155?] is the appropriate approximation. This work focuses on heavy nuclei so we will not dwell on NEO further, but there are other considerations regarding the nuclear model for the heavy nuclei regime.

There are several forms of the stationary, classical nuclear charge potential Eq. (1.29) in common use in quantum chemistry codes. The two we will consider are the point charge model,

$$V_{\text{point}}(\boldsymbol{r}) = -\sum_{A} \frac{Z_A}{|\boldsymbol{r} - \boldsymbol{r}_A|} \tag{1.33}$$

which is the form familiar to a physics student solving the Schrödinger equation for the hydrogen atom for the first time. This potential is non-differentiable at the location of the nuclei $\boldsymbol{r}_A$, which introduces a cusp into the exact wavefunction. A slightly more accurate model is to approximate the nuclear charge as a "Gaussian smear" of charge,

$$V_{\text{gauss}}(\boldsymbol{r}) = -\sum_{A} \mathcal{N}_A Z_A \exp[-\alpha_A (\boldsymbol{r} - \boldsymbol{r}_A)^2] \tag{1.34}$$

where $\mathcal{N}_A$ are normalization factors and α_A are model-dependent parameters. The point charge model can be considered as the limit of the Gaussian charge model where $\alpha_A \to 0$. The Gaussian nuclear charge model produces wavefunctions that are both more realistic and numerically well-behaved[144]. For the rest of this text, we will assume that the nuclear charge potential Eq. (1.29) is Gaussian.

1.2.1 The Electronic Hamiltonian

Let's write the time-independent Schrödinger equation for multiple electrons in a molecule. The electrons are attracted to the nuclei and repelled from each other via the Coulomb interaction. Writing the equation in atomic units ($\hbar = m_\mathrm{e} = 4\pi\epsilon_0 = q_e = 1$),

$$\hat{H}\Psi = (\hat{T} + \hat{V}_\mathrm{nuc} + \hat{V}_{e-e})\Psi = E\Psi \tag{1.35}$$

This Hamiltonian acts on an N-particle wavefunction,

$$\left(-\frac{1}{2}\nabla_i^2 + V_\mathrm{nuc}(\boldsymbol{r}_i) + \frac{1}{2}\sum_{i\neq j}\frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}\right)\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) = E\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) \tag{1.36}$$

The first term in the Hamiltonian on the left is the kinetic energy for the ith particle $\hat{T}_i$, the second term is the nuclear attraction potential, and the third is the Coulombic interaction between electrons. The wavefunction has been written as a capital Ψ to denote a multi-particle wavefunction, which is anti-symmetric under electron label swaps. Indeed, the Hamiltonian itself acts on each particle symmetrically (as it must in the case of indistinguishable particles), so it commutes with the antisymmetrizer

$$[\hat{H}, \mathcal{A}] = 0 \tag{1.37}$$

Taking a nod from perturbation theory, if we assume the interaction term is small and ignore it, we can write the non-interacting Hamiltonian

$$\hat{H}_1 = \sum_i^N \hat{h}_1(i) = \sum_i^N \left[-\frac{1}{2}\nabla_i^2 - \sum_A \mathcal{N}_A Z_A \exp(-\alpha_A(\boldsymbol{r}_i - \boldsymbol{r}_A)^2)\right] \tag{1.38}$$

where the subscript 1 underscores that the Hamiltonian is comprised of only one-body terms. The exact solution to $\hat{H}_1\Psi = E\Psi$ is in fact a single Slater determinant. To see this, note

that $\hat{H}_1$ is the sum of identical one-body terms. Then we can write just one as

$$\hat{h}_1\psi_m = \epsilon_m\psi_m \tag{1.39}$$

where the spin label is suppressed and ψ_m is the mth solution to the eigenvalue problem ($m < M$). Consider the product ansatz

$$\theta = \psi_{m_1}(\boldsymbol{r}_1)\psi_{m_2}(\boldsymbol{r}_2)\cdots\psi_{m_N}(\boldsymbol{r}_N) \tag{1.40}$$

which diagonalizes the N body Hamiltonian

$$\hat{H}_1\theta = \sum_i^N \hat{h}_1(i)\theta = \sum_i^N \epsilon_{m_i}\theta \tag{1.41}$$

so the energy of $\hat{H}_1$ is

$$E = \sum_i^N \epsilon_{m_i} \tag{1.42}$$

Since $[\hat{H}_1, \mathcal{A}] = 0$,

$$\mathcal{A}\hat{H}_1\theta = \mathcal{A}E\theta \rightarrow \hat{H}_1\mathcal{A}\theta = E\mathcal{A}\theta \rightarrow \hat{H}_1\Theta = E\Theta \tag{1.43}$$

where Θ is the Slater determinant generated from the product ansatz θ. Thus a single Slater determinant (as opposed to a linear combination) is the exact solution for the non-interacting Hamiltonian with the eigenenergy $E = \sum_i^N \epsilon_{m_i}$ where each ϵ_{m_i} is obtained from the one-particle eigenproblem $\hat{h}_1\psi_m = \epsilon_m\psi_m$.

If we were to turn on the interaction perturbation, we would find that while the total Hamiltonian still commutes with the antisymmetrizer $[\hat{H}, \mathcal{A}] = 0$, it would not be decomposable into N one-body eigenproblems and a *single* Slater determinant would no longer be the solution.

1.2.2 Configuration Interaction

An expansion of the exact many-body wavefunction in Slater determinants is called *configuration interaction*,

$$\Psi_A(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) = \sum_T C_{TA} \Theta_T(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) \tag{1.44}$$

where A labels the (ground or excited) states and T labels the string of spinor labels including the reference and excited configurations. When all excited states are included, this expansion is called the full configuration interaction (FCI) expansion and is exact for any complete set of one-electron spinor orbitals. The expansion coefficients C_{TA} are called CI coefficients.

Now that we have a basis for many-body wavefunctions, we can simply plug Eq. (1.44) into Eq. (1.36) to get the energy of a state A,

$$E_A = \langle \Psi_A | \left(\hat{h}_1 + \hat{h}_2 \right) | \Psi_A \rangle \tag{1.45}$$

where the one-electron and two-electron parts of the Hamiltonian have been separated into $\hat{h}_1$ and $\hat{h}_2$ [3].

In configuration interaction literature, the basis of Slater determinants is typically decomposed into a *reference* and its *excited* determinants[4]. A reference determinant is chosen to approximate the exact (interacting) wavefunction as best as possible, and excited determinants are *generated* from it by acting on it with the single excitation operator,

$$\hat{E}_{pq} = \sum_{k=1}^{N} |\psi_p(\boldsymbol{r}_k)\rangle \langle \psi_q(\boldsymbol{r}_k)| \tag{1.46}$$

---

[3] In some literature, the one- and two-electron parts are denoted $\hat{h}_0$ and $\hat{h}_1$, deriving the nomenclature of perturbation theory treating the interaction as a parameter λ that is taken from 0 to 1.

[4] Excited determinants have nothing to do with the typical notion of an excited state in physics. It's purely a bookkeeping tool for keeping track of determinant configurations.

The idea is that if the approximation for the reference is good enough, it will have the largest weight in the CI coefficient expansion and the rest of the determinants will have little contribution. This means that the CI expansion is *sparse*, which is a highly desirable condition for a computer code. Considering that in an M-orbital basis of single-particle states, the CI expansion Eq. (1.44) for N electrons takes $\binom{M}{N}$ determinants.

As concrete example, if we have 20 electrons residing in 40 orbitals, the CI expansion would take 1.4×10^{11} determinants. With double precision floating point coefficient storage, that's 1.1TB of data just to store the coefficients. Storing the Hamiltonian would take a number of bytes comparable to Avogadro's number, 1.5 zettabytes or 1.5×10^{11} terabytes. This is why exploiting sparsity is a requirement for writing a computer code to do CI calculations.

Thankfully, the Hamiltonian is sparse too, which also can be exploited to accelerate energy evaluation. Acting on a ket determinant Ψ_I with the one-electron operator $\hat{h}_1$ produces a linear combination of kets that differ by at most one orbital since $\hat{h}_1$ acts on only one electron coordinate. That is, $\hat{E}_{pq} |\Psi_I\rangle$ are the only ones "touched" by $\hat{h}_1 |\Psi_1\rangle$. The two-electron operator $\hat{h}_2$ follows similar logic, except that the surviving determinants have up to two different orbitals. These simple rules, called the *Slater-Condon rules*, can be derived from the antisymmetry of the Slater determinant basis[78]. Since $\hat{H}$ contains only one- and two-electron operators, the structure of the Hamiltonian is quite sparse as well.

1.2.3 Wavefunction Symmetries

Since we've introduced the Hamiltonian Equation (1.36) and Slater determinants, now is a good time to talk about the symmetries of these two objects. The symmetries of the Hamiltonian Eq. (1.36) are multiple. Spatial symmetries, such as inversion symmetry or any point group symmetries, are inherited from the molecular structure in the nuclear attraction term. The spin symmetry of the Hamiltonian Eq. (1.36) are simple, as spin never

enters the Hamiltonian. Therefore,

$$[\hat{H}, \hat{S}^2] = [\hat{H}, \hat{S}_z] = 0 \tag{1.47}$$

the Hamiltonian commutes with the total spin $\hat{S}^2$ and spin projection $\hat{S}_z$ operators. The antisymmetric Slater determinant has the correct fermionic symmetry, but it is also an eigenstate of $\hat{S}_z$, as the eigenvalue can be found by just counting the number of spin up (α) minus spin down (β) spinors in the determinant. It is possible to construct linear combinations of Slater determinants that together are an eigenstate of $\hat{S}^2$, called *configuration state functions* (CSFs)[154]. CSFs are a particularly great ansatz for this Hamiltonian since they are guaranteed to be energy eigenstates. If a basis set of all possible CSFs (all possible values of $\hat{S}^2$ for N electrons) is constructed, then the set is not only complete but the CSFs are orthogonal. These are very nice properties to have for realistic calculations, but in practice, CSFs can be difficult to use. This is largely because the methods one uses to solve for the wavefunction or calculate observables inherit the linear combinatorical form of the CSF, and the expressions can be prohibitively cumberson to handle. However, there are situations common in chemistry where the CSF can be represented by a single determinant, and it depends on the symmetries of the Hamiltonian and the number of electrons.

If there is only one spin arrangement consistent with the $\hat{S}^2$, $\hat{S}_z$ quantum numbers, then the CSF is representable as a single determinant. For example, with two electrons, the singlet state with $\hat{S}^2 = 1$ and $\hat{S}_z = 0$ can be written as either $|\alpha\beta\rangle$ or $|\beta\alpha\rangle$, thus the CSF associated with $\hat{S}^2, \hat{S}_z = 1, 0$ cannot be written as a single determinant. However, the two triplet states with $\hat{S}_z = \pm 1$ can, because there is only one determinant in each case that produces that eigenvalue: $|\alpha\alpha\rangle$ and $|\beta\beta\rangle$. Generally, if the Clebsch-Gordan coefficients[186] for the desired total angular momentum permit multiple composite angular momenta configurations, the corresponding CSF cannot be represented by a single determinant. In chemical terms, only spin states with extremal $\hat{S}_z$ ("stretched" states) are representable this way.

Single determinant references are still used in practice, even if there are multiple valid

spin arrangements. The *restricted Hartree-Fock* (RHF) reference represents the wave function as a single Slater determinant where the α and β spin orbitals have the same spatial parts. The RHF reference is naturally an eigenfunction of both $\hat{S}^2$ and $\hat{S}_z$ and is thus represented by a single CSF. In contrast, the *unrestricted Hartree-Fock* (UHF) reference has completely independent α and β spin orbitals. It's therefore an eigenfunction of $\hat{S}_z$ but not necessarily $\hat{S}^2$, and so it in general is represented by a linear combination of many CSFs with different total spin symmetries. Finally, the *generalized Hartree-Fock* (GHF) reference has total freedom in spin orbitals, that is, each spatial orbital has α and β mixing. This is neither an eigenfunction of $\hat{S}^2$ nor $\hat{S}_z$, and thus is also not representable by a single CSF in general. Table 1.2 summarizes the properties of these references.

These references form a useful trio of single-reference tools for a quantum chemist. While they can violate the intuition that a wavefunction ansatz should obey the same symmetries as the Hamiltonian, they also opens up the possibility of accurately describing multi-reference systems using just one determinant. In the thought experiment of stretched H_2, a GHF reference would be able to smoothly transition from the same-spatial to the separate-spatial orbital structures of the molecule as the bond distance widens. The variational freedom a single GHF determinant has allows it to describe bond-breaking without having to resort to relatively expensive multi-reference techniques.

Of course, it is entirely possible to include new terms in the Hamiltonian that break spin symmetry. For example, including a magnetic field will introduce a term $\boldsymbol{B} \cdot \hat{\boldsymbol{S}}$, which breaks $\hat{S}_z$ symmetry in general (except if $\boldsymbol{B} = B\hat{z}$, of course). A spin-orbit coupling term, important for heavy nuclei, will go as $\boldsymbol{L} \cdot \hat{\boldsymbol{S}}$ and also break $\hat{S}_z$ symmetry. In these cases RHF references will not be able to represent the correct physics. The quantum chemist must be aware of the symmetries of the Hamiltonian in choosing a reference.

Reference	$\hat{S}^2$ efn.	$\hat{S}_z$ efn.	CSF	Orbital type
Restricted HF (RHF)	yes	yes	exactly one	same spatial orbital for α and β
Unrestricted HF (UHF)	no	yes	mixture	different spatial orbitals for α and β
Generalized HF (GHF)	no	no	mixture	general spinors, α and β mixed

Table 1.2. Summary of single-determinant references and their behavior under spin operators $\hat{S}^2$ and $\hat{S}_z$

1.2.4 The Hartree-Fock Reference State

The Hartree-Fock reference state is obtained when the CI expansion is truncated to a single determinant. With such an expansion at hand, how can the interacting Hamiltonian Eq. (1.36) be solved?

In the simplest case, where only a single determinant is included, the energy expression is

$$E_{HF} = \sum_{i}^{N_e} h_1^i + \frac{1}{2}\sum_{ij}^{N} h_2^{ij} \tag{1.48}$$

where the one-electron terms including the kinetic energy and nuclear attraction are

$$h_1^i = \langle \psi_i(1) | \, \hat{h}_1(1) \, | \psi_i(1) \rangle \tag{1.49}$$

and the two-electron terms are

$$h_2^{ij} = \langle \psi_i(1)\psi_j(2) | \, \hat{h}_2(1,2) \, | \psi_i(1)\psi_j(2) \rangle - \langle \psi_i(1)\psi_j(2) | \, \hat{h}_2(1,2) \, | \psi_j(1)\psi_i(2) \rangle = J_{ij} - K_{ij} \tag{1.50}$$

comprised of the Coulomb integral J_{ij} and the exchange integral K_{ij}, where the indices i, j run over orbital indices and $(1), (2)$ indicate which position vector is integrated over. Eq. (1.48) is called the Hartree-Fock (HF) energy[177,205], and a configuration whose spinors minimize this energy is called a Hartree-Fock (reference) state.

The physical interpretation of such states are of importance to physics and chemistry. As noted before, a single determinant is the exact form of a many-body wavefunction with no

interactions. The CI expansion used to derive the HF energy included all these interactions, but was truncated at a single determinant to get the HF reference state. Therefore, this reference corresponds to a fictional *non-interacting* system of electrons whose electronic density is as close as possible to the fully interacting system. By "as close as possible", we mean that it is good as a single-reference representation of the system can get under the variational principle. As discussed in Some systems, such as spin-frustrated systems and stretched covalently bonded diatomic molecules, have multiple degenerate ground states with different spin symmetries. In these cases, a single reference is not even in principle an adequate state, and MRCI techniques are necessary.

A useful form of the Hartree-Fock energy expression is given in terms of density matrices and generalized one and two body terms,

$$h_1^{ij} = \langle \psi_i(1) | \, \hat{h}_1(1) \, | \psi_j(1) \rangle \tag{1.51}$$

$$h_2^{ijkl} = \langle \psi_i(1)\psi_j(2) | \, \hat{h}_2(1,2) \, | \psi_k(1)\psi_l(2) \rangle \tag{1.52}$$

The single-reference terms h_1^i and h_2^{ij} are just special cases of these generalized integrals. This allows us to write the energy expression in terms of density matrices:

$$E_A = \sum_{ij}^{N} \gamma_{ij}^A h_1^{ij} + \sum_{ijkl}^{N} \Gamma_{ijkl}^A h_2^{ijkl} \tag{1.53}$$

where the γ_{ij}^A and Γ_{ijkl}^A are the one- and two-particle reduced density matrices,

$$\gamma^A(1,1') = N \int d(2,3,\ldots,N) \Psi(1,2,3,\ldots,N) \Psi^*(1',2,3,\ldots,N) \tag{1.54}$$

$$\gamma_{ij}^A = \int d1 d1' \psi_i(1) \gamma^A(1,1') \psi_j^*(1') \tag{1.55}$$

$$\gamma^A(1,2;1',2') = N(N-1)\int d(3,\ldots,N)\Psi(1,2,3,\ldots,N)\Psi^*(1',2',3,\ldots,N) \tag{1.56}$$

$$\gamma^A_{ijkl} = \int d(1,1',2,2')\psi_i(1)\psi_j(2)\gamma^A(1,2;1',2')\psi^*_k(1')\psi^*_k(2') \tag{1.57}$$

In the case of a single determinant, the electronic energy simplifies to

$$E^{HF}_A = \sum_i^N h_1^{ii} + \frac{1}{2}\sum_{ij}^N (h_2^{ijij} - h_2^{ijji}) \tag{1.58}$$

where

$$J_{ij} = J_{ijij} = h_2^{ijij} \qquad , \qquad K_{ij} = K_{ijji} = h_2^{ijji}. \tag{1.59}$$

This can be worked out using the Slater Condon rules or by using the second quantization formalism.

1.2.5 The Self-Consistent Field Method

The determine a set of spinors ψ_i for the Hartree-Fock Hamiltonian (or any given expression of the total electronic energy), we can apply the variational principle directly. We can consider the variation of the spinors under the constraint that they remain orthonormal, which can be realized using the method of Lagrange multipliers

$$L[\{\psi_i\},\{\epsilon_{ij}\}] = E_A[\{\psi_i\}] - \sum_{ij}^N \epsilon_{ij}(\langle\psi_i|\psi_j\rangle - \delta_{ij}) \tag{1.60}$$

with ϵ_{ij} representing the Lagrange multipliers. The variation of L with respect to the spinors is written

$$\delta_{\psi_i} L[\{\psi_i\},\{\epsilon_{ij}\}] = \delta_{\psi_i} E_A - \sum_j^N \epsilon_{ij}(\langle\delta\psi_i|\psi_j\rangle + \langle\psi_j|\delta\psi_i\rangle) \tag{1.61}$$

and the variation with respect to the Lagrange multipliers yields a restatement of the spinor orthogonality condition,

$$\frac{\partial L[\{\psi_i\},\{\epsilon_{ij}\}]}{\partial\epsilon_{ij}} = \langle\psi_i|\psi_j\rangle - \delta_{ij}. \tag{1.62}$$

Further simplifying, if we write E_A as the Hartree-Fock energy in terms of the density matrices (for state A, where the label is omitted for brevity)

$$E^{HF} = \sum_{ij}^{N} \gamma_{ij} h_1^{ij} + \frac{1}{2} \sum_{ijkl}^{N} \Gamma_{ijkl} h_2^{ijkl} \tag{1.63}$$

We can write the variation of the Lagrangian functional with respect to the spinors

$$\delta L[\{\psi_i\}] = \sum_{j}^{N} \gamma_{ij} \langle \delta\psi_i | \hat{h}_1 | \psi_j \rangle \tag{1.64}$$

$$+ \frac{1}{2} \sum_{jkl}^{N} \Gamma_{ijkl} \Big(\langle \delta\psi_i(1)\psi_j(2) | \hat{h}_2 | \psi_k(1)\psi_l(2) \rangle - \langle \delta\psi_i(1)\psi_j(2) | \hat{h}_2 | \psi_k(2)\psi_l(1) \rangle \Big) \tag{1.65}$$

$$- \sum_{j}^{N} \epsilon_{ij} \langle \delta\psi_i | \psi_j \rangle + c.c. = 0 \tag{1.66}$$

Since this holds for any arbitrary variation of $\delta\psi_i$, we can factor out the common $\langle \delta\psi_i(1)|$ and write

$$\sum_{j}^{N} \left[\gamma_{ij} h_1 \psi_j + \sum_{kl}^{N} \Gamma_{ijkl} (J_{kl} - K_{kl}) \psi_j - \epsilon_{ij} \psi_j \right] = 0 \tag{1.67}$$

where

$$J_{kl}\psi_j = \psi_j(1) \int d2 \, \langle \psi_k(2) | \hat{h}_2(1,2) | \psi_l(2) \rangle \tag{1.68}$$

$$K_{kl}\psi_j = \psi_l(1) \int d2 \, \langle \psi_k(2) | \hat{h}_2(1,2) | \psi_j(2) \rangle \tag{1.69}$$

Defining the Fock operator f_{ij} [177],

$$f_{ij} = \gamma_{ij} h_1 + \sum_{kl}^{N} [J_{kl}(\boldsymbol{r}) - K_{kl}(\boldsymbol{r})] \tag{1.70}$$

the stationarity condition now can be written

$$\sum_{j}^{N} f_{ij}\psi_j = \sum_{j}^{N} \epsilon_{ij}\psi_j \tag{1.71}$$

or, equivalently,

$$\left[f_{ii} - \epsilon_{ii}\right]\psi_i = \sum_{j\neq i}^{N} [\epsilon_{ij} - f_{ij}]\psi_j \tag{1.72}$$

For a given ψ_i, all the other ψ_j enter into the Fock matrix f_{ij} through the Coulomb and exchange integrals, Eq. (1.68) and Eq. (1.69). The intertwined nature of the equations is through these integrals, hence the adjective "self-consistent". Therefore, the N SCF equations cannot be solved independently of each other; an iterative procedure is required that starts from a guess and converges (hopefully) towards the self-consistent solution.

It can be shown that the Fock matrix is invariant under unitary transformations[177]

$$F \rightarrow U^{\dagger}FU \tag{1.73}$$

which implies that the spinors themselves are not unique, as two sets of spinors related by a unitary transformation give the same total energy. This represents a "gauge freedom" in the choice of spinors. Other representations of the spinors attempt to get closer to the chemical intuition of an "orbital", including the *canonical orbital* representation, wherein the matrix of Lagrange multipliers ($\boldsymbol{\epsilon} = \epsilon_{ij}$)is diagonalized,

$$\mathcal{U}^{\dagger}\boldsymbol{\epsilon}\mathcal{U} = \tilde{\boldsymbol{\epsilon}}_{\text{diag}} \tag{1.74}$$

to obtain to the canonical Hartree-Fock equations,

$$f_{ii}\tilde{\psi}_i = \tilde{\epsilon}_{ii}\tilde{\psi}_i \tag{1.75}$$

These equations are of a more friendly form than the standard HF equations and are often

used in practice. Unless otherwise noted, we will be working with canonical orbitals for the remainder of this work.

1.2.6 The CI Eigenvalue Equation

Here we briefly overview the CI eigenvalue equation, which is used to determine the CI coefficients C_{IA} in Eq. (1.44). The derivation begins with the expectation value for the energy,

$$E_A = \frac{\langle \Psi_A | \hat{H} | \Psi_A \rangle}{\langle \Psi_A | \Psi_A \rangle} = \frac{\sum_{KL} C_{KA}^* C_{LA} \langle \Psi_K | \hat{H} | \Psi_L \rangle}{\sum_K C_{KA}^2} \tag{1.76}$$

where on the bottom line the orthonormality of the determinants has been used. Requiring stationarity with respect to variations in the CI coefficients,

$$\frac{\partial E_A[\{C_{KA}\}]}{\partial C_{IA}} = 0 \tag{1.77}$$

yields, after some simplification,

$$\sum_L \langle \Psi_I | \hat{H} | \Psi_L \rangle C_{LA} = E_A C_{IA} \tag{1.78}$$

as the equation that must hold for any CI coefficient. This may be written in matrix form,

$$\boldsymbol{H}\boldsymbol{C}_A = E_A \boldsymbol{C}_A \tag{1.79}$$

The issue is that the dimensionality of the Hamiltonian scales combinatorically in N, just as the CI coefficients do. It will quickly become too large to store, as Section 1.2.1 demonstrated. However, there are some empirical facts about the Hamiltonian that turn out to be commonly true for most molecular systems in chemistry.

First, the Hamiltonian is diagonally dominant, since the one-body terms are purely diagonal and the interaction terms tend to be mostly on the diagonal. Second, the Hamiltonian is sparse, which stems from the fact that highly excited determinants tend to have little

amplitude in the overall wavefunction. Both of these are empirical rules-of-thumb that tend to be true for most chemical systems, and ultimately stem from the effectiveness of the antisymmetric Slater determinant ansatz. In fact, deviations from these assumptions can be thought of as a measure of how correlated a system is. For example, some of the 2D materials synthesized in the basement of the physics lab would have very significant off-diagonal contributions from the electron-electron interaction in a basis of Slater determinants and strongly multi-reference character, which would make them extremely expensive to treat with the usual quantum chemical means outlined here.

The appropriate matrix diagonalization technique for these intractably large, sparse, diagonally dominant Hamiltonians are Krylov subspace methods, which diagonalizes the Hamiltonian in a linear subspace spanned by CI coefficient vectors $\{C_L^1, C_L^2, \ldots, C_L^N\}$ where L labels determinants.

$$\sigma_K = \sum_{pq} \sum_{L} h_1^{pq} \langle K| \hat{E}_{pq} |L\rangle C_L \tag{1.80}$$

$$+ \frac{1}{2} \sum_{pqrs} \sum_{L} h_2^{pqrs} \langle K| \hat{E}_{pq}\hat{E}_{rs} - \delta_{qr}\hat{E}_{ps} |L\rangle C_L \tag{1.81}$$

$$= \sum_{pq} \sum_{L} (h_1')^{pq} \langle K| \hat{E}_{pq} |L\rangle C_L \tag{1.82}$$

$$+ \frac{1}{2} \sum_{pqrs} \sum_{LJ} h_2^{pqrs} \langle K| \hat{E}_{pq} |J\rangle \langle J| \hat{E}_{rs} |L\rangle C_L \tag{1.83}$$

where

$$(h_1')^{pq} = h_{pq} - \frac{1}{2} \sum_{r} h_2^{prrq} \tag{1.84}$$

The $\boldsymbol{\sigma}-$vector is used to project the Hamiltonian onto a reduced-dimensionality subspace, producing a Ritz residual[179] representing the deviation from the exact full-dimension Hamiltonian,

$$\boldsymbol{r} = \boldsymbol{HC} - E\boldsymbol{C} - (\boldsymbol{C}^\dagger \boldsymbol{HC})\boldsymbol{C} \tag{1.85}$$

The residual is orthogonal to $\boldsymbol{C}$, so it can be used to expand the subspace. Krylov subspace

methods work by using the residual to measure convergence and iteratively expand the subspace.

Evaluating Eq. (1.80) requires enumerating the determinants J connected to L through excitation operators. The lists of determinants connected through excitation operators are called the *excitation lists*. The Davidson procedure refactors $\boldsymbol{\sigma}$ from two-body excitation operators into a contraction of one-body operators (see Eq. (1.80)), drastically reducing the size of the *excitation lists*. Because L is connected to exactly $N_e \times (N_o - N_e + 1)$ determinants, the overall size of the *excitation lists* is $\binom{N_o}{N_e \times N_e (N_o - N_e + 1)}$. This is a much more manageable amount of data to store compared to the $\binom{N_o}{N_e} \times \binom{N_o}{N_e}$ Hamiltonian matrix.

In practice, a software implementation based on excitation lists must store the address of each determinant in the configuration space. In the 1980s, Peter Knowles and Nicholas Handy[102,220] devised an addressing scheme that collects the single-excitation-connected determinants together. In the Knowles-Handy scheme, a one-to-one map between integer addresses and configuration determinants is generated using a graphical representation of the orbital space. The graph vertices are labeled by the number of electrons e and the number of orbitals o and have a weight $w(e, o)$. Edges connecting two vertices also have weights $w((e, o), (e', o'))$. The tail vertex (N_e, N_o) has a weight of 1, and the weights of all other vertices are given by

$$w(e, o) = w(e + 1, o + 1) + w(e, o + 1) \tag{1.86}$$

Repeating this procedure results in a set of vertices related by vertical edges (differing only in orbital number) and diagonal edges (differing in both orbital and electron number) starting from a vertex (determinant) with (N_e, N_o). Vertical edges correspond to a configuration in which an orbital $(o+1)$ is unoccupied and are assigned zero weight. Diagonal edges occupy orbital $(o + 1)$ with electron $(e + 1)$ and are weighted as

$$w((e, o), (e + 1)(o + 1)) = w(e + 1, o + 1) + \cdots + w(e + 1, e + 1) \tag{1.87}$$

The addresses then correspond to walks in the graph $e_{o\,o=0}^{N_o-1}$,

$$|K\rangle \leftrightarrow \sum_{o=0}^{N_o-1} w((e_o, o), (e_{o+1}, o+1)) \tag{1.88}$$

The Knowles-Handy addressing scheme is a critical development for CI codes because it removes a costly search to find the address of a particular determinant, instead allowing a practitioner to directly compute the address. It is amenable to sparse storage, which is a necessity for any efficient CI code where the dimension of the Hamiltonian can explode[88].

1.2.7 Correlation Energy

In chemical applications, a single Slater determinant can recover most of the total energy of the system, say, 95%. The last 5%, however, turns out to be of great importance for accurate relative energies and chemical dynamics. The difference between the full CI energy (which is exact, modulo spatial discretization) and the Hartree-Fock energy E_{HF} is called the *correlation energy*, defined as

$$\Delta E_{\text{corr}} = E_{\text{exact}} - E_{\text{HF}} \tag{1.89}$$

Methods going beyond the Hartree-Fock approximation, such as CI, are said to recover or add in electron correlation. In the literature, there is a somewhat loosely defined "decomposition" of electron correlation into *dynamic* and *static* correlation in the literature[16]. Both are defined with respect to a reference state of a single Hartree-Fock determinant. Dynamic correlation is due to contributions generated from a single reference determinant, i.e. the exact wavefunction includes a large number of excited determinants with the same $\hat{S}^2$ symmetry as the reference. Static correlation is due to the failure of a single reference to capture multiple spin symmetries at the ground state. In the physical chemistry community, it's often repeated that static correlation is required to get a qualitatively correct description of states, and dynamic correlation is required to get a quantitatively correct description. While

this picture is intuitive, it is definitely specialized to "well-behaved" chemical systems. Some intuition and experience is needed to examine the electronic structure of a molecule and determine which type of correlation is dominant, but as mentioned, these "definitions" are in flux. For example, a He atoms' ground state has no other states nearby (in energy), thus the ground state is singly degenerate and has essentially no static correlation contribution.

A stretched H_2 molecule is a classic example of the challenges of describing electron correlation. At any interatomic bond distance, the ground state is a singlet state with both electrons in opposite spin configurations. At the equilibrium bond distance, both electrons occupy different spin states of the same spatial orbital (a linear combination of the $1s^2$ states of each H atom). This is nicely represented by a Hartree-Fock state like

$$|\Psi\rangle_{HF} = |\sigma^\alpha \sigma^\beta\rangle \tag{1.90}$$

where σ^τ represents the τ spin state of the ground state wavefunction of the H_2 molecule. (The σ orbital looks like an ellipsoid shrouding the two nuclei; it is the $l = 0$ state formed by the melding of the two 1s states of the H atoms). The HF state here is not exact since the ansatz is still a non-interacting state, but the correlation to be recovered is dominated by dynamic correlation, since there is no degeneracy in the ground state. In the limit of dissociation (where bond distance goes to ∞), the state should be essentially be two independent electrons bound to their nuclei, but with opposite spins (to preserve the spin symmetry). This introduces a degeneracy into the system, as both spin configurations are equivalent. The correct wavefunction for this system is provided by a CI expansion,

$$|\Psi\rangle_{\mathrm{CI}} = \frac{1}{\sqrt{2}}(|s_A^\alpha s_B^\beta\rangle - |s_A^\beta s_B^\alpha\rangle) \tag{1.91}$$

where s_X^τ represents a $1s^\tau$ orbital of atom X. This form of the wavefunction is a superposition of both spin configurations. Neither determinant alone is enough to capture the correct physics of the situation, in particular, while each separate determinant *looks correct*:

it has the correct spatial localization and opposite spins, but neither determinant alone is an eigenfunction of $\hat{S}^2$! At dissociation, the exact singlet state cannot be captured by a single configuration state alone. Two configurations become degenerate and must be combined to restore the proper spin symmetry, thus in this regime static correlation dominates.

Configuration interaction calculations can recover all of the electron correlation if every possible determinant is included in the CI expansion Eq. (1.44). In practice, the $\binom{N_o}{N_e}$ combinatorial scaling of the method precludes doing so for electron numbers over a few dozen. Approximate CI techniques include only a subset of the orbitals, called a *configuration space* or *correlation space*, in their CI expansion.

1.2.8 Density Functional Theory

Density functional theory (DFT) is an alternative approach to electron correlation. The idea is to consider the electron density,

$$\rho(\boldsymbol{r}) = \langle \Psi(\boldsymbol{r}) | \Psi(\boldsymbol{r}) \rangle \tag{1.92}$$

as the fundamental variable of quantum mechanics rather than the wavefunction. This is a desirable direction to go, since the density is a scalar function of three variables (the directions of space), rather than the exponentially many variables required to describe a wavefunction. This requires a theory of quantum mechanics for an electron density, which was elucidated in 1964 by Hohenberg and Kohn[82]. The first Hohenberg-Kohn theorem is an existence theorem. Quoting directly from the 1964 paper,

> "*the external potential $V_{ext}(\boldsymbol{r})$ is (to within a constant) a unique functional of $\rho(\boldsymbol{r})$; since, in turn $V_{ext}(\boldsymbol{r})$ fixes $\hat{H}$ we see that the full many particle ground state is a unique functional of $\rho(\boldsymbol{r})$*"

In other words, the only information you need to actually discern the exact, fully-

interacting wavefunction for a given molecular structure (V_{ext}) is the electron density. This implies that the ground state wavefunction Ψ_0, and thus the ground state energy E_0, can be written as a functional of Ψ_0

$$\rho_0(\boldsymbol{r}) \Leftrightarrow \Psi_0(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) \Leftrightarrow E_0 \tag{1.93}$$

which implies that each component of the ground state energy is also a functional of ρ_0,

$$E_0[\rho_0] = T[\rho_0] + V_{\text{nuc}}[\rho_0] + V_{\text{e-e}}[\rho_0] \tag{1.94}$$

At this point, it makes sense to decouple the nuclear attraction potential from the rest of the terms. This is because $V_{\text{nuc}}[\rho_0]$ is the only system-dependent term, in fact, it defines the system. The rest of the terms couldn't care less about whether the system is a molecule, atom, plasma, or free electron laser. Collecting the universal terms into a new quantity called the *Hohenberg-Kohn functional* $F_{\text{HK}}[\rho]$, and writing the nuclear attraction potential functional as an integral, we obtain

$$E[\rho_0] = \int \rho_0(\boldsymbol{r}) V_{\text{nuc}}(\boldsymbol{r}) d\boldsymbol{r} + F_{\text{HK}}[\rho_0] \tag{1.95}$$

All the Hohenberg-Kohn functional does is take in a ground state density ρ_0 and spits out the sum of the kinetic and interaction terms $\langle \Psi | \, \hat{T} + \hat{V}_{\text{e-e}} \, | \Psi \rangle$. This functional does not (yet!) have a known analytic form[5].

The second Hohenberg-Kohn theorem (also in the 1964 paper), solves an ambiguity in the formalism as given, which is the question of whether there are "degenerate" densities that lead to the same ground state energy. The answer is no, because there exists a variational

[5]Which isn't all that surprising, since solving for it is essentially solving the quantum many-body problem once and for all.

principle associated with the densities. For any trail density $\tilde{\rho_0}$, it is the case that

$$E_0 \leq E[\tilde{\rho_0}] \tag{1.96}$$

so that the Hohenberg-Kohn functional $F_{HK}[\rho]$ is guaranteed to deliver the lowest energy at the true ground state density only. There are no imposters! This result completes the loop and allows us to formally equate wavefunction theory with density functional theory. It should be noted, however, that these results only hold for the ground state. The treatment of excited states in DFT is another topic and will be omitted from this work, although a good starting point for this is in[28,202].

The foundations of DFT are certainly enticing, since it would in principle allow us to forego the clunky many-body wavefunction and only have to do bookkeeping on a scalar function. However, it's not always easy to write observables of interest as functionals of the density. For example, the kinetic energy functional $T[\rho]$ has no known form, which is particularly disastrous since it's a part of the total energy functional and thus critical to getting the correct density in the first place. Thankfully, there's another scheme that swaps this difficulty for a slightly more tractable one. The Kohn-Sham (KS)[103] scheme involves adding a new one-body potential energy term to the Hartree-Fock Hamiltonian which, in principle, yields the exact same density as the fully-interacting problem with just one determinant. Thus, all the machinery of Hartree-Fock can be re-used with the addition of the right one-body potential.

The noninteracting ansatz is simply assuming that the many-body wavefunction is of the form of a single determinant,

$$\Psi = \Theta_K(\psi_1, \ldots, \psi_N) \tag{1.97}$$

where K is the spin configuration of the orbitals ψ_i. If we denote the orbitals of a Hartree-

Fock determinant as ψ_i, then the overall density is simply

$$\rho_{\mathrm{KS}} = \sum_i |\psi_i(\boldsymbol{r})|^2 \tag{1.98}$$

The noninteracting density ansatz allows us to exactly calculate some terms in the Hohenberg-Kohn functional $F_{\mathrm{HK}}[\rho]$,

$$E[\rho_{\mathrm{KS}}] = T_s[\rho_{\mathrm{KS}}] + V_{\mathrm{nuc}}[\rho_{\mathrm{KS}}] + J[\rho_{\mathrm{KS}}] + E_{xc}[\rho_{\mathrm{KS}}] \tag{1.99}$$

where the noninteracting kinetic energy $T_s[\rho]$ is

$$T_s[\rho] = -\sum_i \langle \Psi_{HF}| \frac{1}{2}\nabla_i^2 |\Psi_{HF}\rangle = -\sum_i \langle \psi_i| \frac{1}{2}\nabla^2 |\psi_i\rangle \tag{1.100}$$

and the classical electron-electron interaction functional is

$$J[\rho] = \frac{1}{2}\int\int \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} d\boldsymbol{r} d\boldsymbol{r}' \tag{1.101}$$

and the last piece, $E_{\mathrm{xc}}[\rho]$, is called the *exchange-correlation* (xc) functional. The exchange correlation functional, unlike F_{XC}, contains the interacting contribution to the kinetic energy, $\langle \Psi| \hat{T} - \hat{T}_s |\Psi\rangle$. To obtain the Kohn-Sham orbitals themselves, we can take the functional derivative of E_{xc} to get the exchange-correlation potential,

$$V_{xc}(\boldsymbol{r}) = \frac{\delta E_{xc}[\rho]}{\delta \rho(\boldsymbol{r})} \tag{1.102}$$

which is of a mean-field form, just like the Coulomb and exchange terms $\hat{J}$ and $\hat{K}$. The Schrödinger equation becomes

$$\left(-\frac{1}{2}\nabla^2 + v^{\mathrm{eff}}(\boldsymbol{r}_i)\right)\psi_i(\boldsymbol{r}_i) = \epsilon_i \psi_i(\boldsymbol{r}_i) \tag{1.103}$$

These are the Kohn-Sham equations[50,103].

Just like F_{XC}, the exact exchange correlation functional is not known, but there are a plethora of approximations used in practice[32,202]. The simplest of these is the *local density approximation*, which has the form of the density multiplied by the analytic exchange correlation energy of a uniform electron gas $\epsilon_{xc}^{\text{unif}}(\rho(\boldsymbol{r}))$,

$$E_{xc}^{LDA}[\rho] = \int d\boldsymbol{r}\rho(\boldsymbol{r})\epsilon_{xc}^{\text{unif}}(\rho(\boldsymbol{r})) \tag{1.104}$$

This functional satisfies a number of exact constraints which are true regardless of the specific density, and it found early success in calculations of lattice constants, phonon frequencies, and other properties in solids in the 1970s[170]. More advanced approximations were later developed that depend not only on the local density but on its gradients as well, incorporating nonlocal information into the exchange-correlation energy in an approximation called the *generalized gradient approximation* (GGA),

$$E_{xc}^{\text{approx}}[\rho] = \int d\boldsymbol{r}\rho(\boldsymbol{r})\epsilon_{xc}^{GGA}(\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r})) \tag{1.105}$$

GGAs are generally more accurate than LDAs, including the popular Perdew-Burke-Ernzerhof (PBE) functional[162], although they come at a cost of higher computational cost. The expressions for the GGA exchange-correlation energy per particle, ϵ_{xc}^{GGA}, are carefully tuned interpolations between exact limits derived for electron gases.

Innovation in DFT functional development continues to this day: in 2025, Microsoft developed a neural-network based functional[129] trained on over 150k high-accuracy reaction energies. Their approach forgoes the prescription of an algebraic form for the functional, instead leveraging the universality of the neural network to approximate the exact functional.

1.2.9 Gaussian Basis Sets

Now, we can develop the final piece of the puzzle, the spatial discretization of the one-electron functions $\psi_{i\alpha}(\boldsymbol{r})$, also known as the basis set. Wisely choosing a basis set is quite

important to quantum chemical calculations, as all spatial wavefunctions and operators will be expressed in terms of them. If the basis functions (which comprise a basis set) are too dissimilar from the exact solution, the numerical optimization of the orbitals in a Hartree-Fock or CI calculation may have difficulty converging. On the other hand, if the functions are difficult to compute (but close to exact), the overall calculation will struggle to scale to large molecules.

A good mix is struck by Gaussian type orbitals (GTOs), which are similar to the non-relativistic solutions of the Schrödinger equation for hydrogenlike atoms. GTOs have the form

$$\chi_{\alpha lm}^{\mathrm{GTO}}(r,\theta,\phi) = N_{\alpha lm} Y_{lm}(\theta,\phi) r^l \exp(-\alpha r^2) \tag{1.106}$$

where $N_{\alpha lm}$ is a normalization constant and $Y_{lm}(\theta,\phi)$ is a solid harmonic. GTOs have a controllable α parameter that controls the width of the Gaussian, or how diffuse the function is. Together with the completeness of the spherical harmonics, most realistic nucleus-centered wavefunctions can be expressed reasonably well with GTOs. However, the Gaussian dependence on r differs from the exact hydrogenlike functions, which have the form

$$\chi_{nlm}^{\mathrm{STO}} \approx r^{n-1} \exp(-\xi r) \tag{1.107}$$

where $n > l$ is the principal quantum number and ξ is derivable from physical constants and the label "STO" stands for *Slater type orbital*[79]. The use of STOs is difficult in molecular calculations because of the exponential dependence on r, which makes an exact expression for two-electron integrals impossible to do analytically. Numerical methods must be used in this case. On the contrary, GTOs have exact expressions for the one- and two-electron integrals[77] which can be quickly evaluated on a computer.

The choice of GTOs also is very compatible with the Gaussian charge nuclear model we adopted in Sec. 1.2. Slater type orbitals feature a cusp singularity at the nuclear center due to the exponential falloff, which is essentially an artifact of the point charge nuclear model.

Gaussian basis sets will never be able to reproduce this cusp by design, even with extremely tight functions. Adoption of a Gaussian nuclear charge model alleviates this difficulty, and increases the accuracy of both the basis set and the Hamiltonian[204].

1.3 Relativity

The theory of special relativity deals with the description of physical phenomena as observed from frames of reference that move at constant velocity relative to each other (inertial frames). An inertial frame of reference is a choice of coordinates that can move at a constant velocity. For example, if we want to do a calculation of how a particle emitted from the Sun moves into a detector on Earth, there are two good choices of coordinates: one is moving along with the particle, and one is moving along with the detector. In the first frame, the detector is moving towards the particle and the particle is stationary, and vice versa for the second frame. A basic tenet of physics is that a physical law should have the same form in all inertial frames. For "normal" speeds encountered in everyday life, the relationship between inertial frames is expressed in terms of Galilean transformations. Mathematically, Galilean transformations are a group generated by translations in space, translation in time, rotations, and rotationless boosts (pure changes in velocity)[69]. Suppose two inertial frames S and S' are moving relative to each other with speed v such that S' is moving in the positive x direction. If the coordinate axes of S' are parallel to those of S, then the Galilean transformation is

$$x' = x - vt \qquad , \qquad t' = t \tag{1.108}$$

Transformations like these are certainly fine for cars and trains passing by, and is appropriate for many situations in classical physics. However, there's trouble ahead, because after applying this transformation to Maxwells equations, they change form. Then the laws of electromagnetism are different for someone passing by down the street! The solution to this paradox was provided by Albert Einstein in 1905[52] with the following postulate: *The laws of physics are identical in all inertial frames.* From this statement follows all of the

laws of special relativity, which require Lorentz transformations to connect inertial frames of reference. With the same S, S' example from before:

$$x' = \gamma(x - vt) \qquad , \qquad t' = \gamma(t - vx/c^2) \tag{1.109}$$

where $\gamma = \frac{1}{\sqrt{1-(v/c)^2}}$ is the *Lorentz factor*, a scalar value that approaches ∞ as the relative speed between frames increases. The invariance of the laws of physics under Lorentz transformations has profound implications, a taste of which can be gleaned just from our example above. Note how the primed time coordinate t' depends on not just t, but x as well. Lorentz transformations that involve a relative velocity change (i.e. not just a rotation in space or translation in time) will mix together space and time coordinates, implying that time does not progress at the same rate as observed from different frames. The transformed x coordinate has a factor of γ, exemplifying length contraction, the phenomena of objects becoming squashed in the direction of relative motion.

1.4 Relativistic Quantum Mechanics

Combining special relativity with quantum mechanics is not an easy task, and its full solution took decades of development throughout the twentieth century. Quantum electrodynamics (QED) is a quantum field theory of electrons and photons and accurately describes arbitrarily high–energy interactions[201]. While useful for physics calculations, chemistry is mostly a low-energy regime where a full field–theoretical treatment is unnecessary. In this section, we will develop the effective equations of relativistic quantum mechanics used in quantum chemistry calculations.

The time-dependent Schrödinger equation is first order in time but second-order in position (from the kinetic energy term). Relativity tells us that space and time coordinates can be mixed together by changing the velocity of the reference frame, so this equation most definitely is not invariant under Lorentz transformations[177,201]. The earliest successful attempt at writing a covariant theory of quantum mechanics was due to Dirac[39,40,188].

1.4.1 The Dirac Equation

Here we will derive the Dirac equation (as it became to be known) following a different derivation than the original. Starting from the energy-momentum relation of a relativistic particle,

$$E^2 = m^2c^4 + p^2c^2 \tag{1.110}$$

we can perform the usual substitution of operators in place of position and momentum coordinates and set $E \to i\frac{\partial}{\partial t}$ to get the quantized version:

$$-\frac{\partial^2}{\partial t^2} = m^2c^4 + \hat{p}^2c^2 \tag{1.111}$$

If we substitute the position-space representation of the squared momentum operator, $\hat{p}^2 \to -\nabla^2$, we obtain the Klein-Gordon equation[201]

$$-\frac{\partial^2}{\partial t^2}\psi(\boldsymbol{r},t) = (-c^2\nabla^2 + m^2c^4)\psi(\boldsymbol{r},t) \tag{1.112}$$

which is second order in both time and space. This is in fact a relativistically invariant equation (as can be proved), but it is not quantum mechanical. The problem is that it leaks probability: the norm of a state $\int d\boldsymbol{r}\psi^*(\boldsymbol{r},t)\psi(\boldsymbol{r},t)$ doesn't remain constant over time. A different approach is needed. Starting back from Eq. (1.111), we can apply the Dirac identity

$$(\boldsymbol{\sigma}\cdot\boldsymbol{A})(\boldsymbol{\sigma}\cdot\boldsymbol{B}) = \boldsymbol{A}\cdot\boldsymbol{B} + i\boldsymbol{\sigma}\cdot(\boldsymbol{A}\times\boldsymbol{B}) \tag{1.113}$$

or, more precisely, a special case of it:

$$(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}}) = \hat{p}^2 \tag{1.114}$$

where $\hat{p}^2$ is a scalar equal to the momentum of the particle squared and the Pauli vector, $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$, where

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{1.115}$$

Applying it to the $\hat{\boldsymbol{p}}^2$, multiplying by a spinor wavefunction ψ_1, and dividing by c^2, we obtain

$$\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \hat{p}^2\right)\psi_1 = \left(\frac{i}{c}\frac{\partial}{\partial t} + (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})\right)\left(\frac{i}{c}\frac{\partial}{\partial t} - (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})\right) = (mc)^2\psi_1 \tag{1.116}$$

With the Dirac identity applied, the dimension of the equation increased to accommodate the Pauli matrices. This equation now only works for spinor wavefunctions ψ_1, and seems to suggest that the spin degree of freedom is "hidden" in the non-relativistic kinetic energy operator. However, let us pause this rumination and boldly continue on. Identifying the rightmost parentheses with a new wavefunction,

$$mc\phi_2 = \left(\frac{i}{c}\frac{\partial}{\partial t} - (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})\right)\phi_1 \tag{1.117}$$

we can write our working equation as two coupled equations

$$\begin{cases} \left(\frac{i}{c}\frac{\partial}{\partial t} - (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})\right)\phi_1 = mc\phi_2 \\ \left(\frac{i}{c}\frac{\partial}{\partial t} + (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})\right)\phi_2 = mc\phi_1 \end{cases} \tag{1.118}$$

which can be further symmetrized by taking the linear combinations $\phi_1 + \phi_2$ and $\phi_2 - \phi_1$,

$$\begin{cases} \frac{i}{c}\frac{\partial}{\partial t}(\phi_1 + \phi_2) - (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})(\phi_1 - \phi_2) = mc(\phi_1 + \phi_2) \\ -\frac{i}{c}\frac{\partial}{\partial t}(\phi_1 - \phi_2) + (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}})(\phi_1 + \phi_2) = mc(\phi_1 - \phi_2) \end{cases} \tag{1.119}$$

Introducing the large and small components, ψ^L and ψ^S, the equation can finally be written in matrix form

$$\begin{bmatrix} \frac{i}{c}\frac{\partial}{\partial t} & -(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}}) \\ (\boldsymbol{\sigma}\cdot\hat{\boldsymbol{p}}) & \frac{i}{c}\frac{\partial}{\partial t} \end{bmatrix} \begin{bmatrix} \psi^L \\ \psi^S \end{bmatrix} = mc \begin{bmatrix} \psi^L \\ \psi^S \end{bmatrix} \tag{1.120}$$

which can also be stated as the conventional form of the Dirac equation[39,40],

$$\left(\beta mc^2 + c(\boldsymbol{\alpha}\cdot\hat{\boldsymbol{p}})\right)\psi = i\partial_t\psi \tag{1.121}$$

where

$$\psi = \begin{bmatrix} \psi^L \\ \psi^S \end{bmatrix} = \begin{bmatrix} \psi^L_\alpha \\ \psi^L_\beta \\ \psi^S_\alpha \\ \psi^S_\beta \end{bmatrix} \tag{1.122}$$

and

$$\boldsymbol{\alpha} = \begin{bmatrix} 0_2 & \boldsymbol{\sigma} \\ \boldsymbol{\sigma} & 0_2 \end{bmatrix} \quad , \quad \begin{bmatrix} I_2 & 0_2 \\ 0_2 & -I_2 \end{bmatrix} \tag{1.123}$$

are the Dirac matrices. The Dirac equation is indeed invariant under Lorentz transformations[166], although the full proof is beyond the scope of this work, we do really have a quantum theory compatible with special relativity! There are some interesting features we picked up along the way, however. It seems that during the procedure of "unfolding" the 2×2 Pauli matrices out of the $\hat{\boldsymbol{p}}^2$ operator, we forced the wavefunction to be a 4-dimensional vector. In other words, we've postulated that a relativistic expansion of the kinetic energy should depend on spin, which is of dimension 2, and we've gotten out twice that. Indeed, the Dirac matrices are 4×4 matrices which inherit an algebra from the Pauli matrices:

$$\{\alpha^j,\alpha^k\}_{ab} = 2\delta^{jk}_{ab} \quad , \quad \{\alpha^j,\beta\}_{ab} = 0 \quad , \quad (\beta^2)_{ab} = \delta_{ab} \tag{1.124}$$

where $\{A, B\} = AB + BA$ is the anticommutator and $\{a, b\}$ are matrix element indices. It turns out that the exact form of the Dirac matrices does not matter, so long as these properties are fulfilled.

Let's examine the structure of this equation and its solutions for a free particle before moving to molecules. First, what do the four components of a Dirac wavefunction correspond to? We jumped the gun a little with the L, S, α, β labelling of the components of ψ in Eq. (1.122), but consider the operator $\hat{H}^2$:

$$(\beta mc^2 + c(\boldsymbol{\alpha} \cdot \hat{\boldsymbol{p}}))^2_{ab} = \hat{H}^2_{ab} = (\hat{\boldsymbol{p}}^2 c^2 + m^2 c^4)\delta_{ab} \tag{1.125}$$

The eigenstates of $\hat{H}^2$ are momentum eigenstates with eigenvalue $p^2c^2 + m^2c^4$, the correct relativistic energy-momentum relation. Well, together with the fact that $\hat{H}$ is traceless, that means that the four free-particle eigenvalues must be $+\sqrt{p^2c^2 + m^2c^4}, -\sqrt{p^2c^2 + m^2c^4}$, each twice degenerate. This means that the spectrum of solutions for the free-particle Dirac equation Eq. (1.121) has a gap of $2mc^2 \approx 1.02$ MeV $\approx 3.7 \times 10^5$ Ha, well into the gamma ray part of the electromagnetic spectrum.

We won't work it out in gory detail here (see[47] for a great discussion), but we will state that the positive energy free particle solutions have significant amplitudes in the top two ("large") components, and the negative energy solutions have significant amplitude in the bottom two ("small") components. Thus, we can associate the top two components with a spinor (a.la. nonrelativistic physics) which we denote ψ^L, and bottom two with ψ^S. The free particle solutions possess "pure" spinor solutions $\binom{1}{0}$ or $\binom{0}{1}$ in the L/S sector, but non-zero occupation in the S/L sector. Therefore, even a positive energy solution always has a little amplitude in the small space. Additionally, with some work, it can be shown the small amplitudes of the free-particle eigenstates go to zero in the non-relativistic ($c \to \infty$) limit. Thus, we can identify the large components with the spinor solution of the typical non-relativistic free-particle Hamiltonian.

1.4.2 Coupling to an external field

To arrive at a form of the Dirac equation that is useful for quantum chemistry, we again make the Born-Oppenheimer approximation to treat the nuclei classically and use minimal substitution[47] to couple Eq. (1.121) to an external electromagnetic field,

$$\hat{H}^D = \left(c\boldsymbol{\alpha}\cdot(-i\nabla + \boldsymbol{A}) + \beta c^2 + \phi\right)\psi = E\psi \tag{1.126}$$

where atomic units have been applied and $\hat{\boldsymbol{p}} = -i\nabla$ written in the position space representation. Special relativity forbids us from cleanly separating space and time variables, because Lorentz transformations mix them. We can perform this separation, however, if we fix the frame of reference. In this case, it is a reasonable choice to choose the frame in which the nuclei are stationary, called the Born-Oppenheimer frame, because then the nuclear potential is simply the static Coulomb potential ($\boldsymbol{A}(\boldsymbol{r}) = 0$, $\phi(\boldsymbol{r}) = V(\boldsymbol{r})$). Shifting the energy scale by $-c^2$ so that the electronic energy spectrum begins at 0 instead of c^2, we can write the modified Dirac equation

$$\begin{pmatrix} V - E & -c(\boldsymbol{\sigma}\cdot\nabla) \\ -c(\boldsymbol{\sigma}\cdot\nabla) & V - E - 2c^2 \end{pmatrix} \begin{pmatrix} \psi^L \\ \psi_S \end{pmatrix} = 0 \tag{1.127}$$

This form of the Dirac equation can be used to obtain single-electron bispinor solutions to atoms and molecules, the structure of which is encoded in V. In stark contrast to the non-relativistic version of this equation, Eq. (1.38), there is no familiar Laplacian kinetic energy term. Some algebraic manipulation will reveal what's going on: If we solve for ψ_S and eliminate it from Eq. (1.127), we obtain

$$(V - E)\psi_L + c^2(\boldsymbol{\sigma}\cdot\boldsymbol{p})\left[2c^2 + E - V\right]^{-1}(\boldsymbol{\sigma}\cdot\boldsymbol{p})\psi_L = 0 \tag{1.128}$$

which, by the identity $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$, is equal to

$$(V-E)\psi_L + \frac{1}{2}(\boldsymbol{\sigma}\cdot\boldsymbol{p})(\boldsymbol{\sigma}\cdot\boldsymbol{p})\psi_L - \frac{1}{2}(\boldsymbol{\sigma}\cdot\boldsymbol{p})\frac{E-V}{2c^2+E-V}(\boldsymbol{\sigma}\cdot\boldsymbol{p})\psi_L = 0 \tag{1.129}$$

This is true as long as $(2c^2+E-V)^{-1}$ is nonzero, which is the case since V is strictly nonpositive and we only are considering the electronic solutions where $E > 0$. Using the Dirac identity Eq. (1.113) on each term, this simplifies to

$$\left[V - E + \frac{1}{2}p^2 - \frac{E-V}{2c^2+E-V}\frac{1}{2}p^2 + \frac{c^2}{(2c^2+E-V)^2}\left((\boldsymbol{p}V)\cdot\boldsymbol{p} + i\boldsymbol{\sigma}\cdot(\boldsymbol{p}V)\times\boldsymbol{p}\right)\right]\psi_L = 0 \tag{1.130}$$

where this whole operator acts on the spinor ψ_L. The first three terms are the familiar nonrelativistic Hamiltonian, and the next three are relativistic corrections. The first correction is a rescaled kinetic energy, the next is a scalar potential term, and the last term involves spin and can be classified as a spin-orbit interaction[47]. This shouldn't be too much of a surprise, since boosting to the (instantaneous) frame of a bound electron, the positively charged nucleus appears to be moving, thus generating a magnetic field which couples to the electrons spin. Evidently, hidden in the structure of the Dirac equation Eq. (1.127) there's a spin-orbit interaction purely due to relativity, even without the explicit inclusion of a magnetic interaction!

1.5 High-Performance Computing

The introduction of computers revolutionized the field of quantum chemistry, starting with Roothaan's developments in the 1950s[183] which formulated the electronic structure problem in terms of matrices and continuing today with massive 10^{15} determinant CI calculations[88]. Quantum chemistry algorithms have co-evolved with computing paradigms. In the early days, limited memory prevented the storage of integrals, driving developments in fast algorithms for on-the-fly integral evaluation[21,141,151]. The introduction of vectorized computation in the 1980s marked the rise of *high-performance computing* and the fledgling

paradigm found application in the Rys quadrature integral algorithm[42]. Vector computers utilize SIMD (single instruction, multiple data) instructions to do calculations at a much higher throughput than scalar (single instruction, single data) processors. The next few decades saw the rise of parallel processing at every scale: not only did individual processors gain multiple cores, but the message passing interface (MPI) let multiple computers (*nodes*) share data and work concurrently on the same task. The shift from vector processing to distributed computing came with new challenges in parallel algorithm design, as there was a communication time associated with moving data between nodes. Simultaneously, there was a shift away from specially-built supercomputers and towards *commodity* clusters; MPI-linked computers built from off-the-shelf parts could be found in every university. New integral and SCF algorithms[9,34,64,118] for distributed-memory systems were developed for these now widely available computational resources.

Yet another paradigm shift came shortly thereafter, this time not from a national lab or university, but from video games. Graphics processing units (GPUs) are specialized processors with many thousands of cores intended for real-time 3D graphics. The development of general–purpose GPU programming languages such as NVIDIA CUDA[149] produced an explosion of GPU-accelerated quantum chemistry starting in the late 2000s, including integrals[2,106,112], DFT exchange-correlation potentials[105,109], SCFs[10], quantum monte carlo[55,198], CI[58], and even full quantum chemistry packages[193]. GPUs are now integrated into every computing cluster, creating a hybrid CPU-GPU paradigm known as *heterogeneous computing*.

1.6 Relativistic Electronic Structure Theory

In this section, we will develop in some more detail the preliminaries for understanding the primary work of this thesis.

Chemists are concerned with the *electronic* solutions of the low-energy limit of QED. The Dirac equation, which is the low-energy limit of QED, describes both electrons and positrons, and is thus overkill for chemical applications. Thankfully, there is a unitary

transformation to the four-component Dirac Hamiltonian which decouples the electronic and positronic degrees of freedom in the Dirac equation, called the exact-two component (X2C) method[41,43,44,60,80,90,107,127]:

$$\mathcal{U}^\dagger \hat{H}_{4C} \mathcal{U} = \begin{pmatrix} \hat{H}_+ & 0 \\ 0 & \hat{H}_- \end{pmatrix} \tag{1.131}$$

where $\hat{H}_\pm$ are the decoupled postive/negative energy sectors of the Dirac Hamiltonian in Eq. (1.121). The exact unitary matrices $\mathcal{U}$ depend on both the one-electron and two-electron parts of the Dirac Hamiltonian. After decoupling, the negative energy Hamiltonian is discarded, resulting in a purely electronic, positive-energy Hamiltonian $\hat{H}_+$ that has spinor solutions as in non-relativistic theory, but that represent the actual electronic solutions of the full Dirac equation. The one-electron X2C method (1eX2C) provides an explicit construction of the unitary transformation $\mathcal{U}$ based on just the one-electron parts of $\hat{H}_{4C}$. Since it is exact for one-electron terms, X2C preserves the nuclear spin-orbit coupling physics of the four-component Hamiltonian exactly. The details of constructing $\mathcal{U}$ are beyond the scope of this work, but the X2C Hamiltonian will be used to describe all relativistic quantum chemistry throughout.

From our experience with the relativistic treatment of the nuclear attraction, we can expect the relativistic version of the two-electron interaction to introduce new and complicated terms. The derivation of a fully Lorentz-invariant interaction is a reasonably intense exercise in quantum electrodynamics[18], where the Lorentz-invariance is guaranteed by expressing the interaction as an exchange of virtual photons. The need for Lorentz-invariant treatment of photons (responsible for *retardation effects*) is obviated by the equal-time approximation, in which all local clocks are assumed to be ticking along at the same speed (and thus one time t can be assigned to each of their wavefunctions, as in non-relativistic physics). Another approximation can be made in which the energy of the exchanged photon is low enough to the point where we do not need to concern ourselves with pair-creation

effects. This is related to the retardation condition, since a high frequency photon will acquire a phase along its path of travel, and is equivalent to treating the interaction only with $t-$ and $u-$channel diagrams.

These two approximations let us write an expression for the interaction between electrons 1 and 2 in the Coulomb gauge as

$$V_{ij} = \frac{1}{r_{ij}} - \frac{\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{2r_{ij}} - \frac{1}{2}\frac{(\boldsymbol{\alpha}_i \cdot \boldsymbol{r}_{ij})(\boldsymbol{\alpha}_j \cdot \boldsymbol{r}_{ij})}{r_{ij}^3} \tag{1.132}$$

The two rightmost terms are called the *Breit terms*, which together with the nonrelativistic Coulomb interaction is exactly equivalent to the classical relativistic interaction potential for two moving charges with velocities $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ in the nonrelativistic limit ($\boldsymbol{u}/c << 1$)

$$V_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 r}\left[1 - \frac{1}{2}\frac{\boldsymbol{u}_1 \cdot \boldsymbol{u}_2}{c^2} - \frac{1}{2}\frac{(\boldsymbol{u}_1 \cdot \boldsymbol{r})(\boldsymbol{u}_2 \cdot \boldsymbol{r})}{c^2 r^2}\right] \tag{1.133}$$

with the substitution $\boldsymbol{u}_i \to c\boldsymbol{\alpha}_i$. Neat!

The Breit term contains all terms of order c^{-2} including complicated spin interactions between electrons, such as the spin-spin and spin-other-orbit interactions. Higher order terms include the other two one-photon diagrams, the self-energy and the vacuum polarization, which lead to the Lamb shift[?] . While the Breit term is important for late-row chemistry and theoretical spectroscopy of actinides[234], they are computationally expensive to include, as they require a full four component wavefunction.

In practical relativistic calculations, the scalar relativistic corrections and nuclear spin-orbit coupling of X2C is enough. Further motivated by the fact the Breit interaction is suppressed by c^{-2}, we will continue our relativistic treatment at the level of the X2C transformed Hamiltonian $\hat{H}_+$. To get relativistic spinor solutions with X2C, the two-electron parts of the non-relativistic Hamiltonian are added to $\hat{H}_+$ before diagonalization.

1.6.1 Relativistic DFT

In Section 1.2.8, density functional theory was introduced using an RHF wavefunction, where spin is essentially integrated out. It is possible to formulate relativistic DFT in terms of the relativistic four-current [47,201], but we will not consider it here due to the lack of accurate and readily available four-current functionals. Nevertheless, DFT can still be formulated in terms of a X2C generalized-spin (GHF) determinant, but multiple densities are required to fully specify the interacting wavefunction as a functional of them. For a GHF determinant, the density has spin indices which can be decomposed like

$$\rho_{\tau\tau'}(\boldsymbol{r}) = \begin{bmatrix} \rho^{\alpha\alpha}(\boldsymbol{r}) & \rho^{\alpha\beta}(\boldsymbol{r}) \\ \rho^{\beta\alpha}(\boldsymbol{r}) & \rho^{\beta\beta}(\boldsymbol{r}) \end{bmatrix}_{\tau\tau'} = \frac{1}{2} n(\boldsymbol{r}) \left[\mathbb{I}_2\right]_{\tau\tau'} + \frac{1}{2} \boldsymbol{m}(\boldsymbol{r}) \circ \left[\boldsymbol{\sigma}\right]_{\tau\tau'} = \sum_{i}^{occ} |\left\langle \psi_{i\tau} | \psi_{i\tau'} \right\rangle|^2 \quad (1.134)$$

where $\circ$ denotes a dot product in (τ, τ') spin indices. Now, $n(\boldsymbol{r})$ (the *number density*) and $\boldsymbol{m}(\boldsymbol{r})$ (the *magnetization density*) form a set of purely spatial functions that fully describe the two-component density. These variables form a set of densities that have their own Hohenberg-Kohn theorems, and a corresponding Kohn-Sham equation [91]. The exchange-correlation potential also gains spin indices,

$$V_{xc}(\boldsymbol{r}) = V_{xc}^{S}\, \mathbb{I}_2 + V_{xc}^{I} \sigma^{I} \quad (1.135)$$

where the indices $I = \{X, Y, Z\}$ and S indicate a quarternion decomposition of V_{xc}.

The most widely used functionals in quantum chemistry implicitly assume the density to be collinear, i.e. the spin magnetization is aligned with some constant direction (conventionally, the z-axis). For the case of GGAs, these functionals depend on the $\uparrow$ and $\downarrow$ spin densities. These functionals are based on an xc energy of the form (for GGAs)

$$E_{xc} = \int d\boldsymbol{r}\, f(n_{\uparrow}, n_{\downarrow}, \gamma_a, \gamma_b, \gamma_c) \quad (1.136)$$

where

$$\gamma_a = \gamma_{\uparrow\uparrow} = \nabla n_\uparrow \cdot \nabla n_\uparrow \tag{1.137}$$

$$\gamma_b = \gamma_{\downarrow\downarrow} = \nabla n_\downarrow \cdot \nabla n_\downarrow \tag{1.138}$$

$$\gamma_c = \gamma_{\uparrow\downarrow} = \nabla n_\uparrow \cdot \nabla n_\downarrow \tag{1.139}$$

$$\tag{1.140}$$

and the spin densities are

$$n_\uparrow(\boldsymbol{r}) = \rho_{\alpha\alpha} + \rho_{\beta\beta} \tag{1.141}$$

$$n_\downarrow(\boldsymbol{r}) = \rho_{\alpha\alpha} - \rho_{\beta\beta} \tag{1.142}$$

$$\tag{1.143}$$

The set of variables upon which a functional depends, the $\{U_i\}$ variables, are defined in terms of the density variables $\{V_i\}$. In this case, $\{U_i\} = \{n_\uparrow, n_\downarrow, \gamma_a, \gamma_b, \gamma_c\}$ and $\{V_i\} = \{n_\uparrow, n_\downarrow, \nabla n_\uparrow, \nabla n_\downarrow\}$. The xc potential is calculated by taking the first derivative of E_{xc} with respect to the density matrix elements which is evaluated with repeated application of the chain rule,

$$[V_{xc}]_{\tau\tau'} = \frac{\partial E_{xc}}{\partial [P_{\mu\nu}]_{\tau\tau'}} = \sum_{ij} \int d\boldsymbol{r} \frac{\partial f(\{U_i(\boldsymbol{r})\})}{\partial U_i(\boldsymbol{r})} \frac{\partial U_i(\boldsymbol{r})}{\partial V_j(\boldsymbol{r})} \frac{\partial V_j(\boldsymbol{r})}{\partial [P_{\mu\nu}]_{\tau\tau'}} \tag{1.144}$$

where the spin indices for U and V variables are implied. The separation into $\{U_i\}$ and $\{V_i\}$ variables is motivated by convenience. The scalar form of the $\{U_i\}$ variables appeal to DFT functional designers and the ease of calculation of the $\{V_i\}$ makes quantum chemistry software designers happy.

The collinear ansatz is a reasonable assumption for many chemical systems, but some have true spin non-collinearity, like spin-frustrated systems. As proven in Sec. 1.4.2, the relativistic density has one-body contributions that include spin-orbit effects. This can introduce a *noncollinear* magnetization density that points in different directions throughout

the molecule, unsuitable for collinear DFT functionals. However, the a thoughtful choice of $\{V_i\}$ variables in this case can transform the GHF density matrix into a set of $\{U_i\}$ variables compatible with collinear functionals. One such set is given by[50]

$$n_{\pm} = \frac{1}{2} \pm \frac{1}{2}\sqrt{\boldsymbol{m} \circ \boldsymbol{m}} \tag{1.145}$$

$$\gamma^{\pm\pm} = \frac{1}{4}\nabla n \cdot \nabla n + \frac{1}{4}\nabla \boldsymbol{m} \cdot \nabla \boldsymbol{m} \pm \frac{f_{\nabla}}{2}\sqrt{\nabla n \cdot \nabla \boldsymbol{m} \circ \boldsymbol{m} \cdot \nabla n} \tag{1.146}$$

$$\gamma^{+-} = \frac{1}{4}\nabla n \cdot \nabla n - \frac{1}{4}\nabla \boldsymbol{m} \circ \nabla \boldsymbol{m} \tag{1.147}$$

$$f_{\nabla} = \text{sign}(\nabla n \cdot \nabla \boldsymbol{m} \circ \boldsymbol{m}) \tag{1.148}$$

which reduces to the collinear set of variables if $\boldsymbol{m} = m_z \hat{z}$. This choice of $\{U_i\}$ variables obeys the zero global torque theorem[26], a restriction on the exact functional that disallows a global torque from arising from an xc potential. However, a local torque may be induced because these variables incorporate both the parallel (as in the collinear case) and the perpendicular components of the magnetization gradient. It is this property that permits the correct non-collinear spin structure when using these variables.

However, Eq. (1.145) is prone to numerical instability when the norm of the magnetization vector $|\boldsymbol{m}(\boldsymbol{r})|$ is small, because of the appearance of $\boldsymbol{m}$ in the denominator in some of the expressions for $\frac{\partial U_i}{\partial V_j}$. If derivatives of the total energy expression (Eq. (1.99)) are required, as in perturbation theory, then the chain rule introduces more magnetization vector singularities into the expression. To avoid having the magnetization length m in the denominator of the functional derivatives, we can use a scalar form of the magnetization

density in regions where the magnetization length is small. Such a substitution yields

$$n_{\pm} = \frac{1}{2}n \pm \frac{1}{2}m_s \tag{1.149}$$

$$\gamma^{\pm\pm} = \frac{1}{4}\nabla n \cdot \nabla n + \frac{1}{4}\nabla \boldsymbol{m} \circ \nabla \boldsymbol{m} \pm \frac{f_{\nabla}}{2}\nabla n \cdot \nabla m_s \tag{1.150}$$

$$\gamma^{+-} = \frac{1}{4}\nabla n \cdot \nabla n - \frac{1}{4}\nabla \boldsymbol{m} \circ \nabla \boldsymbol{m} \tag{1.151}$$

$$m_s = \frac{1}{3}(m_x + m_y + m_z) \tag{1.152}$$

$$u_s = \frac{1}{3}(u_x + u_y + u_z) \tag{1.153}$$

While this definition breaks formal spin-rotation invariance, the numerical impact of this is controlled so long as the magnetization cutoff is sufficiently small ($m < 10^{-12}$ is used in this work). Using these expressions eliminates the near-zero denominators in the chain rule, and their impact on accuracy has been verified to be negligible for a variety of systems [50,105].

1.6.2 Relativistic CI with the STP-DAS Formalism

The most expensive part of a CI calculation is the $\boldsymbol{\sigma}$ vector calculation,

$$\sigma_K = \sigma_K^{1e} + \sigma_K^{2e} \tag{1.154}$$

$$= \sum_{pq}\sum_{L}(h_1')^{pq} \langle K| \hat{E}_{pq} |L\rangle C_L \tag{1.155}$$

$$+ \frac{1}{2}\sum_{pqrs}\sum_{LJ} h_2^{pqrs} \langle K| \hat{E}_{pq} |J\rangle \langle J| \hat{E}_{rs} |L\rangle C_L \tag{1.156}$$

particularly the two electron part $\boldsymbol{\sigma}^{2e}$. For very large CI calculations, the storage of the excitation lists, which scale like

$$\text{Excitation list size} \sim \binom{N_o}{N_e \times N_e(N_o - N_e + 1)} \tag{1.157}$$

will exceed available computer memory. The small tensor product distributed active space (STP-DAS) framework attempts to reduce excitation list storage and accelerate the $2e$ con-

traction in one fell swoop by factorizing Eq. (1.154) into many small tensor products by partitioning the active space. Each of these partitions have their own *intra-space* excitation list that enumerates the determinants inside that partition. If the $2e$ $\boldsymbol{\sigma}$ build can be formulated in terms of small tensor products local to the partitioned space, then one could re-use the excitation lists from one partition to another and additionally store much smaller excitation lists. Such a scheme would not introduce any approximations relative to the full space so long as the partitioned spaces cover the full space. In this section, we will explore the details of the STP-DAS framework and describe one of main results of this thesis: the GPU-accelerated algorithm for the $2e$ $\boldsymbol{\sigma}$ build.

In the distributed active space (DAS) framework, the total correlation space

A DAS is defined by a set of orbitals $\{\psi\}_\mu$, an electron occupation number n_μ, and sub-determinants $\mathbb{K}_\mu$:

$$\mathbb{X}_\mu = \{\{\psi\}_\mu, n_\mu, \mathbb{K}_\mu\} \tag{1.158}$$

Each DAS is a correlation space with

$$N_\mu = \binom{M_\mu}{n_\mu} \tag{1.159}$$

determinants where $M_\mu = |\{\psi\}_\mu|$ is the number of orbitals in $\mathbb{X}_\mu$. As usual, each orbital is a singly occupied spinor and each determinant is an eigenfunction of $\hat{S}_z$.

Given a correlation space comprised of multiple DASes, excitations within the overall space can be decomposed into inter- and intra-DAS excitations. A configuration category $\mathcal{A} = \{\mathbb{X}_\mu^\mathcal{A}, \mathbb{X}_\nu^\mathcal{A}, \dots\}$ is a set of DASes with unique electron occupation numbers. The orbital partitioning is *defined* by the DASes so that each configuration category shares the same orbital partitioning, even though the electron occupations within each category are different.

Inter-space excitations are also referred to as categorical excitations and denoted

$$\hat{\mathcal{E}}_{pq}^{\mathbb{X}_\mu \mathbb{X}_\nu} = \left\{ \hat{E}_{pq} : p \in \mathbb{X}_\mu, q \in \mathbb{X}_\nu \right\} \tag{1.160}$$

$$\hat{\mathcal{E}}_{pqrs}^{\mathbb{X}_\mu \mathbb{X}_\nu} = \left\{ \hat{E}_{pq}\hat{E}_{rs} - \delta_{qr}\hat{E}_{ps} : p \in \mathbb{X}_\mu, q \in \mathbb{X}_\nu, r \in \mathbb{X}_\lambda, s \in \mathbb{X}_\kappa \right\} \tag{1.161}$$

$$\tag{1.162}$$

A full-space determinant can be recovered from the DASes in a category $\mathcal{A}$,

$$|K^{\mathcal{A}}\rangle = |\mathbb{K}_\mu^{\mathcal{A}}\rangle \oplus |\mathbb{K}_\nu^{\mathcal{A}}\rangle \oplus \dots \tag{1.163}$$

$$|K^{\mathcal{A}}\rangle = \bigoplus_\kappa |\mathbb{K}_\kappa^{\mathcal{A}}\rangle \tag{1.164}$$

and the total number of determinants in category $\mathcal{A}$ is

$$N^{\mathcal{A}} = \prod_\kappa N_\kappa^{\mathcal{A}} \tag{1.165}$$

The machinery of DASes and configuration categories is an intuitive hierarchical decomposition of the full correlation space with the following relations for the overal number of determinants N, the total electron count n_e, and the complete set of orbitals $\{\psi\}$,

$$N = \sum_{\mathcal{A}} N^{\mathcal{A}} \tag{1.166}$$

$$n_e = \sum_\mu n_\mu^{\mathcal{A}} \tag{1.167}$$

$$\{\psi\} = \bigcup_\mu \{\psi\}_\mu \tag{1.168}$$

Now that the machinery of categories and DASes is developed, we can see how the partitioning scheme propagates through the definition of $\boldsymbol{\sigma}$.

$$\sigma^{2e}_{L^{\mathcal{A}}} = \frac{1}{2} \sum_{\mathcal{BC}} \sum_{\mathbb{J}^{\mathcal{B}}_{\mu} \oplus \mathbb{J}^{\mathcal{B}}_{\nu}} \sum_{\mathbb{J}^{\mathcal{B}}_{\kappa} \oplus \mathbb{J}^{\mathcal{B}}_{\lambda}} \sum_{\mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda}} \sum_{pqrs} P_{\mu\nu} P_{\kappa\lambda} \tag{1.169}$$

$$\delta_{\bar{\mathbb{X}}^{\mathcal{A}}_{\mu\nu} \bar{\mathbb{X}}^{\mathcal{B}}_{\mu\nu}} \delta_{\bar{\mathbb{X}}^{\mathcal{A}}_{\kappa\lambda} \bar{\mathbb{X}}^{\mathcal{C}}_{\kappa\lambda}} h_2^{pqrs} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{J}^{\mathcal{B}}_{\mu} \oplus \mathbb{J}^{\mathcal{B}}_{\nu} \right\rangle \tag{1.170}$$

$$\left\langle \mathbb{J}^{\mathcal{B}}_{\kappa} \oplus \mathbb{J}^{\mathcal{B}}_{\lambda} \right| \hat{E}_{pq} \left| \mathbb{K}^{\mathcal{C}}_{\kappa} \oplus \mathbb{K}^{\mathcal{C}}_{\lambda} \right\rangle C_{K^{\mathcal{C}}} \tag{1.171}$$

where $p \in \mathbb{X}^{\mathcal{A}}_{\mu}$, $q \in \mathbb{X}^{\mathcal{C}}_{\nu}$, $r \in \mathbb{X}^{\mathcal{B}}_{\lambda}$, $P_{\mu\nu}$ are the global phase factors, h_2^{pqrs} are the two-electron Hamiltonian elements and $C_{K^{\mathcal{X}}}$ are the CI coefficients of determinant K belonging to configuration category $\mathcal{X}$.

Chapter 2

RELATIVISTIC DFT ON GRAPHICS PROCESSING UNITS (GPUS)

2.1 Introduction

Recently, there has been a resurgence of interest to develop and rationally design materials that exploit the properties of electronic spin.[8,22,59,171,231] Applications for these materials range from functional magnetic materials[14,68,111] and quantum computing,[108] to spintronic devices[59,167,235] and catalytic active sites.[31,76] Fundamental to these technological advances are the important physical characteristics of electronic spin and spin-couplings.[70,108,208] As such, there is a strong need to develop theoretical methods which properly describe electronic spin and spin interactions in materials, as well as the computational capability to simulate the large systems where such phenomena manifest.

From a theoretical perspective, a rigorous description of electronic spin and its interactions is fundamentally rooted in relativistic quantum mechanics. This theory uses the Dirac equation to obtain the relativistic wavefunction in place of the non-relativistic Schrödinger equation.[17,47,125,177,201] Here, the wavefunction is represented using 4-vectors, also known as 4-component bispinors. These structures incorporate the electronic wavefunction in both the positive and negative energy domains, corresponding to the large and small components of the wavefunction. Spin is inherently encoded into this wavefunction through the α and β components of the bispinor. In practice though, using 4-component representations of relativistic wavefunctions is computationally cumbersome, requiring additional storage and computational throughput than corresponding non-relativistic calculations. Furthermore, while the negative energy part of the wavefunction is important to many chemically-relevant phenomena[173] such as spin-forbidden transitions, spin-orbit coupling, and the inert-pair effect[27,91,211], most chemical problems do not require an explicit treatment within the 4-component framework. It is often advantageous from a practical point of view to transform the 4-component Hamiltonian into an electron-only effective 2-component form. This is known as a 2-component transformation, which uses unitary operators to fold the small component of the wavefunction into a pseudo-large component form.[41,43,44,60,80,90,107,127]

Variational electronic structure methods developed within the 4- or 2-component Dirac

framework do not treat spin as a good quantum number, *i.e.*, the spin angular momentum is no longer aligned in a single axis. This approach requires maintaining electronic properties in their four-current form or as a complete magnetization vector.[70] Because the spin-orbit operator interacts with the entire angular momentum tensor, all three components of the magnetic density (ρ_x, ρ_y, ρ_z) must be incorporated in the calculations, leading to a significant increase in computational cost compared to non-relativistic one-component calculations.

Relativistic density functional theory (DFT)[134,174,175] encompasses a powerful set of low-cost theoretical tools that can be used to investigate spin-driven chemical and materials processes and properties at large and experimentally relevant scales. Practical applications commonly utilize workhorse methods implemented in the relativistic 2-component and 4-component Kohn-Sham formalism.[53] While the exact functional is unknown, the excellent balance between accuracy and computational cost makes density functional approximations particularly popular for tackling problems in a wide array of contexts, from physical chemistry and material science[117,147,185,235] all the way to biochemistry and structural biology.[33,213] Many popular software packages that implement relativistic density functional theory leverage the latest advances in high performance computing in order to further accelerate calculations, making larger and more complex systems accessible to this theory.[24,145,171,176,190,207]

In DFT, the evaluation of the exchange-correlation energy and potential is performed using a numerical integration scheme[15,156,203]. This numerical integration procedure is highly amenable to parallelization since the integrand of the functional only depends on the density and density gradients at each point. Furthermore, efficient algorithms for integrating the exchange-correlation potential in non-collinear relativistic DFT have been developed for central processing unit (CPU)-based systems.[11,23,50,130,131,159,168,172,187,214,228] Similarly, graphics processing units (GPU) have long been used to perform the numerical integration for the exchange-correlation potential in *non-relativistic* density functional theory, and

have been shown to provide significant speedup over CPU based numerical integration schemes.[139,150,223,225] As the computation of the exchange-correlation potential is done at every step of self-consistent field and real-time time-dependent calculations[73,75,114,140,184] a software paradigm that prioritizes speed and concurrency is highly desirable.

Despite the clear advantage of GPUs for usage in density functional calculations, to the authors' best knowledge, this hardware not been leveraged to perform fully non-collinear relativistic calculations. In this work, we exploit using GPUs for relativistic quantum chemistry calculations incorporating spin and spin couplings. We showcase a GPU implementation of the highly efficient 2-component relativistic DFT integration algorithm and benchmark its performance evaluating large coinage metal nanoparticles against the corresponding CPU implementation. The results presented in this letter demonstrate how GPUs can be used to heavily accelerate relativistic DFT calculations.

2.2 Methods

The GPU implementation of non-collinear relativistic DFT is developed within the exact-two-component (X2C) relativistic framework.[43–46,49–51,60,72,85,90,104,107,115,122–124,127,128,157,158,194,234] In X2C, one electron spin-orbit coupling is included variationally during the optimization of molecular spinor orbitals. Two-electron spin-orbit effects are accounted for by an empirical screened nuclear spin-orbit (SNSO) treatment.[51?]

Due to the spinor nature of molecular orbitals (MOs) in relativistic calculations, the Kohn-Sham and density matrices ($\mathbf{F}$ and $\mathbf{P}$) formally have the spin-blocked structure

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{\alpha\alpha} & \mathbf{X}^{\alpha\beta} \\ \mathbf{X}^{\beta\alpha} & \mathbf{X}^{\beta\beta} \end{bmatrix}, \quad \mathbf{X} \in \{\mathbf{F}, \mathbf{P}\}, \tag{2.1}$$

with

$$P_{\mu\nu}^{\zeta\zeta'} = \sum_i C_{\mu i}^{\zeta} C_{\nu i}^{\zeta'*}, \qquad \zeta, \zeta' \in \{\alpha, \beta\} \tag{2.2}$$

where $\mathbf{C}$ is the MO coefficient matrix and $\{\mu, \nu\}$ are indices for atomic orbital bases. We

Grid Name	N_{rad}	N_{ang}
`Fine`	75	302
`UltraFine`	99	590
`SuperFine`	250	974

Table 2.1. Grid sizes per atom for the set of grid densities used in this work for Cu, Ag, and Au atoms.

cast the rank–2 spinor structure of $\mathbf{F}$ and $\mathbf{P}$ into the Pauli quaternion form

$$\mathbf{F} = \mathbf{F}^S \otimes \mathbf{I}_2 + \sum_{k\in\{x,y,z\}} \mathbf{F}^k \otimes \boldsymbol{\sigma}_k \tag{2.3}$$

$$\mathbf{P} = \mathbf{P}^S \otimes \mathbf{I}_2 + \sum_{k\in\{x,y,z\}} \mathbf{P}^k \otimes \boldsymbol{\sigma}_k \tag{2.4}$$

which gives rise to an exchange-correlation potential in the quaternion structure $\{\mathbf{V}^S_{\rm xc}, \mathbf{V}^z_{\rm xc}, \mathbf{V}^y_{\rm xc}, \mathbf{V}^x_{\rm xc}\}$, defined as

$$V^S_{{\rm xc},\mu\nu} = \frac{\delta E_{xc}}{\delta P^S_{\mu\nu}}, \quad V^x_{{\rm xc},\mu\nu} = \frac{\delta E_{xc}}{\delta P^x_{\mu\nu}}, \quad V^y_{{\rm xc},\mu\nu} = \frac{\delta E_{xc}}{\delta P^y_{\mu\nu}}, \quad V^z_{{\rm xc},\mu\nu} = \frac{\delta E_{xc}}{\delta P^z_{\mu\nu}} \tag{2.5}$$

The reader is referred to[168] for more details on the integration and assembly of the $V_{\rm xc}$ terms of the Kohn–Sham Fock matrix using a quaternion formalism. In this work, we leverage the parallelized computing capability of GPUs to perform the integration in Eq. (2.5). Integration points are batched together to maximize cache utilization and memory contiguity.[168,225]

2.3 Results and Discussion

The X2C DFT method on the GPUs presented in this work is made publicly available via the GauXC[168,224,226] library for exascale Gaussian basis DFT. All numerical results in this study were obtained using GauXC evaluating density matrices generated by ChronusQ,[227] carried out on the HYAK cluster at the University of Washington. A single A100 GPU card was used for the benchmark study.

Basis Type	N_{bf} (Cu)	N_{bf} (Ag)	N_{bf} (Au)
`DZ`	56	65	81
`TZ`	94	103	119
`QZ`	141	150	166

Table 2.2. Number of basis functions (N_{bf}) per atom for the SAPPORO-DKH3 basis set used in this work for Cu, Ag, and Au atoms.

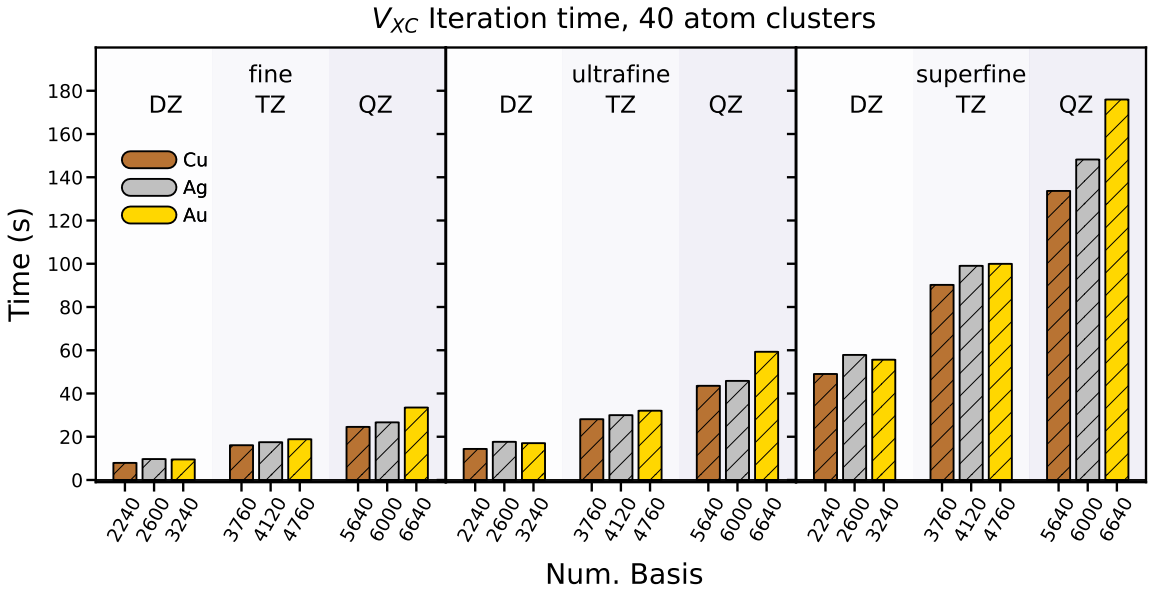

Figure 2.1. X2C V_{xc} computation times for Cu_{40}, Ag_{40}, and Au_{40} clusters for **a)** `Fine`, **b)** `Ultrafine`, **c)** `Superfine` grid sizes on an A100 GPU card using the SAPPORO-DKH3 DZ, TZ, and QZ basis sets. The numbers of basis functions for each cluster are indicated on the x-axis.

We assessed the computational performance of our GPU implementation through all-electron X2C DFT calculations on metal nanoparticles, containing 20 to 40 atoms per cluster. The metal cluster structures were generated using the NanoCrystal suite.[30,163–165] The setup of the grid and the selection of basis functions are detailed in Table 2.1 and Figure 2.1, respectively. These systems are the largest all-electron relativistic calculations performed to date. Our benchmark tests peaked with a calculation on a Au_{40} nanocluster, which consists of 6640 basis functions.

Figure 2.1 shows the V_{xc} integration times using the GPU based numerical integrator for X2C-PBE[160,161] Cu_{40}, Ag_{40}, and Au_{40} nanoparticles. The integration times are plotted over the three different basis sets and three different integration grids. For the `Fine` grid in Figure 2.1.a, all metals display numerical integration times of less than ∼40s irrespective of the basis set. The `UltraFine` grid shows increased integration times, from ∼20s with the DZ basis to ∼60s with the QZ basis. The `SuperFine` grid is the most computationally expensive with integration times of ∼60s with the DZ basis and up to ∼180s with the QZ basis. For relativistic DFT calculations, the `SuperFine` grid is recommended to reliably produce spin and spin-orbit properties on the order of millielectron volts (meV). The timing results in this figure indicate roughly linear scaling with number of basis functions and grid points for all coinage metal clusters. Similar scaling results are observed for smaller clusters.

The second analysis in this work, shown in Figure A.3, plots the V_{xc} integration times for Au_{20-40} nanoparticles, calculated with the restricted, unrestricted, and X2C Kohn-Sham schemes on an A100 GPU. The integration times are plotted for three different integration grids using only the largest, QZ, basis. A complete set of corresponding figures spanning all three basis sets and metal identities is available in the supporting information. The `Fine` grid produces integration times between ∼5-40s and for all three schemes. With the `UltraFine` grid, X2C is shown to be roughly twice as expensive as an unrestricted Kohn-Sham integration, which in turn is twice as expensive as a restricted Kohn-Sham calculation. X2C integration times here are ∼10s for the Au_{20} cluster up to about ∼60s for the Au_{40}

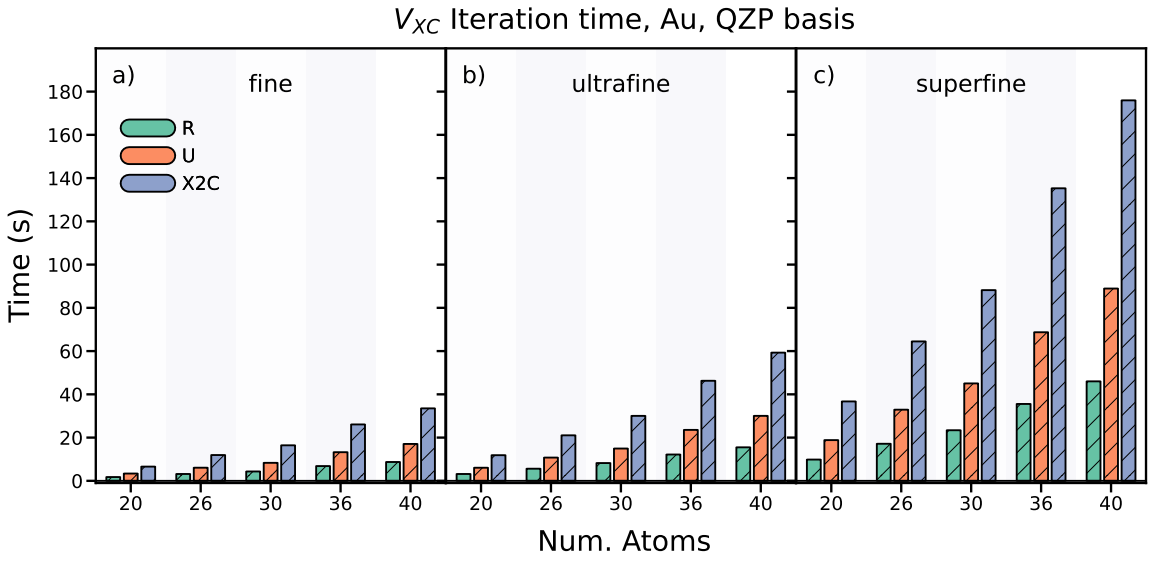

Figure 2.2. Comparison of V_{xc} computation times for Au clusters using restricted, unrestricted, and X2C schemes for **a)** `Fine`, **b)** `Ultrafine`, **c)** `Superfine` grid sizes on an A100 GPU card using the SAPPORO-DKH3 QZ basis set.

cluster. The `SuperFine` grid produces the longest computation times with ∼40s for Au_{20} to ∼180s for Au_{40}.

Finally, in order to demonstrate the computational speedup of the GPU implementation relative to the corresponding CPU based scheme for calculating the exchange-correlation potential in GauXC, Figure 2.3 plots the time required to compute V_{xc} on an NVIDIA A100 GPU against the corresponding CPU implementation running on two Intel XEON Gold 6230 CPUs (total 40 cores) at 2.10GHz. This is performed on Au_{20} with 3320 basis functions up to Au_{40} with 6640 basis functions using the `UltraFine` grid and SAPPORO-DKH3 QZ basis set. Across the board for all examined cluster sizes, the GPU performs about ∼10× faster than the corresponding CPU implementation, with larger speedup for bigger systems. It should be emphasized here that MPI implementations are available for CPU and GPU numerical integration schemes, which could reduce the calculation time. Assuming linear scaling with number of CPU cores, it would take about 400 CPU cores to produce the same computation time as one A100 GPU card. This metric highlights the

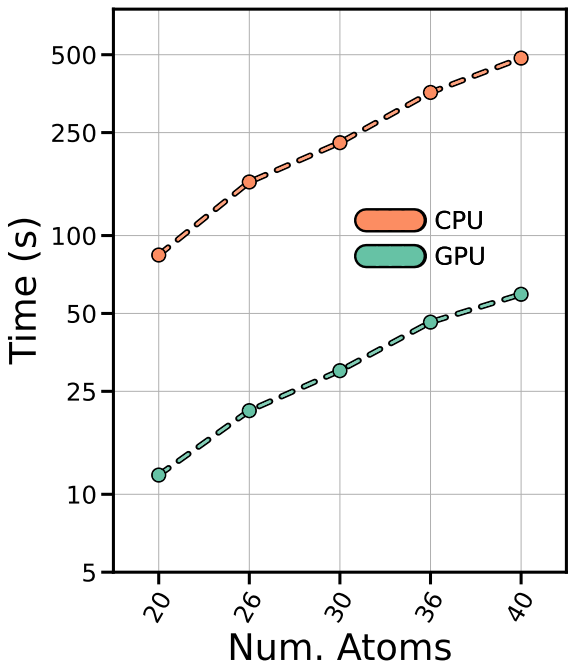

Figure 2.3. Comparison of V_{xc} computation times for Au clusters on a 40-core CPU node and an A100 GPU card, for Au clusters using the SAPPORO-DKH3 QZ basis and `UltraFine` grid.

computational efficiency of the GPU implementation of X2C DFT substantial speedup can be achieved using a single GPU card on a single node, with increasing speedup for larger systems.

This work highlights the acceleration of non-collinear relativistic X2C DFT calculations by exploiting GPU hardware and showcases the exceptionally low integration times on a GPU for X2C calculations on coinage metal nanoclusters. The numerical integration takes between 10-60s for these systems on the GPU. We demonstrate a 10× speedup for V_{xc} calculations of coinage metal nanoparticles relative to a CPU implementation (one A100 vs. 40 CPU cores). It should be emphasized that the systems examined in this letter are large systems composed of many heavy transition metals. Our GPU implementation is able to handle arbitrary-order angular momentum basis functions, making it amenable to heavy element systems of interest to the materials and solid state chemistry fields.

Chapter 3

RELATIVISTIC CI ON GPUS USING THE STP-DAS FRAMEWORK

3.1 Introduction

The formal description of all quantum mechanical systems is governed by the abstract Schrödinger equation,

$$i\hbar\frac{\partial}{\partial t}\left|\psi\right\rangle = \hat{H}\left|\psi\right\rangle, \tag{3.1}$$

where $|\psi\rangle$ is the state of the system and $\hat{H}$ is the Hamiltonian governing it. In the context of the electronic structure theory of molecules in finite basis sets, the solution to Equation (3.1) is given by the full configuration interaction (FCI) method [94,196,199,206,215?] . The many-body state $|\psi\rangle$ is computed in a given basis set by explicitly diagonalizing the corresponding electronic Hamiltonian matrix. Its straightforward nature makes FCI applicable to a wide range of many-body problems. For instance, its adaptation to the two- and four-component Dirac formalisms [48,177] enables the description of relativistic effects such as spin–orbit and spin–spin couplings beyond perturbation theory [3,12,13,86,87,92,93,97–100,128,178,194,200,209,216].

The explicit diagonalization of the Hamiltonian matrix has historically rendered its applicability to small chemical systems. This is because the number of possible configurations increases factorially with system size, causing the dimension of the Hilbert space to explode. As such, even with its special properties (sparse, Hermitian, and diagonally-dominant), the diagonalization of the Hamiltonian matrix quickly becomes intractable.

Due to the explosive growth of the dimension of the space of configurations, practical calculations often rely on approximate configuration interaction (CI) wavefunction methods. The two main families of methods are complete active space CI (CASCI) [61–63,67,96–98,102,153,182,218] and selected CI (SCI) [54,56,71,83,84,89,110,126,191,195,210,232,233]. Both methods work by effectively truncating the Hilbert space of the system to a subspace of significant determinants.

The CASCI method approximates the FCI wavefunction as the CI solution within the configuration space generated by an orbital subspace (called the active space). The outright exclusion of occupied and unoccupied orbitals often leads to an underestimation of dynamic correlation. Therefore, to attain quantitative agreement with experimen-

tal results, it is often required to apply additional methods on top of CASCI to recover the missing correlation. Such methods include multiconfigurational self-consistent field (MCSCF)[12,66,93,95,119,132,135,136,178,200,206,209,217,221,222,230], multireference configuration interaction (MRCI)[81,87,119–121,137,181,206,220], and many-body perturbation theory (MRPT2, CASPT2, NEVPT2, MC-PDFT)[4–6,29,116,128,133,180,200,216].

A recent CI framework[88] remedies the explosive memory footprint of the Hamiltonian matrix by factorizing the active space into small tensor products of distributed active spaces (DASs). In the small tensor product distributed active space (STP-DAS) approach, the computation of the large matrix-vector product associated with the iterative diagonalization of the Hamiltonian matrix is decomposed into the evaluation of a sequence of small tensor products. By employing an addressing scheme compatible with the small-tensor-product approach, the on-the-fly evaluation of matrix-vector products can significantly reuse the address strings local to each DAS to emulate the full excitation list. This reuse leads to dramatic reductions in the memory footprint of the Hamiltonian excitation list. This framework was shown to successfully facilitate CI calculations consisting of over one quadrillion (10^{15}) determinants[197].

The engine that powers CI is the matrix-vector product algorithm within the iterative diagonalization procedure, also known as the $\boldsymbol{\sigma}$-build. As such, CI is amenable to large-scale parallel processing paradigms[7,65,67,88,152,218], so long as subtle issues such as load balancing are properly addressed. Beyond the traditional host-centered paradigms of multi-threading (OpenMP, OpenACC) and message passing interface (MPI), electronic structure method development is increasingly adopting the use of graphics processing units (GPUs) and other accelerators to maximize throughput and parallelism[10,55,57,58,106,138,142,169,193,198,229]. Such accelerators consist of (tens of) thousands of cores purposefully designed to perform vectorized operations with peak efficiency. Their hardware design implements the "single instruction, multiple data" (SIMD) principle, in which all cores execute the exact same operation on different data elements concurrently. While this limits the functionality of GPU cores

compared to traditional central processing unit (CPU) cores, the design drastically reduces the per-core cost of vectorized operations such as matrix-vector products. Despite this remarkable capability, the effective exploitation of the SIMD cores of the accelerator requires careful consideration of the data flow and processing patterns of the underlying algorithm. This includes the clever distribution of the computational workload between the host and the accelerator and the implementation of pipelining. The latter is a technique that hides the overhead associated with data transfers to and from the accelerator by overlapping them with host and accelerator computation. In addition, the adoption of batched processing and coalesced data transfers and memory accesses is often needed to maximize GPU throughput and performance.

Despite their vast computational power, their video random-access memory (VRAM) capacity is significantly more restricted, ultimately limiting their ability to accelerate certain applications. This general restriction does not apply in the case of STP-DAS: the framework decomposes large matrix-vector products into a sequence of small tensor products, each of which can comfortably fit into the limited memory of GPUs! In this work, we accelerate the STP-DAS framework by adapting the two-electron portion of the small-tensor-product $\boldsymbol{\sigma}$-build algorithm to use GPUs. The resulting algorithm implements batched processing and coalesced data transfers within the Knowles-Handy[101,102] approach to contract the Hamiltonian excitation list with CI expansion vectors. It supports multi-GPU and cooperative multi-process execution via MPI and Multi-Process Service (MPS). To showcase its features and performance, we use the accelerated small-tensor-product $\boldsymbol{\sigma}$-build algorithm to conduct exact two-component CASCI (X2C-CASCI) calculations for the relativistic ground states of Au_2 and Xe_2.

3.2 Methods

3.2.1 Accelerating STP-DAS $\boldsymbol{\sigma}$-Build using GPUs

The $\boldsymbol{\sigma}$-build algorithm within STP-DAS[88] implements the evaluation of

$$\sigma_{L^{\mathcal{A}}} = {}^{1\mathrm{e}}\sigma_{L^{\mathcal{A}}} + {}^{2\mathrm{e}}\sigma_{L^{\mathcal{A}}}, \tag{3.2}$$

$$\begin{aligned} {}^{1\mathrm{e}}\sigma_{L^{\mathcal{A}}} = \sum_{\mathcal{B}} \sum_{\mathbb{K}^{\mathcal{B}}_{\mu}\oplus\mathbb{K}^{\mathcal{B}}_{\nu}} \sum_{pq} P_{\mu\nu}\delta_{\bar{\mathbb{X}}^{\mathcal{A}}_{\mu\nu}\bar{\mathbb{X}}^{\mathcal{B}}_{\mu\nu}} \\ h'_{pq} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{K}^{\mathcal{B}}_{\mu} \oplus \mathbb{K}^{\mathcal{B}}_{\nu} \right\rangle C_{K^{\mathcal{B}}}, \end{aligned} \tag{3.3}$$

$$\begin{aligned} {}^{2\mathrm{e}}\sigma_{L^{\mathcal{A}}} = \frac{1}{2} \sum_{\mathcal{CB}} \sum_{\mathbb{J}^{\mathcal{C}}_{\mu}\oplus\mathbb{J}^{\mathcal{C}}_{\nu}} \sum_{\mathbb{J}^{\mathcal{C}}_{\kappa}\oplus\mathbb{J}^{\mathcal{C}}_{\lambda}} \sum_{\mathbb{K}^{\mathcal{B}}_{\kappa}\oplus\mathbb{K}^{\mathcal{B}}_{\lambda}} \sum_{pqrs} P_{\mu\nu}P_{\kappa\lambda} \\ \delta_{\bar{\mathbb{X}}^{\mathcal{A}}_{\mu\nu}\bar{\mathbb{X}}^{\mathcal{C}}_{\mu\nu}} \delta_{\bar{\mathbb{X}}^{\mathcal{C}}_{\kappa\lambda}\bar{\mathbb{X}}^{\mathcal{B}}_{\kappa\lambda}} g_{pqrs} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{J}^{\mathcal{C}}_{\mu} \oplus \mathbb{J}^{\mathcal{C}}_{\nu} \right\rangle \\ \left\langle \mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda} \right| \hat{E}_{rs} \left| \mathbb{K}^{\mathcal{B}}_{\kappa} \oplus \mathbb{K}^{\mathcal{B}}_{\lambda} \right\rangle C_{K^{\mathcal{B}}}, \end{aligned} \tag{3.4}$$

where $p \in \mathbb{X}^{\mathcal{A}}_{\mu}$, $q \in \mathbb{X}^{\mathcal{C}}_{\nu}, r \in \mathbb{X}^{\mathcal{C}}_{\kappa}$, $s \in \mathbb{X}^{\mathcal{B}}_{\lambda}$, $\left\langle \mathbb{X}^{\mathcal{A}}_{\mu} \oplus \mathbb{X}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{X}^{\mathcal{B}}_{\mu} \oplus \mathbb{X}^{\mathcal{B}}_{\nu} \right\rangle$ are categorical one electron excitation lists, $P_{\mu\nu}$ are global phase factors, h'_{pq} and g_{pqrs} are one- and two-body Hamiltonian elements, and $C_{K^{\mathcal{B}}}$ are CI coefficients of determinants K belonging to configuration category $\mathcal{B}$ (see Ref.[88] for definitions and algorithmic details). From the summation structure of Equation (3.4), it is evident that the computational bottleneck is the evaluation of the two-electron portion of the $\boldsymbol{\sigma}$-build. Consequently, it has historically been the focus of algorithmic optimization efforts[101,102,153].

One such optimization is inspired by the Knowles-Handy approach[101,102], which evaluates Equation (3.4) using vectorized operations, thereby enabling the full exploitation of state-of-the-art cache optimization techniques. Vectorization, which comes at the cost of additional memory requirements, is achieved by organizing the Hamiltonian elements and the CI coefficients into *intermediates* and matrix-multiplying them:

$$X^{L^{\mathcal{A}},J^{\mathcal{C}}}_{rs,pq} \leftarrow g_{pqrs} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{J}^{\mathcal{C}}_{\mu} \oplus \mathbb{J}^{\mathcal{C}}_{\nu} \right\rangle \left\langle \mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda} \right| \hat{E}_{rs} \left| \mathbb{K}^{\mathcal{B}}_{\kappa} \oplus \mathbb{K}^{\mathcal{B}}_{\lambda} \right\rangle, \tag{3.5}$$

$$\Omega^{J^{\mathcal{C}},K^{\mathcal{B}}}_{rs,\bar{\mathbb{K}}^{\mathcal{B}}_{\kappa\lambda}} \leftarrow C_{K^{\mathcal{B}}} \text{ such that } \left\langle \mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda} \right| \hat{E}_{rs} \left| \mathbb{K}^{\mathcal{B}}_{\kappa} \oplus \mathbb{K}^{\mathcal{B}}_{\lambda} \right\rangle \neq 0, \tag{3.6}$$

$$\Lambda^{L^{\mathcal{A}}}_{pq,\bar{\mathbb{K}}^{\mathcal{A}}_{\kappa\lambda}} \leftarrow \left(X^{L^{\mathcal{A}},J^{\mathcal{C}}}_{rs,pq} \right)^{\intercal} \times \Omega^{J^{\mathcal{C}},K^{\mathcal{B}}}_{rs,\bar{\mathbb{K}}^{\mathcal{B}}_{\kappa\lambda}}, \tag{3.7}$$

where $\bar{\mathbb{K}}$ and $\bar{\mathbb{L}}$ are the *local* addresses of the portion of the determinant K not belonging to the DASs numbered κ or λ. The locality of the addresses renders the column dimensions of the intermediates in Equations (3.6) and (3.7) small, making the evaluation of Eq. (3.4) on GPUs possible despite memory limitations. This small dimensionality, however, presents challenges in maximizing GPU computational throughput and frequent data transfers between the CPU and the accelerator.

Computational throughput can be maximized by the incorporation of *batching* in the evaluation of Equation (3.4). Instead of successively forming Equations (3.5) and (3.6) and computing Equation (3.7) for every connecting determinant $J^{\mathcal{C}}$, one can form a batch $\left\{ X^{L^{\mathcal{A}},J^{\mathcal{C}}}_{rs,pq}, \Omega^{J^{\mathcal{C}},K^{\mathcal{B}}}_{rs,\bar{\mathbb{K}}^{\mathcal{B}}_{\kappa\lambda}} \right\}_{J^{\mathcal{C}}}$ of intermediates and evaluate Equation (3.7) for the entire batch concurrently. While this approach has a larger memory footprint than naive successive evaluation, its computational advantages are twofold: the increased utilization of the compute cores of the accelerator, and the coalescence of data transfers between the CPU and the accelerator[113].

Because the batching technique organizes the evaluation of Equations (3.5) and (3.6) into large computational tasks, it also enables the implementation of GPU *pipelining*, which hides the cost of large data transfers by overlapping them with significant CPU and GPU computation. To this end, we adopted the use asynchronous operations in our implementation, in which the CPU submits a task to be executed by the GPU without needing to wait for the GPU to complete it. Consequently, upon forming Equations (3.5) and (3.6) for batch n, the CPU can form intermediates for batch $(n+1)$ while, *concurrently*, the inter-

mediates of n are transferred to the GPU, the GPU evaluates Equation (3.7) for batch n, and the computed contractions for batch n are transferred back to the CPU for processing into the $\boldsymbol{\sigma}$-vector, as prescribed in Equation (3.4). This pipelined processing of batches of intermediates not only effectively hides the latency associated with large and frequent data transfers between the CPU and the accelerator, but also dramatically shrinks the idle time the CPU spends waiting for the GPU to complete the contractions of intermediates. The precise execution pattern that yields these optimizations is depicted in Algorithm 1.

The computational load distribution structure within the STP-DAS algorithm [88] naturally enables the multi-GPU acceleration of the evaluation of Equation (3.4). By assigning each MPI rank to a unique GPU, Algorithm 1 can be executed on multiple accelerators to compute and populate distinct sectors of the $\boldsymbol{\sigma}$-vector concurrently. Cooperative multi-process execution can also be realized by assigning multiple MPI ranks to the same GPU and partitioning it by utilizing technologies such as Nvidia's MPS and multi-instance GPU (MIG). This can maximize the performance of Algorithm 1 on multi-node, single-accelerator systems. Notably, this approach to multi-GPU multi-process execution does not require communication among the accelerators. Therefore, in contrast with other implementations [2], Algorithm 1 does not require the use of GPU-aware MPI libraries.

3.3 Results

To showcase the runtime of Algorithm 1 we conducted a sequence of calculations for the X2C-CASCI ground state of PuO_2 with the def2-mTZVP basis set to examine the runtime dependence of Algorithm 1 on intermediates batch size and number of nodes (strong scaling). The calculations were executed on the University of Washington's AI-accelerated research computing platform Tillicum, a small-sized cluster where each node is equipped with an Intel Emerald Rapids CPU, eight NVIDIA Hopper–H200 SXM GPUs, and a 400 GB/s network interface card.

In any GPU algorithm, the time cost of data movement must be amortized by the time savings of fast computation to see speedup. As traditional CPUs also benefit from

Algorithm 1: The GPU-accelerated two-electron STP-DAS $\boldsymbol{\sigma}$-build. Bold text represents algorithmic logic and typewritten text represents comments. Curly brackets represent code-interfacing operations.

`// Process first batch` B_1 `of` $J^{\mathcal{C}}$`'s`

$B_1 \leftarrow \left\{0, \ldots, \min\left(S_{\text{batch}}, \left|\left\{J^{\mathcal{C}}\right\}\right|\right)\right\}$;

`// Form Equations (3.5) and (3.6) for all` $J^{\mathcal{C}} \in B_1$

for $J^{\mathcal{C}} \in B_1$ **do**

$$X^{B_1} \leftarrow X^{B_1} \cup \left\{\left\{ g_{pqrs} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{J}^{\mathcal{C}}_{\mu} \oplus \mathbb{J}^{\mathcal{C}}_{\nu} \right\rangle \left\langle \mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda} \right| \hat{E}_{rs} \left| \mathbb{K}^{\mathcal{B}}_{\kappa} \oplus \mathbb{K}^{\mathcal{B}}_{\lambda} \right\rangle \right\}\right\}$$

$$\Omega^{B_1} \leftarrow \Omega^{B_1} \cup \left\{\left\{ C_{K^{\mathcal{B}}} : \left\langle \mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda} \right| \hat{E}_{rs} \left| \mathbb{K}^{\mathcal{B}}_{\kappa} \oplus \mathbb{K}^{\mathcal{B}}_{\lambda} \right\rangle \neq 0 \right\}\right\}$$

{Asynchronously task GPU with contracting intermediates in B_1}

`// Continue processing next batch` B_n `without waiting for the GPU`

for $n = 2, \ldots, \left|\left\{J^{\mathcal{C}}\right\}\right| / S_{\text{batch}}$ **do**

$B_n \leftarrow \{(n-1) \times S_{\text{batch}}, \ldots, n \times S_{\text{batch}}\}$

$X^{B_n}, \Omega^{B_n}, \Lambda^{B_n} \leftarrow \emptyset$

`// Form Equations (3.5) and (3.6) for all` $J^{\mathcal{C}} \in B_n$

for $J^{\mathcal{C}} \in B_n$ **do**

$$X^{B_n} \leftarrow X^{B_n} \cup \left\{\left\{ g_{pqrs} \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{J}^{\mathcal{C}}_{\mu} \oplus \mathbb{J}^{\mathcal{C}}_{\nu} \right\rangle \left\langle \mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda} \right| \hat{E}_{rs} \left| \mathbb{K}^{\mathcal{B}}_{\kappa} \oplus \mathbb{K}^{\mathcal{B}}_{\lambda} \right\rangle \right\}\right\}$$

$$\Omega^{B_n} \leftarrow \Omega^{B_n} \cup \left\{\left\{ C_{K^{\mathcal{B}}} : \left\langle \mathbb{J}^{\mathcal{C}}_{\kappa} \oplus \mathbb{J}^{\mathcal{C}}_{\lambda} \right| \hat{E}_{rs} \left| \mathbb{K}^{\mathcal{B}}_{\kappa} \oplus \mathbb{K}^{\mathcal{B}}_{\lambda} \right\rangle \neq 0 \right\}\right\}$$

{Asynchronously task GPU with contracting intermediates in B_n}

{Wait for the GPU to finish processing batch B_{n-1} } `// Process completed batch` $\Lambda^{B_{n-1}}$ `into` $^{2e}\sigma$

for $J^{\mathcal{C}} \in B_{n-1}$ **do**

$$^{2e}\sigma_{L^{\mathcal{A}}} += \Lambda^{L^{\mathcal{A}}}_{pq,\bar{\mathbb{K}}^{\mathcal{A}}_{\kappa\lambda}} \text{ such that } \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{J}^{\mathcal{C}}_{\mu} \oplus \mathbb{J}^{\mathcal{C}}_{\nu} \right\rangle$$

$B_{\text{leftover}} \leftarrow \left\{\left(\left|\left\{J^{\mathcal{C}}\right\}\right| / S_{\text{batch}} - 1\right) \times S_{\text{batch}}, \ldots, \left|\left\{J^{\mathcal{C}}\right\}\right|\right\}$

{Wait for the GPU to finish processing batch B_{leftover} } `// Process the leftover batch into` $^{2e}\sigma$

for $J^{\mathcal{C}} \in B_{\text{leftover}}$ **do**

$$^{2e}\sigma_{L^{\mathcal{A}}} += \Lambda^{L^{\mathcal{A}}}_{pq,\bar{\mathbb{K}}^{\mathcal{A}}_{\kappa\lambda}} \text{ such that } \left\langle \mathbb{L}^{\mathcal{A}}_{\mu} \oplus \mathbb{L}^{\mathcal{A}}_{\nu} \right| \hat{E}_{pq} \left| \mathbb{J}^{\mathcal{C}}_{\mu} \oplus \mathbb{J}^{\mathcal{C}}_{\nu} \right\rangle$$

the vectorization provided by the Knowles-Handy paradigm of Algorithm 1, we prepared a version of the algorithm which performs the GEMM on the CPU, obviating the need for expensive data transfers. Table 3.1 shows timing data of the GEMM and total contraction time for the CPU and GPU algorithms on a 86M determinant $\binom{30}{18}$ PuO_2 calculation on a 16 rank, 2 node cluster. For both calculations, 7 CPUs and 1 GPU was assigned to each rank (the GPU was not used for the CPU-only calculations). The GPU version of Algorithm 1

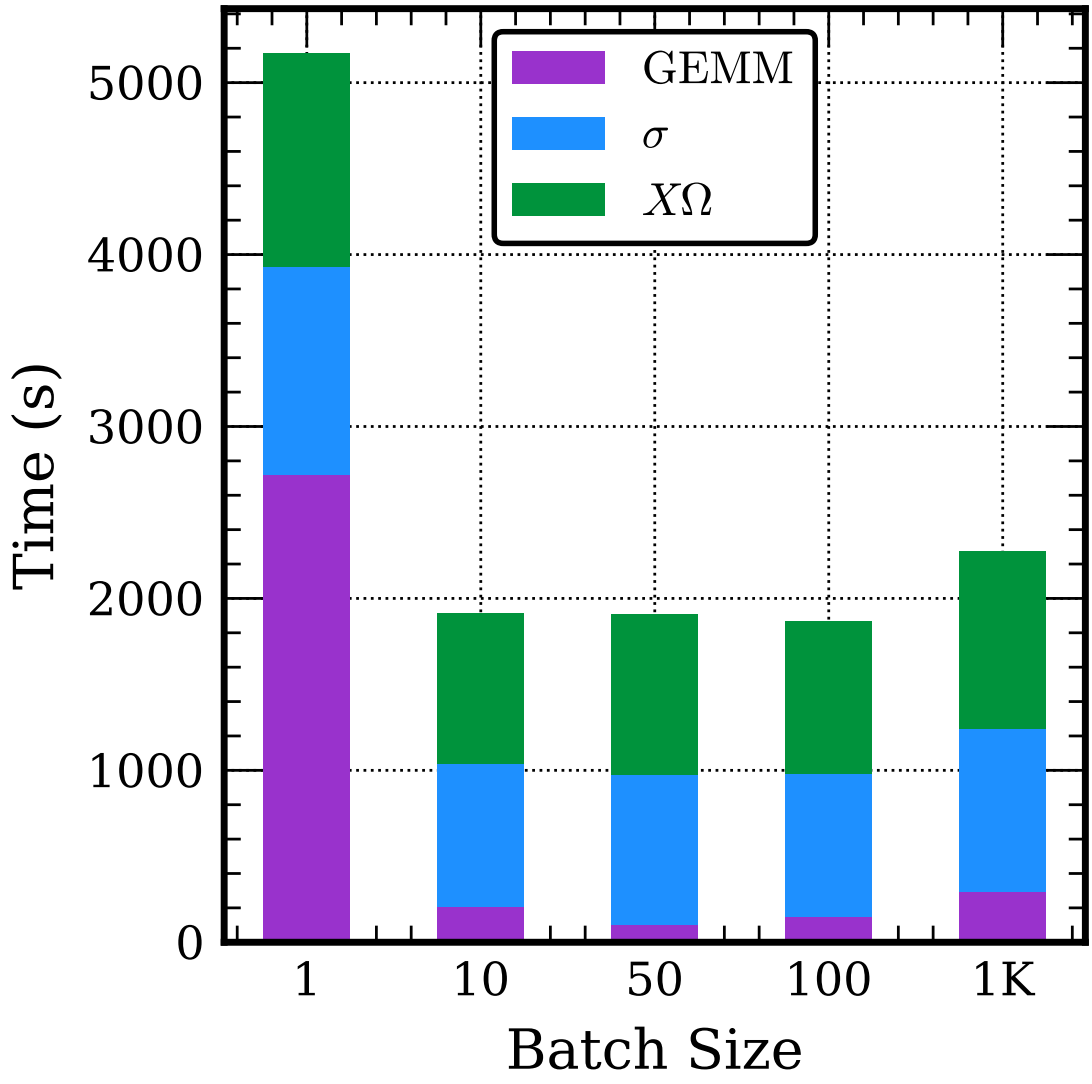

Figure 3.1. Cumulative time spent in the $X\Omega$-gather, $\boldsymbol{\sigma}$-scatter, and GEMM (GPU synchronization) steps of Algorithm 1 for a ground state calculation of $\binom{30}{18}$ PuO_2 (86 million determinants) on a single NVIDIA H200. Batch sizes range from one to one thousand determinants. The GPU-based GEMM reaches a minimum time with a batch size of 50 determinants, indicating maximal overlap of communication and computation.

spends only 8 seconds waiting on the GEMM to complete compared to the CPU algorithm, a speedup of approximately 16×. Including the $X\Omega$-gather/$\boldsymbol{\sigma}$-scatter operations and MPI

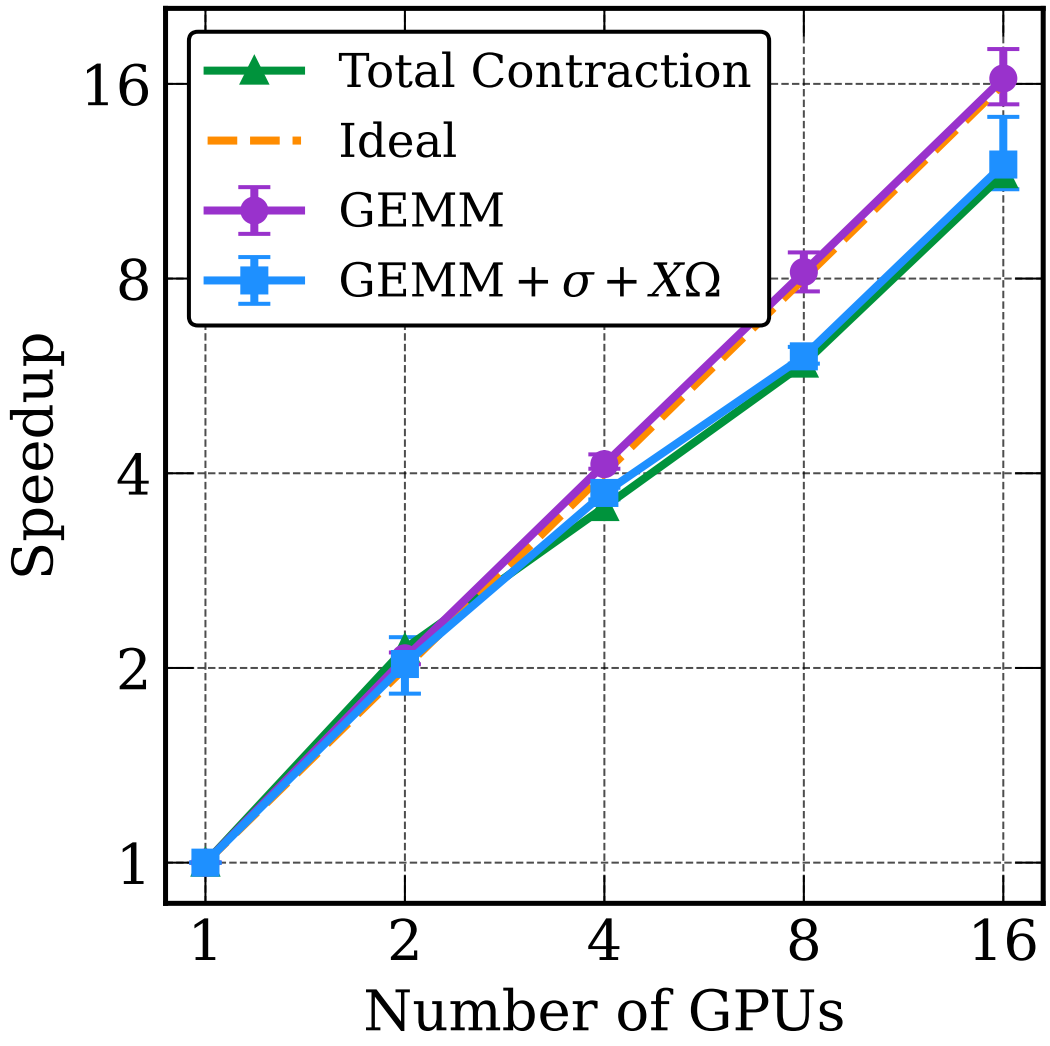

Figure 3.2. Strong scaling of Algorithm 1 for a $\binom{30}{18}$ PuO_2 system on 1–16 NVIDIA H200 GPUs with a batch size of 100. The error bars represent the range of data acquired on each MPI rank. "Total Cont." is the total contraction wall time including MPI reductions and load balancing, "GEMM + σ + $X\Omega$" is the subset of that time taken by Algorithm 1. "GEMM" is solely the synchronization time for the GPU GEMM and memory transfer operations in Algorithm 1.

synchronization time, the total contraction time is 28% faster when the GEMM is offloaded to the GPU. Evidently, the concurrent buffer approach successfully hides the latency of

Algorithm	GEMM (s)	Contraction (s)
CPU	131.44	464.31
GPU	8.08	362.20

Table 3.1. Comparison of evaluation times between CPU and GPU implementations of Algorithm 1, including total time (right column) and the GEMM step (left column). For the GPU algorithm, GEMM is the GPU synchronization time, which includes memory transfer and computation time *not* hidden by overlapping with CPU work. These data are cumulative times through the entire Davidson diagonalization procedure for a $\binom{30}{18}$ PuO_2 system with a batch size of 100.

data transfers with computational throughput, leading to the elimination of the GEMM as a bottleneck.

The GPU accelerated STP-DAS $\boldsymbol{\sigma}$-build algorithm has a batch size parameter that should be tailored to the computational resources and the system size. We ran Algorithm 1 for the X2C-CASCI ground state of a $\binom{30}{18}$ PuO_2 system with 86M determinants and measured the contributions of different steps in the contraction algorithm using batch sizes spanning 3 orders of magnitude from 1 to 1K determinants. Figure 3.1 presents the cumulative time spent in the σ-scatter, $X\Omega$-gather, and GEMM synchronization steps of Algorithm 1. The GEMM time is at a maximum at 1 determinant, falling sharply to a minimum at 50 determinants, then gradually rises through 1 thousand determinants. At the low end, the slowdown is due to the small batch size initiating many small memory transfer operations between the host and device. This prevents saturating memory bandwidth and thus the host spends much of its time waiting for memory operations, in other words, the algorithm is latency-limited. The host-only gather/scatter operations are also slower with a small batch size due to the inefficient cache use and frequent context switching. At the high end of 1 thousand determinants, memory transfers begin to dominate the overall synchronization time. Even with a coalesced memory layout, the sheer amount of data takes long enough to transfer that the host must wait for the transfer to complete. This destroys overlapping of host and device computation which results in a higher overall time to completion. Unlike the small batch size limit, this regime is limited by device memory

throughput, not memory operation latency. These data confirm the intuition that the best batch size is between these two extremes, balancing memory latency and throughput to maximize overlap of host and device work.

To study the strong-scaling behavior of Algorithm 1, we conducted compression-compatible STP-DASCI[197] calculations for the X2C-CASCI ground state of PuO_2 with a varying number of MPI ranks, assigning each a unique GPU and 7 CPUs. The three lines in Figure 3.2 correspond to contributions from different scales of synchronization: intra-node GPU synchronization (labeled "GEMM"), intra-node CPU-GPU synchronization (labeled "GEMM + σ + $X\Omega$"), and inter-node MPI synchronization (labeled "Total Contraction"). As reflected by the overlapping of the intra-node GPU time with the ideal scaling line, Algorithm 1 does not require inter-GPU communication resulting in excellent scaling of the GPU GEMM step. This implies that the optimal batch size on one node can be extended to many nodes without consideration of node-number-dependent effects. The total contraction wall time scales excellently, running 11.6× faster on 16 GPUs than on 1 GPU. As depicted in the upper end of Figure 3.2, the deviation from ideal scaling is predominantly due to intra-node σ-scatter and $X\Omega$-gather operations. The range of speedup for these data points, reflected in the error bars, are due to imperfect load balancing manifested by the choice of active space partitioning: 6 DASes of 6 orbitals were chosen. A higher number of smaller DASes would likely produce superior scaling results at the cost of increased categorical pre-computation. The effect of DAS size on strong scaling is inherited from the STP-DAS framework[88] and not due to the GPU offloading of the GEMM step, which is demonstrated to have excellent strong scaling.

3.3.1 Conclusions

In this work, we showcased the acceleration of the STP-DAS $\boldsymbol{\sigma}$-build algorithm using GPUs. By adopting traditional GPU optimization techniques such as batching and pipelined execution, the algorithm significantly speeds up the evaluation of matrix-vector contractions

and enables rapid CI calculations. The algorithm naturally supports multi-GPU processing and exhibits excellent strong scaling behavior.

Our benchmark studies demonstrate the ability of the GPU-accelerated STP-DAS $\boldsymbol{\sigma}$-build algorithm to massively accelerate the execution time of compressed CI calculations. The batched processing of intermediates within the algorithm implementation delivers excellent throughput and maximizes GPU utilization. Such workload distribution between the GPUs and the hosting CPU is demonstrated to effectively hide the latency associated with data transfers, making the use of accelerators highly efficient. The accelerated algorithm is also shown to exhibit excellent strong-scaling behavior, similarly to the original STP-DAS $\boldsymbol{\sigma}$-build procedure[88]. This enables multi-GPU processing of compressed CI calculations with minimal communication overhead, capable of facilitating large-scale calculations at previously-untenable speed.

Together with categorical compression within the STP-DAS framework[197], the ability to carry accelerated large-scale CI calculations in an inherently-relativistic level-of-theory facilitates the accurate simulation of electronic structure properties and observables for strongly-relativistic systems such as transition-metal, rare-earth, and heavy-element complexes.

Chapter 4

CONCLUSIONS AND OUTLOOK

With thoughtfully designed algorithms, GPUs can be harnessed to massively accelerate quantum chemistry calculations. Since GPUs can process huge amounts of data at once, they are particularly well-suited for highly parallelizable workloads such as exchange-correlation potential integration in density functional theory calculations. The GPU-accelerated DFT integration and assembly algorithm in Chapter 2 gave a 10× speedup compared to CPU-only benchmarks, enabling the study of quantum behavior of large molecules such as those important to biochemistry. It also supports non-collinear spin effects, accurately reproducing the complex spin textures generated from the interplay of spin-orbit coupling and exchange-correlation effects, critical to the electronic structure of late-row elements. If the problem calls for a CASCI calculation, the STP-DAS algorithm handles larger numbers of determinants than other CI algorithms by trading memory usage for computational workload. However, the increased workload is offset by the contraction algorithm in Chapter 3, which splits the workload into parts optimized for the CPU and parts optimized for the GPU. In alignment with the paradigm of heterogeneous computing, the CPU and GPU work on their respective workloads concurrently, resulting in an excellent improvement over CPU-only calculations. Together, the GPU-accelerated DFT integration and STP-DAS contraction algorithms presented in this dissertation extend the boundaries of correlated method calculations, but constitute just a small part of the quantum chemistry communities' efforts toward optimal computational efficiency. Due to the myriad applications in the industrial and scientific realms, quantum chemistry will remain at the forefront of new computing paradigms as they emerge. Looking to the future, quantum computing offers a promising new frontier for quantum chemistry. Experience suggests that it will not replace

today's heterogeneous computing infrastructure, but rather be integrated into it.

Appendix A

SUPPLEMENTARY INFORMATION FOR RELATIVISTIC DFT ON GRAPHICS PROCESSING UNITS

Comparisons of Computation Times

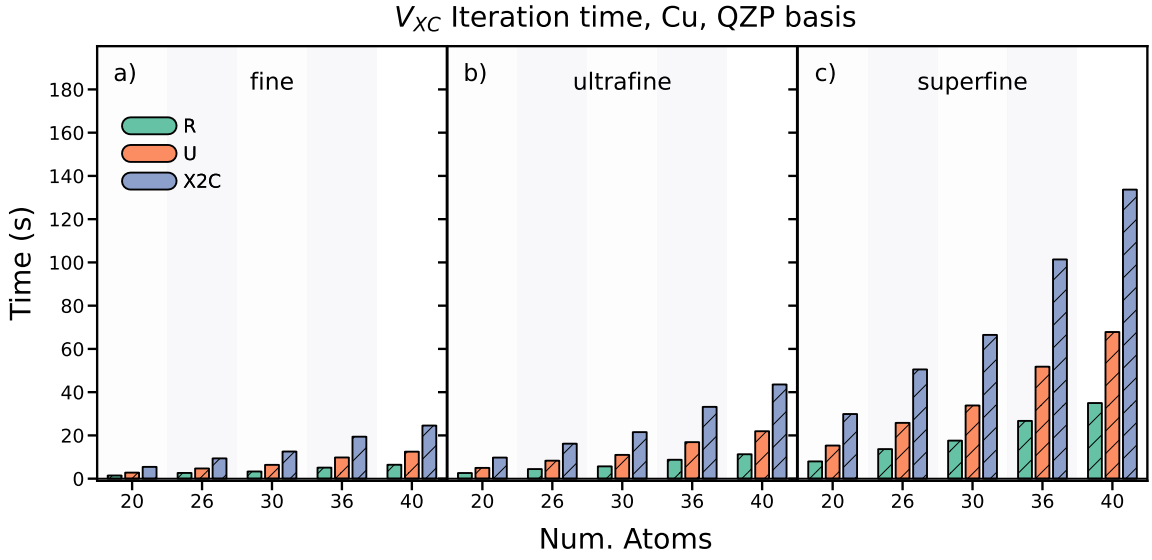

Figure A.1. Comparison of V_{xc} computation times for Cu clusters using restricted, unrestricted, and X2C schemes for **a)** `Fine`, **b)** `Ultrafine`, **c)** `Superfine` grid sizes on an A100 GPU card using the SAPPORO-DKH3 QZ basis set.

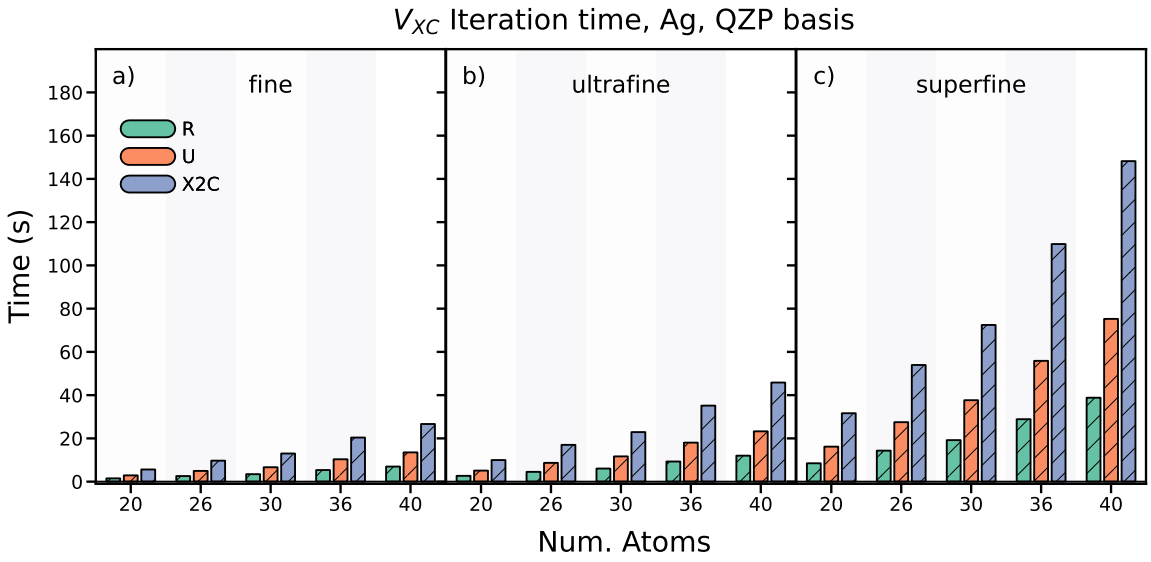

Figure A.2. Comparison of V_{xc} computation times for Ag clusters using restricted, unrestricted, and X2C schemes for **a)** `Fine`, **b)** `Ultrafine`, **c)** `Superfine` grid sizes on an A100 GPU card using the SAPPORO-DKH3 QZ basis set.

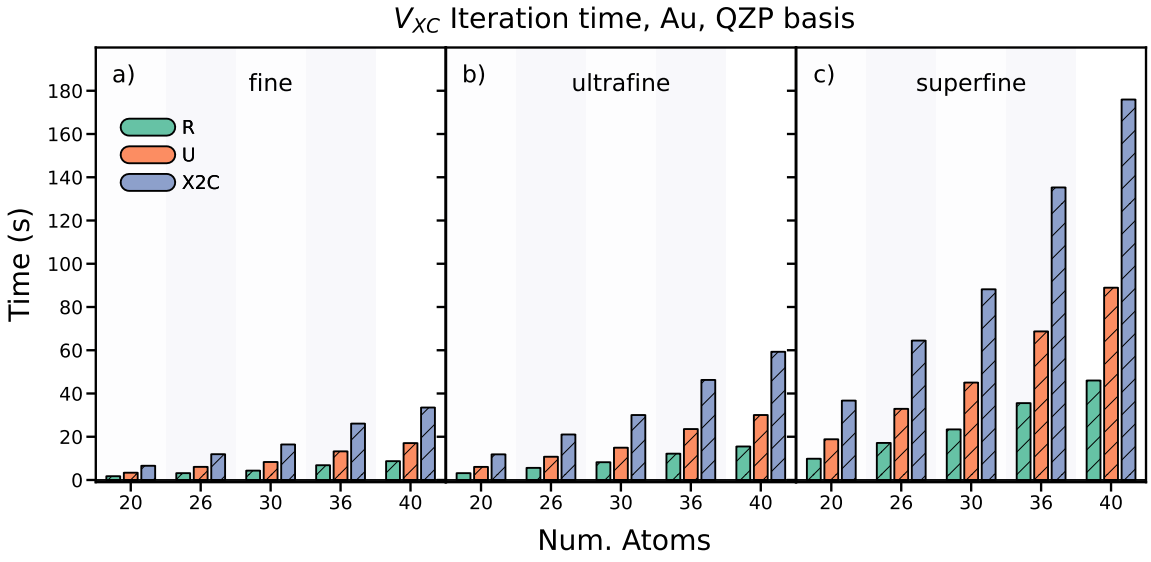

Figure A.3. Comparison of V_{xc} computation times for Au clusters using restricted, unrestricted, and X2C schemes for **a)** `Fine`, **b)** `Ultrafine`, **c)** `Superfine` grid sizes on an A100 GPU card using the SAPPORO-DKH3 QZ basis set.

Table A.1. Comparison of V_{xc} computation times for Au clusters on a 40-core CPU node and an A100 GPU card, spanning the fine, ultrafine, and superfine grids and DZ, TZ, and QZ basis sets. All times are in seconds.

		DZ		TZ		QZ	
	n_atom	gpu_time	cpu_time	gpu_time	cpu_time	gpu_time	cpu_time
fine	20	2.35	11.72	4.50	27.78	6.59	50.90
	26	4.25	23.34	7.07	53.62	11.93	98.45
	30	5.27	33.86	9.57	76.72	16.43	143.56
	36	7.44	54.33	14.62	83.31	26.09	234.30
	40	9.51	71.19	18.86	109.27	33.52	308.52
ultrafine	20	5.12	20.56	7.71	46.47	11.87	84.09
	26	7.21	39.26	12.63	88.66	21.05	161.11
	30	9.21	56.55	17.00	92.53	30.05	228.56
	36	13.50	87.52	25.36	148.23	46.28	357.36
	40	17.04	115.42	32.04	196.73	59.29	485.07
superfine	20	11.89	69.40	22.80	145.29	36.73	261.79
	26	20.44	127.98	40.69	270.40	64.48	488.79
	30	27.82	180.56	54.35	382.70	88.20	693.39
	36	43.38	280.90	78.26	489.28	135.29	1086.55
	40	55.63	369.08	99.96	646.79	175.96	1426.50

GPU Implementation

We present an brief overview of how the exchange-correlation energy, E_{xc}, and the exchange-correlation potential, $\mathbf{V}_{\mathrm{xc}}$, are computed on GPUs. Our work extends the previous GPU implementation of collinear spin Kohn-Sham density functional theory in Ref. 225 to the Pauli spinor basis. The details of relativistic two-component Kohn-Sham density functional theory in the Pauli spinor basis are described in Ref. 168. For the algorithms presented herein, some additional definitions are as follows: $\mathbf{P}$ is the density matrix, and ρ is the corresponding electronic density. Operators in the Pauli spin basis are decomposed as:

$$\mathbf{V}_{\mathrm{xc}} = \mathbf{V}_{\mathrm{xc}}^{S} \otimes \mathbf{I}_2 + \sum_k \mathbf{V}_{\mathrm{xc}}^{k} \otimes \boldsymbol{\sigma}_k \tag{A.1}$$

$$\mathbf{P} = \mathbf{P}^{S} \otimes \mathbf{I}_2 + \sum_k \mathbf{P}^{k} \otimes \boldsymbol{\sigma}_k \tag{A.2}$$

Where $\boldsymbol{\sigma}$ is a Pauli matrix and $k \in \{x, y, z\}$. For a general operator, $\mathbf{X}$, the S, x, y and z components are defined from the α and β spin blocks as:

$$2\mathbf{X}^{S} = \mathbf{X}^{\alpha\alpha} + \mathbf{X}^{\beta\beta} \tag{A.3}$$

$$2\mathbf{X}^{x} = \mathbf{X}^{\alpha\beta} + \mathbf{X}^{\beta\alpha} \tag{A.4}$$

$$-2i\mathbf{X}^{y} = \mathbf{X}^{\alpha\beta} - \mathbf{X}^{\beta\alpha} \tag{A.5}$$

$$2\mathbf{X}^{z} = \mathbf{X}^{\alpha\alpha} - \mathbf{X}^{\beta\beta} \tag{A.6}$$

The problem to solve in density functional theory is a numerical integration of the density over a pre-defined grid of points. Efficient usage of processor caches and parallelization are core design principles of our implementation in GauXC. The high-level scheme is the same between CPU and GPU implementations, but the low-level kernels differ. The outermost layer of parallelization is done by batching grid points, since the density can be independently evaluated on each grid point.

The procedure is outlined in Algorithm 2 and Algorithm 3. At the start of the calcula-

tion, GauXC divides the grid points into batches, each of which contains all the data needed to perform the integration for those grid points. To accommodate different architectures, the batch size can be tuned by the user in the GauXC interface. For each batch of points, basis functions with insignificant overlap are screened out. Reduced density matrices corresponding to the screened basis set are formed, and memory for intermediates, such as the $\mathbf{X}$ and $\mathbf{Z}$ matrices and the U and V variables, are allocated contiguously alongside the reduced basis set and reduced density matrices. The $\mathbf{X}$ matrix is an intermediate used to form the density, and the $\mathbf{Z}$ matrix is an intermediate used to form the exchange-correlation potential. The V variables are the density and density gradients in the Pauli spinor basis. The U variables are the transformed density and density gradients which are fed into the exchange-correlation kernel.

When GauXC is finished with host-device memory transfer, numerical integration can begin. First, the $\mathbf{X}$ matrix and V variables are computed. The $\mathbf{X}$ matrix is computed with a `gemm` call on the density matrix and screened basis functions, Φ on a batch of grid points, performed with a `blas` library:

$$\mathbf{X} = \mathbf{P}\mathbf{\Phi} \tag{A.7}$$

GauXC also supports the `MAGMA` and `CUTLASS` libraries to perform this task with highly efficient batched linear algebra operations. The V variables are evaluated in a low-level kernel which assigns each thread to a single basis function and grid point, and then performs a reduction along the basis functions. The U variable kernel assigns each thread to a grid point. The XC energy and functional derivatives are evaluated with the parallel implementation of `libxc`. The $\mathbf{Z}$ matrix kernel follows the same parallelization scheme as the V variable kernel, and finally V_{xc} is formed by performing a `syr2k` call (performed with `blas`):

$$\mathbf{V}_{\text{xc}} = \mathbf{Z}\mathbf{\Phi}^T + \mathbf{\Phi}\mathbf{Z}^T \tag{A.8}$$

The final step is to distribute the batched indices back to the original indices.

Note that the U variables and XC functional evaluation require all spin densities, while the corresponding V variables and V_{xc} can be evaluated separately for each spin component. While it is possible to parallelize across spin densities within a batch for the V variables and V_{xc}, it would have considerable memory requirements since the $\mathbf{X}$ and $\mathbf{Z}$ intermediates, which are $N_{\text{bf}}^2 \times N_{\text{pts}}$, would need to be allocated for each term. Since the target of two-component GPU-accelerated GauXC is large systems and GPU memory is limited, we opt to perform the calculation for each spin-density sequentially so the memory for the intermediates can be re-used.

Algorithm 2: Scheme for evaluating $\mathbf{V}_{\text{xc}}$ and E_{xc}

1	Form local batches $\mathcal{B}_{\text{local}}$ according to Algorithm 3 of Ref.[225]	(host)
2	Allocate $\mathbf{P}^I$ and $\mathbf{V}^I$ for $I \in \{S, x, y, z\}$	(host/device)
3	Send density matrices $\mathbf{P}^S$, $\mathbf{P}^x$, $\mathbf{P}^y$, $\mathbf{P}^z$ to device	(host/device)
4	Zero out $\mathbf{V}^I_{\text{local}}$ and E_{local} for $I \in \{S, x, y, z\}$	(device)
5	**while** $\mathcal{B}_{\text{local}}$ *is not empty* **do**	
6	$\quad \mathcal{B}_{\text{device}} \leftarrow$ subset of $\mathcal{B}_{\text{local}}$ to saturate device memory	(host)
7	$\quad$ Pack $\mathcal{B}_{\text{device}}$ on host and send to device	(host/device)
8	$\quad$ Update V^I_{local} and E_{local} using Algorithm 3	(device)
9	Retrieve $\mathbf{V}^I_{\text{local}}$ and E_{local} from device for $I \in \{S, x, y, z\}$	(host/device)
10	Reduce $E \leftarrow E_{\text{local}}$ and $V^I_{\text{xc}} \leftarrow V^I_{\text{local}}$ for $I \in \{S, x, y, z\}$	(device)

Algorithm 3: Scheme for evaluating quadrature batches on a GPU device

for $\mathcal{B}_j \in \mathcal{B}$ **do**
  $\mathbf{P}_j^I \leftarrow$ Compress batch local density matrix from $\mathbf{P}^I$ for $I \in \{S, x, y, z\}$;
  $(\mathbf{\Phi}_j, \nabla\mathbf{\Phi}_j) \leftarrow$ Evaluate local batch collocation matrix and its gradient given a batch of basis functions.
  for $I \in \{S, x, y, z\}$ **do**
    $\mathbf{X}_j^I \leftarrow$ Evaluate $\mathbf{X}$ matrix;
    $(\rho_j^I, \nabla\rho_j^I) \leftarrow$ Evaluate density and density gradients;
  $(n^+, n^-, \gamma^{++}, \gamma^{+-}, \gamma^{--}) \leftarrow$ Transform 2-component densities and gradients $(\rho_j^I, \nabla\rho_j^I)$ for $I \in \{S, x, y, z\}$ into 1-component variables. See Eq. 48 of Ref. 168
  $(\varepsilon_j, \frac{\partial\varepsilon}{\partial n}, \frac{\partial\varepsilon}{\partial\gamma}) \leftarrow$ Evaluate xc functional and its derivatives;
  Update E_{xc};
  for $I \in \{S, x, y, z\}$ **do**
    $\mathbf{Z}_j^I \leftarrow$ Evaluate $\mathbf{Z}$ matrix. See Eqns. 57–58 of Ref. 168;
    Increment V_{xc}^I by $\mathbf{Z}_j^I\mathbf{\Phi}_j^T + \mathbf{\Phi}_j(\mathbf{Z}_j^I)^T$;

BIBLIOGRAPHY

[1] Seihwan Ahn, Mannkyu Hong, Mahesh Sundararajan, Daniel H. Ess, and Mu-Hyun Baik. Design and optimization of catalysts based on mechanistic insights derived from quantum chemical reaction modeling. *Chem. Rev.*, 119(11):6509–6560, 2019.

[2] Melisa Alkan, Buu Q. Pham, Daniel Del Angel Cruz, Jeff R. Hammond, Taylor A. Barnes, and Mark S. Gordon. LibERI—A Portable and Performant Multi-GPU Accelerated Library for Electron Repulsion Integrals via OpenMP Offloading and Standard Language Parallelism. *J. Chem. Phys.*, 161(8):082501, 08 2024.

[3] Adel Almoukhalalati, Stefan Knecht, Hans Jørgen Aa. Jensen, Kenneth G. Dyall, and Trond Saue. Electron correlation within the relativistic no-pair approximation. *J. Chem. Phys.*, 145(7):074104, 2016.

[4] Kerstin. Andersson, Per Aake. Malmqvist, Bjoern O. Roos, Andrzej J. Sadlej, and Krzysztof. Wolinski. Second-Order Perturbation Theory with a CASSCF Reference Function. *J. Chem. Phys.*, 94(14):5483–5488, 1990.

[5] Kerstin Andersson, Per-Åke Malmqvist, and Björn O. Roos. Second-Order Perturbation Theory with a Complete Active Space Self-Consistent Field Reference Function. *J. Chem. Phys.*, 96(2):1218–1226, 01 1992.

[6] C. Angeli, R. Cimiraglia, S. Evangelisti, T. Leininger, and J. P. Malrieu. Introduction of n-Electron Valence States for Multireference Perturbation Theory. *J. Chem. Phys.*, 114:10252–10264, 2001.

[7] Roberto Ansaloni, Gian Luigi Bendazzoli, Stefano Evangelisti, and Elda Rossi. A Parallel Full-CI Algorithm. *Comp. Phys. Comm.*, 128(1):496–515, 2000.

[8] David D Awschalom, Lee C Bassett, Andrew S Dzurak, Evelyn L Hu, and Jason R Petta. Quantum spintronics: Engineering and manipulating atom-like spins in semiconductors. *Science*, 339(6124):1174–1179, 2013.

[9] Jon Baker and Matt Shirel. Ab initio quantum chemistry on pc-based parallel supercomputers. *Parallel Computing*, 26(7):1011–1024, 2000.

[10] Giuseppe M. J. Barca, Melisa Alkan, Jorge L. Galvez-Vallejo, David L. Poole, Alistair P. Rendell, and Mark S. Gordon. Faster Self-Consistent Field (SCF) Calculations on GPU Clusters. *J. Chem. Theory Comput.*, 17(12):7486–7503, 2021.

[11] Radovan Bast, Hans J. Aa. Jensen, and Trond Saue. Relativistic adiabatic time-dependent density functional theory using hybrid functionals and noncollinear spin magnetization. *Int. J. Quant. Chem.*, 109:2091–2112, 2009.

[12] Jefferson E. Bates and Toru Shiozaki. Fully Relativistic Complete Active Space Self-Consistent Field for Large Molecules: Quasi-second-order Minimax Optimization. *J. Chem. Phys.*, 142(4):044112, 2015.

[13] Stefano Battaglia, Sebastian Keller, and Stefan Knecht. Efficient relativistic density-matrix renormalization group implementation in a matrix-product formulation. *J. Chem. Theory Comput.*, 14(5):2353–2369, 2018.

[14] Ryan A. Beck, Lixin Lu, Peter V. Sushko, Xiaodong Xu, and Xiaosong Li. Defect-induced magnetic skyrmion in a two-dimensional chromium triiodide monolayer. *JACS Au*, 1(9):1362–1367, 2021.

[15] Axel D Becke. A multicenter numerical integration scheme for polyatomic molecules. *J. Chem. Phys.*, 88(4):2547–2553, 1988.

[16] Carlos L. Benavides-Riveros, Nektarios N. Lathiotakis, and Miguel A. L. Marques. Towards a formal definition of static and dynamic electronic correlations. *Phys. Chem. Chem. Phys.*, 19:12655–12664, 2017.

[17] Vladimir B. Beresteckij, Evgenij M. Lifšic, Lev P. Pitaevskij, Vladimir B. Beresteckij, and Lev Davidovič Landau. *Quantum Electrodynamics*. Course of Theoretical Physics. Butterworth-Heinemann, Oxford, 2nd, reprint edition, 2008.

[18] V. B. Berestetskii, E. M. Lifshitz, and L. P. Pitaevskii. *QUANTUM ELECTRODYNAMICS*, volume 4 of *Course of Theoretical Physics*. Pergamon Press, Oxford, 1982.

[19] Max Born and J. Robert Oppenheimer. Zur Quantentheorie Der Molekeln. *Ann. Phys.*, 389(20):457–484, 1927.

[20] Robert Boyle. *The sceptical chymist or, Chymico-physical doubts paradoxes*. Printed by J. Cadwell for J. Crooke, 1661.

[21] S. F. Boys. Electronic wave functions - i. a general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 200(1063):542–554, 02 1950.

[22] Anatoly L Buchachenko and Vitaly L Berdinsky. Electron spin catalysis. *Chem. Rev.*, 102(3):603–612, 2002.

[23] Ireneusz W. Bulik, Giovanni Scalmani, Michael J. Frisch, and Gustavo E. Scuseria. Noncollinear density functional theory having proper invariance and local torque properties. *Phys. Rev. B*, 87:035117, 2013.

[24] Kieron Burke. Perspective on density functional theory. *J. Chem. Phys.*, 136(15), 2012.

[25] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. Quantum chemistry in the age of quantum computing. *Chemical reviews*, 119(19):10856–10915, 2019.

[26] K. Capelle, G. Vignale, and B. L. Györffy. Spin currents and spin dynamics in time-dependent density-functional theory. *Phys. Rev. Lett.*, 87:206403, Oct 2001.

[27] Klaus Capelle, Giovanni Vignale, and Carsten A Ullrich. Spin gaps and spin-flip energies in density-functional theory. *Phys. Rev. B*, 81(12):125114, 2010.

[28] Mark E Casida. Time-dependent density-functional theory for molecules and molecular solids. *J. Mol. Struct. Theochem*, 914(1-3):3–18, 2009.

[29] Paolo Celani and Hans-Joachim Werner. Multireference Perturbation Theory for Large Restricted and Selected Active Space Reference Wave Functions. *J. Chem. Phys.*, 112(13):5546–5557, 04 2000.

[30] Alexios Chatzigoulas, Konstantina Karathanou, Dimitris Dellis, and Zoe Cournia. Nanocrystal: A web-based crystallographic tool for the construction of nanoparticles based on their crystal habit. 58(12):2380–2386, 2018.

[31] Benjamin WJ Chen, Lang Xu, and Manos Mavrikakis. Computational methods in heterogeneous catalysis. *Chem. Rev.*, 121(2):1007–1048, 2020.

[32] Aron J. Cohen and Paula Mori-Sánchezand Weitao Yang. Challenges for density functional theory. *Chemical Reviews*, 112(1):289–320, 2012.

[33] Daniel J Cole and Nicholas DM Hine. Applications of large-scale density functional theory in biology. *J. Phys. Condens. Matter*, 28(39):393001, 2016.

[34] Lawrence A. Covick and Kenneth M. Sando. Four-index transformation on distributed-memory parallel computers. *Journal of Computational Chemistry*, 11(10):1151–1159, 1990.

[35] Filippo De Angelis. Modeling materials and processes in hybrid/organic photovoltaics: from dye-sensitized to perovskite solar cells. *Acc. Chem. Res.*, 47(11):3349–3360, 2014.

[36] Peter Deglmann, Ansgar Schäfer, and Christian Lennartz. Application of quantum calculations in the chemical industry—an overview. *International Journal of Quantum Chemistry*, 115(3):107–136, 2015.

[37] P.A.M. Dirac. *The Principles of Quantum Mechanics*. Oxford, 1958.

[38] Paul Adrien Maurice Dirac. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 123(792):714–733, 04 1929.

[39] Paul Adrien Maurice Dirac and Ralph Howard Fowler. The quantum theory of the electron. 117:610–624, 1928.

[40] Paul Adrien Maurice Dirac and Ralph Howard Fowler. The quantum theory of the electron. part ii. 118:351–361, 1928.

[41] Marvin Douglas and Norman M. Kroll. Quantum electrodynamical corrections to the fine structure of helium. *Ann. Phys.*, 82:89–155, 1974.

[42] Michel Dupuis, John Rys, and Harry F. King. Evaluation of molecular integrals over gaussian basis functions. *J. Chem. Phys.*, 65(1):111–116, 07 1976.

[43] K. G. Dyall. Interfacing relativistic and nonrelativistic methods. i. normalized elimination of the small component in the modified dirac equation. *J. Chem. Phys.*, 106(23):9618–9626, 1997.

[44] Kenneth G. Dyall. Interfacing relativistic and nonrelativistic methods. II. investigation of a low-order approximation. *J. Chem. Phys.*, 109:4201–4208, 1998.

[45] Kenneth G. Dyall. Interfacing relativistic and nonrelativistic methods. IV.. One- and two-electron scalar approximations. *J. Chem. Phys.*, 115:9136–9143, 2001.

[46] Kenneth G. Dyall and Thomas Enevoldsen. Interfacing relativistic and nonrelativistic methods. III. atomic 4-spinor expansions and integral approximations. *J. Chem. Phys.*, 111:10000–10007, 1999.

[47] Kenneth G. Dyall and Knut Fægri, Jr. *Introduction to Relativistic Quantum Chemistry*. Oxford University Press, 2007.

[48] Kenneth G. Dyall and Knut Fægri, Jr. *Introduction to Relativistic Quantum Chemistry*. Oxford University Press, 2007.

[49] Franco Egidi, Joshua J. Goings, Michael J. Frisch, and Xiaosong Li. Direct atomic-orbital-based relativistic two-component linear response method for calculating excited-state fine structures. *J. Chem. Theory Comput.*, 12(8):3711–3718, 2016.

[50] Franco Egidi, Shichao Sun, Joshua J. Goings, Giovanni Scalmani, Michael J. Frisch, and Xiaosong Li. Two-component non-collinear time-dependent spin density functional theory for excited state calculations. *J. Chem. Theory Comput.*, 13(6):2591–2603, 2017.

[51] Jordan Ehrman, Ernesto Martinez-Baez, Andrew J. Jenkins, and Xiaosong Li. Improving one-electron exact-two-component relativistic methods with the dirac–

coulomb–breit-parameterized effective spin–orbit coupling. *J. Chem. Theory Comput.*, 19:5785–5790, 2023.

[52] Albert Einstein. Zur elektrodynamik bewegter körper. *Annalen der Physik*, 322(10):891–921, 1905.

[53] E Engel. Relativistic density functional theory: Foundations and basic formalism. volume 11, pages 523–621. Elsevier, 2002.

[54] Janus J. Eriksen. The Shape of Full Configuration Interaction to Come. *J. Phys. Chem. Lett.*, 12(1):418–432, January 2021.

[55] Kenneth Esler, Jeongnim Kim, David Ceperley, and Luke Shulenburger. Accelerating Quantum Monte Carlo Simulations of Real Materials on GPU Clusters. *Comput. Sci. Eng.*, 14(1):40–51, 2012.

[56] Stefano Evangelisti, Jean-Pierre Daudey, and Jean-Paul Malrieu. Convergence of an Improved CIPSI Algorithm. *Chem. Phys.*, 75(1):91–102, 1983.

[57] B. Scott Fales, Ethan R. Curtis, K. Grace Johnson, Dean Lahana, Stefan Seritan, Yuanheng Wang, Hayley Weir, Todd J. Martínez, and Edward G. Hohenstein. Performance of Coupled-Cluster Singles and Doubles on Modern Stream Processing Architectures. *J. Chem. Theory Comput.*, 16(7):4021–4028, 2020.

[58] B. Scott Fales and Todd J. Martínez. Efficient Treatment of Large Active Spaces through Multi-GPU Parallel Implementation of Direct Configuration Interaction. *J. Chem. Theory Comput.*, 16(3):1586–1596, 2020.

[59] Yuan Ping Feng, Lei Shen, Ming Yang, Aizhu Wang, Minggang Zeng, Qingyun Wu, Sandhya Chintalapati, and Ching-Ray Chang. Prospects of spintronics based on 2d materials. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 7(5):e1313, 2017.

[60] Michael Filatov and Dieter Cremer. A new quasi-relativistic approach for density functional theory based on the normalized elimination of the small component. *Chem. Phys. Lett.*, 351:259–266, 2002.

[61] Timo Fleig, Hans Jørgen Aa. Jensen, Jeppe Olsen, and Lucas Visscher. The generalized active space concept for the relativistic treatment of electron correlation. iii. large-scale configuration interaction and multiconfiguration self-consistent-field four-component methods with application to uo2. *The Journal of Chemical Physics*, 124:104106, 2006.

[62] Timo Fleig, Jeppe Olsen, and Christel M. Marian. The generalized active space concept for the relativistic treatment of electron correlation. i. kramers-restricted two-component configuration interaction. *The Journal of Chemical Physics*, 114:4775–4790, 2001.

[63] Timo Fleig, Jeppe Olsen, and Lucas Visscher. The generalized active space concept for the relativistic treatment of electron correlation. ii. large-scale configuration interaction implementation based on relativistic 2- and 4-spinors and its application. *The Journal of Chemical Physics*, 119:2963–2971, 2003.

[64] Thomas R. Furlani and Harry F. King. Implementation of a parallel direct scf algorithm on distributed memory computers. *Journal of Computational Chemistry*, 16(1):91–104, 1995.

[65] Zhengting Gan and R.J. Harrison. Calibrating Quantum Chemistry: A Multi-Teraflop, Parallel-Vector, Full-Configuration Interaction Program for the Cray-X1. In *SC '05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*, pages 22–22, 2005.

[66] Dmitry Ganyushin and Frank Neese. A Fully Variational Spin-orbit Coupled Complete Active Space Self-Consistent Field Approach: Application to Electron Paramagnetic Resonance g-tensors. *J. Chem. Phys.*, 138(10):104113, 2013.

[67] Hong Gao, Satoshi Imamura, Akihiko Kasagi, and Eiji Yoshida. Distributed implementation of full configuration interaction for one trillion determinants. *Journal of Chemical Theory and Computation*, 20:1185–1192, 2024.

[68] Dylan C Gary, Alessio Petrone, Xiaosong Li, and Brandi M Cossairt. Investigating the role of amine in inp nanocrystal syntheses: Destabilizing cluster intermediates by z-type ligand displacement. *Chem. Commun.*, 53:161–164, 2017.

[69] Robert Gilmore. Lie groups, lie algebras, and some of their applications., 1974.

[70] J. J. Goings, F. Egidi, and X Li. Current development of non-collinear electronic structure theory. *Int. J. Quant. Chem.*, 118:e25398, 2018.

[71] Joshua J. Goings, Hang Hu, Chao Yang, and Xiaosong Li. Reinforcement learning configuration interaction. *J. Chem. Theory Comput.*, 17(9):5482–5491, 2021.

[72] Joshua J. Goings, Joseph M. Kasper, Franco Egidi, Shichao Sun, and Xiaosong Li. Real time propagation of the exact two component time-dependent density functional theory. *J. Chem. Phys.*, 145(10):104107, 2016.

[73] Joshua J. Goings, Patrick J. Lestrange, and Xiaosong Li. Real-time time-dependent electronic structure theory. *WIREs Comput. Mol. Sci.*, 8(1):e1341, 2018.

[74] Herbert Goldstein. *Classical Mechanics*. Addison-Wesley, 1980.

[75] a EKU Gross, JF Dobson, and M Petersilka. Density functional theory of time-dependent phenomena. *Density Funct. Theory Ii Relativistic Time Depend. Ext.*, pages 81–172, 2005.

[76] Jürgen Hafner, Christopher Wolverton, and Gerbrand Ceder. Toward computational materials design: The impact of density functional theory on materials research. *Mr Bull.*, 31(9):659–668, 2006.

[77] Martin Head-Gordon and John. A. Pople. A method for two-electron gaussian integral and integral derivative evaluation using recurrence relations. *J. Chem. Phys.*, 89:5777–5786, 1988.

[78] Trygve Helgaker, Poul Jørgensen, and Jeppe Olsen. *Molecular Electronic-Structure Theory*. John Wiley Sons Ltd, 2000.

[79] Trygve Helgaker and Peter R. Taylor. *Gaussian Basis Sets and Molecular Integrals*, pages 725–856. World Scientific Publishing Company, 1995.

[80] Bernd A. Hess. Relativistic electronic-structure calculations employing a two-component no-pair formalism with external-field projection operators. *Physical Review A*, 33:3742–3748, 1986.

[81] K Hirao, editor. *Recent Advances in Multireferences Methods*. World Sientific, 1999.

[82] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[83] Adam A. Holmes, Norm M. Tubman, and C. J. Umrigar. Heat-Bath Configuration Interaction: An Efficient Selected Configuration Interaction Algorithm Inspired by Heat-Bath Sampling. *J. Chem. Theory Comput.*, 12(8):3674–3680, August 2016.

[84] Adam A. Holmes, C. J. Umrigar, and Sandeep Sharma. Excited States Using Semistochastic Heat-Bath Configuration Interaction. *J. Chem. Phys.*, 147(16):164111, October 2017.

[85] Chad E. Hoyer, Hang Hu, Lixin Lu, Stefan Knecht, and Xiaosong Li. Relativistic kramers-unrestricted exact-two-component density matrix renormalization group. *J. Phys. Chem. A*, 126(30):5011–5020, 2022.

[86] Chad E. Hoyer, Lixin Lu, Hang Hu, Kirill D. Shumilov, Shichao Sun, Stefan Knecht, and Xiaosong Li. Correlated dirac–coulomb–breit multiconfigurational self-consistent-field methods. *J. Chem. Phys.*, 158(4):044101, 2023.

[87] Hang Hu, Andrew J. Jenkins, Hongbin Liu, Joseph M. Kasper, Michael J. Frisch, and Xiaosong Li. Relativistic two-component multireference configuration interaction method with tunable correlation space. *J. Chem. Theory Comput.*, 16(5):2975–2984, 2020.

[88] Hang Hu, Shiv Upadhyay, Lixin Lu, Andrew J. Jenkins, Tianyuan Zhang, Agam Shayit, Stefan Knecht, and Xiaosong Li. Small Tensor Product Distributed Active Space (STP-DAS) Framework for Relativistic and Non-relativistic Multiconfiguration Calculations: Scaling from 10^9 on a Laptop to 10^{12} Determinants on a Supercomputer. *Chem. Phys. Rev.*, 5(4):041404, 10 2024.

[89] B. Huron, J. P. Malrieu, and P. Rancurel. Iterative Perturbation Calculations of Ground and Excited State Energies from Multiconfigurational Zeroth-Order Wavefunctions. *J. Chem. Phys.*, 58(12):5745–5759, June 1973.

[90] Miroslaŵ Ilias and Trond Saue. An infinite-order relativistic hamiltonian by a simple one-step transformation. *J. Chem. Phys.*, 126:064102, 2007.

[91] Christoph R Jacob and Markus Reiher. Spin in density-functional theory. *Int. J. Quantum Chem.*, 112(23):3661–3684, 2012.

[92] Andrew J. Jenkins, Hang Hu, Lixin Lu, Michael J. Frisch, and Xiaosong Li. Two-component multireference restricted active space configuration interaction for the computation of l-edge x-ray absorption spectra. *J. Chem. Theory Comput.*, 18(1):141–150, 2022.

[93] Hans J Aa. Jensen, Kenneth G. Dyall, Trond Saue, and Knut Fægri. Relativistic four-component multiconfigurational self-consistent-field theory for molecules: Formalism. *J. Chem. Phys.*, 104(11):4083–4097, 1996.

[94] J. A. Karwowski and I. Shavitt. Configuration Interaction. In S. Wilson, editor, *Handbook of Molecular Physics and Quantum Chemistry*. John Wiley & Sons, Ltd, 2003.

[95] Michael Klene, Michael A. Robb, Lluís Blancafort, and Michael J. Frisch. A New Efficient Approach to the Direct Restricted Active Space Self-Consistent Field Method. *J. Chem. Phys.*, 119:713–728, 2003.

[96] Michael Klene, Michael A. Robb, Michael J. Frisch, and Paolo Celani. Parallel implementation of the ci-vector evaluation in full ci/cas-scf. *The Journal of Chemical Physics*, 113:5653–5665, 2000.

[97] Stefan Knecht, Hans Jørgen Aa. Jensen, and Timo Fleig. Large-scale parallel configuration interaction. i. nonrelativistic and scalar-relativistic general active space implementation with application to (rb–ba)+. *The Journal of Chemical Physics*, 128:014108, 2008.

[98] Stefan Knecht, Hans Jørgen Aa. Jensen, and Timo Fleig. Large-scale parallel configuration interaction. ii. two- and four-component double-group general active space

implementation with application to bih. *The Journal of Chemical Physics*, 132:014108, 2010.

[99] Stefan Knecht, Hans Jørgen Aa. Jensen, and Trond Saue. Relativistic Quantum Chemical Calculations Show that the Uranium Molecule U_2 has a Quadruple Bond. *Nat. Chem.*, 11(1):40–44, 2019.

[100] Stefan Knecht, Ös Legeza, and Markus Reiher. Communication: Four-component density matrix renormalization group. *J. Chem. Phys.*, 140(4):041101, 2014.

[101] Peter J. Knowles and Nicholas C. Handy. A Determinant Based Full Configuration Interaction Program. *Comp. Phys. Comm.*, 54(1):75–83, April 1989.

[102] P.J. Knowles and N.C. Handy. A new determinant-based full configuration interaction method. *Chemical Physics Letters*, 111:315–321, 1984.

[103] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[104] Lukas Konecny, Marius Kadek, Stanislav Komorovsky, Olga L. Malkina, Kenneth Ruud, and Michal Repisky. Acceleration of relativistic electron dynamics by means of X2C transformation: Application to the calculation of nonlinear optical properties. *J. Chem. Theory Comput.*, 12(12):5823–5833, Dec 2016.

[105] Mikael Kovtun, Eleftherios Lambros, Aodong Liu, Diandong Tang, David B. Williams-Young, and Xiaosong Li. Accelerating relativistic exact-two-component density functional theory calculations with graphical processing units. *J. Chem. Theory Comput.*, 20(18):7694–7699, 2024.

[106] Jörg Kussmann and Christian Ochsenfeld. Hybrid CPU/GPU Integral Engine for Strong-Scaling *Ab Initio* Methods. *J. Chem. Theory Comput.*, 13(7):3153–3159, 2017.

[107] Werner Kutzlenigg and Wenjian Liu. Quasirelativistic theory equivalent to fully relativistic theory. *J. Chem. Phys.*, 123:241102, 2005.

[108] Thaddeus D Ladd, Fedor Jelezko, Raymond Laflamme, Yasunobu Nakamura, Christopher Monroe, and Jeremy Lloyd O'Brien. Quantum computers. *Nature*, 464(7285):45–53, 2010.

[109] Susi Lehtola, Conrad Steigemann, Micael J.T. Oliveira, and Miguel A.L. Marques. Recent developments in libxc — a comprehensive library of functionals for density functional theory. *SoftwareX*, 7:1–5, 2018.

[110] Junhao Li, Matthew Otten, Adam A. Holmes, Sandeep Sharma, and C. J. Umrigar. Fast Semistochastic Heat-Bath Configuration Interaction. *J. Chem. Phys.*, 149(21):214110, December 2018.

[111] Nan Li, Zonglong Zhu, Chu-Chen Chueh, Hongbin Liu, Bo Peng, Alessio Petrone, Xiaosong Li, Liduo Wang, and Alex K.-Y. Jen. Mixed cation $fa_xpea_{1-x}pbi_3$ with enhanced phase and ambient stability toward high-performance perovskite solar cells. *Adv. Energy Mater.*, 7:1601307, 2016.

[112] Rui Li, Qiming Sun, Xing Zhang, and Garnet Kin-Lic Chan. Introducing gpu acceleration into the python-based simulations of chemistry framework. *The Journal of Physical Chemistry A*, 129(5):1459–1468, 2025.

[113] Ruihao Li, Bagus Hanindhito, Sanjana Yadav, Qinzhe Wu, Krishna Kavi, Gayatri Mehta, Neeraja J. Yadwadkar, and Lizy K. John. Performance Implications of Pipelining the Data Transfer in CPU-GPU Heterogeneous Systems. *ACM Trans. Archit. Code Optim.*, 22(3), September 2025.

[114] Xiaosong Li, Niranjan Govind, Christine Isborn, A. Eugene DePrince, and Kenneth Lopata. Real-time time-dependent electronic structure theory. *Chem. Rev.*, 120(18):9951–9993, 2020.

[115] Zhendong Li, Yunlong Xiao, and Wenjian Liu. On the spin separation of algebraic two-component relativistic hamiltonians. *J. Chem. Phys.*, 137:154114, 2012.

[116] Giovanni Li Manni, Rebecca K. Carlson, Sijie Luo, Dongxia Ma, Jeppe Olsen, Donald G. Truhlar, and Laura Gagliardi. Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.*, 10(9):3669–3680, 2014.

[117] Can Liao, Manzhou Zhu, De-en Jiang, and Xiaosong Li. Manifestation of the interplay between spin–orbit and jahn–teller effects in au_{25} superatom uv-vis fingerprint spectra. *Chem. Sci.*, 14:4666–4671, 2023.

[118] Hans Lischka, Holger Dachsel, Ron Shepard, and Robert J. Harrison. Parallel computing in quantum chemistry — message passing and beyond for a general ab initio program system. *Future Generation Computer Systems*, 11(4):445–450, 1995. High Performance Computing and Networking.

[119] Hans Lischka, Thomas Müller, Péter G. Szalay, Isaiah Shavitt, Russell M. Pitzer, and Ron Shepard. Columbus—a Program System for Advanced Multireference Theory Calculations. *WIREs Comput. Mol. Sci.*, 1:191–199, 2011.

[120] Hans Lischka, Ron Shepard, Franklin B. Brown, and Isaiah Shavitt. New Implementation of the Graphical Unitary Group Approach for Multireference Direct Configuration Interaction Calculations. *Int. J. Quant. Chem.*, 20(S15):91–100, 1981.

[121] B. Liu. *Ab Initio* Potential Energy Surface for Linear H_3. *J. Chem. Phys.*, 58(5):1925–1937, 03 1973.

[122] Junzi Liu and Lan Cheng. Relativistic coupled-cluster and equation-of-motion coupled-cluster methods. *WIREs Comput. Mol. Sci.*, 11(6):1536, 2021.

[123] W. Liu and D. Peng. Exact two-component hamiltonians revisited. *J. Chem. Phys.*, 131(3):031104, 2009.

[124] Wenjian Liu. Ideas of relativistic quantum chemistry. *Mol. Phys.*, 108:1679–1706, 2010.

[125] Wenjian Liu. *Handbook of Relativistic Quantum Chemistry.* Springer-Verlag Berlin Heidelberg, 2017.

[126] Wenjian Liu and Mark R. Hoffmann. iCI: Iterative CI Toward Full CI. *J. Chem. Theory Comput.*, 12(3):1169–1178, March 2016.

[127] Wenjian Liu and Daoling Peng. Infinite-order quasirelativistic density functional method based on the exact matrix quasirelativistic theory. *J. Chem. Phys.*, 125:044102, 2006.

[128] Lixin Lu, Hang Hu, Andrew J. Jenkins, and Xiaosong Li. Exact-two-component relativistic multireference second-order perturbation theory. *J. Chem. Theory Comput.*, 18(5):2983–2992, 2022.

[129] Giulia Luise, Chin-Wei Huang, Thijs Vogels, Derk P Kooi, Sebastian Ehlert, Stephanie Lanius, et al. Accurate and scalable exchange-correlation with deep learning. *arXiv preprint arXiv:2506.14665*, 2025.

[130] Sijie Luo, Ivan Rivalta, Victor Batista, and Donald G. Truhlar. Noncollinear spins provide a self-consistent treatment of the low-spin state of a biomimetic oxomanganese synthetic trimer inspired by the oxygen evolving complex of photosystem ii. *J. Phys. Chem. Lett.*, 2:2629–2633, 2011.

[131] Sijie Luo and Donald G. Truhlar. Noncollinear spin states for density functional calculations of open-shell and multi-configurational systems: Dissociation of mno and nio and barrier heights of o3, beh2, and h4. *J. Chem. Theory Comput.*, 9:5349–5355, 2013.

[132] Dongxia Ma, Giovanni Li Manni, and Laura Gagliardi. The Generalized Active Space Concept in Multiconfigurational Self-Consistent Field Methods. *J. Chem. Phys.*, 135:044128, 2011.

[133] Dongxia Ma, Giovanni Li Manni, Jeppe Olsen, and Laura Gagliardi. Second-Order Perturbation Theory for Generalized Active Space Self-Consistent-Field Wave Functions. *J. Chem. Theory Comput.*, 12(7):3208–3213, 2016.

[134] A. H. MacDonald and S. H. Vosko. A relativistic density functional formalism. *J. Phys. C Solid State*, 12:2977, 1979.

[135] Per Åke Malmqvist. Calculation of transition density matrices by nonunitary orbital transformations. *Int. J. Quant. Chem.*, 30(4):479–494.

[136] Per Åke Malmqvist and Björn O. Roos. The CASSCF State Interaction Method. *Chem. Phys. Lett.*, 155(2):189–194, 1989.

[137] Per-Åke Malmqvist and Björn O." Roos. The Restricted Active Space Self-Consistent-Field Method, Implemented with a Split Graph Unitary Group Approach. *J. Phys. Chem.*, 94:5477–5482, 1990.

[138] Fionn D. Malone, Shuai Zhang, and Miguel A. Morales. Accelerating Auxiliary-Field Quantum Monte Carlo Simulations of Solids with Graphical Processing Units. *J. Chem. Theory Comput.*, 16(7):4286–4297, 2020.

[139] Madushanka Manathunga, Yipu Miao, Dawei Mu, Andreas W Goötz, and Kenneth M Merz Jr. Parallel implementation of density functional theory methods in the quantum interaction computational kernel program. *J. Chem. Theory Comput.*, 16(7):4315–4326, 2020.

[140] Miguel AL Marques and Eberhard KU Gross. Time-dependent density functional theory. *Annu. Rev. Phys. Chem.*, 55:427–455, 2004.

[141] L. E. McMurchie and E. R. Davidson. One- and two-electron integrals over cartesian gaussian functions. *J. Chem. Phys.*, 26(2):218–231, 1978.

[142] Andor Menczer, Maarten van Damme, Alan Rask, Lee Huntington, Jeff Hammond, Sotiris S. Xantheas, Martin Ganahl, and Örs Legeza. Parallel Implementation of the Density Matrix Renormalization Group Method Achieving a Quarter petaFLOPS Performance on a Single DGX-H100 GPU Node. *J. Chem. Theory Comput.*, 20(19):8397–8404, 2024.

[143] Kenneth M Merz Jr. Using quantum mechanical approaches to study biological systems. *Accounts of Chemical Research*, 47(9):2804–2811, 2014.

[144] A. Mohanty and E. Clementi. Dirac-fock self-consistent field method for closed-shell molecules with kinetic balance and finite nuclear size. *Int. J. Quant. Chem.*, 39:487–517, 1991.

[145] Stephan Mohr, Laura E Ratcliff, Luigi Genovese, Damien Caliste, Paul Boulanger, Stefan Goedecker, and Thierry Deutsch. Accurate and efficient linear scaling dft calculations with universal applicability. *Phys. Chem. Chem. Phys.*, 17(47):31360–31370, 2015.

[146] S. Navas et al. Review of particle physics. 110(3):030001, 2024.

[147] Jörg Neugebauer and Tilmann Hickel. Density functional theory in materials science. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3(5):438–448, 2013.

[148] Isaac Newton. *Philosophiæ Naturalis Principia Mathematica*. Jussu Societatis Regiæ ac Typis Josephi Streater, Londini, 1687.

[149] John R. Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming. *2008 IEEE Hot Chips 20 Symposium (HCS)*, pages 40–53, 2008.

[150] Matías A Nitsche, Manuel Ferreria, Esteban E Mocskos, and Mariano C Gonzalez Lebrero. Gpu accelerated implementation of density functional theory for hybrid qm/mm simulations. *J. Chem. Theory Comput.*, 10(3):959–967, 2014.

[151] S. Obara and A. Saika. Efficient recursive computation of molecular integrals over cartesian gaussian functions. *J. Chem. Phys.*, 84(7):3963–3974, 04 1986.

[152] Jeppe Olsen, Poul Jørgensen, and Jack Simons. Passing the One-Billion Limit in Full Configuration-Interaction (FCI) Calculations. *Chem. Phys. Lett.*, 169(6):463–472, 1990.

[153] Jeppe Olsen, Björn O. Roos, Poul Jo/rgensen, and Hans Jo/rgen Aa. Jensen. Determinant based configuration interaction algorithms for complete and restricted configuration interaction spaces. *The Journal of Chemical Physics*, 89:2185–2192, 1988.

[154] Ruben Pauncz. *Spin Eigenfunctions*. Plenum Press, New York, 1979.

[155] Fabijan Pavošević, Tanner Culpitt, and Sharon Hammes-Schiffer. Multicomponent quantum chemistry: Integrating electronic and nuclear quantum effects via the nuclear–electronic orbital method. *Chem. Rev.*, 120(9):4222–4253, 2020.

[156] Mark R Pederson and Koblar A Jackson. Variational mesh for quantum-mechanical simulations. *Phys. Rev. B*, 41(11):7453, 1990.

[157] Daoling Peng, Wenjian Liu, Yunlong Xiao, and Lan Cheng. Making four- and two-component relativistic density functional methods fully equivalent based on the idea of from atoms to molecule. *J. Chem. Phys.*, 127:104106, 2007.

[158] Daoling Peng, Nils Middendorf, Florian Weigend, and Markus Reiher. An efficient implementation of two-component relativistic exact-decoupling methods for large molecules. *J. Chem. Phys.*, 138:184105, 2013.

[159] Juan E. Peralta, Gustavo E. Scuseria, and Michael J. Frisch. Noncollinear magnetism in density functional calculations. *Phys. Rev. B*, 75:125119, 2007.

[160] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77(18):3865–3868, 1996.

[161] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.*, 78:1396–1396, Feb 1997.

[162] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, Oct 1996.

[163] Kristin Persson. Materials data on ag (sg:225) by materials project, 11 2014.

[164] Kristin Persson. Materials data on au (sg:225) by materials project, 4 2016.

[165] Kristin Persson. Materials data on cu (sg:225) by materials project, 4 2016.

[166] Michael Edward Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press, 1995. Addison-Wesley.

[167] Alessio Petrone, Joshua J. Goings, and Xiaosong Li. Quantum confinement effects on optical transitions in nanodiamonds containing nitrogen vacancies. *Phys. Rev. B*, 94:165402, 2016.

[168] Alessio Petrone, David B. Williams-Young, Shichao Sun, Torin F. Stetina, and Xiaosong Li. An efficient implementation of two-component relativistic density functional theory with torque-free auxiliary variables. *Euro. Phys. J. B*, 91(7):169, 2018.

[169] Johann V. Pototschnig, Anastasios Papadopoulos, Dmitry I. Lyakh, Michal Repisky, Loïc Halbert, André Severo Pereira Gomes, Hans Jørgen Aa Jensen, and Lucas Visscher. Implementation of relativistic coupled cluster theory for massively parallel gpu-accelerated computing architectures. *J. Chem. Theory Comput.*, 17(9):5509–5529, 2021.

[170] John P.Perdew. Climbing the ladder of density functional approximations. *MRS Bulletin*, 38(9):743–750, 2013.

[171] Joseph CA Prentice, Jolyon Aarons, James C Womack, Alice EA Allen, Lampros Andrinopoulos, Lucian Anton, Robert A Bell, Arihant Bhandari, Gabriel A Bramley, Robert J Charlton, et al. The onetep linear-scaling density functional theory program. *J. Chem. Phys.*, 152(17):174111, 2020.

[172] Zhichen Pu, Hao Li, Ning Zhang, Hong Jiang, Yiqin Gao, Yunlong Xiao, Qiming Sun, Yong Zhang, and Sihong Shao. Noncollinear density functional theory. *Phys. Rev. Res.*, 5:013036, Jan 2023.

[173] Pekka Pyykkö. Relativistic effects in chemistry: More common than you thought. *Annu. Rev. Phys. Chem.*, 63:45–64, 2012.

[174] A. K. Rajagopal. Inhomogeneous relativistic electron gas. *J. Phys. C Solid State*, 11:L943, 1978.

[175] A. K. Rajagopal and J. Callaway. Inhomogeneous electron gas. *Phys. Rev. B*, 7:1912–1919, 1973.

[176] Laura E Ratcliff, Stephan Mohr, Georg Huhs, Thierry Deutsch, Michel Masella, and Luigi Genovese. Challenges in large scale quantum mechanical calculations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 7(1):e1290, 2017.

[177] Markus Reiher and Alexander Wolf. *Relativistic Quantum Chemistry*. Wiley-VCH, second edition, 2015.

[178] Ryan D. Reynolds, Takeshi Yanai, and Toru Shiozaki. Large-scale Relativistic Complete Active Space Self-Consistent Field with Robust Convergence. *J. Chem. Phys.*, 149(1):014106, 2018.

[179] Walter Ritz. ber eine methode zur løsung gewisser variationsprobleme der mathematischen physik. *J. Reine Angew. Math.*, 135:1–61, 1909.

[180] Björn O Roos, Kerstin Andersson, Markus P Fülscher, Per-âke Malmqvist, Luis Serrano-Andrés, Kristin Pierloot, and Manuela Merchán. Multiconfigurational Perturbation Theory: Applications in Electronic Spectroscopy. *Adv. Chem. Phys.*, 93:219–331, 1996.

[181] Björn O. Roos, Peter R. Taylor, and Per E. M. Sigbahn. A Complete Active Space SCF Method (CASSCF) using a Density Matrix Formulated Super-CI Approach. *Chem. Phys.*, 48:157–173, 1980.

[182] Björn O. Roos. The Complete Active Space Self-Consistent Field Method and Its Applications in Electronic Structure Calculations. *Adv. Chem. Phys.*, 69:399–445, 1987.

[183] C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Mod. Phys.*, 23:69–89, Apr 1951.

[184] Erich Runge and Eberhard KU Gross. Density-functional theory for time-dependent systems. *Phys. Rev. Lett.*, 52(12):997, 1984.

[185] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

[186] J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. Cambridge University Press, 3 edition, 2020.

[187] Pawel Salek, Trygve Helgaker, and Trond Saue. Linear response at the 4-component relativistic density-functional level: Application to the frequency-dependent dipole polarizability of Hg, AuH and PtH_2. *Chem. Phys.*, 311:187–201, 2005.

[188] Trond Saue. Relativistic hamiltonians for chemistry: A primer. *ChemPhysChem*, 12:3077–3094, 2011.

[189] Trond Saue. Does chemistry need more physics? *Pure and Applied Chemistry*, 97(10):1255–1276, 2025.

[190] Trond Saue and Trygve Helgaker. Four-component relativistic kohn–sham theory. *J. Comput. Chem.*, 23:814–823, 2002.

[191] Jeffrey B. Schriber and Francesco A. Evangelista. Communication: An Adaptive Configuration Interaction Approach for Strongly Correlated Electrons with Tunable Accuracy. *J. Chem. Phys.*, 144(16):161106, April 2016.

[192] Edwin Schrödinger. Quantisierung als eigenwertproblem. *Annalen der Physik*, 384(4):361–376, 1926.

[193] Stefan Seritan, Christoph Bannwarth, Bryan S. Fales, Edward G. Hohenstein, Christine M. Isborn, Sara I. L. Kokkila-Schumacher, Xin Li, Fang Liu, Nathan Luehr, James W. Snyder Jr., Chenchen Song, Alexey V. Titov, Ivan S. Ufimtsev, Lee-Ping Wang, and Todd J. Martinez. TeraChem: A Graphical Processing Unit-Accelerated Electronic Structure Package for Large-Scale *Ab Initio* Molecular Dynamics. *WIREs Comput. Mol. Sci.*, 11(2):e1494, 2021.

[194] Prachi Sharma, Andrew J. Jenkins, Giovanni Scalmani, Michael J. Frisch, Donald G. Truhlar, Laura Gagliardi, and Xiaosong Li. Exact-two-component multiconfiguration pair-density functional theory. *J. Chem. Theory Comput.*, 18(5):2947–2954, 2022.

[195] Sandeep Sharma, Adam A. Holmes, Guillaume Jeanmairet, Ali Alavi, and C. J. Umrigar. Semistochastic Heat-Bath Configuration Interaction Method: Selected Configuration Interaction with Semistochastic Perturbation Theory. *J. Chem. Theory Comput.*, 13(4):1595–1604, April 2017.

[196] Isaiah Shavitt. The History and Evolution of Configuration Interaction. *Mol. Phys.*, 94(1):3–17, 1998.

[197] Agam Shayit, Can Liao, Shiv Upadhyay, Hang Hu, Tianyuan Zhang, A. Eugene DePrince III, Chao Yang, and Xiaosong Li. Numerically Exact Configuration Interaction at Quadrillion-Determinant Scale. *Nat. Commun.*, 16(1):11016, 2025.

[198] James Shee, Evan J. Arthur, Shiwei Zhang, David R. Reichman, and Richard A. Friesner. Phaseless Auxiliary-Field Quantum Monte Carlo on Graphical Processing Units. *J. Chem. Theory Comput.*, 14(8):4109–4121, 2018.

[199] C. David Sherrill and Henry F. Schaefer. The Configuration Interaction Method: Advances in Highly Correlated Approaches. *Adv. Quantum Chem.*, 34:143–269, 1999.

[200] Toru Shiozaki and Wataru Mizukami. Relativistic Internally Contracted Multireference Electron Correlation Methods. *J. Chem. Theory Comput.*, 11(10):4733–4739, 2015.

[201] Mark Srednicki. *Quantum Field Theory*. Cambridge University Press, 2007.

[202] John F. Stanton. *A Chemist's Guide to Density Functional Theory*. American Chemical Society, 2001.

[203] R Eric Stratmann, Gustavo E Scuseria, and Michael J Frisch. Achieving linear scaling in exchange-correlation density functional quadratures. *Chem. Phys. Lett.*, 257(3-4):213–223, 1996.

[204] S. Sun, T. F. Stetina, T. Zhang, and X. Li. volume 1388, chapter Chapter 10 - On the Finite Nuclear Effect and Gaussian Basis Set for Four-Component Dirac Hartree-Fock Calculations., pages 207–218. American Chemical Society, 2021.

[205] Bertha Swirles and Douglas Rayner Hartree. The relativistic self-consistent field. 152:625–649, 1935.

[206] Péter G. Szalay, Thomas Müller, Gergely Gidofalvi, Hans Lischka, and Ron Shepard. Multiconfiguration Self-Consistent Field and Multireference Configuration Interaction Methods and Applications. *Chem. Rev.*, 112:108–181, 2012.

[207] Andrew M Teale, Trygve Helgaker, Andreas Savin, Carlo Adamo, Bálint Aradi, Alexei V Arbuznikov, Paul W Ayers, Evert Jan Baerends, Vincenzo Barone, Patrizia

Calaminici, et al. Dft exchange: Sharing perspectives on the workhorse of quantum chemistry and materials science. *Phys. Chem. Chem. Phys.*, 24(47):28700–28781, 2022.

[208] Gergő Thiering and Adam Gali. Ab initio calculation of spin-orbit coupling for an nv center in diamond exhibiting dynamic jahn-teller effect. *Phys. Rev. B*, 96(8):081115, 2017.

[209] Jørn Thyssen, Timo Fleig, and Hans Jørgen Aa Jensen. A Direct Relativistic Four-component Multiconfiguration Self-Consistent-Field Method for Molecules. *J. Chem. Phys.*, 129(3):034109, 2008.

[210] Norm M. Tubman, Joonho Lee, Tyler Y. Takeshita, Martin Head-Gordon, and K. Birgitta Whaley. A Deterministic Alternative to the Full Configuration Interaction Quantum Monte Carlo Method. *J. Chem. Phys.*, 145(4):044112, July 2016.

[211] Andrew J. S. Valentine and Xiaosong Li. Intersystem crossings in late-row elements: A perspective. *J. Phys. Chem. Lett.*, 13(13):3039–3046, 2022.

[212] Marc W. van der Kamp and Jaida Begum. *QM / MM for Structure-Based Drug Design: Techniques and Applications*, chapter 6, pages 119–156. John Wiley Sons, Ltd, 2024.

[213] Tanja Van Mourik, Michael Bühl, and Marie-Pierre Gaigeot. Density functional theory across chemistry physics and biology, 2014.

[214] Christoph van Wüllen. Spin densities in two-component relativistic density functional calculations: Noncollinear versus collinear approach. *J. Comput. Chem.*, 23:779–785, 2002.

[215] Petr Čársky. Configuration Interaction. In P. v. R. Schleyer, editor, *Encyclopedia of Computational Chemistry*. John Wiley & Sons, Ltd, 2002.

[216] Bess Vlaisavljevich and Toru Shiozaki. Nuclear Energy Gradients for Internally Contracted Complex Active Space Second-Order Perturbation Theory: Multistate Extensions. *J. Chem. Theory Comput.*, 12:3781–3787, 2016.

[217] Konstantinos D. Vogiatzis, Giovanni Li Manni, Samuel J. Stoneburner, Dongxia Ma, and Laura Gagliardi. Systematic Expansion of Active Spaces beyond the CASSCF Limit: A GASSCF/SplitGAS Benchmark Study. *J. Chem. Theory Comput.*, 11:3010–3021, 2015.

[218] Konstantinos D. Vogiatzis, Dongxia Ma, Jeppe Olsen, Laura Gagliardi, and Wibe A. De Jong. Pushing configuration-interaction to the limit: Towards massively parallel mcscf calculations. *The Journal of Chemical Physics*, 147:184111, 2017.

[219] Shuang Wan, Shunshun Zhao, Weiting Ma, and Shimou Chen. Computational approaches to electrolyte design for advanced lithium-ion batteries. *Chem. Commun.*, 61:7019–7034, 2025.

[220] Hans-Joachim Werner and Peter J. Knowles. An efficient internally contracted multiconfiguration–reference configuration interaction method. *The Journal of Chemical Physics*, 89:5803–5814, 1988.

[221] Hans-Joachim Werner and Wilfried Meyer. A Quadratically Convergent MCSCF Method for the Simultaneous Optimization of Several States a Quadratically Convergent MCSCF Method for the Simultaneous Optimization of Several States. *J. Chem. Phys.*, 74:5794, 1981.

[222] Oskar Weser, Kai Guther, Khaldoon Ghanem, and Giovanni Li Manni. Stochastic

Generalized Active Space Self-Consistent Field: Theory and Application. *J. Chem. Theory Comput.*, 18:251–272, 2022.

[223] David B Williams-Young, Andrey Asadchev, Doru Thom Popovici, David Clark, Jonathan Waldrop, Theresa L Windus, Edward F Valeev, and Wibe A de Jong. Distributed memory gpu accelerated fock construction for hybrid gaussian basis density functional theory. *J. Chem. Phys.*, 158(23), 2023.

[224] David B Williams-Young, Abhishek Bagusetty, Wibe A de Jong, Douglas Doerfler, Hubertus JJ van Dam, Álvaro Vázquez-Mayagoitia, Theresa L Windus, and Chao Yang. Achieving performance portability in gaussian basis set density functional theory on accelerator based architectures in nwchemex. *Parallel Comput.*, 108:102829, 2021.

[225] David B Williams-Young, Wibe A De Jong, Hubertus JJ Van Dam, and Chao Yang. On the efficient evaluation of the exchange correlation potential on graphics processing unit clusters. *Front. Chem.*, 8:581058, 2020.

[226] David B. Williams-Young, Wibe A. de Jong, Hubertus J.J. van Dam, and Chao Yang. On the efficient evaluation of the exchange correlation potential on graphics processing unit clusters. *Front. Chem.*, 8:581058, 2020.

[227] David B Williams-Young, Alessio Petrone, Shichao Sun, Torin F Stetina, Patrick Lestrange, Chad E Hoyer, Daniel R Nascimento, Lauren Koulias, Andrew Wildman, Joseph Kasper, Joshua J. Goings, Feizhi Ding, A. Eugene DePrince III, Edward F. Valeev, and Xiaosong Li. The chronus quantum (chronusq) software package. *WIREs Comput. Mol. Sci.*, 10:e1436, 2020.

[228] Xuefei Xu, Ke R. Yang, and Donald G. Truhlar. Testing noncollinear spin-flip, collinear spin-flip, and conventional time-dependent density functional theory for pre-

dicting electronic excitation energies of closed-shell atoms. *J. Chem. Theory Comput.*, 10:2070–2084, 2014.

[229] Victor Wen-zhe Yu and Marco Govoni. GPU Acceleration of Large-Scale Full-Frequency GW Calculations. *J. Chem. Theory Comput.*, 18(8):4690–4707, 2022.

[230] Boyi Zhang, Jonathon E. Vandezande, Ryan D. Reynolds, and Henry F. Schaefer. Spin-Orbit Coupling via Four-Component Multireference Methods: Benchmarking on p-Block Elements and Tentative Recommendations. *J. Chem. Theory Comput.*, 14(3):1235–1246, 2018.

[231] Hongbin Zhang. High-throughput design of magnetic materials. *Electron. Struct.*, 3(3):033001, 2021.

[232] Ning Zhang, Wenjian Liu, and Mark R. Hoffmann. Iterative Configuration Interaction with Selection. *J. Chem. Theory Comput.*, 16(4):2296–2316, April 2020.

[233] Ning Zhang, Wenjian Liu, and Mark R. Hoffmann. Further Development of iCIPT2 for Strongly Correlated Electrons. *J. Chem. Theory Comput.*, 17(2):949–964, February 2021.

[234] Tianyuan Zhang, Samragni Banerjee, Lauren N. Koulias, Edward F. Valeev, A. Eugene III DePrince, and Xiaosong Li. Dirac–coulomb–breit molecular mean-field exact-two-component relativistic equation-of-motion coupled-cluster theory. *J. Phys. Chem. A*, 128(17):3408–3418, 2024.

[235] Yinggan Zhang, Zhou Cui, Baisheng Sa, Naihua Miao, Jian Zhou, and Zhimei Sun. Computational design of double transition metal mxenes with intrinsic magnetic properties. *Nanoscale Horiz.*, 7(3):276–287, 2022.