Extending the Capability of Classical Quantum Many-Body Methods

Thesis by
Yang Gao

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

Caltech

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended Dec 08, 2021

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor Garnet Chan. Throughout my time here at Caltech, Garnet has never ceased to provide tons of knowledge, ideas, and perspectives across a wide range of topics. He encouraged me to follow the topics that I genuinely feel passionate about, and none of this would have been possible without him.

I am also grateful to Austin Minnich, who advised my early years here. Austin brought me to the field of classical quantum simulation, and it has been great pleasure to discuss research with him over the years.

I would like to express my gratitude for the entire Chan group and Minnich group. A special thanks to Qiming Sun, Mario Motta, and Johnnie Gray. Getting started in a new research field is always hard, but these amazing people helped me navigate through these most difficult times.

It goes without saying that there are many friends from Caltech and beyond who have made my journey much more pleasant. To my ski buddies, tennis partners, board game friends, and so on, I appreciate all the fun times we shared and thank you all for helping me maintain my sanity.

Last and most importantly, I would like to thank my parents for their unconditional love and unwavering support over all these years. I owe them for everything I have.

ABSTRACT

This thesis discusses several topics in extending the capability of conventional quantum many-body methods. The first project focuses on extending quantum chemical methods, namely coupled cluster theory, to the correlated systems in the condensed phase. We consider bulk nickel oxide and manganese oxide, which are two paradigmatic correlated electron materials that pose challenges to traditional density functional theory-based simulation framework. We adapted molecular coupled cluster singles and doubles theory using Gaussian basis sets with translational symmetry and norm-conserving pseudopotential. This allowed us to carry a detailed study on the ground and excited states of the two materials.

The second project investigates numerical optimization techniques for Abelien group symmetric tensor contractions. In many-body quantum simulations, group symmetries in states and operators often lead to block sparse structure in the representing tensors. Exploiting this opportunity can significantly reduce the computation cost and memory footprint in tensor contractions. We consider cyclic group symmetry and introduce an efficient remapping scheme to express the sparse tensor contractions almost fully in terms of dense tensor operations.

The third project is devising a wavefunction-based method for coupled electrons and phonons. We are interested in simulating the interacting electrons and phonons at the same footing using coupled cluster methods. The ground state and excited state of two types of systems are investigated in this work: the Hubbard Holstein model and diamond crystal in ab initio setting.

Finally, the fourth project is to develop a generic framework for tensor network simulation on fermionic systems. Tensor network methods are powerful tools to study strongly correlated physical systems. However, traditionally these methods have been developed with commutative algebraic rules, which are commensurate with bosons but not compatible with anti-symmetric fermions. Our approach encodes the fermion statistics directly in the block sparse tensor backend so the tensors behave just like anti-commuting fermion operators.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Yang Gao et al. “Automatic transformation of irreducible representations for efficient contraction of tensors with cyclic group symmetry”. In: *arXiv Prepr. arXiv2007.08056* (2020).
Y. G. contributed to the design of the algorithm and the implementation of the numerical library. ISSN: 2331-8422. arXiv: 2007.08056. URL: http://arxiv.org/abs/2007.08056.

[2] Yang Gao et al. “Electronic structure of bulk manganese oxide and nickel oxide from coupled cluster theory”. In: *Phys. Rev. B* 101.16 (2020).
Y. G. contributed to the implementation of the algorithm, performed the simulation and results analysis. DOI: 10.1103/PhysRevB.101.165138.

[3] Qiming Sun et al. “Recent developments in the PySCF program package”. In: *J. Chem. Phys.* 153.2 (2020).
Y. G. contributed to the periodic boundary condition module within the PySCF package. DOI: 10.1063/5.0006074.

[4] Alec F. White et al. “A coupled cluster framework for electrons and phonons”. In: *J. Chem. Phys.* 153.22 (2020).
Y. G. contributed to the implementation of the theory and the ab initio Hamiltonian parameterization. ISSN: 10897690. DOI: 10.1063/5.0033132.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Chapter 1

INTRODUCTION

Computer programs are powerful tools to help physicists to obtain a deep understanding of how matters interact. However, despite the ever-growing computational power, exact solution of quantum many-body problems is in principle classically intractable. Therefore, to achieve an accurate description of the system at a reasonable computational cost, one must carefully balance efforts in many aspects: approximation within theory, numerical optimization, extrapolation towards continuum limit, to name a few. In this thesis, we consider four topics related to extending the capability of conventional quantum simulation methods. In this chapter we first describe our motivation followed by an overview for each topic that will be discussed in subsequent chapters. The detailed background for each topic will be elaborated in the corresponding chapter.

1.1 Quantum Many-Body Problem

Material and chemical properties are governed by the many-body Schrödinger equation which describes the quantum mechanical interaction of the constituent electrons under the the electric potential field from the nuclei. At the heart of what we do is to solve the time-independent Schrödinger equation for a given Hamiltonian H:

$$H|\Psi\rangle = E|\Psi\rangle. \tag{1.1}$$

In the context of non-relativistic electronic structure theory, the Hamiltonian can be expressed as:

$$\begin{aligned} H =& -\sum_i \frac{\hbar^2}{2m_i}\nabla_i^2 - \sum_A \frac{Z_A e^2}{|\vec{r}_i - \vec{r}_A|} + \sum_{i\neq j} \frac{e^2}{2|\vec{r}_i - \vec{r}_j|} \\ & - \sum_A \frac{\hbar^2}{2M_A}\nabla_A^2 + \sum_{A\neq B} \frac{Z_A Z_B e^2}{2|\vec{r}_A - \vec{r}_B|}, \end{aligned} \tag{1.2}$$

where m_i and M_A denote the mass of electron i and nucleus A respectively. $\vec{r}$ represents the spatial coordinate while $\hbar$ and e are the reduced Planck constant and the charge of electron.

To make the problem a bit tractable, one typically adopts the Bohr-Oppenheimer approximation to treat the electrons and nuclei separately. Due to the large difference between the mass of the electron and the nucleus, we assume that the full wavefunction $|\Psi(\{r\},\{R\})\rangle$ can be approximated by a product of the nucleus wavefunction $|\Psi(\{R\})\rangle$ and the electron wavefunction $|\Psi(\{r\})\rangle_{\{R\}}$ at the fixed nucleus coordinate. This so called “clamped-nuclei" approximation leads to the simplified electronic Schrödinger equation $H_e|\Psi(\{r\})\rangle_{\{R\}} = E_e|\Psi(\{r\})\rangle_{\{R\}}$, where

$$H_e = -\sum_i \frac{\hbar^2}{2m_i}\nabla_i^2 - \sum_A \frac{Z_A e^2}{|\vec{r}_i - \vec{r}_A|} + \sum_{i\neq j} \frac{e^2}{2|\vec{r}_i - \vec{r}_j|}. \tag{1.3}$$

The Bohr-Oppenheimer approximation hugely reduces computational complexity and is widely applicable to a broad spectrum of molecules and materials. Detailed discussion of its validity can be widely found in literature. We here note that the approaximation can break down when the lowest lying electronic states change very rapidly with nuclear position and approach each other in energy.

Nevertheless, solving the many-body Schrödinger equation on just the electronic part is by no means a simple task. The problem is further complicated by the Pauli exclusion principle, and numerical solution is a “non-polynomial hard" problem. Although the exact solution comes with exponential cost in system size, we can numerically solve these equations in an approximate manner if we are smart about it. In this regard, approximation must be treated with care at each stage of our simulation. Here we discussion some of the main considerations.

First, what kind of approximate theory should be adopted in our calculation? There is a vast array of numerical methods developed to approximately solve the quantum many-body problem, each with different traits in terms of representability, level of empiricism, computational cost, numerical stability, and so on. As of yet, there is no obvious winner that can consistently fit into any physical system of interest. Some of the key for deciding on the theory include: Are we looking for quantitative or qualitative answer? How much correlation effect is expected in the system of interest? Is the computational cost manageable with the available resource?

Once the approximate theory is decided, we need to further map the Schrödinger equation onto a discretized basis. The Hamiltonian can then be written in the second quantized form as

$$H = \sum_{pq} t_{pq} a_p^\dagger a_q + \sum_{pqrs} v_{pqrs} a_p^\dagger a_q^\dagger a_s a_r, \tag{1.4}$$

where t_{pq} and v_{pqrs} are the one-body and two-body Hamiltonian tensors, respectively. Our goal is to compute the wavefunction, or more often, the properties of interest in this setting. Computational solution typically amounts to linear algebra operations on the wavefunction tensor and the Hamiltonian tensors, which reflected on the computational cost. To this end, the development of computer hardware and high-performance linear algebra libraries has greatly aided modern electronic structure theory, making it possible to run simulations that were once deemed out of reach on just consumer laptops.

Going back to basis discretization, two different philosophies have been adopted for computational quantum many-body problem:

1. Basis size are systematically increased in an effort to obtain quantitative results in the complete basis set limit. This is typically adopted in ab initio quantum chemistry and computational materials community.
2. Only the minimal Hamiltonian terms that capture the essential physics are preserved. This is more often adopted in the quantum physics community where high-level numerical theory is used to obtain insight into the underlying mechanism.

The two philosophies are essentially two different strategies to leverage computational cost against problem complexity. As our computational power keeps growing, there is an increasing trend in the research community to blur the boundary between them.

As we have described above, developing many-body simulation tools is an intricate task that requires careful leverage over theory formulation, numerical setting, algorithm optimization, implementation, and so on. With the ever-growing computational power and electronic structure infrastructure, modern quantum many-body simulation is rapidly evolving as well. In the remaining part of this chapter, we will briefly introduce the opportunities we identified in this dynamics.

1.2 Precision Simulation for the Condensed Phase

In computational materials science, density functional theory has been the workhorse over the fast few decades. This framework maps the many-body problem into an auxiliary system of non-interacting electrons where each electron feels an average potential generated by exchange correlation functionals acting on the 3-dimensional electron density. The mean field approximation yields a set of single particle Kohn-Sham equation that can be solved self-consistently:

$$\begin{aligned} H_{KS}\phi_i(r) &= \epsilon_i\phi_i(r), && (1.5)\\ H_{KS} &= -\frac{\hbar^2}{2m}\nabla^2 + v_s(r), && (1.6)\\ v_s(r) &= v_{ext}(r) + \int \frac{n(r')}{|r-r'|} + v_{xc}(r), && (1.7) \end{aligned}$$

where v_s, v_{ext}, and v_{xc} represent the total effective potential, the external potential, and the exchange correlation potential. The electron denisty $n(r)$ can be computed by

$$n(r) = \sum_i \phi_i^*(r)\phi_i(r). \qquad (1.8)$$

By reducing the many-body problem to non-interacting problem, a significant cost reduction is achieved. However, although the theory is exact in nature, the exchange correlation functional that contains all the many-body effects can only be approximated, and there exists no systematic route to improve it. In fact, development of the exchange correlation functionals is very slow, and rarely without ambiguity. As a result, materials simulations can only provide a more qualitative and rough guide rather than definitive and quantitative solution.

So how do we move beyond this approach to achieve precision simulation for materials? Within the density functional framework, much effort has been devoted to develop methods based on low-order time-dependent perturbation theory including GW approximations and Bethe-Salpeter equations. These methods generally achieve better results in the weak to intermediate correlated region, but the success is not consistent, especially in the strongly correlated region, and it is unclear how to further improve these methods in a systematic manner.

Alternatively, we can consider a paradigm shift to the wavefunction-based approaches widely adopted in the quantum chemistry community. Rather than working on the electron density, a reduced quantity, quantum chemical methods generally aim to approximate the full wavefunction through expansion in the many-body bases, which are constructed as Slater determinants from a set of single-particle orbitals $\phi_i(x_i)$:

$$|\Phi_I\rangle = |\phi_0(x_0)\phi_1(x_1)\dots\phi_n(x_n)\rangle. \tag{1.9}$$

Here for simplicity, we use x to represent both the spatial coordinate and the spin part $\sigma(\omega) \in \{\alpha, \beta\}$.

At the Hartree-Fock (lowest) level, the ground state wavefunction is assumed to be a single Slater determinant ($|\Psi_0\rangle \approx |\Phi_0\rangle$). Based on the variational principle, we can arrive at the Hartree-Fock equations much similar to the Kohn-Sham equations:

$$\begin{aligned} h(x_1)\phi_i(x_1) + \sum_{j\neq i}\int & dx_2|\phi_j(x_2)|^2 r_{12}^{-1}\phi_i(x_1) \\ - \sum_{j\neq i}\int & dx_2\phi_j^*(x_2)\phi_i(x_2)r_{12}^{-1}\phi_j(x_1) = \epsilon_i\phi_i(x_1). \end{aligned} \tag{1.10}$$

One can immediately see that only Coulomb and exchange interactions are included in this mean field approximation. Nevertheless, solution of these equations yields a set of $\phi_i(x_i)$ ($2M$ in total where M is the size of spatial atomic basis we choose) that can be a good starting point to move towards higher level theories. In this context, the most straightforward way to improve is the so-called Configuration Interaction (CI) ansatz where the wavefunction is expanded with all the single determinants that can be generated from all $\phi_i(x_i)$,

$$|\Psi\rangle = |\Phi_0\rangle + \sum_{ia} C_i^a|\Phi_i^a\rangle + \sum_{i<j,a<b} C_{ij}^{ab}|\Phi_{ij}^{ab}\rangle + \dots \tag{1.11}$$

Here $|\Phi_i^a >$ refers to the singly excited determinant formed by replacing the orbital ϕ_i with orbital ϕ_a in the Hartree-Fock wavefunction Φ_0. Assuming a closed-shell system, the entire set of ϕ_i amounts to $\binom{M}{\frac{N}{2}}$ configurations where N is the number of electrons, which is prohibitively demanding for computational cost. Fortunately, not all Slater determinants are equally important to represent the low-lying states we are most interested in. In fact, we can often aggressively truncate the CI series and yet achieve an accurate approximation of the true ground state. In this regard, numerous methods have been developed with different strategies on how to truncate the Hilbert space and how to compute the reduced wavefunction coefficients.

These quantum chemical methods generally form a well-established hierarchical framework to compute properties and spectral functions with high accuracies. While these methods obviously hold great potential in accurate materials simulation, they mostly come with a higher-order polynomial scaling with respect to system size and have thus not been fully explored in the condensed phase.

When one thinks about porting the wavefunction-based quantum chemical methods to the condensed phase, multiple factors must be carefully examined. For instance, what type of correlation can be propertly handled by the theory? How much reference dependence is expected, and how does computational cost scale? We here highlight another important factor—size extensivity of the theory, which describes whether the energy grows linearly with respect to the number of particles. This is particularly important in the condensed phase, as a size-inextensive method predicts the correlation energy per unit cell would be zero in the limit of an infinite number of unit cells.

Upon balancing all the factors above, coupled cluster theory is arguably the most promising candidate for application in the solid state. The theory is size-extensive with manageable polynomial cost, and more importantly, the ansatz inherently offers a systematically improvable framework by tuning the level of excitation operators to include. In fact, in recent years, pioneering work in the field has suggested promising results from applying low-level coupled cluster theory to weakly correlated crystalline materials. More excitingly, the so-called equation of motion formalism can be further applied on top of the ground state coupled cluster wavefunction to extract information on excited states. We are interested in investigating the performance of coupled cluster in the correlated solid-state region. Our target is the first-row transition metal oxide, a class of materials that posed a significant challenge for conventional density functional framework.

1.3 Tensor Contraction with Symmetry Groups

As mentioned in Chapter 1.1, in numerical quantum many-body simulations, the Schrödinger equation is mapped onto a discretized basis, and tensors are ubiquitously used to represent states or operators under this basis. Formulation of quantum many-body theory generally consists of two parts, one on how to obtain the wavefunction coefficient and the other on how to compute expectation values of interest. In both cases, algebraic equations are formed between the wavefunction tensors and the Hamiltonian tensors, mostly in the form of tensor contractions between these quantities, thus making it one of the most important computational primitives in quantum many-body simulations.

Simply getting the computer program to perform these tensor contractions can be straightforward. An example for a naive implementation of matrix-matrix multiplication kernel is shown in Algorithm 1, and the algorithm can be easily extended to arbitrary tensor contraction.

Algorithm 1 Loop nest to perform matrix multiplication $C_{mn} = \sum_k A_{mk} B_{kn}$

```
for $m = 1, \ldots, M$ do
    for $n = 1, \ldots, N$ do
        for $k = 1, \ldots, K$ do
            $C[m, n] + = A[m, k] B[k, n]$
        end for
    end for
end for
```

However, such implementation is extremely inefficient in practice. Fortunately, computer hardware vendors and the open-source research community have spent decades developing highly optimized math libraries. For instance, level 3 BLAS routines offer efficient kernels for various matrix-matrix operations including general matrix mulitplication (GEMM). These libraries are implemented with sophisticated optimization in cache usage, memory access pattern, vectorized instruction, parallelization, and so on.

Although there currently exists no unified application programming interface (API) among vendors and the open-source community for tensor-level operations, tensor contraction can be viewed as a high-dimensional generalization of matrix-matrix or matrix-vector multiplication. Therefore, a widely adopted approach to perform tensor contraction is to transpose the inputs into proper matrix form followed by a GEMM call and transposition on the output.

Besides taking advantage of highly optimized math libraries, in the context of quantum many-body simulation, there exist more opportunities to boost the performance of our implementation. Here we highlight the effect of the symmetry group inherent to the system, which means that the states or operators are often invariant under certain symmetry transformations. Such a symmetry group translates into the representing tensors, implying a block-sparse structure where a significant portion of the tensors is zero and makes no contribution to the final outcome. Therefore, this block-sparse structure can be exploited to reduce the computational cost and memory footprint.

Taking matrix multiplication in Algorithm 1 as an example, in the presence of block sparsity, only a small subset of $\{m\}$, $\{n\}$, and $\{k\}$ needs to be stored and iterated over explicitly in the multiplication operation. A widely adopted approach in a variety of state-of-the-art tensor contraction software libraries to exploit this opportunity is to either iterate over blocks or adopt general block-sparse tensor format. In principle, these strategies can reduce the computation scaling and memory footprint of nontrivial contractions (any contraction with a cost that is superlinear in input/output size) by a factor of $O(G^2)$ and $O(G)$ respectively, where G is the size of the symmetry group. However in practice, such a claim comes with caveats: unlike dense tensor operations where coalesced memory access pattern and better cache usage can often be exploited, the handling of block sparsity introduces additional code complexity and overhead for managing and scheduling (potentially small) blocks. Meanwhile, these blocks may not be contiguous in memory, thus leading to a performance drop which is often highly sensitive to the size of the blocks and the number of blocks to contract. Therefore, the treatment of block sparsity must be taken with care in order to actually speed up our simulation.

We are interested in a special scenario where the contraction is constrained by a cyclic group and the block sizes are equal for all symmetry sectors. This structure is commonly identified in systems with cyclic point group symmetry, solid-state crystals with translation invariance for example. As a concrete example, efficient handling of this symmetry can lead to a computational saving of N_k^2 for our coupled cluster calculation in materials where N_k is the number of sampling in the Brillouin Zone. We here propose a technique, irreducible representation alignment, to efficiently handle this block-sparse contraction by using contraction-specific compressed forms. As a result, our algorithm amounts to only dense tensor operations and shows high efficiency and parallel scalability.

1.4 Beyond Bohr-Oppenheimer Approximation

We mentioned in Chapter 1.1 that the full Schödinger equation is often decoupled into the pure electronic part and nucleus part using the Bohr-Oppenheimer approximation. The electronic Schödinger equation assumes the nuclei to be located at fixed locations, which is in principle not true, even at $0K$. Therefore, the approximation can break down when the electron nucleus coupling becomes large and no longer negligible. In fact, even away from the strong coupling region, the coupling can give rise to a wide range of phenomena in the condensed phase, for instance, “unexpected" phase transition, the temperature dependence of electronic transport, and optical properties. In the strong coupling region, they are the key interaction underpinning the Bardeen-Cooper-Schrieffer type of superconductivity, which has attracted significant research interest.

Therefore it is crucial to address the coupling between the electronic and nuclear degree of freedom. In the solid state, crystal vibrations are typically characterized by phonons which represent the collective excitations of the underlying lattice. Traditionally the computational study of electron-phonon interaction can be mostly divided into two categories:

1. Simplified lattice models with semi-empirical Hamiltonians are constructed and then solved nearly exactly using high-level theory to capture the essential physics.
2. Ab initio Hamiltonians that model realistic materials are first extracted, typically with the use of density functional theory framework on a fine grid, and then treated with more simplified theory due to the complexity with large system size.

Again, this is another manifestation of circumstances where we have to leverage the level of many-body theory against the problem size. In this context, method (1) generally treats the electronic and nuclear degree of freedom at the same footing in an effort to obtain qualitative answers. The results provide insightful interpretation of the underlying mechanism but lack predictive power for realistic materials. On the other side of the spectrum, method (2) often treats the coupling as a perturbative term to the electronic problem as a computational compromise. Consequently, the simulation can be performed at a large basis setting, yielding good quantitative results, especially for systems with lower level of correlation. However, the success does not translate to systems with a complex interplay of electron-electron correlation and electron-phonon coupling.

Our goal is to bridge the gap between the two different paradigms by devising a cost-efficient theory that treats the interacting problem at a correlated level. Our starting point is the coupled cluster theory which has long been one of the most reliable computational methods in electronic structure theory. We demonstrate that given a Hamiltonian with electron phonon coupling, we can develop a systematically-improvable framework to describe the coupled system by combining the electronic and vibrational coupled cluster theory.

1.5 Tensor Network Methods for Fermion Simulation

In previous sections of the chapter, our attention has been focused on quantum chemical wavefunction methods where the Hilbert space is explicitly truncated at the ansatz level. Commensurate to the development of these numerical tools is a set of diagrammatic tools to describe different types of interactions, for example, Feymann diagrams and Goldstone diagrams. Although these diagrams are constructed in a systematic manner with an exact mathematical form, they are not intuitive in providing information on the entanglement structure of the system. Looking beyond, a few questions naturally arise:

1. What are the other ways to efficiently represent the low-lying states?
2. Nature is always local and entanglement is often structured. Can we use that information to refine the wavefunction ansatz for better representability?

To address these problems, tensor network theory has recently emerged as a novel mathematical language to describe quantum many-body systems. The theory amounts to breaking up the huge wavefunction coefficients into smaller tensors that are connected to each other based on certain geometry. The geometry can be chosen to reflect the entanglement structure of the system. We provide here a concrete example based on a four-site system in a 1-dimension (1D) geometry. The tensor network states can be expressed as Equation 1.12 while the diagrammatic representation is provided in Figure 1.1:

$$\begin{aligned}|\Psi\rangle &= \sum_{i_0 i_1 i_2 i_3} C_{i_0 i_1 i_2 i_3} |i_0\rangle|i_1\rangle|i_2\rangle|i_3\rangle \\ &\approx \sum_{i_0 i_1 i_2 i_3} \sum_{jkl} A^0_{j i_0} A^1_{jk i_1} A^2_{kl i_2} A^3_{l i_3} |i_0\rangle|i_1\rangle|i_2\rangle|i_3\rangle,\end{aligned} \tag{1.12}$$

where we use vertices and lines to represent tensors and indices respectively. The size of the shared index j, k, and l is termed as the bond dimension and can be viewed as a single parameter to control the level of approximation in the theory. We can immediately see that tensor network theory is systematically improvable in that

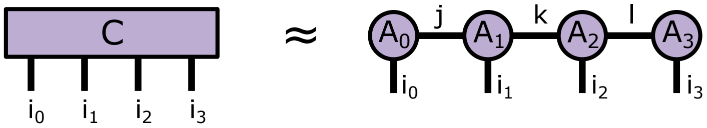

Figure 1.1: 1D tensor network states representation for a four-site quantum system.

.

we can tune both the geometry and the bond dimension to control the truncation on the entire Hilbert space. Meanwhile, we can clearly identify a simplified graphical framework to understand and classify the state of matter for complex systems.

The field of tensor network kicked off with Steve White's invention of density matrix renormalization group (DMRG) in 1992, and has been undergoing extremely rapid developments in the last two decades.

At the core of DMRG's success is the underlying matrix product states, which is just the simplest variant of tensor network states as shown in Figure 1.1. Despite such simplicity, the density matrix renormalizaton group has dominated computational study of one dimensional (1D) lattice models including a 1D Hubbard model and Heisenberg model, yielding extremely accurate ground-state properties.

However, the remarkable success of DMRG is not universal: matrix product states are only best suited for ground states of gapped 1D Hamiltonians due to the 1D entanglement entropy area law. Unfortunately, the essential physics of many interesting quantum many-body problems is often beyond one dimension, the superconductivity manifested in 2D Hubbard model for example. It is thus crucial to move beyond the 1D matrix product states and explore the capabilities of tensor network states with more sophisticated geometries.

However, porting the success of DMRG to higher dimensional fermionic systems is by no means a straightforward translation. The increase in dimension and geometric complexity entails numerous challenges that have seeded a plethora of extensive research. Following is a list of challenges that we will discuss here:

1. How to compute the individual tensors to optimize the wavefunction?
2. How to efficiently contract tensor network graphs in arbitrary geometry?
3. How to efficiently represent the anti-symmetric wavefunction with minimal bond dimension?

For question (1), there are generally two types of methods to compute the tensors, one by variational optimization such as DMRG and the other through time-evolution block-decimation (TEBD), an approximate version of imaginary time evolution. In both cases, one may inevitably need to contract the entire graph, which is complicated by itself in two regards. On the one hand, exact contraction on an arbitrary graph generally requires exponential resource. On the other hand, the exponential resource can potentially be lowered by a huge prefactor with an optimal contraction path, but searching the optimal contraction path itself is a “non-polynomial hard" problem. Therefore, such computational cost constraint means that one may have to use certain approximation in optimizing tensors and contracting the network.

Finally, the central topic that will be discussed later in the thesis is related to question (3) on how to efficiently account for fermion statistics in tensor network theory. For 1D MPS, the anti-commuting fermion statistics is mostly avoided by mapping the fermion operators to hard-core boson operators with Pauli strings. Beyond 1D, such transformation can no longer preserve the locality of the Hamiltonian and a “fictitious" long-range interaction could be introduced into the system, resulting in an unnecessarily higher requirement on the bond dimension. At high dimension, the computational cost of contracting the graph scales much higher than the matrix product states and such an increase could easily make the cost unmanageable.

We are interested in achieving the most efficient representation of fermion wavefunction on tensor network with arbitrary geometry. Our strategy is to render the numerical tensor library so that the fermionic statistics are explicitly accounted for in the backend. This approach allows us to directly inherit much of the pre-existing tensor network infrastructure. Using all the machinery we have built, we perform a benchmark study on various types of Hubbard models to evaluate the strength of our method and the accuracy of our approximate contraction schemes.

Chapter 2

ELECTRONIC STRUCTURE OF CRYSTALLINE MATERIALS FROM COUPLED CLUSTER THEORY

2.1 Abstract

We present the ground- and excited-state electronic structure of two prototypical transition metal oxides, MnO and NiO using coupled cluster in its sinlges and doubles approximation (CCSD and EOM-CCSD). Since the ground states are magnetically ordered, we use the spin unrestricted CC formalism. Fundamental gaps of MnO and NiO are determined to be 3.46eV and 4.83eV respectively based on a 16-unit supercell simulation. Amid finite-size error from coarse Brillouin zone sampling, our results show clear improvement compared with standard mean field methods. Additionally, our wavefunction representation allows for a detailed analysis of the charge-transfer/Mott-insulating character and atomic nature of the electronic bands of the two materials.

The work in this chapter is presented in the paper [1].

2.2 Introduction

Solids with correlated electrons pose a long-standing challenge in modern condensed matter physics. One of the prominent examples is the first-row transition metal oxides such as MnO and NiO. While the partially filled *d* band suggest metallicity in these materials, experimentally they turned out to be insulators with large gaps [2–4]. Two important explanations shown in Figure 2.1 have been developed to account for this discrepancy: the first one was proposed by Mott arguing that large electron repulsion could prohibit conduction, forming a so-called Mott insulator [5]; the second one gained insights from correlating model cluster calculations and experimental spectra [2, 6] to highlight the effect of the ligand-to-metal charge transfer process. Since then, the electronic characters of these transition metal oxides have been a fertile topic of study.

In principle, these questions could be unambiguously resolved through accurate first-principles calculation on the bulk material. However, achieving quantitative accuracy in computing properties for transition metal oxides has been difficult. For example, local and gradient density functional theories (DFT) typically underesti-

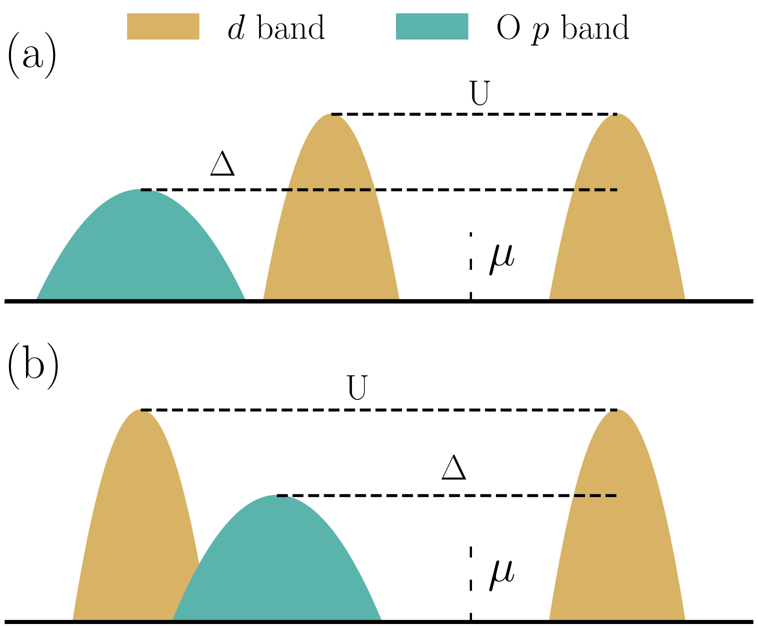

Figure 2.1: Schematic diagrams of the insulating mechanism for (a) Mott-Hubbard type where the charge transfer energy Δ is larger than the on-site Coulomb repulsion U and (b) charge-transfer type with $\Delta < U$. Here μ denotes the chemical potential.

mate both the insulating gap and order parameters, such as the magnetic moment [7]. While hybrid functionals can give better gaps, this success does not always translate to better properties and the performance is not consistent across materials [8–10]. Quantum Monte Carlo methods can provide higher accuracy at greater cost [11, 12], but do not allow access to the full spectrum. Low-order diagrammatic approaches such as the GW approximation [13–16] have also been applied to these systems, with mixed success. Finally, while DFT with a Hubbard U (DFT+U) [17–19] and dynamical mean-field theory (DMFT) calculations [20–28] have provided a practical approach to obtain important insights, these methods contain a degree of empiricism that introduces uncertainty into the interpretations.

Coupled cluster (CC) theory is a theoretical framework originating in quantum chemistry and nuclear physics [29, 30], which has recently emerged as a new way to treat electronic structure in solids at the many-body level [31, 32]. The method is systematically improvable in terms of particle-hole excitation levels, giving rise to the coupled cluster with singles, doubles, triples and higher approximations. While the earliest formulation was for ground states, excited states can be computed via the equation of motion (EOM) formalism [29, 33–35]. Recent single-particle spectra computed for the electron gas [36], and simple covalent solids [31] demonstrate that high accuracies can be achieved at the level of coupled cluster singles and doubles (CCSD). Note that at the ground-state CCSD level, the coupled cluster energy includes all ladders and ring diagrams, some of the couplings between the two, as well as partially self-consistently renormalized propagators. Thus compared to approximate GW methods, the CCSD equations are less sensitive to the single-particle starting point, while the inclusion of ladders (which are entirely omitted in GW) provides for some ability to treat stronger correlations. A detailed comparison of the diagrammatic content of GW and excited-state EOM-CCSD can be found in [37].

The rest of the chapter is organized as follows. In Section 2.3 we formulate the ground-state coupled cluster theory and equation of motion (EOM) ansatz for excited states in periodic systems. In Section 2.4 we present the CCSD results on NiO and MnO where a detailed analysis on the numerical convergence, ground-state properties, as well as the nature of the insulating states is provided. We conclude with Section 2.5.

2.3 Theory

Coupled Cluster Theory for Fermions

In this chapter we will use $a^\dagger$ and a to represent fermion creation and annihilation operators respectively.

The coupled cluster wavefunction is parameterized in an exponential fashion

$$|\Psi_{CC}\rangle = e^{T}|\Phi_0\rangle, \tag{2.1}$$

where $|\Phi_0\rangle$ is a single determinant reference. The T-operator here is defined in some space of excited configurations such that

$$T = \sum_{ia} t_i^a a_a^\dagger a_i + \frac{1}{4}\sum_{ijab} t_{ij}^{ab} a_a^\dagger a_b^\dagger a_j a_i + \dots \tag{2.2}$$

where i and a index occupied (hole) and virtual (particle) spin orbitals respectively.

Generally, the T-operator is truncated at some finite excitation level. For example, letting $T = T_1 + T_2$ yields the coupled cluster singles and doubles (CCSD) approximation. The coupled cluster energy and amplitudes are then determined from a projected Schrödinger equation:

$$\langle \Phi_0 | e^{-T} H e^{T} | \Phi_0 \rangle = E_{\mathrm{HF}} + E_{\mathrm{CC}} \tag{2.3}$$

$$\langle \Phi_\mu | e^{-T} H e^{T} | \Phi_0 \rangle = 0. \tag{2.4}$$

Here Φ_μ denotes a single determinant with particle-hole excitations, and $e^{-T}He^{T}$ is commonly referred to as the similarity transformed Hamiltonian $\bar{H}$.

The most computationally expensive step in CCSD formally scales as $O(N_{occ}^2 N_{vir}^4)$ where N_{occ} and N_{vir} are the size of occupied orbitals and virtual orbitals respectively.

Equation of Motion Coupled Cluster

Excited states can be computed within the EOM formalism which parameterizes a neutral or charged excitation by applying an excitation operator to the CC ground-state:

$$|\Psi_{ex}\rangle = R|\Psi_{\mathrm{CC}}\rangle = Re^{T}|\Phi_0\rangle. \tag{2.5}$$

Because the excitation operator, R, commutes with the excitation operators in T, solving this eigenvalue problem is equivalent to finding the right eigenvector of the similarity transformed Hamiltonian:

$$\langle \mu | \bar{H} R^n | \Phi_0, 0 \rangle = E_n R^n_\mu. \tag{2.6}$$

Here, E_n is the energy of the nth excited state, and μ indexes an element of the excitation operator R.

The excitation operator, R, can be constructed to access charged or neutral excitations:

$$R_{\mathrm{IP}} = \sum_i r_i a_i + \frac{1}{2}\sum_{ija} r^a_{ij} a^\dagger_a a_i a_j + \dots \tag{2.7}$$

$$R_{\mathrm{EA}} = \sum_a r^a a^\dagger_a + \frac{1}{2}\sum_{iab} r^{ab}_i a^\dagger_a a^\dagger_b a_i + \dots \tag{2.8}$$

$$R_{\mathrm{EE}} = \sum_{ia} r^a_i a^\dagger_a a_i + \frac{1}{4}\sum_{ijab} r^{ab}_{ij} a^\dagger_a a^\dagger_b a_j a_i + \dots \tag{2.9}$$

For the excited state calculations in this chapter, we will use R_{IP} and R_{EA} in conjunction with CCSD wavefunction to compute the ionized states (IP-EOM-CCSD) and electron attached states (EA-EOM-CCSD). The excitation space for the ionized states is restricted to the space of 1-hole (1h) and 2-hole, 1-particle (2h1p) states while the electron attached states lie in the space of 1-particle (1p) and 2-particle, 1-hole (2p1h) states. Note that although the ground-state CCSD theory exhibits a computational scaling of N^6, the subsequent IP-EOM-CCSD and EA-EOM-CCSD calculations come with a reduced scaling of N^5.

Extension to Periodic Systems

For periodic systems, we choose an underlying single-particle basis of crystalline Gaussian-based atomic orbitals (AOs) for compact representation of the Hilbert space. These are translational-symmetry-adapted linear combinations of Gaussian AOs of the form

$$\phi_{\mu,k}(r) = \sum_{T} e^{ik\cdot T} \tilde{\phi}_{\mu}(r - T), \tag{2.10}$$

where k is a crystal momentum vector in the first Brillouin zone and T is a lattice translation vector. In presence of translational symmetry, each AO and molecular orbital (MO) carries an addition label for the crystal momentum, and all quantities must conserve crystal momentum. For instance, the two electron integrals, which are defined (per unit cell Ω) as

$$(pk_p qk_q | rk_r sk_s) = \int_{\Omega} dr_1 \int dr_2 \phi^*_{pk_p}(r_1)\phi_{qk_q}(r_1) v_{12} \phi^*_{rk_r}(r_2)\phi_{sk_s}(r_2), \tag{2.11}$$

are only non-zero if $k_p + k_r - k_q - k_s = G$ where G is a reciprocal lattice vector. Since particle and holes states now all carry a net crystal momentum, adaptations are also required at CCSD and EOM-CCSD level:

$$\begin{aligned}
t_1 &= \sideset{}{'}\sum_{k_i} \sum_{ia} t^{ak_a}_{ik_i} a^{\dagger}_{ak_a} a_{ik_i}, && (2.12)\\
t_2 &= \frac{1}{4} \sideset{}{'}\sum_{k_i k_j k_a k_b} \sum_{ijab} t^{ak_a bk_b}_{ik_i jk_j} a^{\dagger}_{ak_a} a^{\dagger}_{bk_b} a_{jk_j} a_{ik_i}, && (2.13)\\
R^k_{IP} &= \sum_{i} r_{ik} a_{ik} + \frac{1}{2} \sideset{}{'}\sum_{k_a k_i k_j} \sum_{aij} r^{ak_a}_{ik_i jk_j} a^{\dagger}_{ak_a} a_{ik_i} a_{jk_j}, && (2.14)\\
R^k_{EA} &= \sum_{a} r^{ak} a^{\dagger}_{ak} + \frac{1}{2} \sideset{}{'}\sum_{k_a k_b k_i} \sum_{abi} r^{ak_a bk_b}_{ik_i} a^{\dagger}_{ak_a} a^{\dagger}_{bk_b} a_{ik_i}. && (2.15)
\end{aligned}$$

Here the primed sum indicates crystal momentum conservation. For t_2, this amounts to $k_a + k_b - k_i - k_j = G$, while for $\hat{R}^k_{IP}$ and $\hat{R}^k_{EA}$, the primed sum suggests a net change of momentum k, ie., $k_i + k_j - k_a = k + G$ for IP and $k_a + k_b - k_i = k + G$ for EA. With the introduction of crystal momentum, the formal computational scaling of periodic CCSD and EOM-CCSD are increased by a prefactor of N_k^4 and N_k^3 respectively where N_k is the number of k-point sampling in the first Brillouin zone. For detailed equations of our periodic unrestricted equation-of-motion CC, readers are encouraged to refer to [1].

2.4 Results

Computational Details

NiO and MnO both crystallize in a rocksalt structure with alternating ferromagnetic (111) planes stacked along the [111] direction. To host this antiferromagnetic (AFM) order, our calculations used a rhombohedral supercell with two units of XO (X=Mn or Ni). The lattice constants are taken to be the experimental values at 300 K , i.e. a = 4.43 Å and a = 4.17 Å for MnO and NiO respectively [38].

All of our methods are implemented within, and calculations performed using the PySCF package [39, 40]. In our calculations, GTH pseudopotential and corresponding single-particle basis [41] from the CP2K package [42] are used. In order to assess basis set convergence, we first performed a set of calcuations using GTH-SZV/DZVP/TZVP-MOLOPT(-SR) (SZV/DZVP/TZVP for short) for the metal and oxygen respectively [41]. For all other calculations, we used DZVP basis, which amounts to 78 orbitals per rhombohedral unit cell. Electron repulsion integrals were generated by periodic Gaussian density fitting with an even-tempered Gaussian auxiliary basis [43] and our initial mean field reference for CC calculations were generated from unrestricted Hartree-Fock calculations. The CC reduced density matrices are approximately computed using the right eigenvector of $\bar{H}$ [44] for subsequent observable calculations. Population analysis in the crystalline intrinsic atomic orbital basis [45, 46] is performed for atomic character analysis and local magnetic moments calculation.

Convergence

In order for calculations to carry predictive power, one must try to converge the simulations towards complete basis set limit (CBS) and the thermodynamic limit (TDL). However, due to the steep scaling of CC calculations, it is not possible to compute properties at a fully converged setting. We thus first assess the convergence of the theory.

We first focus on the basis set effect by computing the CC total energies, local magnetic moments on the metal and single-particle gaps as a function of increasing basis size for a 1x1x1 rhombohedral cell. The results are summarized in Table 2.1. Note the single-particle gaps here include both the direct gap at Γ and a gap for an indirect transition from $\Lambda_{\frac{1}{2}}(Z)$ (mid-point of the Λ symmetry direction of the primitive cell, equivalent to the Z high-symmetry point of the rhombohedral cell, see Figure 2.5) to Γ [47]. This transition is presumed to be where the fundamental gap is from and we will refer the fundamental gap to this specific transition throughout the rest of the chapter.

System	Basis	E_{CC}/eV	μ_B	Δ_{Γ}^{cc}/eV	Δ_{ind}^{cc}/eV
MnO	SZV	-2.66	4.29	0.36	1.04
	DZVP	-12.16	4.61	2.49	1.48
	TZVP	-14.23	4.61	2.40	1.42
NiO	SZV	-3.36	0.46	2.49	2.13
	DZVP	-13.40	1.18	3.22	2.62
	TZVP	-15.68	1.19	3.21	2.49

Table 2.1: Basis set convergence of CCSD total energy, local magnetic moment on metal, direct Γ gap Δ_{Γ} and indirect fundamental gap Δ_{ind} for a 1x1x1 cell.

From Table 2.1, we found that as basis size increases from SZV to DZVP to TZVP, the magnetic moment is already well converged at DZVP level, while the CC energies still changes drastically, as expected. We also find the single-particle direct gaps Δ_{Γ} to be well converged while the indirect fundamental gaps Δ_{ind} are slightly less converged with a change by more than 0.1 eV in NiO moving from DZVP to TZVP.

While the remaining basis error is estimated to be of several tenths of an eV, we will use DZVP basis for all the remaining calculations due to the computational cost.

We then turn our focus onto finite size error. The same quantities presented in Table 2.1 are shown in Table 2.2 for a 2x2x2 supercell. Note here that for the magnetic moments calculations, we used twist average [48, 49] technique with another 2x2x2 grid to achieve an effective 4x4x4 sampling of the Brillouin zone.

Compared with the 1x1x1 cell, we found significant change in both the magnetic moments and gaps. Notably, although from Table 2.1 we found the basis error converging the gap from above, the (larger) finite size error here converges the gap from below. To account for the finite-size scaling of the fundamental gap, we performed a rough extrapolation of the CC and UHF gaps assuming a $N_k^{-\frac{1}{3}}$ scaling. The results are shown in Figure 2.2. When extrapolated to TDL, the CC gaps are increased by $\sim$ 2 eV. Again, if we further take the basis error into account, the converged EOM-CCSD gaps at TDL are estimated to be 1-2 eV larger than the 2x2x2 results reported here.

System	Property	UHF	PBE	CCSD	exp
MnO	μ_B	4.86	4.56	4.76	4.58, 4.79
	Δ_{ind}/eV	8.05(12.09)	1.09(1.21)	3.46(5.44)	3.6-3.9
	Δ_Γ/eV	8.72(13.05)	1.77(1.84)	4.26(5.91)	-
NiO	μ_B	1.85	1.34	1.72	1.77, 1.90
	Δ_{ind}/eV	9.51(13.95)	1.19(1.38)	4.83(7.04)	4.3
	Δ_Γ/eV	9.89(14.80)	2.45(2.62)	5.56(7.90)	-

Table 2.2: Local magnetic moment, fundamental gap, and direct Γ gap from UHF, PBE, and CCSD with a 2x2x2 k-point mesh (DZVP basis). Extrapolated TDL gap is listed in parentheses. Experimental gaps and moments are also reported (see main text for a discussion of the comparison). The experimental magnetic moments are taken from Refs. [50] and [38]. The measured experimental gaps are taken from Refs. [3] and [2] for MnO and NiO respectively.

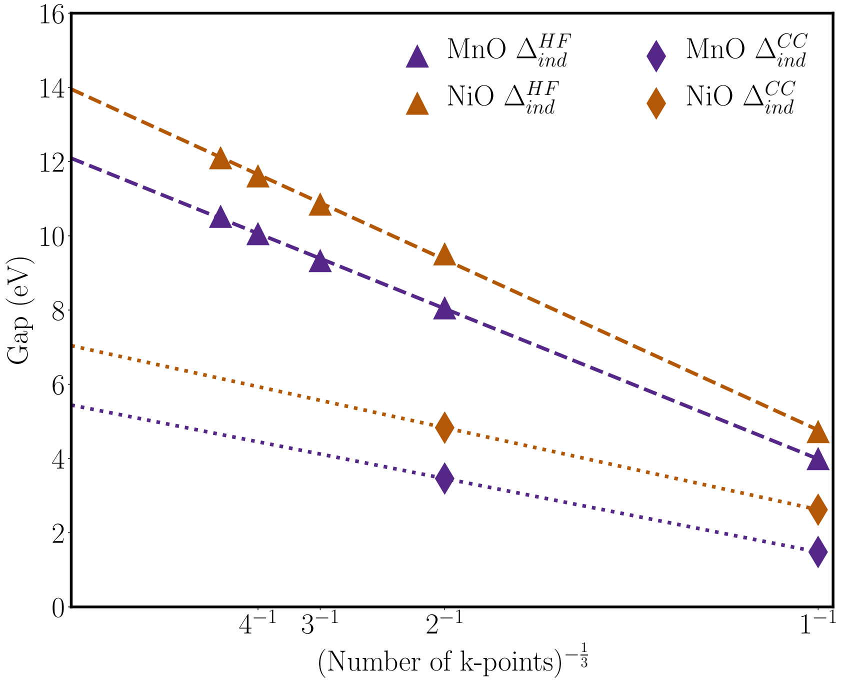

Figure 2.2: Band gap extrapolation for MnO and NiO. The purple and brown triangles denote the UHF indirect gap for MnO and NiO respectively. The purple and brown diamonds denote the CC indirect gap. The dashed lines and dotted lines give the linear extrapolation to the TDL for HF and CC respectively.

Ground-State Properties

We now present a more detailed analysis of the ground-state CCSD wavefunctions for NiO and MnO.

The CC ground-state moments reported in Table 2.2 are significantly reduced from those of the UHF reference. This is consistent with the well-known observation that Hartree-Fock tends to overestimate spin polarization. Conversely, PBE severely underpolarizes in NiO. Note that theoretical results for the magnetic moment have some variation depending on the definition of the atomic decomposition, while the experimental error bars are themselves relatively large, approximately 0.2 μ_B [51]. Therefore the direct comparison between theory and experiment for this quantity should be taken with a degree of caution.

Figure 2.3 shows the spin density distribution of the two materials in the (100) surface. For MnO, an isotropic spin density is observed around the metal site, reflecting all $3d$ orbitals partially occupied. However, for NiO we find a clear e_g symmetry pattern around the Ni atom. Meanwhile, a weakly induced spin density is also observed around the ligand oxygen site. Note that the O $2p$ spin density is aligned in the [110] instead of [100] direction, thus allowing maximal superexchange between the nearest Ni sites.

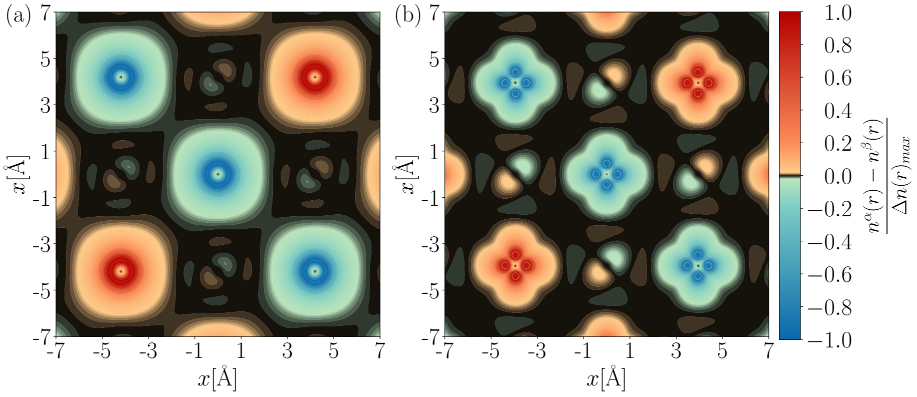

Figure 2.3: Normalized spin density on the (100) surface for (a) MnO and (b) NiO. The transition metal atom is located at (0, 0) in the xy-plane.

To further analyze the ground-state correlation, we computed the T_1, $|t_1|_{max}$ and $|t_2|_{max}$ diagnostics for the CCSD wavefunction. The results are shown in Figure 2.4.

The T_1 metric is the Frobenius norm (normalized by the number of correlated electrons) of the t_1 amplitudes. Previous studies have suggested that values of these diagnostics larger than ~ 0.1 can be considered "large" [52, 53]. The T_1 and $|t_1|_{max}$ metrics measure the importance of orbital relaxation from the mean-field reference while $|t_2|_{max}$ measures the true many-particle correlations. As seen from Figure 2.4, the effect of orbital relaxation is greater in NiO than in MnO, consistent with the greater degree of overpolarization of the Ni moment in the starting HF reference, than is seen for Mn. The small $|t_2|_{max}$ values (0.009 for MnO and 0.013 for NiO) however, indicate that both materials are reasonably described by the broken-symmetry mean-field reference.

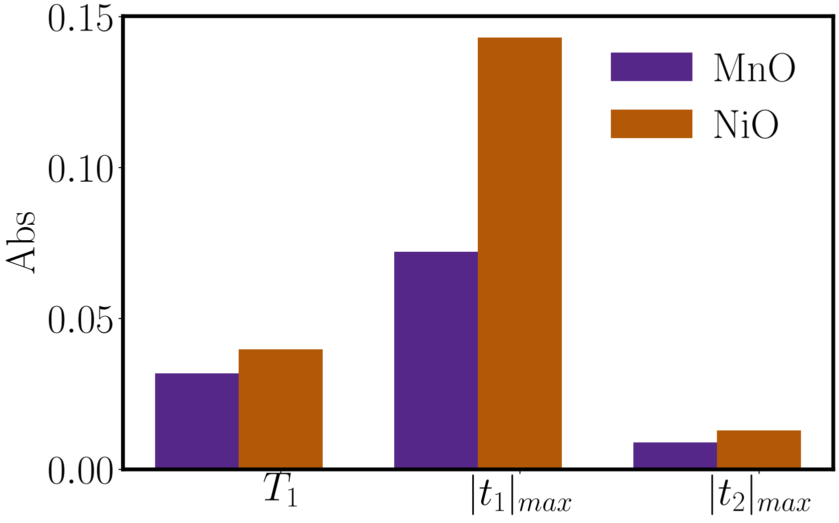

Figure 2.4: CCSD amplitude diagnostics for MnO and NiO. Purple columns are for MnO and brown are for NiO. T_1 is the Frobenius norm of the t_1 amplitudes normalized by the number of correlated electrons. $|t_1|_{max}$ and $|t_2|_{max}$ are the maximum absolute value for t_1 and t_2, respectively.

Charged Excitations

We next turn to discussion on the excited states from EOM-CCSD.

From Table 2.2 we see that the fundamental gaps obtained by PBE and UHF for MnO are 1.09 eV and 8.05 eV, respectively, both far from the experimental estimate of 3.6–3.9 eV [3]. In contrast, EOM-CCSD with a 2x2x2 supercell finds the indirect gap to be 3.46 eV. This is similar to the 3.5 eV gap found in prior quasiparticle self-consistent GW (QPscGW) calculations by Faleev and co-workers [13]. In NiO we observe an indirect gap of 9.51 eV, 1.19 eV, and 4.83 eV with HF, PBE, and EOM-CCSD (2x2x2 supercell) respectively. The EOM-CCSD gap is much larger than the 2.9 eV gap found by GGA-based GW [14] and close to the 4.8 eV gap found by QPscGW [13] as well as the experimental estimate of 4.3 eV [2]. However, as discussed in Chapter 2.4, the estimated finite size and basis effects in the EOM-CC calculations are quite large (TDL extrapolations are shown in parentheses in Table 2.2) thus the final basis set limit and TDL EOM-CCSD gaps are overestimated by 1–2 eV. The sizable T_1 diagnostics in the ground state suggest that this error may arise from differential orbital relaxation between the ground and excited states.

The nature of the insulating gap in MnO and NiO is of some interest. Figure 2.5 plots the correlated band structure at discrete points in reciprocal space from EOM-CC, with the atomic characters labeled by the colors and symbols. Quasiparticle weights are indicated for selected excitations as the normalized weight of the entire 1h (IP) or 1p (EA) sector, i.e. $\sum_i |r_{ik}|^2$ and $\sum_a |r^{ak}|^2$. We appproximated the k-resolved density of states (DOS) by summing over all the computed EOM-CCSD roots at each momentum and the DOS at selected points in the Brillouin zone is shown in Figure 2.6.

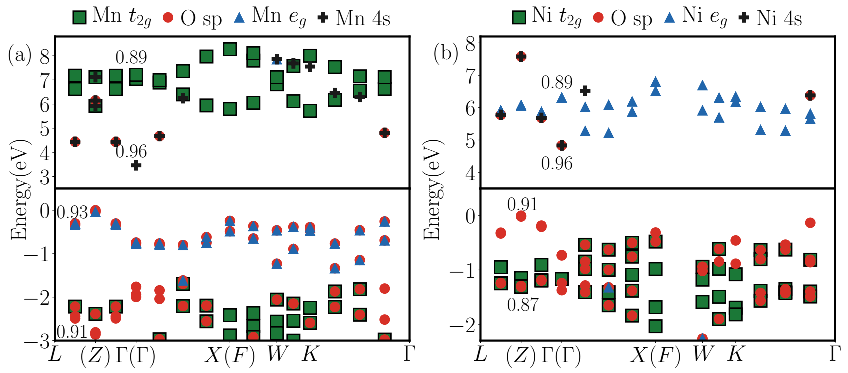

Figure 2.5: Electronic structure and quasiparticle weight analysis of (a) MnO and (b) NiO. The labels for the high-symmetry points are those defined by the primitive FCC cell; symmetry labels for the AFM rhombohedral cell are provided in brackets when the special points coincide. The upper panel is for the conduction band and the lower one for the valence band. Valence band maxima (VBM) are shifted to 0 eV. Atomic character with weight larger than 30% is shown by the indicated symbols. Quasiparticle weights are shown for the highest and lowest root computed at Γ and $\Lambda_{\frac{1}{2}}(Z)$.

Figure 2.5 shows that the top of the valence band in MnO is hybridized between the Mn e_g states and the O $2p$ states, while the conduction band minimum (CBM) consists mainly of non-dispersive t_{2g} character, except near the Γ point (CBM) where we found significant contributions from s character. In NiO, the valence band

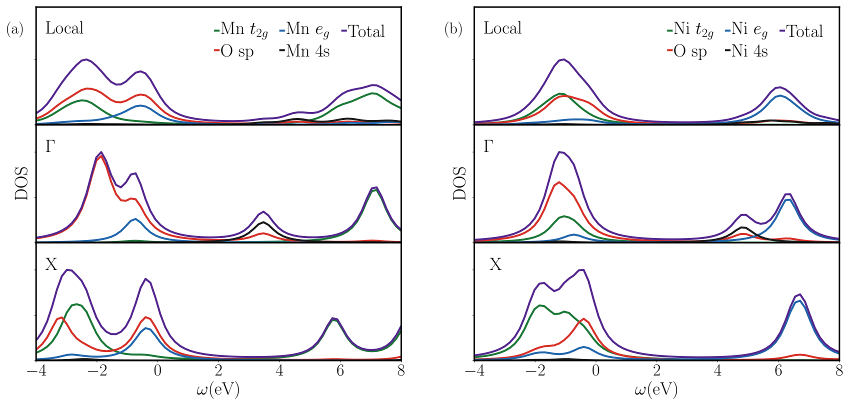

Figure 2.6: Approximate density of states of (a) MnO and (b) NiO computed by summing over the EOM-CC roots. The first panel is the local DOS, and the two panels below are the DOS at high-symmetry points Γ and X. The spectral functions are computed with a Lorentzian broadening factor $\eta = 0.4$ eV.

near the VBM is dominated by O $2p$ states (81% at VBM), while the picture for the conduction band is similar to that in MnO, including the s character near the CBM. The above picture is complemented by the DOS in Figure 2.6 where in MnO, near the Fermi level, the O $2p$ states contribute slightly more weight to the valence bands than the Mn e_g states, and the two appear at nearly identical peak positions at around -0.7 eV (relative to the VBM). The relative positions of the valence e_g and t_{2g} bands (-0.7 eV, -2.3 eV) are similar to what is seen in QPscGW (-0.5 eV and -2.2 eV respectively).

Similarly, in NiO, there is little e_g weight (peak around -0.4 eV) near the VBM, and the first peak for t_{2g} is found to be around -1.0 eV. Compared with QPscGW, our calculation suggests less weight for e_g around VBM and the location of t_{2g} is similar to their finding (~-1.0 eV). Note that additional valence e_g peaks in NiO are expected to lie deeper in the spectrum [13] and thus do not appear in Figure 2.5. Quasiparticle weights at the CBM and VBM in both materials are large (~ 0.9).

The observed s character of the CBM in MnO and NiO is also found in some earlier GGA-based GW calculations [14], but not others [15, 16]. This feature was missed in early DMFT impurity model calculations where the Ni impurity was defined using only the $3d$ shell [20–22] although it has been seen in more careful treatments in very recent DMFT calculations [25, 27, 28]. The orbital character of the CBM and VBM, including the s character, can be visualized explicitly in real space by defining quasiparticle orbitals for the CBM/VBM excitation,

$$|\psi_k^-\rangle = \sum_i r_{ik}|\phi_{ik}\rangle, \tag{2.16}$$

$$|\psi_k^+\rangle = \sum_a r^{ak}|\phi_{ak}\rangle. \tag{2.17}$$

where ϕ_{ik}, ϕ_{ak} are occupied and virtual mean-field orbitals with crystal momentum k. Real-space density plots of the quasiparticle orbitals at the VBM and CBM are shown in Figure 2.7.

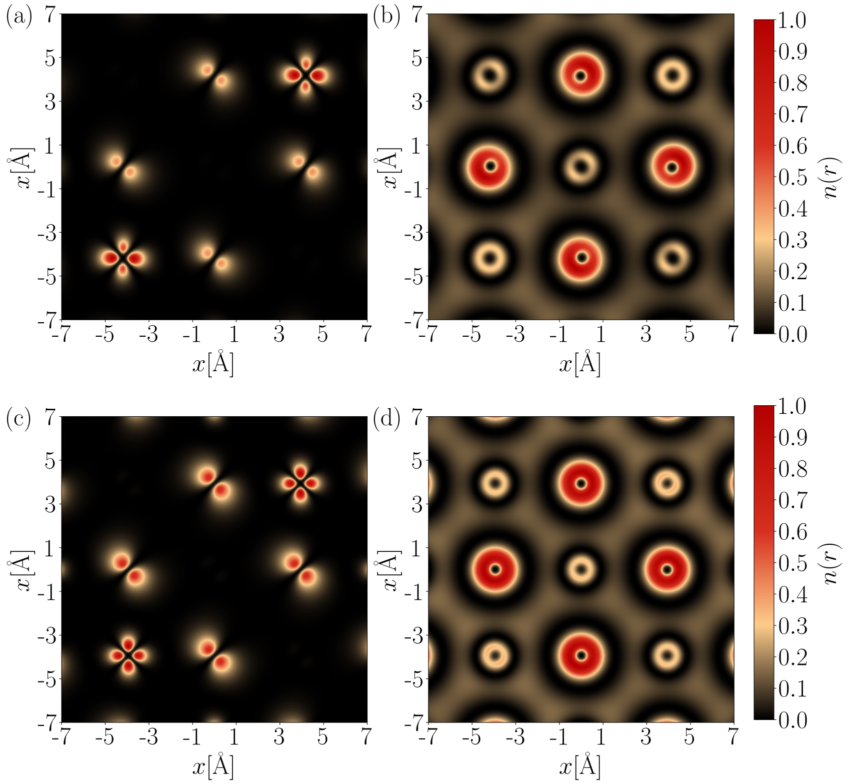

Figure 2.7: Spatial density distribution of quasiparticle orbitals on the (100) surface for (a) MnO VBM, (b) CBM, (c) NiO VBM, and (d) CBM. For MnO, we show the xy plane where the projected ionization charge shows e_g symmetry and for NiO, the quasiparticle orbitals are projected onto the xz plane.

From the analysis above, both MnO and NiO appear as insulators of mixed charge-transfer/Mott character. However, this picture is not uniform across the Brillouin zone. In particular, when only the fundamental gap is examined, NiO is clearly a charge-transfer insulator while MnO remains of mixed character. Thus the nature of the insulating state in these systems should be regarded as momentum-dependent.

2.5 Conclusion

In conclusion, we have carried out a detailed study of the ground and excited states of MnO and NiO using coupled cluster theory. While the description of the spectrum is significantly improved over mean-field methods, and quantitatively accurate at the level of 2x2x2 supercells, the gaps in the thermodynamic limit remain somewhat overestimated, likely due to orbital relaxation effects and lack of higher-order excitations. Unfortunately, we are not yet able to provide a quantitative estimate of the effect of triples due to the prohibitive cost. Nonetheless, coupled cluster offers interesting new insights into the qualitative nature of the insulating state in these materials, allowing for a detailed analysis of the charge-transfer/Mott-insulating character, atomic character of the bands (which indicates the important participation of *s* character states in the conduction band minima), and quasiparticle weights. Most intriguingly, our results show that the charge-transfer Mott nature of the insulating state should be considered to be a momentum-dependent quantity. Our work marks a significant first step towards the application of periodic coupled cluster methods to understand correlated electronic materials.

References

[1] Yang Gao et al. “Electronic structure of bulk manganese oxide and nickel oxide from coupled cluster theory”. In: *Phys. Rev. B* 101.16 (2020).
Y. G. contributed to the implementation of the algorithm, performed the simulation and results analysis. DOI: 10.1103/PhysRevB.101.165138.

[2] GA Sawatzky and JW Allen. “Magnitude and origin of the band gap in NiO”. In: *Phys. Rev. Lett* 53.24 (1984), p. 2339.

[3] J Van Elp et al. “Electronic structure of MnO”. In: *Phys. Rev. B* 44.4 (1991), p. 1530.

[4] A Fujimori et al. “Electronic structure of MnO”. In: *Phys. Rev. B* 42.12 (1990), p. 7580.

[5] Nevill F Mott. “Prc”. In: *R. Soc. London*. Vol. 62. 1949, p. 416.

[6] Atsushi Fujimori, Fujio Minami, and Satoru Sugano. “Multielectron satellites and spin polarization in photoemission from Ni compounds”. In: *Phys. Rev. B* 29.9 (1984), p. 5225.

[7] Kl Terakura et al. “Transition-metal monoxides: band or Mott insulators”. In: *Phys. Rev. Lett* 52.20 (1984), p. 1830.

[8] Martijn Marsman et al. “Hybrid functionals applied to extended systems”. In: *J. Phys. Condens. Matter* 20.6 (2008), p. 064201.

[9] Thomas Bredow and Andrea R Gerson. “Effect of exchange and correlation on bulk properties of MgO, NiO, and CoO”. In: *Phys. Rev. B* 61.8 (2000), p. 5194.

[10] Cesare Franchini et al. “Density functional theory study of MnO by a hybrid functional approach”. In: *Phys. Rev. B* 72.4 (2005), p. 045132.

[11] Fengjie Ma et al. “Quantum Monte Carlo calculations in solids with downfolded Hamiltonians”. In: *Phys. Rev. Lett* 114.22 (2015), p. 226401.

[12] Chandrima Mitra et al. “Many-body ab initio diffusion quantum Monte Carlo applied to the strongly correlated oxide NiO”. In: *J. Chem. Phys.* 143.16 (2015), p. 164710.

[13] Sergey V Faleev, Mark Van Schilfgaarde, and Takao Kotani. “All-Electron Self-Consistent G W Approximation: Application to Si, MnO, and NiO”. In: *Phys. Rev. Lett* 93.12 (2004), p. 126406.

[14] Je-Luen Li, G-M Rignanese, and Steven G Louie. “Quasiparticle energy bands of NiO in the G W approximation”. In: *Phys. Rev. B* 71.19 (2005), p. 193102.

[15] S Massidda et al. “Quasiparticle energy bands of transition-metal oxides within a model GW scheme”. In: *Phys. Rev. B* 55.20 (1997), p. 13494.

[16] F Aryasetiawan and O Gunnarsson. “Electronic structure of NiO in the GW approximation”. In: *Phys. Rev. Lett* 74.16 (1995), p. 3221.

[17] Vladimir I Anisimov, Jan Zaanen, and Ole K Andersen. “Band theory and Mott insulators: Hubbard U instead of Stoner I”. In: *Phys. Rev. B* 44.3 (1991), p. 943.

[18] Lei Wang, Thomas Maxisch, and Gerbrand Ceder. “Oxidation energies of transition metal oxides within the GGA+ U framework”. In: *Phys. Rev. B* 73.19 (2006), p. 195107.

[19] Giancarlo Trimarchi, Zhi Wang, and Alex Zunger. “Polymorphous band structure model of gapping in the antiferromagnetic and paramagnetic phases of the Mott insulators MnO, FeO, CoO, and NiO”. In: *Phys. Rev. B* 97.3 (2018), p. 035107.

[20] J Kuneš et al. “Local correlations and hole doping in NiO: A dynamical mean-field study”. In: *Phys. Rev. B* 75.16 (2007), p. 165115.

[21] J Kuneš et al. “NiO: correlated band structure of a charge-transfer insulator”. In: *Phys. Rev. Lett* 99.15 (2007), p. 156404.

[22] X Ren et al. “LDA+ DMFT computation of the electronic spectrum of NiO”. In: *Phys. Rev. B* 74.19 (2006), p. 195114.

[23] Krzysztof Byczuk et al. “Quantification of correlations in quantum many-particle systems”. In: *Phys. Rev. Lett* 108.8 (2012), p. 087004.

[24] Patrik Thunström, Igor Di Marco, and Olle Eriksson. “Electronic entanglement in late transition metal oxides”. In: *Phys. Rev. Lett* 109.18 (2012), p. 186401.

[25] Subhasish Mandal et al. “Influence of magnetic ordering on the spectral properties of binary transition metal oxides”. In: *Phys. Rev. B* 100.24 (2019), p. 245109.

[26] Ivan Leonov et al. “Magnetic collapse and the behavior of transition metal oxides at high pressure”. In: *Phys. Rev. B* 94.15 (2016), p. 155135.

[27] Long Zhang et al. “DFT+ DMFT calculations of the complex band and tunneling behavior for the transition metal monoxides MnO, FeO, CoO, and NiO”. In: *Phys. Rev. B* 100.3 (2019), p. 035104.

[28] Tianyu Zhu, Zhi-Hao Cui, and Garnet Kin Chan. “Efficient Implementation of Ab Initio Quantum Embedding in Periodic Systems: Dynamical Mean-Field Theory”. In: *arXiv preprint arXiv:1909.08592* (2019).

[29] Isaiah Shavitt and Rodney J Bartlett. *Many-body methods in chemistry and physics: MBPT and coupled-cluster theory*. Cambridge university press, 2009.

[30] Rodney J Bartlett and Monika Musiał. “Coupled-cluster theory in quantum chemistry”. In: *Rev. Mod. Phys.* 79.1 (2007), p. 291.

[31] James McClain et al. “Gaussian-based coupled-cluster theory for the ground-state and band structure of solids”. In: *J. Chem. Theory Comput.* 13.3 (2017), pp. 1209–1218.

[32] Thomas Gruber et al. “Applying the coupled-cluster ansatz to solids and surfaces in the thermodynamic limit”. In: *Phys. Rev. X* 8.2 (2018), p. 021043.

[33] Hendrik J Monkhorst. “Calculation of properties with the coupled-cluster method”. In: *Int. J. Quantum Chem.* 12.S11 (1977), pp. 421–432.

[34] Anna I Krylov. “Equation-of-motion coupled-cluster methods for open-shell and electronically excited species: The hitchhiker’s guide to Fock space”. In: *Annu. Rev. Phys. Chem.* 59 (2008), pp. 433–462.

[35] John F. Stanton and Rodney J. Bartlett. “The equation of motion coupled-cluster method. A systematic biorthogonal approach to molecular excitation energies, transition probabilities, and excited state properties”. In: *J. Chem. Phys.* 98.9 (1993), pp. 7029–7039. DOI: `10.1063/1.464746`. eprint: `https://doi.org/10.1063/1.464746`.

[36] James McClain et al. “Spectral functions of the uniform electron gas via coupled-cluster theory and comparison to the G W and related approximations”. In: *Phys. Rev. B* 93.23 (2016), p. 235139.

[37] Malte F Lange and Timothy C Berkelbach. "On the relation between equation-of-motion coupled-cluster theory and the GW approximation". In: *J. Chem. Theory Comput.* 14.8 (2018), pp. 4224–4236.

[38] AK Cheetham and DAO Hope. "Magnetic ordering and exchange effects in the antiferromagnetic solid solutions Mn x Ni 1- x O". In: *Phys. Rev. B* 27.11 (1983), p. 6964.

[39] Qiming Sun et al. "PySCF: the Python-based simulations of chemistry framework". In: *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 8.1 (2018), e1340.

[40] Qiming Sun et al. *PySCF: the Python-based simulations of chemistry framework*. 2017. DOI: `10.1002/wcms.1340`.

[41] Joost VandeVondele and Juerg Hutter. "Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases". In: *J. Chem. Phys.* 127.11 (2007), p. 114105.

[42] Joost VandeVondele et al. "Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach". In: *Comput. Phys. Commun.* 167.2 (2005), pp. 103–128. ISSN: 0010-4655. DOI: `https://doi.org/10.1016/j.cpc.2004.12.014`.

[43] Qiming Sun et al. "Gaussian and plane-wave mixed density fitting for periodic systems". In: *J. Chem. Phys.* 147.16 (2017), p. 164119.

[44] John F Stanton. "Why CCSD (T) works: a different perspective". In: *Chem. Phys. Lett.* 281.1-3 (1997), pp. 130–134.

[45] Gerald Knizia. "Intrinsic atomic orbitals: An unbiased bridge between quantum theory and chemical concepts". In: *J. Chem. Theory Comput.* 9.11 (2013), pp. 4834–4843.

[46] Zhi-Hao Cui, Tianyu Zhu, and Garnet Kin Chan. "Efficient Implementation of Ab Initio Quantum Embedding in Periodic Systems: Density Matrix Embedding Theory". In: *arXiv preprint arXiv:1909.08596* (2019).

[47] Mark van Schilfgaarde, Takao Kotani, and Sergey Faleev. "Quasiparticle self-consistent g w theory". In: *Phys. Rev. Lett* 96.22 (2006), p. 226402.

[48] Claudius Gros. "The boundary condition integration technique: results for the Hubbard model in 1D and 2D". In: *Z. Phys. B* 86.3 (1992), pp. 359–365.

[49] C Lin, FH Zong, and David M Ceperley. "Twist-averaged boundary conditions in continuum quantum Monte Carlo algorithms". In: *Phys. Rev. E* 64.1 (2001), p. 016702.

[50] BEF Fender, AJ Jacobson, and FA Wedgwood. "Covalency Parameters in MnO, α-MnS, and NiO". In: *J. Chem. Phys.* 48.3 (1968), pp. 990–994.

[51] V Fernandez et al. "Observation of orbital moment in NiO". In: *Phys. Rev. B* 57.13 (1998), p. 7870.

[52] Nathan J DeYonker et al. “Quantitative computational thermochemistry of transition metal species”. In: *J. Phys. Chem. A* 111.44 (2007), pp. 11269–11277.

[53] David R Lide. *CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data*. CRC press, 1995.

Chapter 3

EFFICIENT CONTRACTION SCHEME FOR TENSORS WITH CYCLIC GROUP SYMMETRY

3.1 Abstract

Tensor contractions are ubiquitous in numerical simulation of quantum many-body problems. In this work, we describe how to accelerate tensor contractions involving block sparse tensors whose structure is induced by a cyclic group symmetry or a product of cyclic group symmetries. Tensors of this kind naturally arise in quantum systems with intrinsic symmetry groups. With intuition aided by tensor diagrams, we present our irreducible representation alignment technique, which enables efficient handling of such block sparsity structure via only dense tensors operations. Our proposed algorithm is generally applicable to arbitrary order group symmetric contractions. The algorithm is implemented in Python, and we perform benchmark calculations on a variety of representative contractions from quantum chemistry and tensor network methods. As a consequence of relying on only dense tensor contractions, we can easily make use of efficient batched matrix multiplication via Intel's MKL and distributed tensor contraction via the Cyclops library, achieving good efficiency and parallel scalability on up to 4096 Knights Landing cores of a supercomputer.

3.2 Introduction

Tensors and Tensor Contraction

A tensor $\mathcal{T}$ is defined by a set of real or complex numbers indexed by tuples of integers (indices) $i, j, k, l, \ldots$, where the indices take integer values $i \in 1 \ldots D_i$, $j \in 1 \ldots D_j, \ldots$ etc., and a single tensor element is denoted $t_{ijkl\ldots}$. We will refer to the number of indices of the tensor as its *rank* and the sizes of their ranges as its *dimensions* ($D_i \times D_j \times \cdots$). We call the set of indices *modes* of the tensor. In this context, tensors can been viewed as a generalization of vectors and matrices as we can clearly see that scalars, vectors, and matrices are just rank 0, 1, and 2 tensors respectively.

We now introduce the graphical notation for tensors. As shown in Figure 3.1, a tensor is represented by a vertex with edges sticking out of it, each corresponding to one of its mode.

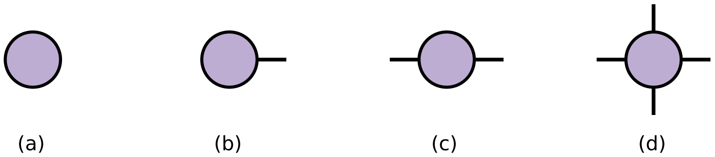

Figure 3.1: Graphical notation of tensors: (a) scalar, (b) vector, (c) matrix, (d) rank-4 tensor.

In theoretical quantum chemistry and physics, tensors generally represent states or operators and contractions express the algebra of these quantities. Tensor contractions are represented by a sum over indices of two tensors. In the case of matrices and vectors, the only possible contractions correspond to matrix and vector products. For higher rank tensors, there are more possibilities, and an example of a contraction of two rank-4 tensors is

$$w_{abij} = \sum_{k,l} u_{abkl} v_{klij}. \tag{3.1}$$

The structure of such a contraction can be also be illustrated by extending the diagrammatic notation introduced above. Here lines that represent the contracted mode are joined between vertices, and an example for Equation 3.1 is displayed in Figure 3.2.

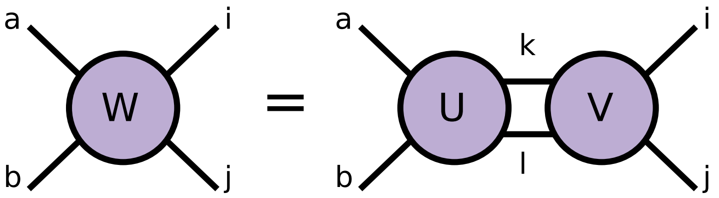

Figure 3.2: Graphical notation for tensor contraction between two rank-4 tensors.

In general, tensor contractions can be reduced to matrix multiplication (or a simpler matrix/vector operation) after appropriate transposition of the data to interchange the order of modes.

Tensors in Quantum Chemistry and Physics

As we have seen in Chapter 2, in many-body theories, tensors represent everything from wavefunction coefficients to one- and two-electron integrals. The main challenges of handling tensors in this context include large data storage, data sparsity from symmetry, and many different types of contractions required in a typical calculation. In order to achieve best performance, practical implementation must make use of all available properties of tensors that can reduce the computational cost and memory footprint.

One of such opportunities comes from the underlying symmetry group of the system. The presence of such symmetry groups enforces constraints on the relevant computations. Specifically, under the operations of the group, the computational objects (e.g. the tensors) are transformed by a matrix representation of the group, which can be decomposed into irreducible representations (irreps) of the group. Computationally, the elements of the tensors are thus constrained, and each tensor can be stored in a compressed form, referred to as its *reduced form*.

A special structure that often appears is one that is associated with a cyclic group. If each index transforms as an irrep of such a group and the overall tensor transforms as the symmetric representation, this constraint can be expressed by a sparsity structure defined on the indices, e.g.

$$t_{\mathbf{ijk}\dots} = 0 \quad \text{if} \quad \lfloor \mathbf{i}/G_1 \rfloor + \lfloor \mathbf{j}/G_2 \rfloor + \lfloor \mathbf{k}/G_3 \rfloor + \cdots \neq 0 \pmod{G}, \tag{3.2}$$

where the offset G_i denotes the size of the symmetry group for the ith index.

For tensors, such sparsity would lead to blocked structure where each block is the same size. Figure 3.3 shows such sparsity pattern for a square matrix with $G_1 = G_2 = 3$. We refer to such tensors as tensors with *cyclic group symmetry*, or cyclic group tensors for short.

In some applications, the block sizes are non-uniform, but this can be accommodated in a cyclic group tensor by padding blocks with zeros to a fixed size during initialization. With this assumption, the original tensor indices can then be unfolded into symmetry modes and the symmetry blocks, where the symmetry modes fully

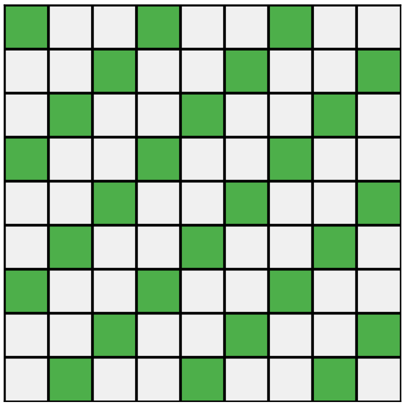

Figure 3.3: Block sparsity structure for cyclic group matrix with $G = 3$. The non-zero blocks are marked with green color.

express the block sparse structure,

$$t_{iI,jJ,kK\ldots} = 0 \quad \text{if} \quad I + J + K \cdots \neq 0 \pmod{G}. \tag{3.3}$$

Here we use the convention that the uppercase indices are the symmetry modes and the lowercase letters index into the symmetry blocks. The relationship between the symmetry modes is referred to as a symmetry conservation rule.

Given a number of symmetry sectors G, cyclic group symmetry can reduce tensor contraction cost by a factor of G for some simple contractions and G^2 for most contractions of interest (any contraction with a cost that is superlinear in input/output size).

State-of-the-art sequential and parallel libraries for handling cyclic group symmetry, both in specific physical applications and in domain-agnostic settings, typically iterate over non-zero blocks stored in a block-sparse tensor format [1–12]. There are mainly two drawbacks with this approach: (1) The use of explicit looping (over possibly small blocks) makes it difficult to reach theoretical peak performance. (2) Parallelization in distributed-memory setting is challenging due to the potentially sophisticated communication and redistribution needed for scheduling block-wise multiplication.

We introduce a general transformation of cyclic group symmetric tensors, *irreducible representation alignment*, which allows all contractions between such tensors to be transformed into a series of dense tensor contractions with optimal cost and memory footprint. In our construction, the two input reduced forms as well as the output are indexed by a new auxiliary index. This transformation provides three advantages:

1. It avoids the need for data structures to handle block sparsity or scheduling over blocks,
2. It makes possible an efficient software abstraction to contract tensors with cyclic group symmetry,
3. It enables effective use of parallel libraries for dense tensor contraction and batched matrix multiplication.

Our approach is closely related to the previous work on direct product decomposition (DPD) [13, 14], which similarly seeks an aligned representation of the two tensor operands. However, the unfolded structure of cyclic group tensors in Equation (3.3) allows for a much simpler conversion to an aligned representation, both conceptually and in terms of implementation complexity. In particular, our approach can be implemented efficiently with existing dense tensor contraction primitives.

We develop a software library, *Symtensor*, that implements the irrep alignment algorithm and contraction. We study the efficacy of this new method for cyclic group tensor contractions arising in periodic coupled cluster theory and tensor network methods. We demonstrate that across a variety of tensor contractions, the library achieves orders of magnitude improvements in parallel performance and a matching sequential performance relative to the manual loop-over-blocks approach. The resulting algorithm may also be easily and automatically parallelized for distributed-memory architectures. Using the Cyclops Tensor Framework (CTF) library [15] as the contraction backend to Symtensor, we demonstrate good strong and weak scalability with up to at least 4096 Knights Landing cores of the Stampede2 supercomputer.

The rest of the chapter is organized as follows. In Section 3.3 we describe our irreducible representation alignment algorithm with an intuitive example. The implementation of our algorithm is described in Section 3.4. Then we provide benchmark results on a variety of representative contractions from coupled cluster and tensor network methods in Section 3.5. We conclude with Section 3.6.

3.3 Theory

We now describe our proposed approach. We first describe the algorithm on an example contraction and provide intuition for correctness based on conservation of flow in a tensor diagram graph. These arguments are analogous to the conservation arguments used in computations with Feynman diagrams (e.g. momentum and energy conservation) [16] or with quantum numbers in tensor networks [17], although the notation we use is slightly different.

We consider a contraction of rank-4 tensors $\mathcal{U}$ and $\mathcal{V}$ into a new rank-4 tensor $\mathcal{W}$, where all tensors have cyclic group symmetry. We can express this cyclic group symmetric contraction as a contraction of tensors of rank 8, by separating indices into symmetry-block (lower-case) indices and symmetry-mode (upper-case) indices, so

$$w_{aA,bB,iI,jJ} = \sum_{k,K,l,L} u_{aA,bB,kK,lL} v_{kK,lL,iI,jJ}. \tag{3.4}$$

Here and later we use commas to separate index groups for readability.

The input and output tensors are assumed to transform as symmetric irreps of a cyclic group, which implies the following relationships between the symmetry modes and associated block structures,

$$w_{aA,bB,iI,jJ} \neq 0 \text{ if } A + B - I - J \equiv 0 \pmod{G}, \tag{3.5}$$

$$u_{aA,bB,kK,lL} \neq 0 \text{ if } A + B - K - L \equiv 0 \pmod{G}, \tag{3.6}$$

$$v_{kK,lL,iI,jJ} \neq 0 \text{ if } K + L - I - J \equiv 0 \pmod{G}. \tag{3.7}$$

Ignoring the symmetry, this tensor contraction would have cost $O(N^4G^4)$ for memory footprint and $O(N^6G^6)$ for computation, where N is the dimension of each symmetry sector.

With the use of symmetry, the cost for memory and computation can be reduced to $O(N^4G^3)$ and $O(N^6G^4)$ respectively. This can be achieved by first representing the original tensor in a reduced dense form indexed by just 3 symmetry modes. In particular, we refer to the reduced form indexed by 3 symmetry modes that are a subset of the symmetry modes of the original tensors, as the standard reduced form. The equations below show the mapping from the original tensor to one of its standard reduced forms:

$$\bar{w}_{aA,b,iI,jJ} = w_{aA,b,I+J-A \bmod G,iI,jJ}, \tag{3.8}$$

$$\bar{u}_{aA,b,kK,lL} = u_{aA,b,K+L-A \bmod G,kK,lL}, \tag{3.9}$$

$$\bar{v}_{kK,l,iI,jJ} = v_{kK,l,I+J-K \bmod G,iI,jJ}. \tag{3.10}$$

As an example, the associated graphical notation for $w_{aA,bB,iI,jJ}$ is shown in Figure 3.4. Note here the graphical notation is slightly different from that in Figure 3.2:

arrows are now used to indicate the sign associated with the symmetry mode in the symmetry conservation rule for the tensor, and the legs for the lower letters are omitted in the graph for simplicity (they are always stored).

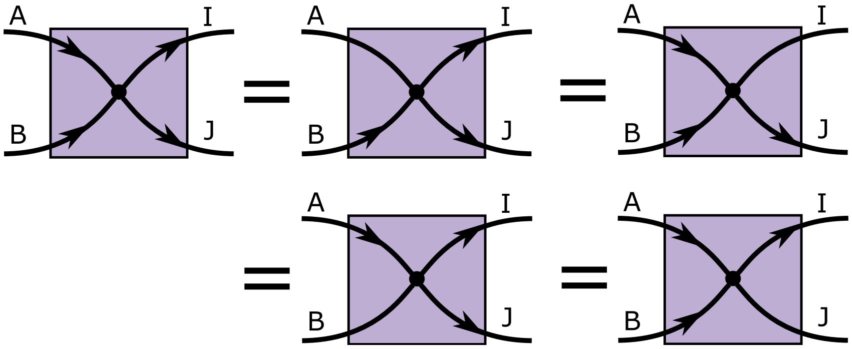

Figure 3.4: Tensor diagrams of the standard reduced form. Arrows on each leg represent the corresponding symmetry indices that are explicitly stored. Symmetry indices on legs without arrows are not stored but are implicitly represented with the symmetry conservation law $(A + B = I + J \pmod{G})$.

The standard reduced form provides an implicit representation of the unstored symmetry mode due to symmetry conservation and can be easily used to implement the block-wise contraction approach prevalent in many libraries. This is achieved via manual loop nest over the appropriate symmetry modes of the input tensors, as shown in Algorithm 2. All elements of $\mathcal{W}$, $\mathcal{U}$, and $\mathcal{V}$ in the standard reduced form can be accessed with 4 independent nested-loops to perform the multiplication and accumulation operation in Algorithm 2. The other two implicit symmetry modes can be obtained inside these loops using symmetry conservation, reducing the computation cost to $O(G^4)$.

However, the indirection needed to compute L and J within the innermost loops prevents expression of the contraction in terms of standard library operations for a single contraction of dense tensors. Furthermore, the need to parallelize general block-wise tensor contraction operations in the nested loop approach above, creates a significant software-engineering challenge and computational overhead for tensor contraction libraries.

The main idea in the irreducible representation alignment algorithm is to first transform (reindex) the tensors using an auxiliary symmetry mode which subsequently allows a dense tensor contraction to be performed without the need for any indirection. In the above contraction, we define the auxiliary mode index as $Q \equiv I + J \equiv A + B \equiv K + L \pmod{G}$ and thus obtain a new reduced form for each

Algorithm 2 Loop nest to perform group symmetric contraction $w_{aA,bB,iI,jJ} = \sum_{k,K,l,L} u_{aA,bB,kK,lL} v_{kK,lL,iI,jJ}$ using standard reduced forms $\bar{w}_{aA,bB,iI,j}$, $\bar{u}_{aA,bB,kK,l}$, and $\bar{v}_{kK,lL,iI,j}$.

```
for $A = 1, \ldots, G$ do
    for $B = 1, \ldots, G$ do
        for $I = 1, \ldots, G$ do
            $J = A + B - I \bmod G$
            for $K = 1, \ldots, G$ do
                $L = A + B - K \bmod G$
                $\forall a, b, i, j, \quad \bar{w}_{aA,bB,iI,j} = \bar{w}_{aA,bB,iI,j} + \sum_{k,l} \bar{u}_{aA,bB,kK,l} \bar{v}_{kK,lL,iI,j}$
            end for
        end for
    end for
end for
```

tensor. The relations of this reduced form with the sparse form are as follows:

$$\hat{w}_{aA,b,i,jJ,Q} = w_{aA,b,Q-A \bmod G,i,Q-J \bmod G,jJ}, \tag{3.11}$$

$$\hat{u}_{aA,b,k,lL,Q} = u_{aA,b,Q-A \bmod G,k,Q-L \bmod G,lL}, \tag{3.12}$$

$$\hat{v}_{k,lL,i,jJ,Q} = v_{k,Q-L \bmod G,lL,i,Q-J \bmod G,jJ}. \tag{3.13}$$

This reduced form is displayed in Figure 3.5. Note here that though the Q arrow does not stick out of the vertex, it is stored explicitly. This is slightly different from the general graphical notation for tensors.

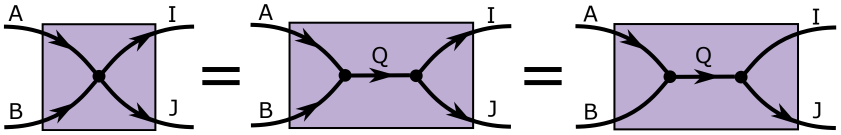

Figure 3.5: The symmetry aligned reduced form is defined by introducing the Q symmetry mode. Each of the two vertices defines a symmetry conservation relation: $A + B = Q \pmod G$ and $Q = I + J \pmod G$, allowing two of the arrows to be removed in the 3rd diagram (implicitly stored).

The Q symmetry mode is chosen so that it can serve as part of the reduced forms of each of $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$. An intuition for why this alignment is possible is given via tensor diagrams in Figure 3.6. The new auxiliary indices (P and Q) of the two contracted tensors satisfy a conservation law $P = Q$, and so can be reduced to a single index.

As shown in Figure 3.7, given the aligned reduced forms of the two operands, we can contract them directly to obtain a reduced form for the output that also has the

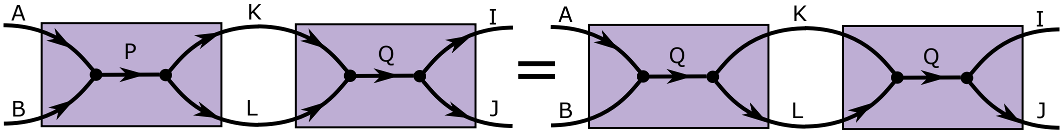

Figure 3.6: By defining conservation laws on the vertices, we see that $P = K + L \pmod{G}$ and $K + L = Q \pmod{G}$. Consequently, the only non-zero contributions to the contraction must have $P = Q$.

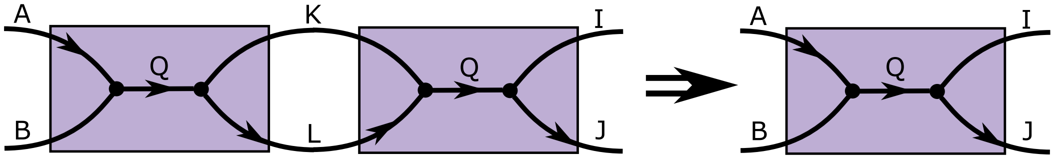

Figure 3.7: The reduced forms allows an efficient contraction operation to compute the output reduced forms, namely through `einsum` operation `W=einsum("AQL,LQJ->AQJ",U,V)` (intra-block indices ignored).

additional symmetry mode Q. Specifically, it suffices to perform the dense tensor contraction below:

$$\hat{w}_{aA,b,i,jJ,Q} = \sum_{L,k,l} \hat{u}_{aA,b,k,lL,Q}\hat{v}_{k,lL,i,jJ,Q}. \tag{3.14}$$

This contraction can be expressed as a single `einsum` operation (available via NumPy, CTF, etc.) and can be done via a batched matrix multiplication (available in Intel's MKL). Once $\hat{\mathcal{W}}$ is obtained in this reduced form, it can be remapped to any other desired reduced form.

The remaining step is to define how to carry out the transformations between the aligned reduced forms and the standard reduced form. These can be performed via contraction with a Kronecker delta tensor defined on the symmetry modes, constructed from symmetry conservation, e.g, $\hat{u}_{aA,b,k,lL,Q} = \sum_B \bar{u}_{aA,bB,IL,j}\delta_{A,B,Q}$, where

$$\delta_{A,B,Q} = 0 \quad \text{if} \quad A + B - Q \neq 0 \pmod{G}. \tag{3.15}$$

Using this approach, all steps in our algorithm can be expressed fully in terms of single dense, or batched dense, tensor contractions. Extending this algorithm to arbitrary rank tensor contractions is straightforward, and the details for the derivation are provided in Appendix A.

3.4 Implementation

We implement the irrep alignment algorithm as a Python library, Symtensor[1]. The library automatically selects the appropriate reduced form to align the irreps for the contraction, constructs the generalized Kronecker deltas to convert input and output tensors to the target forms, and performs the batched dense tensor contractions that implement the numerical computation. The dense tensor contraction is interfaced to different contraction backends. Besides the default NumPy `einsum` backend, we also provide a backend that leverages MKL's batched matrix-multiplication routines [18] to obtain good threaded performance, and employ an interface to Cyclops [15] for distributed-memory execution.

```python
import numpy as np
from symtensor import array, einsum

# Define Z3 Symmetry
irreps = [0,1,2]
G = 3
total_irrep = 0
z3sym = ["++--", [irreps]*4, total_irrep, G]

# Initialize two sparse tensors as input
N = 10
Aarray = np.random.random([G,G,G,N,N,N,N])
Barray = np.random.random([G,G,G,N,N,N,N])

# Initialize symtensor with raw data and symmetry
u = array(Aarray, z3sym)
v = array(Barray, z3sym)

# Compute output symtensor
w = einsum('abkl,klij->abij', u, v)
```

Figure 3.8: Symtensor library example for contraction of two group symmetric tensors.

In Figure 3.8, we provide an example on how to perform the contraction of two cyclic group tensors with Z_3 (cyclic group with $G = 3$) symmetry for each index using Symtensor library. In the code, the Symtensor library initializes the rank-4 cyclic group symmetric tensor using an underlying rank-7 dense reduced representation. Once the tensors are initialized, the subsequent `einsum` operation implements the

[1] `https://github.com/yangcal/symtensor`

contraction shown in Figure 3.6 without referring to any symmetry information in its interface. While the example is based on a simple cyclic group for a rank-4 tensor, the library supports arbitrary orders, as well as products of cyclic groups and infinite cyclic groups (e.g. $U(1)$ symmetries).

As introduced in Section 3.3, the main operations in our irrep alignment algorithm consist of transformation of the reduced form and the contraction of reduced forms. Symtensor chooses the least costly version of the irrep alignment algorithm from a space of variants defined by different choices of the implicitly represented modes of the three tensors in symmetry aligned reduced form. This choice is made by enumerating all valid variants.

After choosing the best reduced forms, the required generalized Kronecker deltas are generated as dense tensors. This permits both the transformations and the reduced form contraction to be done as `einsum` operations of dense tensors with the desired backend.

3.5 Benchmarks

Computational Details

As a testbed for this approach, we survey a few group symmetric tensor contractions that arise in quantum chemistry and quantum many-body physics methods.

The first set of contractions come from the periodic coupled cluster introduced in Chapter 2. In crystalline (periodic) materials, translational symmetry amounts to a product of cyclic symmetry groups along each lattice dimension. For a three-dimensional crystal, the size of the resulting symmetry group takes the form $G = G_1 \times G_2 \times G_3$, where G_1, G_2, and G_3 are the number of k points sampling along each dimension. In periodic CCSD, three common expensive tensor contractions can be written as

$$w_{iI,jJ,aA,bB} = \textstyle\sum_{cC,dD} u_{iI,jJ,cC,dD} v_{aA,bB,cC,dD}, \tag{3.16}$$

$$w_{iI,jJ,aA,kK} = \textstyle\sum_{bB,cC} u_{bB,cC,iI,jJ} v_{bB,cC,kK,aA}, \tag{3.17}$$

$$w_{iI,jJ,mM,nN} = \textstyle\sum_{aA,bB} u_{aA,bB,mM,nN} v_{aA,bB,iI,jJ}, \tag{3.18}$$

where each symmetry mode of the tensors is associated with the translational symmetry group. Although all contractions above have an asymptotic scaling of G^4N^6, the size of virtual orbital (a, b, c, d) and occupied orbital (i, j, m, n, k) can differ significantly in practice. This makes the perfect test case to evaluate our algorithm in a realistic setting.

The second set of contractions come from tensor network algorithms. In this context, the emergence of cyclic group symmetric tensors can be traced to conservation of particle and spin quantum numbers. We consider the two contractions below:

$$\begin{aligned} w_{iI,jJ,lL,mM} &= \sum_{kK} u_{iI,jJ,kK} v_{kK,lL,mM}, && (3.19) \\ w_{iI,jJ,mM,nN} &= \sum_{kK,lL} u_{iI,jJ,kK,lL} v_{kK,lL,mM,nN}, && (3.20) \end{aligned}$$

where the first contraction is encountered when optimizing a single matrix product states (MPS) tensor and the second one arises during the computation of the normalization of the projected entangled pair states (PEPS).

Table 3.1 summarizes the details for all the contractions above.

Table 3.1: Summary of coupled cluster and tensor network contractions used in our benchmark test and their costs. We include matrix multiplication (MM) as a point of reference. The three CC contractions described in Equation 3.16 are labeled CC_1,CC_2, and CC_3 respectively.

Label	Contraction	Symmetric Cost
MM	$w_{iI,kK} = \sum_{jJ} u_{iI,jJ} v_{jJ,kK}$	$O(GN^3)$
CC_1	$w_{iI,jJ,aA,bB} = \sum_{cC,dD} u_{iI,jJ,cC,dD} v_{aA,bB,cC,dD}$	$O(G^4N^6)$
CC_2	$w_{iI,jJ,aA,kK} = \sum_{bB,cC} u_{bB,cC,iI,jJ} v_{bB,cC,kK,aA}$	$O(G^4N^6)$
CC_3	$w_{iI,jJ,aA,bB} = \sum_{kK,lL} u_{iI,jJ,kK,lL} v_{mM,nN,kK,lL}$	$O(G^4N^6)$
MPS	$w_{iI,jJ,lL,mM} = \sum_{kK} u_{iI,jJ,kK} v_{kK,lL,mM}$	$O(G^3N^5)$
PEPS	$w_{iI,jJ,mM,nN} = \sum_{kK,lL} u_{iI,jJ,kK,lL} v_{kK,lL,mM,nN}$	$O(G^4N^6)$

Performance experiments were carried out on the Stampede2 supercomputer. Each Stampede2 node is a Intel Knight's Landing (KNL) processor, on which we use up to 64 of 68 cores by employing up to 64 threads with single-node NumPy/MKL and 64 MPI processes per node with 1 thread per process with Cyclops. We use the Symtensor library together with one of three external contraction backends: Cyclops, default NumPy, or a batched BLAS backend for NumPy arrays (this backend leverages HPTT [19] for fast tensor transposition and dispatches to Intel's MKL BLAS for batched matrix multiplication). We also compare against the loop-over blocks algorithm as illustrated in Algorithm 2. This implementation performs each block-wise contraction using MKL, matching state of the art libraries for tensor contractions with cyclic group symmetry [20, 21].

Sensitivity to Size of Symmetry Group

We first examine the performance of the irrep alignment algorithm for three contractions as a function of increasing G. The results are displayed in Figure 3.9 with the left, center, and right plots showing the scaling for the contractions labeled MM, CC_1, and PEPS in Table 3.1.

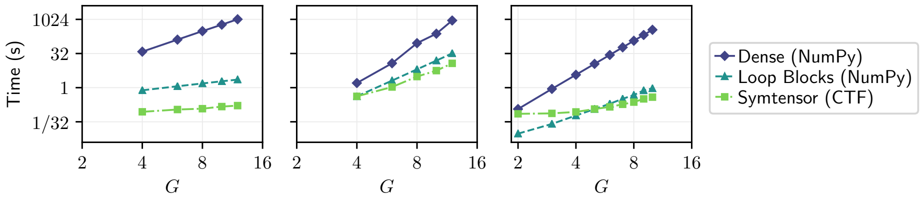

Figure 3.9: Comparison of the execution times for contractions on a single thread using three different algorithms: a dense, non-symmetric contraction, loops over symmetry blocks, and our Symtensor library. From left to right, the plots show the scaling for matrix multiplication (MM), a coupled cluster contraction (CC_1), and a tensor network contraction (PEPS). The dense and loop-over blocks calculations use NumPy as a contraction backend, while the Symtensor library here uses Cyclops as the contraction backend.

Here we compare scaling relative to two conventional approaches: a dense contraction without utilizing symmetry and loops over symmetry blocks, both using NumPy's `einsum` function. The dimensions of the tensors considered are, for matrix multiplication, $N = 500$ and $G \in [4, 12]$, for the CC contraction, $N_i = N_j = N_k = N_l = 8$, $N_a = N_b = N_c = N_d = 16$, with $G \in [4, 12]$, and for the PEPS contraction, $N_{\text{mps}} = 16$, $N_{\text{peps}} = 4$, with $G \in [2, 10]$. We found that our symtensor approach consistently improves the performance for all but the smallest contractions. Additionally, a comparison of the slopes of the lines in each of the three plots indicates that the dense tensor contraction scheme results in a higher order asymptotic scaling in G than either of the symmetric approaches.

Then we perform a comprehensive study on the absolute performance of our algorithm on all contractions in Table 3.1. The resulting performance with 1 thread and 64 threads is shown in Figure 3.10. For each contraction, we consider one with a large number of symmetry sectors (G) with small block size (N) (labeled with a subscript a) and another with fewer symmetry sectors and larger block size (labeled with a subscript b). The specific dimensions of all tensors studied are provided in Table 3.2. For each of these cases, we compare the execution time, in seconds, using loops over blocks dispatching to NumPy contractions, the Symtensor library with

NumPy arrays and batched BLAS as the contraction backend, and the Symtensor library using Cyclops as the array and contraction backend.

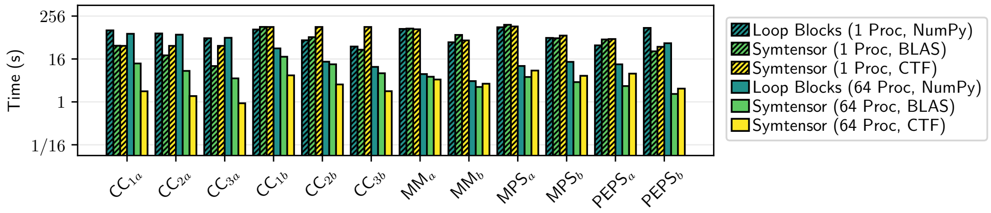

Figure 3.10: Comparison of contraction times using the Symtensor library (using Cyclops for the array storage and contraction backend, or NumPy as the array storage with batched BLAS contraction backend) and loops over blocks using NumPy as the contraction backend. The different bars indicate both the algorithm and backend used and the number of threads used on a single node.

A clear advantage in parallelizability of Symtensor is evident in Figure 3.10. With 64 threads, Symtensor outperforms manual looping by a factor of at least 1.4X for all contraction benchmarks, and the largest speed-up, 69X, is obtained for the CC_{3a} contraction.

Table 3.2: Dimensions of the tensors used for contractions in Figure 3.10 and Figure 3.11.

Label	Specifications
CC_{1a}	$G = 8, N_a = N_b = N_c = N_d = 32, N_i = N_j = 16$
CC_{2a}	$G = 8, N_a = N_b = N_c = 32, N_i = N_j = N_k = 16$
CC_{3a}	$G = 8, N_a = N_b = 32, N_i = N_j = N_k = N_l = 16$
CC_{1b}	$G = 16, N_a = N_b = N_c = N_d = 16, N_i = N_j = 8$
CC_{2b}	$G = 16, N_a = N_b = N_c = 16, N_i = N_j = N_k = 8$
CC_{3b}	$G = 16, N_a = N_b = 16, N_i = N_j = N_k = N_l = 8$
MM_a	$G = 2, N = 10000$
MM_b	$G = 100, N = 2000$
MPS_a	$G = 2, N_i = N_k = N_m = 3000, N_j = 10, N_l = 1$
MPS_b	$G = 5, N_i = N_k = N_m = 700, N_j = 10, N_l = 1$
PEPS_a	$G = 2, N_i = N_j = 400, N_k = N_l = N_m = N_n = 20$
PEPS_b	$G = 10, N_i = N_j = 64, N_k = N_l = N_m = N_n = 8$

We also observed a significant difference between the contractions labeled to be of type a (large G and small N) and type b (large N and small G), with the geometric mean speedup for these two being 11X and 2.8X respectively on 64 threads. This discrepancy is also observed on a single thread, though less drastically, with respective geometric mean speedups of 1.9X and 1.2X. This can be traced to

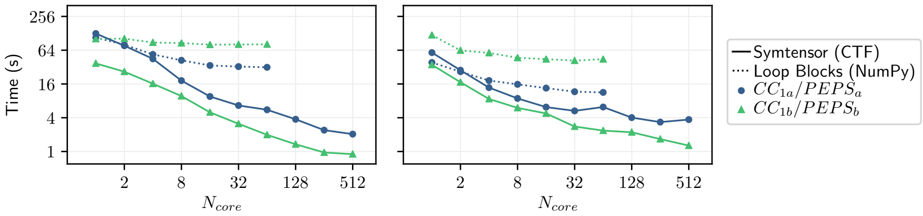

Figure 3.11: Strong scaling behavior for the CC contractions (left) labeled CC_{1a} (blue circles) and CC_{1b} (green triangles) and the PEPS contractions (right) labeled PEPS_a (blue circles) and PEPS_b (green triangles). The dashed lines correspond to calculations done using a loop over blocks algorithm with a NumPy backend while the solid lines correspond to Symtensor calculations using the irrep alignment algorithm, with a Cyclops backend.

the larger number of symmetry blocks in type b cases, which amplifies the overhead of manual looping.

Multi-Node Performance

We now illustrate the parallelizability of the irrep alignment algorithm by studying scalability across multiple nodes with distributed memory. All parallelization in Symtensor is handled via the Cyclops library in this case.

The solid lines in Figure 3.11 show the strong scaling (fixed problem size) behavior of the Symtensor implementation on up to eight nodes. As a reference, we provide comparison to strong scaling on a single node for the loop over blocks method using NumPy as the backend.

We again observe that the Symtensor irrep alignment implementation provides a significant speedup over the loop over blocks strategy, which is especially evident when there are many symmetry sectors in each tensor. For example, using 64 threads on a single node, the speedup achieved by Symtensor over the loop over blocks implementation is 41X for CC_{1a}, 5.7X for CC_{1b}, 4.1X for PEPS_a, and 27X for PEPS_b. We additionally see that the contraction times continue to scale with good efficiency when the contraction is spread across multiple nodes.

Finally, in Figure 3.12 we display weak scaling performance, where the dimensions of each tensor are scaled with the number of nodes (starting with the problem size reported in Table 3.2 on 1 node) used so as to fix the tensor size per node. Thus, in this experiment, we utilize all available memory and seek to maximize performance rate.

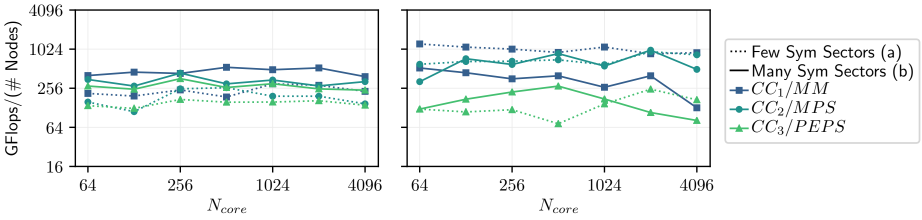

Figure 3.12: Weak scaling behavior for CC (left) and TN (right) contractions. The dashed lines correspond to contractions with a small symmetry group (small G), previously labeled (a), while solid lines correspond to contractions with a large symmetry group (large G), labeled (b). The blue squares correspond to the CC_1 and matrix multiplication performance, the dark green circles correspond to the CC_2 and MPS performance, and the light green triangles correspond to the CC_3 and PEPS performance.

Figure 3.12 displays the performance rate per node, which varies somewhat across contractions and node counts, but generally does not fall off with increasing node count, demonstrating good weak scalability. When using 4096 cores, the overall performance rate approaches 4 Teraflops/s for some contractions, but can be lower in other contractions with less arithmetic intensity.

3.6 Conclusion

The irrep alignment algorithm leverages symmetry conservation rules implicit in cyclic group symmetry to provide a contraction method that is efficient across a wide range of tensor contractions. This technique is applicable to many numerical methods for quantum-level modeling of physical systems that involve tensor contractions. The automatic handling of group symmetry with dense tensor contractions provided via the Symtensor library provides benefits in productivity, portability, and parallel scalability for such applications.

References

[1] David Ozog et al. “Inspector-executor load balancing algorithms for block-sparse tensor contractions”. In: *2013 42nd International Conference on Parallel Processing*. IEEE. 2013, pp. 30–39.

[2] Pai-Wei Lai et al. “A framework for load balancing of tensor contraction expressions via dynamic task partitioning”. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. 2013, pp. 1–10.

[3] Chong Peng et al. “Massively parallel implementation of explicitly correlated coupled-cluster singles and doubles using TiledArray framework”. In: *The Journal of Physical Chemistry A* 120.51 (2016), pp. 10231–10244.

[4] Khaled Z Ibrahim et al. “Analysis and tuning of libtensor framework on multicore architectures”. In: *2014 21st International Conference on High Performance Computing (HiPC)*. IEEE. 2014, pp. 1–10.

[5] Ricky A. Kendall et al. “High performance computational chemistry: An overview of NWChem a Distributed parallel application”. In: *Computer Physics Communications* 128.1-2 (2000), pp. 260–283.

[6] So Hirata. “Tensor Contraction Engine: Abstraction and Automated Parallel Implementation of Configuration-Interaction, Coupled-Cluster, and Many-Body Perturbation Theories”. In: *The Journal of Physical Chemistry A* 107.46 (2003), pp. 9887–9897. DOI: `10.1021/jp034596z`.

[7] Jaroslaw Nieplocha, Robert J. Harrison, and Richard J. Littlefield. “Global Arrays: A nonuniform memory access programming model for high-performance computers”. In: *The Journal of Supercomputing* 10 (2 1996), pp. 169–189.

[8] In: *ITensor Library (version 2.0.11) http://itensor.org* ().

[9] Garnet Kin-Lic Chan and Martin Head-Gordon. “Highly correlated calculations with a polynomial cost algorithm: A study of the density matrix renormalization group”. In: *The Journal of chemical physics* 116.11 (2002), pp. 4462–4476.

[10] Sandeep Sharma and Garnet Kin-Lic Chan. “Spin-adapted density matrix renormalization group algorithms for quantum chemistry”. In: *The Journal of chemical physics* 136.12 (2012), p. 124121.

[11] Yuki Kurashige and Takeshi Yanai. “High-performance ab initio density matrix renormalization group method: Applicability to large-scale multireference problems for metal compounds”. In: *The Journal of chemical physics* 130.23 (2009), p. 234114.

[12] Ying-Jer Kao, Yun-Da Hsieh, and Pochung Chen. “Uni10: An open-source library for tensor network algorithms”. In: *Journal of Physics: Conference Series*. Vol. 640. 1. IOP Publishing. 2015, p. 012040.

[13] John F Stanton et al. “A direct product decomposition approach for symmetry exploitation in many-body methods. I. Energy calculations”. In: *J. Chem. Phys.* 94.6 (1991), pp. 4334–4345.

[14] Devin A Matthews. “On extending and optimising the direct product decomposition”. In: *Molecular Physics* 117.9-12 (2019), pp. 1325–1333.

[15] Edgar Solomonik et al. “A massively parallel tensor contraction framework for coupled-cluster computations”. In: *Journal of Parallel and Distributed Computing* 74.12 (2014), pp. 3176–3190.

[16] Alexander L Fetter and John Dirk Walecka. *Quantum theory of many-particle systems*. Courier Corporation, 2012.

[17] Ulrich Schollwöck. “The density-matrix renormalization group in the age of matrix product states”. In: *Annals of physics* 326.1 (2011), pp. 96–192.

[18] Ahmad Abdelfattah et al. “Performance, design, and autotuning of batched GEMM for GPUs”. In: *International Conference on High Performance Computing*. Springer. 2016, pp. 21–38.

[19] Paul Springer, Tong Su, and Paolo Bientinesi. “HPTT: A High-Performance Tensor Transposition C++ Library”. In: *Proceedings of the 4th ACM SIGPLAN International Workshop on Libraries, Languages, and Compilers for Array Programming*. ARRAY 2017. Barcelona, Spain: ACM, 2017, pp. 56–62. ISBN: 978-1-4503-5069-3. DOI: 10.1145/3091966.3091968. URL: http://doi.acm.org/10.1145/3091966.3091968.

[20] Evgeny Epifanovsky et al. “New implementation of high-level correlated methods using a general block-tensor library for high-performance electronic structure calculations”. In: *Journal of Computational Chemistry* (2013). ISSN: 1096-987X.

[21] Ryan Levy, Edgar Solomonik, and Bryan Clark. “Distributed-Memory DMRG via Sparse and Dense Parallel Tensor Contractions”. In: *2020 SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. IEEE Computer Society, pp. 319–332.

Chapter 4

COUPLED CLUSTER ANSATZ FOR ELECTRONS AND PHONONS

4.1 Abstract

We describe a coupled cluster framework for coupled systems of electrons and harmonic phonons. Charged excitations are computed via the equation of motion version of the theory. Benchmarks on the Hubbard-Holstein model allow us to assess the strengths and weaknesses of different coupled cluster approximations which generally perform well for weak to moderate coupling. Finally, we report progress towards an implementation for ab initio calculations on solids, and present some preliminary results on finite-size models of diamond with a linear electron-phonon coupling. We also report the implementation of electron-phonon coupling matrix elements from crystalline Gaussian type orbitals (cGTO) within the PySCF program package.

4.2 Introduction

Electron-phonon interactions (EPIs) are ubiquitous in materials science and condensed matter physics. For instance, they underpin the temperature dependence of electronic transport and optical absorption in semiconductors. Additionally, observation of characteristic kinds and Hohn anomalies in photoemission and Rama and neutron spectra can be largely traced to these interactions. They are also the critical interaction that give rise to the Bardeen-Cooper-Schrieffer type of superconductivity.

The phenomenology surrounding EPIs has been extensively studied in the context of various lattice models and semi-empirical Hamiltonians. For example, the Hamiltonians of Frölich[1] and Holstein[2] type capture the limits of non-local and local electron-phonon interactions respectively. The Su, Schreiffer, and Heeger (SSH) model was introduced as a simplified model of 1-dimensional polyacetylene that contains EPIs[3], and is now commonly used as a simple example of a 1-dimensional system with topological character[4]. In addition to the EPIs, these models tend to also feature an electron-interaction term (usually in the form of a Hubbard interaction) to allow study on the regime where both EPI and electronic correlation are important. The Hubbard-Holstein (HH) model is one such simple model that has been extensively studied and its well-studied phase diagram clearly displays the rich structure that can result from the interplay of electron-phonon and electron-electron interactions[5–8].

Complementary to the study of model Hamiltonians is the development of ab initio theory of EPIs. Within this framework, density functional theory is generally used as the base electronic structure theory and the EPIs are computed either through finite difference differentiation (the “supercell approach") [9–11] or density functional perturbation theory (DFPT)[12–14]. While the expense of these calculations often necessitates the use of DFT, there have been some attempts to move beyond the DFT framework[15–21]. These results suggest that going beyond DFT quasiparticle energies can change the effects of the EPI significantly. Going further, converging these ab initio calculations towards TDL requires the development of specialized interpolation schemes so that the EPI matrix elements may be represented on a very dense grid in the Brillouin zone[22]. Due to the large size of the EPI matrix, property calculations have been limited to relatively simplified theories as a computational compromise. This is in stark contrast to the situation for model systems where the coupled problem with small system size can be mostly solved nearly exactly. For more in-depth review on the topic, readers are encouraged to refer to Refs. [23, 24].

We are interested in eventually bridging the gap between the sophisticated treatment of simplified EPIs typical in model problems and simple treatments using ab initio EPIs. Our tool will be coupled cluster theory which, as introduced in Chapter 2, is a reliable ansatz to treat electronic structure of both molecules and periodic solids. CC theory has also been extended to study the vibrational structure of molecules including anharmonicity[25–34]. Our work in this chapter is inspired by Monkhorst's early proposal on a “molecular coupled cluster" method[35] which seeks to use CC theory for coupled electrons and nuclei in molecules when the Born-Oppenheimer approximation breaks down. In this work we describe a coupled cluster theory and corresponding equation of motion extension for interacting electrons and phonons. This theory is similar to some coupled cluster theories for cavity polaritons that have been independently developed around the same time[36, 37].

The rest of the chapter is organized as follows. In Section 4.3, we describe ground state electron-phonon coupled cluster theory and EOM formalism for excited states for coupled electron-phonon systems. In Section 4.4 we provide benchmark results on Hubbard-Holstein models. In Section 4.5 we discuss the theory for ab initio electron phonon Hamiltonian, provide the details in our implementations and show our results for diamond calculation. We finish the chapter with a conclusion in

Section 4.6.

4.3 Theory

In Section 2.3 we introduced coupled cluster and equation of motion coupled cluster methods for fermion systems. In this section we will first review CC theory for bosons and then formulate the theory for coupled electrons (fermions) and phonons (bosons). In addition to the fermionic operators denoted by $a^\dagger$ and a, we will use $b^\dagger$ (b) to represent bosonic creation (annihilation) operators.

Coupled Cluster for Bosons

Bosonic coupled cluster theory adopts the same exponential form as fermions. However, two different flavors of bosonic CC theory have been proposed with different treatment on the vibrational excitations:

1. Excitations in each mode are treated as bosons such that the nth excited state is an occupation of n bosons[25, 38].
2. Each excited state in each mode is treated as a separate bosonic degree of freedom with the constraint that exactly one state in each mode is occupied[32].

When formulating coupled cluster theory, (1) has the advantage that no truncation of the excitation space beyond the truncation of the T operator is necessary. This means that e^T acting on the vacuum creates up to infinite order excitations with just a finite set of excitation operators. On the other hand, (2) has the advantage to host more general “modal", or, to put in another way, the reference need not be harmonic. Both formulations have been used in vibrational coupled cluster theories[25, 32, 38], and both pictures have been used recently in independent works on coupled cluster methods for molecules interacting with cavity photons[36, 37]. Since we will assume harmonic phonons in this work anyway, we will use second quantization of type (1):

$$|\Psi_{\text{CC}}\rangle = e^T|0\rangle \tag{4.1}$$

$$T = \sum_x t_x b_x^\dagger + \frac{1}{2}\sum_{xy} t_{xy} b_x^\dagger b_y^\dagger + \dots \tag{4.2}$$

where we have used $x, y, \dots$ to index the bosonic modes and $|0\rangle$ represents the bosonic vacuum.

To construct a coupled cluster formalism for electron-phonon systems, we use an exponential ansatz on top of a product reference:

$$|\Psi_{\text{CC}}\rangle = e^{T}|\Phi_0\rangle|0\rangle. \tag{4.3}$$

We will refer to theories of this type as electron-phonon coupled cluster (ep-CC).

Coupled Cluster Models for Electron-phonon Systems

For coupled electron phonon systems, the T operator in CC theory consists of a purely electronic part, purely phononic part, and a coupled part:

$$T = T_{\text{el}} + T_{\text{ph}} + T_{\text{ep}}. \tag{4.4}$$

For each of these pieces, we can independently truncate the level of excitations to create different approximate models of the theory. We will follow the convention in electronic CC to use SDT. . . to specify the electronic amplitudes. Numbers, 123. . ., are used to indicate the purely phononic amplitudes that we include. A combination of letters and numbers are used to denote the coupled amplitudes. The theories considered in this work are summarized in Table 4.1.

model	T_{ph}	T_{ep}
ep-CCSD-1-S1	$t_x b_x^\dagger$	$t^a_{i,x} b_x^\dagger a_a^\dagger a_i$
ep-CCSD-12-S1	$t_x b_x^\dagger + \frac{1}{2} t_{xy} b_x^\dagger b_y^\dagger$	$t^a_{i,x} b_x^\dagger a_a^\dagger a_i$
ep-CCSD-12-S12	$t_x b_x^\dagger + \frac{1}{2} t_{xy} b_x^\dagger b_y^\dagger$	$t^a_{i,x} b_x^\dagger a_a^\dagger a_i + \frac{1}{2} t^a_{i,xy} b_x^\dagger b_y^\dagger a_a^\dagger a_i$

Table 4.1: The names, phonon, and electron-phonon excitation operators for the theories considered in this chapter. All the theories include singles and doubles for the pure electronic part of the T_{el} operator (not shown), and we have omitted the $\sum$ here for simplicity.

Assuming that the number of occupied orbitals, virtual orbitals, and phonon modes all scale linearly with system size N, aall variants of the theory in Table 4.1 have a computational scaling of N^6. Furthermore, only the ep-CCSD-12-S12 method carries additional N^6 steps compared with CCSD. This means that, in theory, the ep-CCSD-1-S1 and ep-CCSD-12-S1 methods have effectively the same cost as CCSD. However in practice, as more amplitudes are included, there are significantly more terms in the amplitude equations, and this creates a practical barrier for efficient implementation. Effectively leveraging spin symmetry, point-group symmetry, k-point symmetry, and reuse of intermediates becomes increasingly difficult. Here we employed the irreducible alignment algorithm developed in Chapter 3 to achieve an implicit spin-unrestricted implementation with our computer-derived equation in general spin orbital basis. Note that our ep-CCSD-1-S1 method is the same as the QED-CCSD-1 method presented in Ref. [37].

Equation of Motion Coupled Cluster

For the coupled system, EOM formalism for electronic excitation follows the same construction as in Section 2.3 with small modification required to account for the coupled excitation with phonons:

$$\begin{aligned} R_{\mathrm{IP}} &= \textstyle\sum_i r_i a_i + \frac{1}{2} \sum_{ija} r_{ij}^{a} a_a^\dagger a_i a_j + \sum_{ix} r_{ix} b_x^\dagger a_i, && (4.5) \\ R_{\mathrm{EA}} &= \textstyle\sum_a r^a a_a^\dagger + \frac{1}{2} \sum_{iab} r_i^{ab} a_b^\dagger a_a^\dagger a_i + \sum_{ax} r_x^a b_x^\dagger a_a^\dagger. && (4.6) \end{aligned}$$

In practice, the eigenvalue problem is solved by iterative diagonalization.

4.4 Benchmark on Hubbard-Holstein Model

In order to understand the strengths of our method, we first perform a benchmark study on the Hubbard-Holstein (HH) model, a simple lattice model of correlated electrons and phonons. The Hubbard-Holstein Hamiltonian is

$$H = -t \sum_{j\sigma} \left(a_{(j+1)\sigma}^\dagger a_{j\sigma} + \text{h.c.} \right) + U \sum_j n_{j\uparrow} n_{j\downarrow} + \omega \sum_J b_J^\dagger b_J + g \sum_j n_j (b_J + b_J^\dagger). \quad (4.7)$$

where the lowercase and capital indices run over the fermionic and bosonic degrees of freedom at each lattice site and σ represents the spin degrees of freedom of the fermions. The fermionic part of the Hamiltonian is a Hubbard model with hopping t and on-site repulsion U. We use n_i to represent the fermionic density at site i. The bosonic part of the Hamiltonian is an independent oscillator at each site with frequency ω, and the final term couples the fermionic density at a given site with a linear displacement in the oscillator at that site. This coupling is controlled by g.

The HH model is an important model in condensed matter physics as it captures both antiferromagnetic order due to electron correlation and pairing from the electron-phonon interaction[5–8, 39–53]. As a minimal model of electron correlation and electron-phonon coupling, it is an ideal benchmark testbed with which we can evaluate the performance of our coupled cluster models in different parameter regimes.

The electron-phonon coupling strength,

$$\lambda \equiv \frac{g^2}{\omega}, \tag{4.8}$$

provides a measure of the effective strength of the electron-phonon interaction. Using a path-integral method, the phonon degrees of freedom can be integrated out to yield an effective electron-electron interaction, the static limit of which becomes attractive when

$$\lambda = \frac{U}{2}. \tag{4.9}$$

Note that our definition of λ may differ by a factor of 2 from some other common definitions. For the large coupling regime, the effective electron-electron interaction is attractive, and we would not expect our coupled cluster methods to perform well for such an attractive interaction. The extension to this regime should be possible by breaking particle number symmetry[54, 55], but this is beyond the scope of our work in this chapter.

Benchmark of Ground-state Methods

The four-site (linear) HH model at half-filling is numerically solvable by exact diagonalization. We thus first computed the correlation energy from three CC methods and compare them with exact correlation energy. In all cases, we use an unrestricted Hartree-Fock (UHF) reference and a generalized coherent state reference for the phonons,

$$\tilde{b}_I = b_I + g\frac{\langle \Phi_0 | n_i | \Phi_0 \rangle}{\omega}, \tag{4.10}$$

where Φ_0 is the electronic UHF reference. In terms of these transformed boson operators, the Hamiltonian has the same form except that the interaction term appears as

$$g\sum_j (n_j - \langle n_j \rangle)(\tilde{b}_J + \tilde{b}_J^\dagger), \tag{4.11}$$

and there is an energy shift of

$$-g^2\sum_i \frac{\langle n_i \rangle^2}{\omega}. \tag{4.12}$$

This transformation diagonalizes the effective phononic Hamiltonian obtained by normal ordering the electronic part of the EPI term.

In addition to the coupled cluster methods, we also show the energy computed by adding a second-order perturbation theory (PT2) correction to the fermionic CCSD

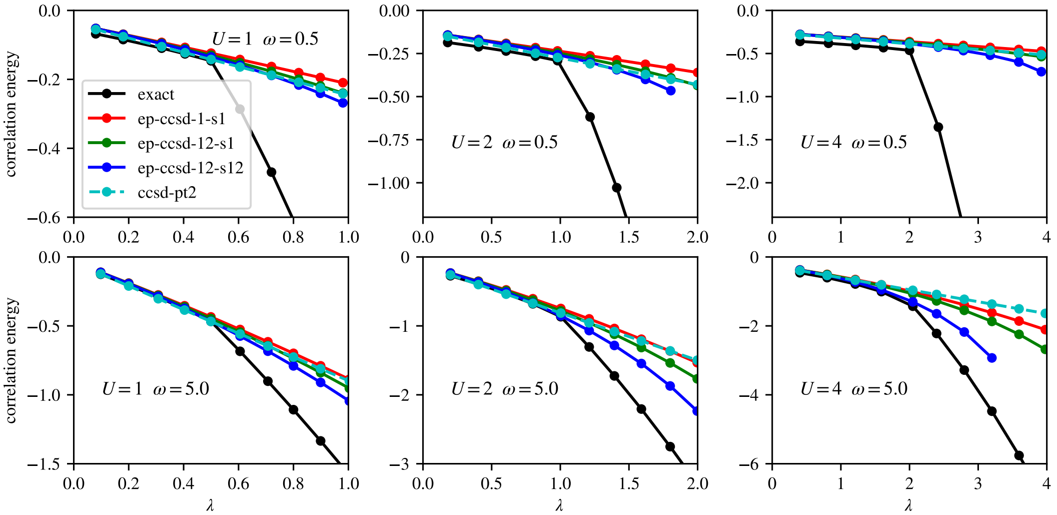

Figure 4.1: Correlation energy of the four-site HH model for $U = 1, 2, 4$ and $\omega = 0.5, 5.0$. In all cases we found a qualitative change at $\lambda = 0.5U$ which is not captured by the approximate methods presented here. Both the energy and the coupling strength λ are plotted in units of the hopping, t.

energy. This correlation energy is given, in the UHF orbital basis, as

$$E_{\text{pt2}} = -\sum_{ia,I} \frac{|g^{a}_{i,I}|^2}{\varepsilon_a - \varepsilon_i + \omega}, \tag{4.13}$$

where i (a) are occupied (virtual) orbitals and I runs over the oscillators. Note that the interaction, g, becomes a generally non-diagonal tensor in the UHF orbital basis. We label the elements of this tensor as $g^{a}_{i,I}$ where I labels an oscillator and i, a label occupied and virtual UHF orbitals respectively.

The correlation energy computed from these methods is compared to the exact results in Figure 4.1 for $U = 1, 2, 4$ and $\omega = 0.5, 5.0$. The values of U are chosen to be low enough that CCSD should provide qualitatively correct results in the limit of zero EPI, while the two values of ω are chosen to show approximately the limits of low frequency (adiabatic) and high frequency (anti-adiabatic). The transition to an attractive effective potential at $\lambda = U/2$ is evident in all cases, and the approximate methods described here fail qualitatively above this transition as expected.

For $\lambda < U/2$, all the methods shown here provide qualitatively correct results in the adiabatic and anti-adiabatic limits. The coupled cluster methods are systematic in that ep-CCSD-12-S12 outperforms ep-CCSD-12-S1 which outperforms ep-CCSD-1-S1 in all cases. This is one of the primary advantages of coupled cluster theory. The CCSD-PT2 method performs surprisingly well on this problem, possibly due to error cancellation as Equation 4.13 tends to overestimate the electron-phonon correlation energy while CCSD generally underestimates the electronic correlation energy.

EOM-ep-CCSD-1-S1 for Charge Gap

We then turn our focus to the accuracy of excited states properties from EOM variant of the theory.

In the thermodynamic limit, the 1-dimensional HH model at half-filling has a well-studied phase diagram: a Mott phase at small λ/U, a Peierls phase at large λ/U, and a metallic phase in between[7, 47–51, 56].

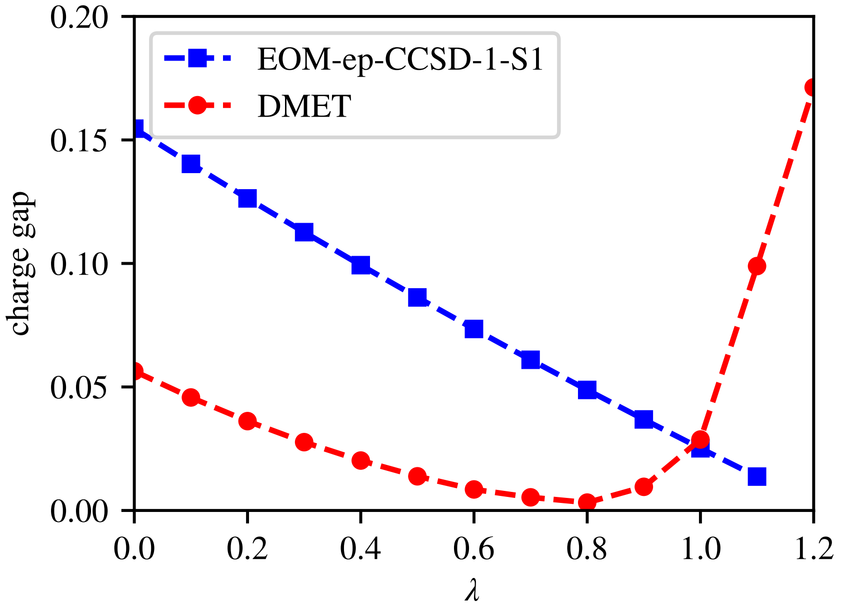

Figure 4.2: Charge gap of the HH model in the thermodynamic limit for $U = 1.6$ and $\omega = 0.5$. At $\lambda = 0.8$ ep-CCSD-1-S1 will break down, and we would not expect correct results for $\lambda > 0.8$. The density matrix embedding theory (DMET) results are from Ref. [57].

In Figures 4.2 and 4.3 we show the charge gap computed by IP/EA-EOM-ep-CCSD-1-S1 in the adiabatic case and the anti-adiabatic case respectively. In Figure 4.2 we present the extrapolated EOM-ep-CCSD-1-S1 band gap for $\omega = 0.5$ and $U = 1.6$. We used calculations results on the $L = 64$ and $L = 128$ systems with periodic boundary conditions for extrapolation assuming an asymptotica scaling of $1/L$. At $\lambda = 0$ (the $U = 1.6$ Hubbard model, our results agrees poorly with the DMET reference [57]. This is consistent with the general observation that EOM coupled

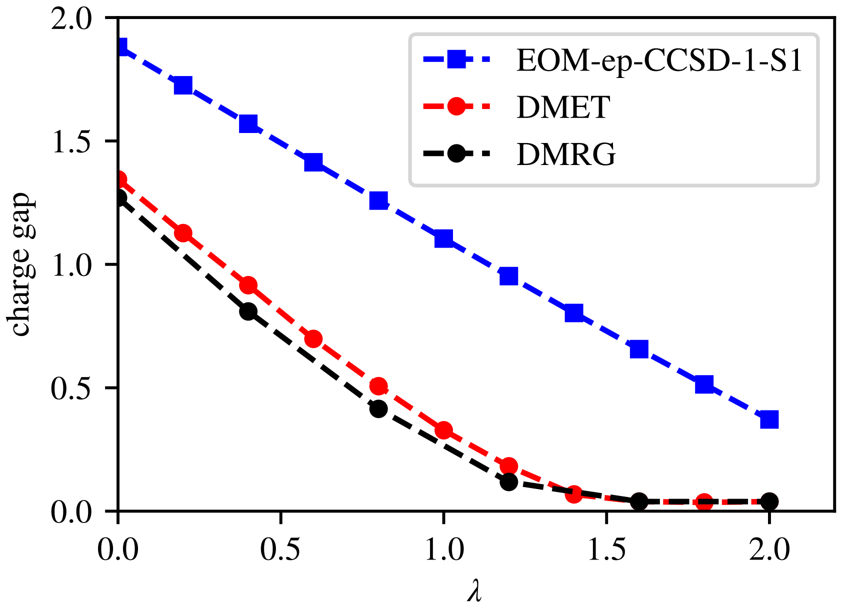

Figure 4.3: Charge gap of the HH model in the thermodynamic limit for $U = 4$ and $\omega = 5.0$. At $\lambda = 2.0$ ep-CCSD-1-S1 will break down, and we would not expect correct results for $\lambda > 2.0$. The density matrix embedding theory (DMET) and density matrix renormalization group (DMRG) calculations are from Ref. [57].

cluster tends to perform poorly on nearly metallic systems. As λ increases, we find the results to be qualitatively correct though the closing of the gap at $\lambda = 0.6$ and the Peierls insulating state at $\lambda > 0.6$ are not captured by this approximation. In particular, note that EOM-ep-CCSD-1-S1 for the HH model does not perform worse than EOM-CCSD for the Hubbard model. We find a similar performance in the anti-adiabatic case (as shown in Figure 4.3: $\omega = 5.0$ and $U = 4$). In this case, finite-size effects are less pronounced, and an extrapolation from calculations on $L = 32$ and $L = 64$ lattices is sufficient to estimate the TDL. Because of the larger U, the Hamiltonian has a larger gap at $\lambda = 0$ and it is less severely overestimated by EOM. Again, similar to the adabatic case, we qualitatively correct results for small λ, but EOM breaks down as the system becomes metallic.

4.5 Application to Periodic Solids

To extend our theory to ab initio problems, a Hamiltonian of the following form is required:

$$H = H_{\text{el}} + H_{\text{ph}} + H_{\text{ep}}, \tag{4.14}$$

where H_{ep} is both detailed enough to capture the physics of electron-phonon coupling from first principles and yet simple enough so that the matrix elements can be easily computed in the relevant basis. As we will describe later this chapter, this is already quite a challenge. This is further complicated by the cost of controlling finite-size errors. In this section we will first discuss our frozen-phonon implementation of phonon frequencies and EPI matrix elements using cGTO of the PySCF package. Then we will show our preliminary results for the zero-point renormalization of diamond. We will conclude this section with a summary of the challenges and future plans for addressing them.

Ab Initio Electron Phonon Coupling

Nearly all ab initio calculations includes only the linear coupling of the EPI term:

$$\sum_{\mathbf{kq}mnx} g^{\mathbf{q}x}_{(\mathbf{k}+\mathbf{q})n,\mathbf{k}m} c^{\dagger}_{(\mathbf{k}+\mathbf{q})n} c_{\mathbf{k}m} \left(b_{\mathbf{q}x} + b^{\dagger}_{-\mathbf{q}x} \right). \tag{4.15}$$

Here, m and n label the electronic bands and x labels the phonon modes. The EPI matrix elements are, in practice, computed as

$$g^{x}_{pq} = \sum_{\alpha,s} \sqrt{\frac{\hbar}{2m_s\omega_x}} \epsilon^{x}_{s\alpha} \left\langle p \middle| \frac{dV_{KS}}{dR_{s\alpha}} \middle| q \right\rangle \tag{4.16}$$

where we have suppressed the momentum indices in this expression. Here, V_{KS} is the Kohn-Sham (or Hartree-Fock) potential, s labels a particular atom, α labels a Cartesian direction, m_s is the mass of the sth atom, ω_x is the frequency of the xth phonon mode, and the ϵ tensor transforms between Cartesian displacements and displacements in the phonon basis. One thing to note here is that when using a Hamiltonian of this form, two approximations are made implicitly: (1) Higher order coupling, like the term quadratic in displacements are ignored. In principle, this approximation can be relaxed by including the higher order terms. (2) The phonon modes come from a calculation that already includes, to some extent, the response of the ground state electronic energy to changes in the nuclear positions. Relaxing this approximation is difficult. One option would be to work within the self-consistent field-theoretic framework of the Hedin-Baym equations[58, 59]. This issue is discussed in more detail in Ref. [60]. Overall, the construction of second-quantized model Hamiltonians for coupled electron-nuclear dynamics in molecules, including investigations of the validity of a linear coupling, has been an area of recent interest,[61–63] though we are not aware of similar work on solids. As a starting point, we will use the standard linear coupling.

Implementation

In order to test the performance of this coupled cluster method in ab initio setting, we have implemented the first-order electron phonon matrix for molecules and extended systems in the PySCF program package[64]. The molecular implementation computes the analytical EPI matrix through the coupled-perturbed self-consistent field (CPSCF) formalism, similar to the implementation in FHI-AIMS[65]. The periodic system implementation is based on a finite difference approach and currently supports only a single k-point. Specifically, finite differentiation is first performed on analytical nuclear gradients to yield the mass weighted hessian (dynamical matrix). Phonon modes are then obtained by diagonalizing this matrix.

Throughout this work, we have used GTH-Pade pseudopotentials[66, 67] and the corresponding GTH Gaussian bases[68]. All integrals are generated by Fast Fourier transform-based density fitting (FFTDF)[69]. In Table 4.2 we compare the optical phonon frequency computed at the Γ point using different basis sets. Note that for our TZVP calculations, basis Gaussians with exponents less than 0.1 are discarded due to the diffuse nature of the functions. Amid the discrepancies in basis sets, pseudopotentials, and other numerical cutoffs, our results show overall good agreement with the implementations in CP2K[70] and the plane-wave (PW) code Quantum Espresso (QE)[71].

ω_{Γ}^{OP}	PySCF	CP2K	QE
GTH-SZV(LDA)	2385.56	2393.30	-
GTH-DZVP(LDA)	2207.67	2214.58	-
GTH-TZVP(LDA)	2290.95	2221.85	2262.67(PW)
GTH-SZV(PBE)	2379.07	2384.69	-
GTH-DZVP(PBE)	2202.70	2209.07	-
GTH-TZVP(PBE)	2288.15	2212.95	2255.60(PW)

Table 4.2: A comparison of the Γ point optical phonon mode (cm^{-1}) from our implementation in PySCF against those computed from CP2K and QE. Note that PySCF and CP2K use the same GTH pseudopotentials, while Hartwigsen-Goedeker-Hutter (HGH) pseudopotentials[67] were used for QE. The QE calculations use a kinetic energy cutoff of 60 Rydberg. To ensure that the QE and PySCF numbers can be directly compared, the electron density used for the QE DFPT computation is from a Γ point DFT calculation (unconverged with respect to Brillouin zone sampling).

Experimentally, the optical phonons of diamond appear around 1300 cm^{-1} and this is consistent with calculations using large supercells (see for example Refs. [72, 73]). While this does not affect the comparison between different implementations, it clearly suggest a significant finite size error associated with the 1x1x1 cell.

The evaluation of the Kohn-Sham response matrix is broken into three terms:

$$\left\langle p \middle| \frac{dV_{KS}}{dR_{s\alpha}} \middle| q \right\rangle = \frac{d}{dR_{s\alpha}} \left\langle p \middle| V_{KS} \middle| q \right\rangle - \left\langle \frac{dp}{dR_{s\alpha}} \middle| V_{KS} \middle| q \right\rangle - \left\langle p \middle| V_{KS} \middle| \frac{dq}{dR_{s\alpha}} \right\rangle. \tag{4.17}$$

The first term is evaluated by finite difference. The second and third terms are obtained analytically as part of normal nuclear gradient routine. In our implementation, the response matrix is first evaluated in the AO basis and then transformed to the MO basis when needed. This is to avoid problems arising from different MO gauges that can occur in finite-difference calculations. Our implementation differs from standard PW codes in that the electron density and MO basis are converged in the same SCF procedure.

To allow for easier comparison of our implementation against PW-based codes, we take the occupied block of the potential response matrix as

$$Z_{ij}^{s\alpha} = \left\langle i \middle| \frac{dV_{KS}}{dR_{s\alpha}} \middle| j \right\rangle \tag{4.18}$$

and define a gauge- and basis-independent z metric for comparisons:

$$z = \mathrm{Tr}\, Z^{\dagger} Z. \tag{4.19}$$

In Table 4.3 we compare our results for the z metric with those from a PW implementation. For the PW reference, DFT/DFPT results from QE are used by Perturbo[74] to extract the potential response matrix. For our Gaussian basis implementation, a slow basis convergence behavior is observed moving from DZVP to TZVP, but again, given the differences in many numerical choices, our results in the TZVP basis are qualitatively similar to those from the PW reference.

z	GTH-SZV	GTH-DZVP	GTH-TZVP	PW
LDA	0.0864	0.1639	0.1768	0.2278
PBE	0.0841	0.1631	0.1739	0.2260

Table 4.3: z metric (E_h) of diamond computed in a cGTO basis (PySCF) compared to results from QE/Perturbo computed in a PW basis.

In order to enable large simulations using ab initio Hamiltonians, the following strategies are adopted to optimize our ep-CC Python implementation:

1. We take advantage of the irreducible alignment algorithm we developed in Chapter 3 to obtain an implicitly unrestricted implementation starting from generalized spin-orbital equations from our code generator.
2. We used the Cyclops Tensor Framework[75] as our numerical backend, and this allows efficient tensor contractions in distributed parallel setting.

Results on Diamond

Diamond has emerged as a paradigmatic example in the field of ab initio electron-phonon computation, and the accurate computation of relatively simple quantities, like the zero-point renormalization (ZPR) (the shift of the bandgap due to phonon effects) remains a challenge. Experimental values based on isotopic shifts suggest a ZPR of the indirect gap of -364 meV[76]. Calculations of the ZPR of the direct gap suggest that it is higher, closer to -600 meV[18, 77, 78]. Importantly, it has been shown that many-body electronic effects are important to the ZPR of the direct gap[18] and that dynamical effects are important to capture some qualitative features of the EPI[79]. We here provide a summary of previous theoretical and experimental results in Table 4.4.

ZPR	EPI	electronic structure	ZPR	gap	reference
-700	-	tight-binding	PIMC	direct	[77]
-615	LDA	LDA	AHC	direct	[78]
-628	LDA	GW	AHC	direct	[18]
-334	-	LDA	Ref. [80]	indirect	[80]
-345	-	LDA	WL	indirect	[81]
-337	-	GW	MC	indirect	[82]
-364	-	Experiment	-	indirect	[76]

Table 4.4: Selected literature results for the ZPR of diamond. Monte Carlo is abbreviated as MC. Path integral molecular dynamics is abbreviated as PIMD, Allen-Heine-Cordona[83, 84] theory is abbreviated as AHC, and the theory of Williams[85] and Lax[86] is abbreviated as WL. The method used to get the ZPR in Ref. [80] does not have a commonly used name, but it is clearly described in given reference.

In Table 4.5 we present the ZPR of diamond computed by IP/EA-EOM-ep-CCSD-1-S1 and IP/EA-EOM-CCSD-PT2. The EPI matrix elements and phonon frequencies are computed from Hartree-Fock calculations. It was necessary to remove the most diffuse s orbital from the GTH-DZVP basis and the most diffuse s and p orbitals from the GTH-TZVP basis in order to eliminate numerical instabilities in the calculation of the EPI matrix elements.

The experimental lattice constant of diamond, 3.566Å, is used throughout. For EOM-CCSD-PT2, the electronic CCSD amplitudes are used along with a PT2 estimate of the electron-phonon amplitudes:

$$t_{i,x}^{a} = -\frac{g_{i,x}^{a}}{\varepsilon_a - \varepsilon_i + \omega_x}. \tag{4.20}$$

The quantities in Table 4.5 are directly comparable to the ZPR of the direct gap which has recently been reported to be in the range of -600 to -700 meV[18, 77, 78]. However, the very small size of our simulation cell means that these numbers require some estimate of the finite-size error for a meaningful comparison with experiments. In diamond, the finite size effects are significant. However, the strength of coupled cluster methods is that they explicitly treat many-body electronic effects as well as dynamical electron-phonon correlation in a consistent framework. Thus recomputation using the approximate literature treatments within the same smaller cells would allow for the magnitude of higher-order many-body effects to be estimated from these CC calculations.

Basis	CCSD-1-S1		CCSD-PT2	
	full	no-VV	full	no-VV
GTH-SZV	-671	-366	-671	-366
GTH-DZVP*	-831	-617	-826	-512
GTH-TZVP*	-1343	-767	-1115	-645

Table 4.5: Band gap renormalization (meV) at the Γ point (direct gap) for a 1x1x1 unit cell in different basis sets. Note that the most diffuse s orbital was removed from the GTH-DZVP basis and the most diffuse s and p orbitals were removed from the GTH-TZVP basis. In the “no-VV" columns, the unoccupied-unoccupied EPI matrix elements were discarded, which forms a more direct comparison with typical treatments of band-gap renormalization.

We can draw two conclusions from these finite-size ep-CC calculations. First, using the PT2 estimate of the coupled amplitudes provides EOM results that are comparable to the converged CC results, though the results from converged CC amplitudes are consistently lower. This suggests that the converged CC ground state is probably not necessary to obtain reasonable excited-state properties of typical large-gap insulators. Second, we find that the band-gap renormalization becomes unexpectedly large as the size of the basis set is increased. This affect can be mostly traced to the unoccupied-unoccupied (virtual-virtual, or VV) block of the electron-phonon matrix elements which do not appear in the widely used Allen-Heine-Cordona (AHC) treatment[83, 84]. This could indicate that the Hamiltonian of Equation 4.15 does not properly describe the electron-phonon coupling between unoccupied bands which does not enter into typical calculations in the material community. Alternatively, it could be due to the small finite size of the simulation.

Results for a larger supercell are shown in Table 4.6.

supercell	CCSD-1-S1	CCSD-PT2
1x1x1	-671	-671
2x2x2	-134	-142
3x3x3	-	-42

Table 4.6: ZPR (meV) of diamond supercells in the GTH-SZV basis set. The 2x2x2 and 3x3x3 supercells provide estimates of the indirect band gap renormalization. In the 3x3x3 supercell, we were unable to obtain converged CCSD-1-S1 amplitudes.

These results are not constrained to compute the direct gap, so the results for 2x2x2 and 3x3x3 supercells should be viewed as finite-size approximations to the ZPR of the indirect bandgap. These results affirm that using CCSD-PT2 amplitudes in the EOM calculation is a reasonable approximation. The ZPR is smaller for larger supercells which is consistent with the smaller ZPR for the indirect gap. Though the simulation cell is still too small for a reliable extrapolation, the numbers are consistent in magnitude with results that have been reported in the literature. The slow and oscillatory convergence of the ZPR of diamond with supercell size is a well-known problem[80, 81, 87].

Future Directions for Ab Initio Calculations

In the previous section, we identified two significant sources of error in our CC calculations which explicitly include EPI: (1) The finite-size error which is difficult to control since the CC equations must be solved simultaneously for all electronic and phononic degrees of freedom. (2) The form of the EPI term, which may be insufficient, especially for the unoccupied bands.

We intend to address the finite-size error by using a perturbative correction to EOM-CCSD eigenvalues which can be interpolated to denser k-point grids as is usually done in traditional calculations of EPI. The coupled cluster framework presented here will be useful in evaluating the validity of these perturbative approximations.

The validity of the linear EPI term also needs to be investigated further as does the omission of anharmonic effects. This requires very accurate calculations on small systems or model systems, and we expect this coupled cluster framework to be useful in that it can provide more systematic results for such problems.

4.6 Conclusion

We have presented a coupled cluster framework for a systematic, correlated treatment of interacting electrons and phonons. The theory is a straightforward combination of fermionic (electronic) and bosonic (phononic) coupled cluster ansatz. Despite the formal simplicity of the ansatz, sophisticated diagrammatic techniques and automated operator algebra were necessary to efficiently implement the equations. These techniques are described in the appendices. In order to benchmark these methods, we have applied them to the Hubbard-Holstein model. Calculations on the four-site HH model, which can be exactly solved numerically, reveal that all the CC methods discussed here perform well for small to moderate coupling. Calculations of the excited-state properties of the model suggest that the EOM-ep-CC methods can provide excited-state energies with an accuracy comparable to EOM-CC for electronic excitations.

Finally we have discussed the details of an ab initio implementation in the context of crystalline Gaussian-type orbitals. Preliminary calculations on the ZPR of diamond are consistent with values reported in the literature, but a better treatment of finite-size error is necessary for truly quantitative calculations. This motivates the future development of more approximate theories that can utilize EPI matrix elements interpolated onto a very fine momentum-space grid. We found unexpectedly large values for the ZPR when coupling between virtual bands was included in the calculations which suggests that the approximate, linear form of the EPI may not be sufficient in the more sophisticated many-body treatments of electron-phonon effects where these states must enter.

References

[1] H. Fröhlich. "Electrons in lattice fields". In: *Adv. Phys.* 3.11 (1954), pp. 325–361. ISSN: 14606976. DOI: 10.1080/00018735400101213.

[2] T. Holstein. "Studies of polaron motion. Part I. The molecular-crystal model". In: *Ann. Phys.* 8.3 (1959), pp. 325–342. ISSN: 1096035X. DOI: 10.1016/0003-4916(59)90002-8.

[3] W. P. Su, J. R. Schrieffer, and A. J. Heeger. "Solitons in polyacetylene". In: *Phys. Rev. Lett.* 42.25 (1979), pp. 1698–1701. ISSN: 00319007. DOI: 10.1103/PhysRevLett.42.1698.

[4] M. Z. Hasan and C. L. Kane. “Colloquium: Topological insulators”. In: *Rev. Mod. Phys.* 82.4 (2010), pp. 3045–3067. ISSN: 00346861. DOI: `10.1103/RevModPhys.82.3045`.

[5] G Beni, P. Pincus, and J. Kanamori. “Low-temperature properties of the one-dimensional polaron band. I. Extreme-band-narrowing regime”. In: *Phys. Rev. B* 10.5 (1974), pp. 1896–1901.

[6] E. Berger, P. Valáek, and W. Von Der Linden. “Two-dimensional Hubbard-Holstein model”. In: *Phys. Rev. B* 52.7 (1995), pp. 4806–4814. ISSN: 01631829. DOI: `10.1103/PhysRevB.52.4806`.

[7] Johannes Bauer and Alex C. Hewson. “Competition between antiferromagnetic and charge order in the Hubbard-Holstein model”. In: *Phys. Rev. B* 81.23 (2010), pp. 1–17. ISSN: 10980121. DOI: `10.1103/PhysRevB.81.235113`.

[8] E. A. Nowadnick et al. “Competition between antiferromagnetic and charge-density-wave order in the half-filled Hubbard-Holstein model”. In: *Phys. Rev. Lett.* 109.24 (2012). ISSN: 00319007. DOI: `10.1103/PhysRevLett.109.246404`. arXiv: `1209.0829`.

[9] Michel M Dacorogna, Marvin L Cohen, and Pui K. Lam. “Self-Consistent Calculation of the q Dependence of the Electron-Phonon Coupling in Aluminum”. In: *Phys. Rev. Lett.* 55.8 (1985), pp. 837–840. ISSN: 00319007. DOI: `10.1103/PhysRevLett.55.837`.

[10] Michel M. Dacorogna, K. J. Chang, and Marvin L. Cohen. “Pressure increase of the electron-phonon interaction in superconducting hexagonal silicon”. In: *Phys. Rev. B* 32.3 (1985), pp. 1853–1855. ISSN: 01631829. DOI: `10.1103/PhysRevB.32.1853`.

[11] Pui K. Lam, Michel M. Dacorogna, and Marvin L. Cohen. “Self-consistent calculation of electron-phonon couplings”. In: *Phys. Rev. B* 34.8 (1986), pp. 5065–5069. ISSN: 01631829. DOI: `10.1103/PhysRevB.34.5065`.

[12] Stefano Baroni, Paolo Giannozzi, and Andrea Testa. “Green’s-function approach to linear response in solids”. In: *Phys. Rev. Lett.* 58 (18 1987), pp. 1861–1864. DOI: `10.1103/PhysRevLett.58.1861`. URL: `https://link.aps.org/doi/10.1103/PhysRevLett.58.1861`.

[13] Xavier Gonze, Douglas C. Allan, and Michael P. Teter. “Dielectric tensor, effective charges, and phonons in -quartz by variational density-functional perturbation theory”. In: *Phys. Rev. Lett.* 68.24 (1992), pp. 3603–3606. ISSN: 00319007. DOI: `10.1103/PhysRevLett.68.3603`.

[14] S. Y. Savrasov. “Linear-Response Calculations of Lattice Dynamics Using Muffin-Tin Basis Sets”. In: *Phys. Rev. Lett.* 69.19 (1992), pp. 2819–2822.

[15] Michele Lazzeri et al. "Impact of the electron-electron correlation on phonon dispersion: Failure of LDA and GGA DFT functionals in graphene and graphite". In: *Phys. Rev. B* 78.8 (2008), pp. 8–11. ISSN: 10980121. DOI: 10.1103/PhysRevB.78.081406. arXiv: 0808.2285.

[16] Carina Faber et al. "Electron-phonon coupling in the C60 fullerene within the many-body GW approach". In: *Phys. Rev. B* 84.15 (2011), pp. 30–34. ISSN: 10980121. DOI: 10.1103/PhysRevB.84.155104. arXiv: 1109.0885.

[17] Z. P. Yin, A. Kutepov, and G. Kotliar. "Correlation-enhanced electron-phonon coupling: Applications of GW and screened hybrid functional to bismuthates, chloronitrides, and other high-Tc superconductors". In: *Phys. Rev. X* 3.2 (2013), pp. 1–20. ISSN: 21603308. DOI: 10.1103/PhysRevX.3.021011. arXiv: 1110.5751.

[18] G. Antonius et al. "Many-body effects on the zero-point renormalization of the band structure". In: *Phys. Rev. Lett.* 112.21 (2014), pp. 1–5. ISSN: 10797114. DOI: 10.1103/PhysRevLett.112.215501.

[19] C. Faber et al. "Exploring approximations to the GW self-energy ionic gradients". In: *Phys. Rev. B* 91.15 (2015), pp. 1–9. ISSN: 1550235X. DOI: 10.1103/PhysRevB.91.155109. arXiv: 1501.07058.

[20] Bartomeu Monserrat. "Correlation effects on electron-phonon coupling in semiconductors: Many-body theory along thermal lines". In: *Phys. Rev. B* 93.10 (2016), pp. 1–6. ISSN: 24699969. DOI: 10.1103/PhysRevB.93.100301. arXiv: 1603.00551.

[21] Zhenglu Li et al. "Electron-Phonon Coupling from Ab Initio Linear-Response Theory within the GW Method: Correlation-Enhanced Interactions and Superconductivity in Ba1-x KxBiO3". In: *Phys. Rev. Lett.* 122.18 (2019), p. 186402. ISSN: 10797114. DOI: 10.1103/PhysRevLett.122.186402. URL: https://doi.org/10.1103/PhysRevLett.122.186402.

[22] Feliciano Giustino, Marvin L. Cohen, and Steven G. Louie. "Electron-phonon interaction using Wannier functions". In: *Phys. Rev. B* 76.16 (2007), pp. 1–19. ISSN: 10980121. DOI: 10.1103/PhysRevB.76.165108.

[23] Feliciano Giustino. "Electron-phonon interactions from first principles". In: *Rev. Mod. Phys.* 89.1 (2017), pp. 1–63. ISSN: 15390756. DOI: 10.1103/RevModPhys.89.015003. arXiv: 1603.06965.

[24] Marco Bernardi. "First-principles dynamics of electrons and phonons". In: *European Physical Journal B* 89 (2016), p. 239. ISSN: 14346036. DOI: 10.1140/epjb/e2016-70399-4. arXiv: 1607.00080.

[25] M. Durga Prasad. "Time-dependent coupled cluster method: A new approach to the calculation of molecular absorption spectra". In: *J. Chem. Phys.* 88.11 (1988), pp. 7005–7010. ISSN: 00219606. DOI: 10.1063/1.454399.

[26] V. Nagalakshmi et al. “Coupled cluster description of anharmonic molecular vibrations. Application to O3 and SO2”. In: *Chem. Phys. Lett.* 217.3 (1994), pp. 279–282. ISSN: 00092614. DOI: 10.1016/0009-2614(93)E1380-Y.

[27] G. Madhavi Sastry and M. Durga Prasad. “The time-dependent coupled cluster approach to molecular photodissociation dynamics”. In: *Chem. Phys. Lett.* 228.1-3 (1994), pp. 213–218. ISSN: 00092614. DOI: 10.1016/0009-2614(94)00934-1.

[28] G. Sree Latha and M. Durga Prasad. “Time-dependent coupled cluster approach to multimode vibronic dynamics”. In: *J. Chem. Phys.* 105.8 (1996), pp. 2972–2977. ISSN: 00219606. DOI: 10.1063/1.472170.

[29] M. Durga Prasad. “Time dependent coupled cluster approach to Resonance Raman excitationprofiles from general anharmonic surfaces”. In: *Int. J. Mol. Sci.* 3.5 (2002), pp. 447–458. ISSN: 14220067. DOI: 10.3390/i3050447.

[30] Subrata Banik, Sourav Pal, and M. Durga Prasad. “Calculation of vibrational energy of molecule using coupled cluster linear response theory in bosonic representation: Convergence studies”. In: *J. Chem. Phys.* 129.13 (2008). ISSN: 00219606. DOI: 10.1063/1.2982502.

[31] Subrata Banik, Sourav Pal, and M. Durga Prasad. “Calculation of dipole transition matrix elements and expectation values by vibrational coupled cluster method”. In: *J. Chem. Theory Comput.* 6.10 (2010), pp. 3198–3204. ISSN: 15499618. DOI: 10.1021/ct1003669.

[32] Ove Christiansen. “Vibrational coupled cluster theory”. In: *J. Chem. Phys.* 120.5 (2004), pp. 2149–2159. ISSN: 00219606. DOI: 10.1063/1.1637579.

[33] Peter Seidler and Ove Christiansen. “Vibrational excitation energies from vibrational coupled cluster response theory”. In: *J. Chem. Phys.* 126.20 (2007). ISSN: 00219606. DOI: 10.1063/1.2734970.

[34] Peter Seidler, Eduard Matito, and Ove Christiansen. “Vibrational coupled cluster theory with full two-mode and approximate three-mode couplings: The VCC[2pt3] model”. In: *J. Chem. Phys.* 131.3 (2009). ISSN: 00219606. DOI: 10.1063/1.3158946.

[35] Hendrik J. Monkhorst. “Chemical physics without the Born-Oppenheimer approximation: The molecular coupled-cluster method”. In: *Phys. Rev. A* 36.4 (1987), pp. 1544–1561. ISSN: 10502947. DOI: 10.1103/PhysRevA.36.1544.

[36] Uliana Mordovina et al. “Polaritonic coupled-cluster theory”. In: *Phys. Rev. Res.* 2.2 (2020), pp. 1–8. ISSN: 2643-1564. DOI: 10.1103/physrevresearch.2.023262. arXiv: 1909.02401.

[37] Tor S. Haugland et al. “Coupled Cluster Theory for Molecular Polaritons: Changing Ground and Excited States”. In: *arXiv:2005.04477 [physics.chem-ph]* (2020), pp. 1–18. arXiv: 2005.04477. URL: http://arxiv.org/abs/2005.04477.

[38] Jacob A. Faucheaux and So Hirata. “Higher-order diagrammatic vibrational coupled-cluster theory”. In: *J. Chem. Phys.* 143.13 (2015). ISSN: 00219606. DOI: 10.1063/1.4931472.

[39] F. Guinea. “Local many-body effects in one dimension”. In: *J. Phys. C.* 16.22 (1983), pp. 4405–4413. ISSN: 00223719. DOI: 10.1088/0022-3719/16/22/015.

[40] E Hirsch and Eduardo Fradkin. “Phase diagram of one-dimensional electron phonon systems”. In: *Phys. Rev. B* 27.7 (1983), pp. 4302–4316.

[41] Laurent G Caron and Claude Bourbonnais. “Two-cutoff renormalization and quantum versus classical aspects for the one-dimensional electron-phonon system”. In: *Phys. Rev. B* 29.8 (1984), pp. 4230–4241. ISSN: 01631829. DOI: 10.1103/PhysRevB.29.4230.

[42] J. E. Hirsch. “Phase diagram of the one-dimensional molecular-crystal model with Coulomb interactions: Half-filled-band sector”. In: *Phys. Rev. B* 31.9 (1985), pp. 6022–6031.

[43] H. Zheng, D. Feinberg, and M. Avignon. “Quantum lattice fluctuations in the one-dimensional Peierls-Hubbard model”. In: *Phys. Rev. B* 41.16 (1990), pp. 11557–11563. ISSN: 01631829. DOI: 10.1103/PhysRevB.41.11557.

[44] K. Yonemitsu and M. Imada. “Spin-gap phase in nearly half-filled one-dimensional conductors coupled with phonons”. In: *Phys. Rev. B* 54.4 (1996), pp. 2410–2420. ISSN: 1550235X. DOI: 10.1103/PhysRevB.54.2410.

[45] A. La Magna and R. Pucci. “Mobile intersite bipolarons in the discrete Holstein-Hubbard model”. In: *Phys. Rev. B* 55.22 (1997), pp. 14886–14891. ISSN: 1550235X. DOI: 10.1103/PhysRevB.55.14886.

[46] C. Pao and H. Schüttler. “Superconducting instability in the Holstein-Hubbard model: A numerical renormalization-group study”. In: *Phys. Rev. B* 57.9 (1998), pp. 5051–5054. ISSN: 1550235X. DOI: 10.1103/PhysRevB.57.5051.

[47] R. T. Clay and R. P. Hardikar. “Intermediate phase of the one dimensional half-filled Hubbard-Holstein model”. In: *Phys. Rev. Lett.* 95.9 (2005), pp. 1–4. ISSN: 00319007. DOI: 10.1103/PhysRevLett.95.096401.

[48] W. Koller et al. “Phase diagram and dynamic response functions of the Holstein-Hubbard model”. In: *Physica B* 359-361.SPEC. ISS. (2005), pp. 795–797. ISSN: 09214526. DOI: 10.1016/j.physb.2005.01.230. arXiv: 0406241 [cond-mat].

[49] R. P. Hardikar and R. T. Clay. “Phase diagram of the one-dimensional Hubbard-Holstein model at half and quarter filling”. In: *Phys. Rev. B* 75.24 (2007), pp. 1–10. ISSN: 10980121. DOI: 10.1103/PhysRevB.75.245103.

[50] Masaki Tezuka, Ryotaro Arita, and Hideo Aoki. “Phase diagram for the one-dimensional Hubbard-Holstein model: A density-matrix renormalization group study”. In: *Phys. Rev. B* 76.15 (2007), pp. 1–11. ISSN: 10980121. DOI: 10.1103/PhysRevB.76.155114. arXiv: 0707.3197.

[51] Martin Hohenadler and Fakher F. Assaad. “Excitation spectra and spin gap of the half-filled Holstein-Hubbard model”. In: *Phys. Rev. B* 87.7 (2013), pp. 1–10. ISSN: 10980121. DOI: 10.1103/PhysRevB.87.075149.

[52] Yuta Murakami et al. “Ordered phases in the Holstein-Hubbard model: Interplay of strong Coulomb interaction and electron-phonon coupling”. In: *Phys. Rev. B* 88.12 (2013), pp. 1–14. ISSN: 10980121. DOI: 10.1103/PhysRevB.88.125126. arXiv: 1305.5771.

[53] Natanael C. Costa et al. “Phase diagram of the two-dimensional Hubbard-Holstein model”. In: *Communications Physics* 3.1 (2020), pp. 1–6. ISSN: 23993650. DOI: 10.1038/s42005-020-0342-2. URL: http://dx.doi.org/10.1038/s42005-020-0342-2.

[54] T Duguet and A Signoracci. “Symmetry broken and restored coupled-cluster theory: II. Global gauge symmetry and particle number”. In: 44.1 (2016), p. 015103. DOI: 10.1088/0954-3899/44/1/015103. URL: https://doi.org/10.1088/0954-3899/44/1/015103.

[55] Y. Qiu et al. “Particle-number projected Bogoliubov-coupled-cluster theory: Application to the pairing Hamiltonian”. In: *Phys. Rev. C* 99.4 (2019), pp. 1–18. ISSN: 24699993. DOI: 10.1103/PhysRevC.99.044301. arXiv: 1810.11245.

[56] Philipp Werner and Andrew J. Millis. “Efficient dynamical mean field simulation of the holstein-hubbard model”. In: *Phys. Rev. Lett.* 99.14 (2007), pp. 1–4. ISSN: 00319007. DOI: 10.1103/PhysRevLett.99.146404.

[57] Teresa E. Reinhard et al. “Density-Matrix Embedding Theory Study of the One-Dimensional Hubbard-Holstein Model”. In: *J. Chem. Theory Comput.* 15.4 (2019), pp. 2221–2232. ISSN: 15499626. DOI: 10.1021/acs.jctc.8b01116. arXiv: 1811.00048.

[58] Gordon Baym. “Field-theoretic approach to the properties of the solid state”. In: *Ann. Phys.* 14 (1961), pp. 1–42. ISSN: 00034916. DOI: 10.1006/aphy.2000.6009.

[59] L. Hedin and S. Lundqvist. “Effects of electron-electron and electron-phonon interactions on the one-electron states of solids”. In: *Solid State Physics*. Ed. by F. Seitz, D. Turnbull, and H. Ehrenreich. Academic Press, 1970, pp. 1–181. ISBN: 9780126077230.

[60] Robert Van Leeuwen. "First-principles approach to the electron-phonon interaction". In: *Phys. Rev. B* 69.11 (2004). ISSN: 1550235X. DOI: 10.1103/PhysRevB.69.115110.

[61] Sudip Sasmal and Oriol Vendrell. "Non-adiabatic quantum dynamics without potential energy surfaces based on second-quantized electrons: Application within the framework of the MCTDH method". In: *J. Chem. Phys.* 153.15 (2020). ISSN: 10897690. DOI: 10.1063/5.0028116. arXiv: 2010.00366. URL: https://doi.org/10.1063/5.0028116.

[62] Marat Sibaev et al. "Molecular second-quantized Hamiltonian: Electron correlation and non-adiabatic coupling treated on an equal footing". In: *J. Chem. Phys.* 153.12 (2020). ISSN: 10897690. DOI: 10.1063/5.0018930. URL: https://doi.org/10.1063/5.0018930.

[63] Thomas Dresselhaus et al. "Coupling electrons and vibrations in molecular quantum chemistry". In: (2020), pp. 1–33. arXiv: arXiv:2010.04654. URL: http://arxiv.org/abs/2010.04654.

[64] Qiming Sun et al. "Recent developments in the PySCF program package". In: *J. Chem. Phys.* 153.2 (2020).
Y. G. contributed to the periodic boundary condition module within the PySCF package. DOI: 10.1063/5.0006074.

[65] Honghui Shang et al. "Lattice dynamics calculations based on density-functional perturbation theory in real space". In: *Comput. Phys. Commun.* 215 (2017), pp. 26–46. ISSN: 00104655. DOI: 10.1016/j.cpc.2017.02.001. arXiv: 1610.03756. URL: http://dx.doi.org/10.1016/j.cpc.2017.02.001.

[66] S. Goedecker, M. Teter, and J. Hutter. "Separable dual-space Gaussian pseudopotentials". In: *Phys. Rev. B* 54 (3 1996), pp. 1703–1710. DOI: 10.1103/PhysRevB.54.1703. URL: https://link.aps.org/doi/10.1103/PhysRevB.54.1703.

[67] C. Hartwigsen, S. Goedecker, and J. Hutter. "Relativistic separable dual-space Gaussian pseudopotentials from H to Rn". In: *Phys. Rev. B* 58 (7 1998), pp. 3641–3662. DOI: 10.1103/PhysRevB.58.3641. URL: https://link.aps.org/doi/10.1103/PhysRevB.58.3641.

[68] Joost VandeVondele and Jürg Hutter. "Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases". In: *J. Chem. Phys.* 127.11 (2007), p. 114105. ISSN: 00219606. DOI: 10.1063/1.2770708.

[69] Joost VandeVondele et al. "Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach". In: *Comput. Phys. Commun.* 167.2 (2005), pp. 103–128.

[70] Thomas D. Kühne et al. "CP2K: An electronic structure and molecular dynamics software package -Quickstep: Efficient and accurate electronic structure calculations". In: *J. Chem. Phys.* 152.19 (2020). ISSN: 10897690. DOI: 10.1063/5.0007045. arXiv: 2003.03868. URL: https://doi.org/10.1063/5.0007045.

[71] Paolo Giannozzi et al. "QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials". In: *J. Phys. Condens. Matter* 21.39 (2009). ISSN: 09538984. DOI: 10.1088/0953-8984/21/39/395502. arXiv: 0906.2569.

[72] Tomokatsu Watanabe et al. "Monte Carlo simulations of electron transport properties of diamond in high electric fields using full band structure". In: *J. Appl. Phys.* 95.9 (2004), pp. 4866–4874. ISSN: 00218979. DOI: 10.1063/1.1682687.

[73] Kunie Ishioka et al. "Coherent optical phonons in diamond". In: *Applied Physics Letters* 89.23 (2006). ISSN: 00036951. DOI: 10.1063/1.2402231.

[74] Jin-Jian Zhou et al. "Perturbo: a software package for ab initio electron-phonon interactions, charge transport and ultrafast dynamics". In: (2020). arXiv: 2002.02045. URL: http://arxiv.org/abs/2002.02045.

[75] Edgar Solomonik et al. "Cyclops tensor framework: Reducing communication and eliminating load imbalance in massively parallel contractions". In: *Proceedings - IEEE 27th International Parallel and Distributed Processing Symposium, IPDPS 2013* (2013), pp. 813–824. DOI: 10.1109/IPDPS.2013.112.

[76] Manuel Cardona. "Electron-phonon interaction in tetrahedral semiconductors". In: *Solid State Commun.* 133.1 (2005), pp. 3–18. ISSN: 00381098. DOI: 10.1016/j.ssc.2004.10.028.

[77] Rafael Ramírez, Carlos P. Herrero, and Eduardo R. Hernández. "Path-integral molecular dynamics simulation of diamond". In: *Phys. Rev. B* 73.24 (2006), pp. 1–8. ISSN: 10980121. DOI: 10.1103/PhysRevB.73.245202.

[78] Feliciano Giustino, Steven G. Louie, and Marvin L. Cohen. "Electron-phonon renormalization of the direct band gap of diamond". In: *Phys. Rev. Lett.* 105.26 (2010), pp. 1–4. ISSN: 00319007. DOI: 10.1103/PhysRevLett.105.265501.

[79] Elena Cannuccia and Andrea Marini. "Effect of the quantum zero-point atomic motion on the optical and electronic properties of diamond and trans-polyacetylene". In: *Phys. Rev. Lett.* 107.25 (2011), pp. 1–5. ISSN: 00319007. DOI: 10.1103/PhysRevLett.107.255501.

[80] Bartomeu Monserrat and R. J. Needs. "Comparing electron-phonon coupling strength in diamond, silicon, and silicon carbide: First-principles study". In: *Phys. Rev. B* 89.21 (2014), pp. 1–8. ISSN: 1550235X. DOI: 10.1103/PhysRevB.89.214304. arXiv: 1406.0654.

[81] Marios Zacharias and Feliciano Giustino. “One-shot calculation of temperature-dependent optical spectra and phonon-induced band-gap renormalization”. In: *Phys. Rev. B* 94.7 (2016). ISSN: 24699969. DOI: 10.1103/PhysRevB.94.075125. arXiv: 1604.02394.

[82] Ferenc Karsai et al. “Electron-phonon coupling in semiconductors within the GW approximation”. In: *New J. Phys.* 20.12 (2018). ISSN: 13672630. DOI: 10.1088/1367-2630/aaf53f.

[83] P. B. Allen and V. Heine. “Theory of the temperature dependence of electronic band structures”. In: *J. Phys. C.* 9.12 (1976), pp. 2305–2312. ISSN: 00223719. DOI: 10.1088/0022-3719/9/12/013.

[84] P. B. Allen and M. Cardona. “Theory of the temperature dependence of the direct gap of germanium”. In: *Phys. Rev. B* 23.4 (1981), pp. 1495–1505. ISSN: 01631829. DOI: 10.1103/PhysRevB.23.1495.

[85] Ferd E. Williams. “Theoretical low temperature spectra of the thallium activated potassium chloride phosphor [30]”. In: *Phys. Rev.* 82.2 (1951), pp. 281–282. ISSN: 0031899X. DOI: 10.1103/PhysRev.82.281.2.

[86] Melvin Lax. “The franck-condon principle and its application to crystals”. In: *J. Chem. Phys.* 20.11 (1952), pp. 1752–1760. ISSN: 00219606. DOI: 10.1063/1.1700283.

[87] S. Poncé et al. “Temperature dependence of the electronic structure of semiconductors and insulators”. In: *J. Chem. Phys.* 143.10 (2015). ISSN: 00219606. DOI: 10.1063/1.4927081. arXiv: 1504.05992. URL: http://dx.doi.org/10.1063/1.4927081.

Chapter 5

FERMIONIC TENSOR NETWORK SIMULATION WITH ARBITRARY GEOMETRY

5.1 Introduction

Tensor network (TN), a recently developed mathematical tool, has been undergoing rapid developments over the past few decades, enabling major advances in condensed matter physics, atomic physics, quantum information science, and so on. In the context of quantum many-body physics, the first major success of tensor network can be traced to Steve White's invention of density matrix renormalization group (DMRG) algorithm for matrix product states (MPS)[1, 2]. Since then, DMRG has gained wide popularity in studying model Hamiltonians to describe high-temperature superconductivity, quantum spin liquids, and other strongly correlated systems[3–7]. However, due to the one-dimensional (1D) entanglement entropy area law, MPS are only best suited for computing the ground states of gapped, 1D Hamiltonians[8]. Despite such limitation, DMRG is still mostly viewed as the method of reference for some higher dimensional problems due to its robustness.

On the other hand, it was recently realized that MPS is just a special class of a broad family of tensor networks, each with different traits in terms of geometry, representability, computational cost, and so on. For instance, projected entangled pair states (PEPS), the 2D generalization of MPS, is capable of capturing correlation functions with polynomial decay[9, 10], making it a promising candidate to describe both gapped systems and critical states of matter. In addition to better representability, higher dimensional tensor networks are also deeply connected to quantum information science and machine learning, thus attracting great interests in the research community.

In the context of quantum many-body problems, one of the grand challenges for tensor network is how to adapt tensor network methods (potentially with arbitrary geometry) to study fermion systems, where the anti-symmetry from Pauli's exclusion principle prohibits a direction translation. In this chapter, we describe a numerical framework for fermionic tensor network with arbitrary geometry using special unitary groups. Our scheme is a direct extension of Pižorn's work[11] by encoding fermion statistics in the block sparse tensor backend, thus allowing direct inheritance of most of the pre-established tensor network infrastructures. This strategy is inherently equivalent to the other swap gates-based approaches[12–14]. In addition to the specially tailed backend, we introduce additional rules in fermionic tensor network to account for fermion statistics yet maintain a clear graphical representation. We benchmark our framework by investigating the Hubbard model on both a 2D square lattice and a 3D diamond-like lattice. Thanks to the symmetry support in our backend and the various types of approximate contraction methods, we were able to perform simulations on large systems (up to 250 sites) and evaluate the eneriges of these graphs. Our results exhibits competitive accuracy with coupled cluster methods, indicating that the fermionic tensor network is a promising candidate for correlated fermion systems.

The rest of the chapter is organized as follows. In Section 5.2, we describe the formulation for tensor network and its extension to fermion systems. We then describe the algebraic rules for fermionic tensors and the special rules for tensor operations in Section 5.3 and Section 5.4 respectively. In Section 5.5 we introduce approximate contraction methods and present numerical results on Hubbard model on a 2D square lattice and a 3D diamond-like lattice. We conclude with Section 5.6.

5.2 Tensor Network Theory

We begin by considering a quantum many-body system on certain lattice $\mathcal{L}$ made of N sites. These lattices are labeled by $i \in \{1, 2, ..., N\}$ and each site i resides in a complex vector space with basis states $|\mathbf{s_i}\rangle_{\mathbf{s_i}=1,2,...,m}$ where m is the size for each vector space. The wavefunction can then be expressed as:

$$|\Psi\rangle = \sum_{\mathbf{s_1},\mathbf{s_2},...,\mathbf{s_N}} C_{\mathbf{s_1},\mathbf{s_2},...,\mathbf{s_N}} |\mathbf{s_1}\rangle \otimes |\mathbf{s_2}\rangle ... \otimes |\mathbf{s_N}\rangle, \tag{5.1}$$

where the expansion coefficients $C_{\mathbf{s_1},\mathbf{s_2},...,\mathbf{s_N}}$ scale as m^N.

A main goal of quantum many-body theory is to be able to simulate the low-lying energies states of certain Hamiltonian $\hat{H}$ and compute expectation value of interest with $\langle\Psi|\hat{O}|\Psi\rangle$. Due to the steep scaling of $C_{\mathbf{s_1},\mathbf{s_2},...,\mathbf{s_N}}$, it is computationally prohibitive to perform exact diagonalization (ED) on systems beyond a few tens of sites.

Tensor network methods were introduced as an efficient ansatz to approximately represent the wavefunction. The theory amounts to breaking down the huge coefficient tensor into a collection of N smaller tensors $\{A^i\}$ where each tensor carries both the so-called physical index for $\mathbf{s_i}$ and a set of virtual bonds $\mathbf{v_i}$ that are connected to other tensors following the geometry of the lattice $\mathcal{L}$:

$$|\Psi\rangle \approx \sum_{\mathbf{s_1,s_2,...,s_N}}^{\mathbf{v_1,v_2,...,v_n}} A^1_{\mathbf{s_1,v_1}} A^2_{\mathbf{s_2,v_2}} ... A^N_{\mathbf{s_N,v_n}} |\mathbf{s_1}\rangle \otimes |\mathbf{s_2}\rangle ... \otimes |\mathbf{s_N}\rangle. \tag{5.2}$$

Depending on the geometry of $\mathcal{L}$, multiple ansatzs including MPS, PEPS, and multi-scale entanglement renormalization ansatz (MERA) have been developed to simulate different systems based on entanglement area law. The geometries of MPS, PEPS, and MERA are displayed in Figure 5.1.

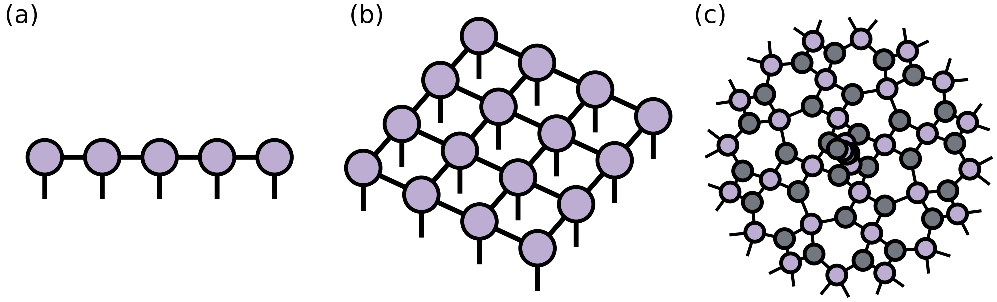

Figure 5.1: Geometries of selected class of tensor network: (a) MPS, (b) PEPS, (c) MERA.

Extension to Fermion Systems

Traditionally tensor network methods have been mostly developed with commutative tensor algebra rules, which can be directly transferable to bosonic systems where such tensor algebras are commensurate with the commuting bosonic operators. Fermionic operators on the other hand, follow the anti-commutation rules due to the anti-symmetry nature. This is manifested as $|\mathbf{s_1}\rangle \otimes |\mathbf{s_2}\rangle = (-1)^{P_{\mathbf{s_1}} P_{\mathbf{s_2}}} |\mathbf{s_2}\rangle \otimes |\mathbf{s_1}\rangle$ where $P_{\mathbf{s_i}}$ denotes the parity of the basis $\mathbf{s_i}$. This prohibits a direct translation of numerical algorithms developed with commutative algebraic rules.

Several methods have been proposed to extend TN methods to fermion systems, which can mostly be divided into three categories:

1. Fermion operators are transformed to hard-core boson operators, typically through encoding schemes such as Jordon-Wigner mapping[15] or Bravyi-Kitaev transformation[16].

2. Swap gates are explicitly introduced into the graph to account for potential phases from anti-symmetry[12–14].

3. Special algebraic rules are enforced in tensor operations while the diagrammatic notation remains unchanged[11].

When formulating fermionic tensor network, (1) can only preserve locality of the Hamiltonian in 1D and introduces “fictitious" long-range interactions into higher dimension systems. (2) is generally applicable to arbitrary graph, but the addition of geometry- and operation-dependent swap gates clutters the intuitive graphical representation. (3) encodes fermion statistics directly in the tensor backend which requires each parity sector to be handled differently. Our approach falls into the category of type (3). This strategy maintains the clean graphical representation with minimal modifications required to reuse pre-established tensor network algorithms. Specifically, we enforce anti-commuting tensor algebras at the backend level so that they behave just like fermionic operators. Intuitively, our “tensor" object now represents a collection of fermionic operators:

$$\hat{A}^i = A^i_{\mathbf{s_i},\mathbf{v_i}} \hat{O}_{\mathbf{s_i}} \hat{O}_{\mathbf{v_i}}, \tag{5.3}$$

where $A^i_{\mathbf{s_i},\mathbf{v_i}}$ is the pure tensor object that obeys commutative algebraic rules and $\hat{O}_{\mathbf{s_i}}$ and $\hat{O}_{\mathbf{v_i}}$ are the fermionic operators that act on the physical space and the virtual space respectively. We can then express the wavefunction using these operators as:

$$|\Psi\rangle \approx \prod \hat{A}^1 \hat{A}^2 ... \hat{A}^N |0\rangle. \tag{5.4}$$

Our work is initially inspired by Pižorn’s work with additional features as below:

1. Tensors are implemented in a block sparse format with symmetry group beyond parity symmetry (Z_2) supported, e.g U_1, $Z_2 \otimes Z_2$ and $U_1 \otimes U_1$.

2. All tensors are constrained with a symmetry conservation rule, allowing efficient representation for quantum states under different symmetry irreducible representation.

5.3 Linear Algebras for Fermionic Tensors

We here introduce the implementation details of our fermion tensor library in Python. For our block sparse implementation, each index **q** (physical or virtual) is unfolded into symmetry modes (P_q) and symmetry blocks (q). For instance, Equation 5.3 is now transformed into:

$$\hat{A}^i = A^i_{s_i P_{s_i}, v_i P_{v_i}} \hat{O}^{P_{s_i}}_{s_i} \hat{O}^{P_{v_i}}_{v_i}. \tag{5.5}$$

Here the symmetry group of P_q can be extended beyond the parity symmetry group (Z_2), e.g U_1, $Z_2 \otimes Z_2$ and $U_1 \otimes U_1$. The size of symmetry sectors can in principle differ from each other (this is different from Chapter 3), and we enforce symmetry conservation here such that the total symmetry for for each block amounts to a fixed quantity P^i, e.g $P_{s_i} + P_{v_i} = P^i$ for all blocks of the tensor.

Data Structure

Since our fermion tensors are designed to mimic fermionic operators, the tensor object must store both the pure tensor part and the operator part of Equation 5.5. The pure tensor data are stored in block sparse format similar to the compressed sparse row format for sparse matrix: (1) The data for each block (subtensor) is flattened and then concatenated into a full 1D array (order of the blocks are unsorted). (2) We keep three sets of markers (stored as Numpy array) to preserve the symmetry structure of the original tensor: one storing the shapes for each block, one storing the irreps of each block (implemented as hash values), and the last as the pointer for data locations in the full array. In addition to the irreps marker, we label each index with plus or minus signs depending on the symmetry conservation relation. Figure 5.2 is a schematic diagram showing data structure for a fermion matrix with block size $D_{U_1(0)} = 1$, $D_{U_1(1)} = 3$, $D_{U_1(2)} = 2$ for both dimensions and a symmetry conservation rule of $P_i + P_j = U_1(2)$ where D_{P_i} represents the block size for the symmetry sector of P_i and i, j are the indices for row and column respectively. In Figure 5.2 (a) we introduced arrows on the indices to denote the algebraic sign in symmetry conservation. This will be suppressed in subsequent figures for simplicity unless noted.

Tensor Transposition

For regular tensors, transposition refers to the permutation of original indices $\mathbf{i_1}, \mathbf{i_2}, \ldots, \mathbf{i_n}$ into a different order $\mathscr{F}(\mathbf{i_1}, \mathbf{i_2}, \ldots, \mathbf{i_n})$. For fermion tensors, transposi-

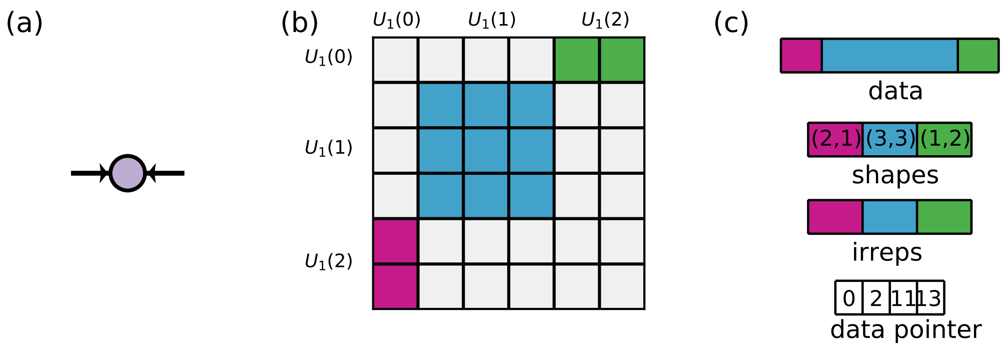

Figure 5.2: Schematic diagram of fermion tensor data structure: (a) graphical notation where arrows on the indices denote algebraic sign in symmetry conservation, (b) sparsity structure in dense matrix format where each non-zero block is marked with different colors, (c) explicitly stored data sets.

tion translates to permuting the indices for both the pure tensor part and the operator part while keeping the underlying operator unchanged. As shown in Equation 5.6, this requires performing block-wise transposition and computing the corresponding phase that arises from permuting the order of operators:

$$\hat{A} = A_{\mathbf{i_1},\ldots,\mathbf{i_n}} \hat{O}_{\mathbf{i_1}} \ldots \hat{O}_{\mathbf{i_n}} = A_{\mathscr{F}(\mathbf{i_1},\ldots,\mathbf{i_n})} \times (-1)^{f(\mathscr{F},P_{i_1},\ldots,P_{i_n})} \mathscr{F}(\hat{O}_{\mathbf{i_1}} \ldots \hat{O}_{\mathbf{i_n}}). \tag{5.6}$$

Here we use $(-1)^{f(\mathscr{F},P_{i_1},\ldots,P_{i_n})}$ to represent the permutation- and parity-dependent phase for each block. This phase is in practice absorbed onto the permuted tensor data to form a new fermion tensor object.

Tensor Contraction

One common strategy for performing regular tensor contraction is to first permute and reshape the tensors into matrices so that the contraction can be mapped to a matrix multiplication. Modern BLAS libraries such as Intel MKL can efficiently perform such matrix multiplication and the overhead from preprocessing is mostly worthwhile. We here adopt the same philosophy for our block sparse tensor contraction with the key difference that the permuting step needs to account for the potential phase. After that, contraction on the block sparse tensors reduces to doing a set of smaller dense matrix multiplications on various pairs of blocks from the input tensors. Our contraction backend leverages HPTT [17] for fast tensor transposition and dispatches to BLAS libraries for each individual matrix multiplication.

The graphical notation of fermion tensor contraction is also slightly modified to account for the order of the two operators. For instance, the contraction $\hat{Y}_{\mathbf{ijkl}} = \hat{A}_{\mathbf{ija}}\hat{B}_{\mathbf{akl}}$ is displayed in Figure 5.3

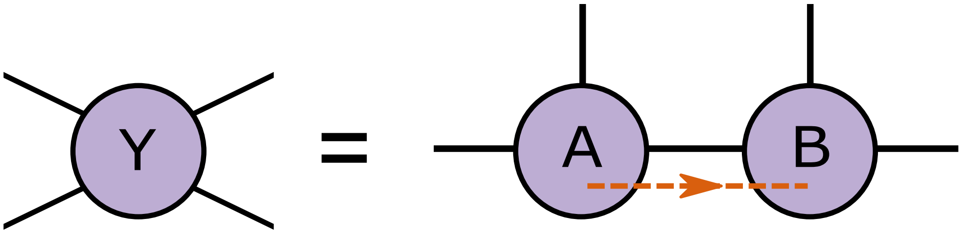

Figure 5.3: Diagramatic representation of fermion tensor contraction. The dashed brown arrow denotes the ordering of operator $\hat{A}$ and $\hat{B}$ (arrows denoting the symmetry conservation rules are suppressed).

where we have introduced the brown arrow to represent the order of the two operators. One of the most basic principles of fermionic tensor network is that only tensors are that adjacently ordered (connected by the brown arrow) can be directly contracted.

Tensor Decomposition

Tensor decomposition routines are heavily used in tensor network algorithms to canonize or compress tensors. The operation is typically done through QR factorization or singular value decomposition (SVD) as shown in Figure 5.4.

Since all fermionic tensors must carry a fixed total symmetry, there is a degree of freedom on how to partition the input total symmetry in the output, i.e $P_Y = P_Q + P_R = P_U + P_S + P_V$. Noticeably, in tensor network theories, decomposition routines are mostly called for canonicalization or compression between two tensors, so it is natural to adopt the convention that the two output tensors each take the same total symmetry as the two input tensors.

In our implementation, once the symmetry partition is determined on the output, we can compute all the potential irreps for the shared index of the outputs tensors. For each different irrep, we gather all relevant blocks and reshape them into a matrix before performing block-wise decomposition. The decomposed outputs are then re-assembled based on their irreps into the output fermionic tensors.

In SVD, one typically truncates the singular values up to some fixed dimension χ

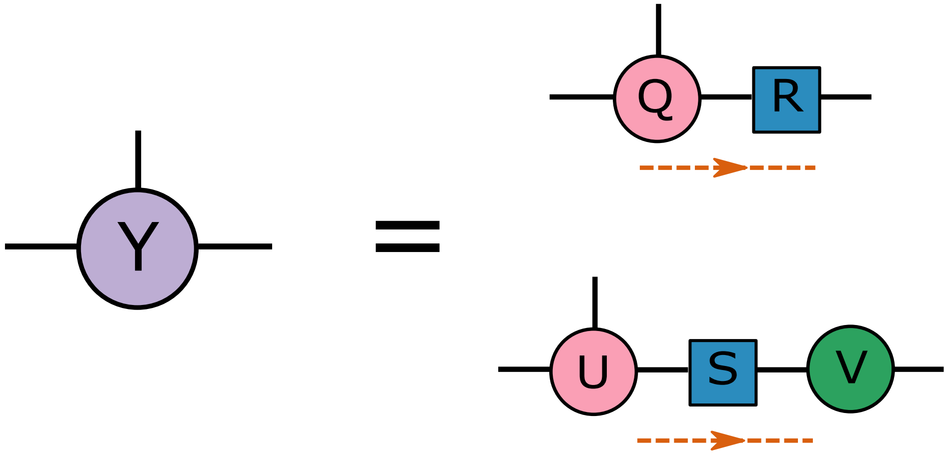

Figure 5.4: Diagramatic representation of fermion tensor decomposition through QR (upper right) and SVD (lower right). The dashed brown arrow denotes the ordering of output operators.

as an approximation. In this case, we gather all singular values from all blocks of $\hat{S}$, sort them, and keep the largest χ values.

5.4 Rules for Fermionic Tensor Network

In Section 5.3 we have introduced how to account for fermionic statistics when performing pair-wise operations, but this is only well defined when the two operators are adjacent. In principle, fermionic operators are placed in some order with respect to the vaccum (as shown in Equation 5.4) and operators with shared indices are not necessarily adjacent. Therefore, we describe here how to perform operations on non-adjacent tensors.

Contraction

It has been shown that the order of fermion tensors can be reversed by introducing a phase into each block of the tensor inputs[11]. For instance, for pair-wise contraction between adjacent fermion operator $\hat{A} = \hat{A}_{\mathscr{F}^a(\{\mathbf{i}\},\{\mathbf{m}\})}$ and $\hat{B} = \hat{B}_{\mathscr{F}^b(\{\mathbf{j}\},\{\mathbf{m}\})}$ where $\mathscr{F}$ denotes the permutation of the set of indices $\{i\}, \{j\}$ (uncontracted) and $\{m\}$ (contracted), the swap rule is manifested as

$$\hat{A}\hat{B} = \sum_{\{m\}} A_{\mathscr{F}^a(\{\mathbf{i}\},\{\mathbf{m}\})}\mathscr{F}^a(\hat{O}_{\{\mathbf{i}\}}\hat{O}^a_{\{\mathbf{m}\}})B_{\mathscr{F}^b(\{\mathbf{j}\},\{\mathbf{m}\})}\mathscr{F}^b(\hat{O}_{\{\mathbf{j}\}}\hat{O}^b_{\{\mathbf{m}\}}) \tag{5.7}$$

$$= \sum_{\{m\}} g_{\{m\}} \times B_{\mathscr{F}^b(\{\mathbf{j}\},\{\mathbf{m}\})}\mathscr{F}^b(\hat{O}_{\{\mathbf{j}\}}\hat{O}^{b*}_{\{\mathbf{m}\}})A_{\mathscr{F}^a(\{\mathbf{i}\},\{\mathbf{m}\})}\mathscr{F}^a(\hat{O}_{\{\mathbf{i}\}}\hat{O}^{a*}_{\{\mathbf{m}\}}) \tag{5.8}$$

where $g_{\{m\}} = (-1)^{P_A P_B + P_{\{m\}}}$ depends on both the total parity of $\hat{A}$, $\hat{B}$ (P_A and P_B) and the parity of each contracted index m (P_m). This block-wise phase can be fully absorbed onto the tensor part of either A or B, and the original contraction can be transformed to $\hat{A}\hat{B} = \hat{\tilde{B}}\hat{\tilde{A}}$ where either $\tilde{A}$ or $\tilde{B}$ contains the phase and the other remains unchanged. Graphically this can be be represented as Figure 5.5.

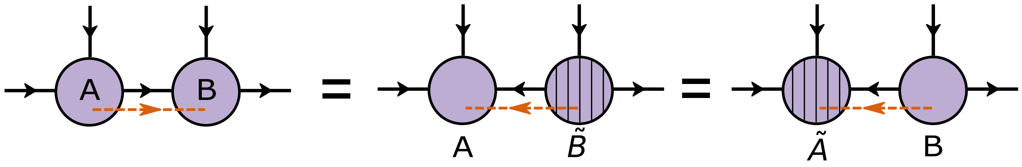

Figure 5.5: When applying the fermionic swap operation, the phase can be fully factorized onto either B (middle plot) or A (right plot). The hatch pattern on the vertices shows where the phase gets factorized onto. The reverse of contraction order is shown via the brown dashed arrow. The flip of the arrow direction on the shared index indicates the hermitian conjugate operation on the operator for the shared index.

The swap rule allows all pair-wise operations to be well defined as we can reverse tensor orders so that the tensor pairs are adjacently ordered. For tensor contractions, this means we can take advantage of the optimized contraction path by swapping tensor orders on the fly so that the operands (potentially with a phase) are always adjacently ordered.

Compression and Canonicalization

Although pair-wise operations are only well defined when operators are adjacently ordered, in practice, not all operations would require an explicit swap operation. As we shall prove here, compression and canonicalization in our construction can be performed in place without the need for a swap. We here provide an intuitive example based on three tensors with initial order ABC, and we wish to compress the index k between non-adjacent A and C. This schematic diagram for a well-defined compression operation is shown in Figure 5.6.

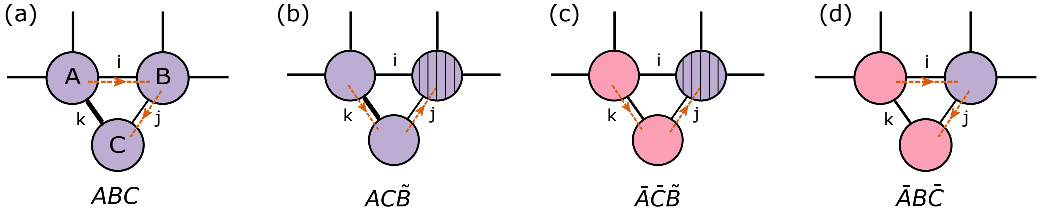

Figure 5.6: Operations required to perform compression between non-adjacent tensors A and C. The four stages are labeled with (a), (b), (c), and (d). Dashed brown arrows are used to represent the relative order of operators. The details for these steps are provided in the main text.

The entire compression operation can be divided into three steps: (1) Swap operation is performed between B and C so that the new state becomes $AC\tilde{B}$. As shown in (b) of Figure 5.6, we can factorize the phase $g_0 = (-1)^{P_C P_B + P_j}$ onto B so that the new state becomes $AC\tilde{B}$. (2) Compression is performed on adjacent AC, leading to the new state $\bar{A}\bar{C}\tilde{B}$. This is manifested as the transition from (b) to (c) in Figure 5.6. As mentioned in Chapter 5.3, we do not alter the total symmetry partition during this step, i.e $P_{\bar{A}} = P_A$ and $P_{\bar{C}} = P_C$. (3) Another swap operation is called between $\bar{C}$ and $\tilde{B}$ to get to the same order as the initial state, during which another phase $g_1 = (-1)^{P_{\bar{C}} P_{\tilde{B}} + P_j}$ arises. It is straightforward to see that $g_1 = g_0$, and we can thus revert $\tilde{B}$ back to B by again absorbing g_1 onto $\tilde{B}$. By a comparison between (a) and (d), we can clearly see that the whole routine is effectively equivalent to performing compression between A and C in place as if they are adjacent. This is only true as our implementation enforces symmetry conservation and no re-partition of total symmetry during compression or canonicalization. We can easily generalize to cases with an arbitrary number of operators between A and C as all the operators in between can be viewed as a large, contracted B.

In principle, the removal of explicit swap in compression and canonicalization reduces a huge amount of overhead from reordering the tensors.

5.5 Results

In this section we will provide our benchmark results on half-filled Hubbard model on two types of finite lattices: a 2D square lattice and a 3D diamond lattice constructed by extending the primitive cell of diamond crystal in 3D with a tetrahedral fashion. An example for a 3x3x3 diamond graph is shown in Figure 5.7

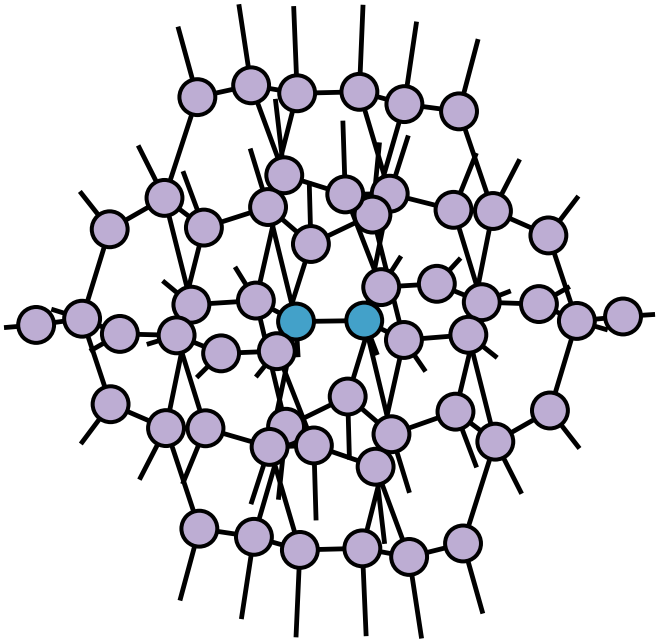

Figure 5.7: Geometry of a 3x3x3 diamond-like lattice. The central sites in blue are extended in 3D with tetrahedral bonding.

The Hamiltonian of Hubbard model can be expressed as:

$$H = -t \sum_{<i,j>\sigma} \left(a_{i\sigma}^{\dagger} a_{j\sigma} + \text{h.c.} \right) + U \sum_{i} n_{i\uparrow} n_{i\downarrow}, \tag{5.9}$$

where $< i, j >$ denotes nearest neighbors within the structure and U characterizes the on-site Coulomb repulsion. Despite the simplicity of its mathematical form, the Hubbard model is one of the most important models of correlation physics.

Computational Details

Our fermionic tensor backend is interfaced to the Quimb package[18] to take advantage of the tensor network infrastructures. The ground-state wavefunctions are optimized by simple update style time-evolution block-decimation (TEBD)[19] implemented in Quimb, and we used $U(1)$ symmetry for fixed the total particle number. For comparison, we benchmark our results against other methods including DMRG, UCCSD, UCCSD(T), and ED when computationally feasible. DMRG calculations are performed using Block2 package[20]. We used a maximum bond dimension of 6000 and extrapolated the sweep energy linearly as a function of the maximum discarded weight. All coupled cluster calculations are performed using

PySCF package[21] with an unrestricted reference.

For expectation (energy) value computations, exact contractions on any of the two types of lattices is prohibitively expensive. Approximate contractions are thus performed as a compromise.

For our approximate PEPS contraction, we used a bilayer MPS-MPO style boundary contraction method in Quimb to form the environment for each pair of neighboring tensors. The level of approximation is tuned by the maximal truncation bond χ. Typically $\chi \approx nD^2$ should reach good accuracy, here we use $\chi = 128$ for all our calculations. The expectation value computation for the diamond graph is even harder than the PEPS. This is due to the fact that tensors in diamond graph are more “dispersed" and it is not clear how to perform approximate contractions efficiently. Therefore, we introduce two types of approximate contraction methods for our diamond graph calculations.

The first type is the so-called compressed contraction implemented in Quimb, which is shown in Figure 5.8. For each term in the energy evaluation, we first use the cotengra package[22] to search for the approximate contraction path with the lowest peak memory footprint. Before each pair-wise contraction, the involved tensors are compressed with neighboring tensors just in time if the size of any shared bonds exceeds the threshold χ. This process is carried out throughout each step in the contraction path. However, despite our aggressive compression, the diamond graph is computationally more expensive than 2d square lattice, and we will use a lower χ in our calculation.

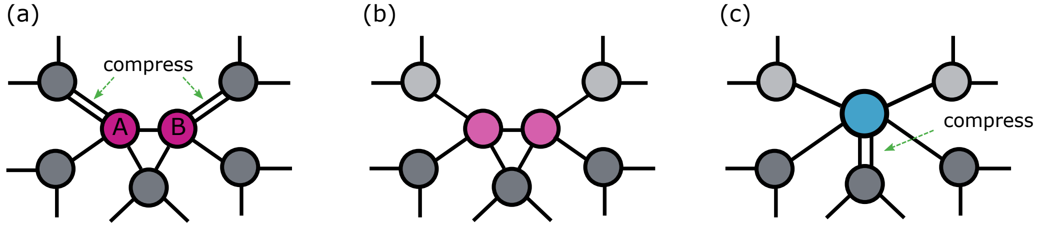

Figure 5.8: Schematic diagrams for compressed contraction: (a) compression before contraction, (b) pairwise contraction on reduced tensors, (c) proceed to next contraction. See main text for details.

The second type is referred to as cluster approximation as shown in Figure 5.9, which is inspired by the cluster update method in TEBD[23]. For each term in the energy evaluation, we construct a cluster by only including the tensors around the central sites (where the Hamiltonian term acts on) up to some radius r. In order to account for the effect from the environment, the gauges on the boundary are first absorbed onto the cluster and then traced out so no dangling bonds remain after the cut. The cluster approximation can further take advantage of the compressed contraction scheme above to further reduce the cost.

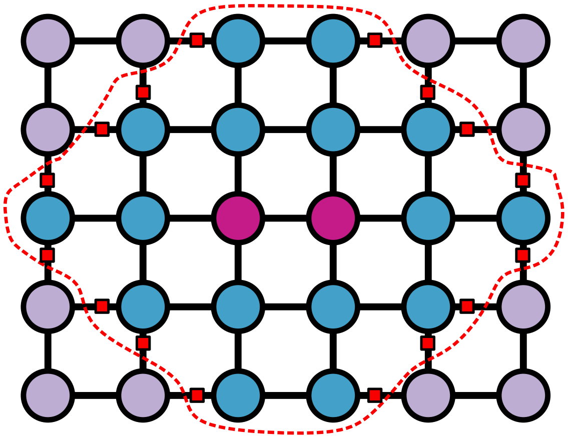

Figure 5.9: Schematic diagrams for cluster approximation with a radius of 2. The cluster encapsulated by the red dashed line serves as the approximate network (with a trace operation on the dangling bonds). This includes the dark pink sites, the blue sites, and the red squares, which represent the central sites, the neighboring sites with a Manhattan distance no larger than 2, and the gauges on the dangling bonds respectively. The remaining light purple sites are thus neglected.

2D Square Lattice

For our 2D Hubbard model calculation, we surveyed different parameter sets (L, U) where L is the width of the lattice and U is the on-site Coulomb repulsion.

We first examine small system size with $L = 4$ to check the convergence behavior of our method. The results together with other reference data are presented in Figure 5.10. As D increases, we found the PEPS energies quickly surpass coupled cluster energies with a slow convergence towards the exact ground state energies. Notably, the minimal D required to bypass coupled cluster energies decreases as U increases. This is consistent with the general observation that tensor network methods are more suited for systems with local interactions.

Figure 5.11 shows the accuracy of our largest PEPS calculations (D=18) at L=4, 6, and 8 compared to other methods. For L=4, we use ED energies as the reference and

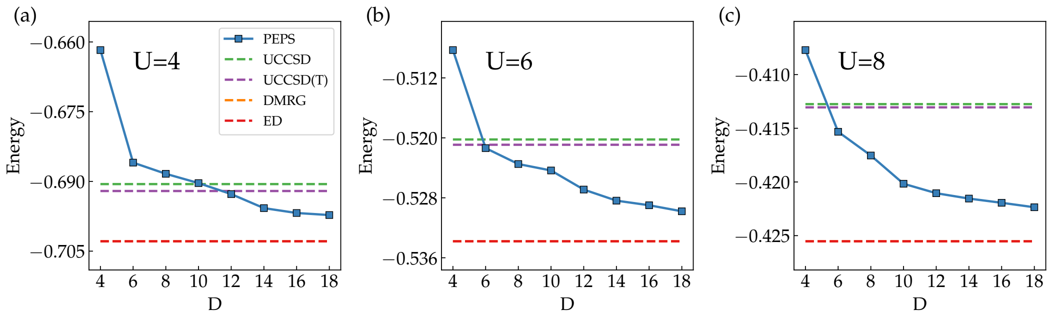

Figure 5.10: Ground-state energies for 2D Hubbard model at $L = 4$ with $U = 4, 6, 8$. The PEPS results are represented by the blue solid lines and the references from UCCSD, UCCSD(T), DMRG, and ED are displayed with dashed lines using different colors.

for the two larger lattices, extrapolated DMRG energies are used as the reference. We can clearly observe that at D=18, PEPS energies can consistently achieve an error rate below 1%, which is much lower than UCCSD and UCCSD(T) in all cases. Previous work has shown that much of the remaining errors can be ascribed to the over-simplification of environment effect during simple update[12, 19, 24]. Another interesting observation is that the relative error is decreasing as system size increases. This not necessarily indicates an improvement from the PEPS side, but rather more like DMRG struggling with larger systems.

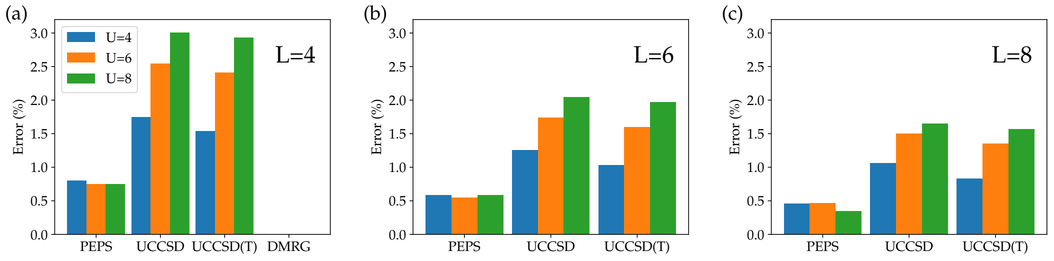

Figure 5.11: Relative error of 2D Hubbard model energies computed from different methods compared to the reference: (a) L=4 with ED as the reference, (b) L=6 with DMRG as the reference, (c) L=8 with DMRG as the reference.

Our largest calculation at L=10 is shown in Figure 5.12. We found a similar convergence behavior system L=4 in Figure 5.10. Notably, even as the lattice expands from L=4 to L=10, the minimal bond dimension for PEPS to outperform UCCSD and UCCSD(T) remains almost the same for each U (D=12 for U=4, D=6 for U=6 and 8). The results for all our 2D Hubbard model calculations are summarized in Table 5.1:

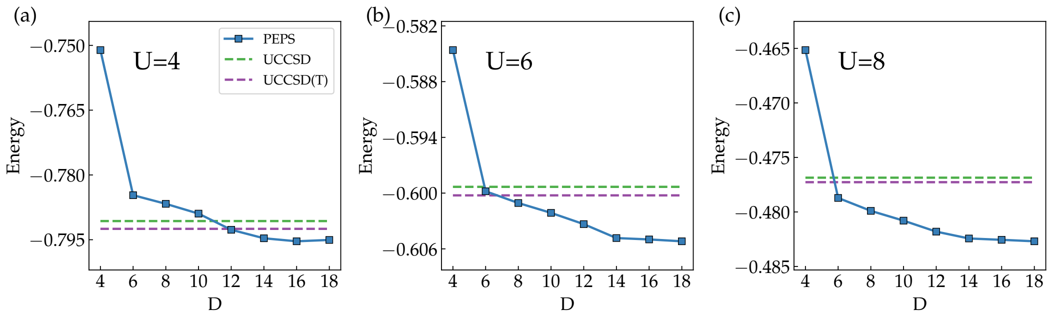

Figure 5.12: Ground-state energies for 2D Hubbard model at $L = 10$ with $U = 4, 6, 8$. The blue solids lines represents the PEPS results. The references from UCCSD and UCCSD(T) are displayed with dashed lines using different colors.

L	U	PEPS	UCCSD	UCCSD(T)	DMRG	ED
4	4	-0.698	-0.691	-0.692	-0.703	-0.703
	6	-0.531	-0.520	-0.521	-0.534	-0.534
	8	-0.423	-0.413	-0.413	-0.426	-0.426
6	4	-0.752	-0.747	-0.749	-0.756	NA
	6	-0.572	-0.564	-0.565	-0.574	NA
	8	-0.456	-0.448	-0.449	-0.458	NA
8	4	-0.779	-0.774	-0.776	-0.783	NA
	6	-0.593	-0.586	-0.587	-0.595	NA
	8	-0.473	-0.466	-0.467	-0.474	NA
10	4	-0.795	-0.791	-0.792	NA	NA
	6	-0.605	-0.599	-0.600	NA	NA
	8	-0.483	-0.477	-0.477	NA	NA

Table 5.1: Ground-state energies of 2D Hubbard model computed from PEPS (with D=18 and SU), UCCSD, UCCSD(T), DMRG, and ED.

3D Diamond Lattice

For our 3D diamond graph calculations, we first benchmark our method in 3x3x3 and 4x4x4 diamond graph. The expectation value computation is performed using the compressed contraction scheme with $\chi = 32$ and $\chi = 16$ for the two graph respectively. The comparison with UCCSD, UCCSD(T), and DMRG is provided in Figure 5.13. Due to the steep scaling of diamond simulation with respect to D, the largest bond dimension is set to 6 for all these calculations. Nevertheless, for all these cases, we found the SU energies to reach competitive accuracy as the coupled cluster at our best simulation. The results are expected to further improve at larger D though the computational cost is already beyond what our platform can handle.

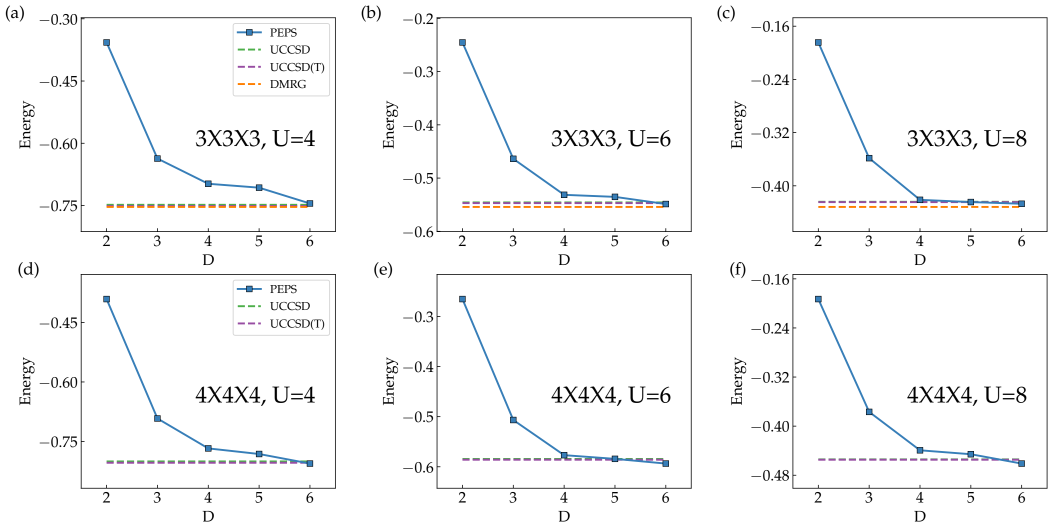

Figure 5.13: Ground-state energies for Hubbard model on 3x3x3 (top panel) and 4x4x4 (bottom panel) diamond lattice with $U = 4, 6, 8$. The blue solid lines represent the PEPS results and the references from UCCSD, UCCSD(T), and DMRG are displayed with dashed lines using different colors.

As a first attempt towards larger systems, we would like to get a rough estimate on how our methods perform in a 5x5x5 diamond graph (250 sites). This is computationally too demanding for our compressed contraction method. Therefore we will be computing the energies through the cluster approximation method shown in Figure 5.9 . We first assess the accuracy of cluster approximation method by computing the energies as a function of increasing radius r and make a comparison to results from the compressed contraction ($\chi = 32$ for 3x3x3 lattice and $\chi = 16$ for 4x4x4 lattice). The results are shown in Figure 5.14. Note here the comparison must be taken with caution as it was not possible to gauge the true accuracy of each method due to the huge cost of exact contraction. Still, we can find that for the 3x3x3 graph, a relative difference less than 1% can be consistently achieved with $r >= 2$. For the 4x4x4 structure however, the smallest relative difference all occurs at $r = 2$. As mentioned earlier, this does not necessary mean that $r = 2$ gives the best approximated results, but rather the references from compressed contraction may not be accurate themselves ($\chi = 16$ being too low).

Nevertheless, from our assessment above, we can still compute the energies of 5x5x5 graph with cluster approximation and expect the approximation error to be around a few percent or lower. We thus used a radius of 3 to compute the approximate energies and the results are shown in Figure 5.15. Similar to what we have observed

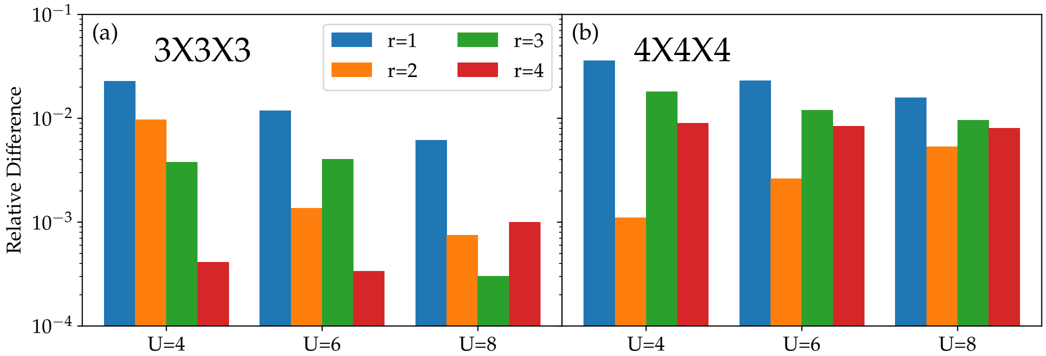

Figure 5.14: Relative difference between energies of different systems computed from cluster approximation and compressed contraction. The different colors represent cluster approximations with different radii r.

in small systems, the SU energies gradually decreases and becomes competitive with coupled cluster at D=6. Again, the performance of tensor network improves as locality (U) increases. The results for all our calculations are summarized in Table 5.2.

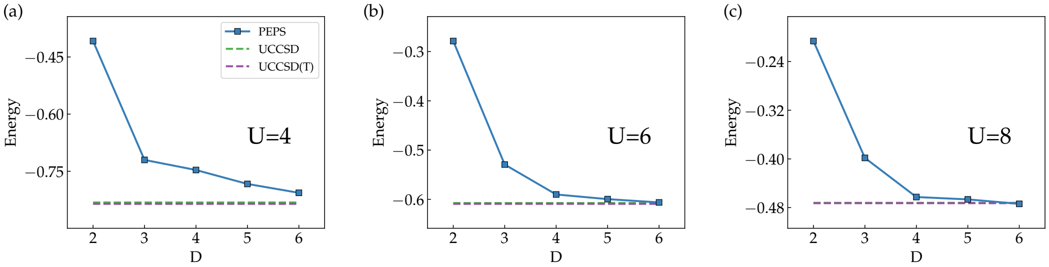

Figure 5.15: Ground-state energies of Hubbard model on 5x5x5 diamond lattice with $U = 4, 6, 8$.

5.6 Conclusion

We have presented a numerical framework for tensor network simulation on fermion systems with arbitrary geometry. The method is based on a tailored block sparse tensor library with support on block symmetries to account for fermion statistics. We introduced several rules when performing operations in fermionic tensor networks. In order to benchmark our methods, we have applied them to the Hubbard model on two types of geometries and introduced approximate methods to contract these networks. Despite the simplicity of our wavefunction optimization method (simple update), our results at all parameter regions exhibit competitive accuracy with coupled cluster, indicting the huge potential of fermionic tensor network in

Size	U	SU (D=6)	UCCSD	UCCSD(T)	DMRG
3x3x3	4	-0.745, -0.745	-0.748	-0.751	-0.753
	6	-0.549, -0.549	-0.546	-0.547	-0.554
	8	-0.427, -0.427	-0.424	-0.425	-0.432
4x4x4	4	-0.806, -0.799	-0.801	-0.804	NA
	6	-0.594, -0.589	-0.584	-0.586	NA
	8	-0.461, -0.457	-0.454	-0.455	NA
5x5x5	4	NA, -0.806	-0.832	-0.835	NA
	6	NA, -0.606	-0.608	-0.609	NA
	8	NA, -0.474	-0.472	-0.473	NA

Table 5.2: Ground-state energies of Hubbard model on diamond lattices. The first entry in the SU column is computed from compressed contraction and the second from cluster approximation.

correlation physics.

References

[1] Steven R White. “Density matrix formulation for quantum renormalization groups”. In: *Physical review letters* 69.19 (1992), p. 2863.

[2] Steven R White. “Density-matrix algorithms for quantum renormalization groups”. In: *Physical Review B* 48.14 (1993), p. 10345.

[3] Steven R White and Douglas J Scalapino. “Density matrix renormalization group study of the striped phase in the 2D t- J model”. In: *Physical review letters* 80.6 (1998), p. 1272.

[4] Steven R White and DJ Scalapino. “Stripes on a 6-leg Hubbard ladder”. In: *Physical review letters* 91.13 (2003), p. 136403.

[5] Stefan Depenbrock, Ian P McCulloch, and Ulrich Schollwöck. “Nature of the spin-liquid ground state of the S= 1/2 Heisenberg model on the kagome lattice”. In: *Physical review letters* 109.6 (2012), p. 067201.

[6] Zhendong Li et al. “Electronic landscape of the P-cluster of nitrogenase as revealed through many-electron quantum wavefunction simulations”. In: *Nature chemistry* 11.11 (2019), pp. 1026–1033.

[7] Bo-Xiao Zheng et al. “Stripe order in the underdoped region of the two-dimensional Hubbard model”. In: *Science* 358.6367 (2017), pp. 1155–1160.

[8] Matthew B Hastings. “Entropy and entanglement in quantum ground states”. In: *Physical Review B* 76.3 (2007), p. 035114.

[9] Frank Verstraete et al. "Criticality, the area law, and the computational power of projected entangled pair states". In: *Physical review letters* 96.22 (2006), p. 220601.

[10] G Scarpa et al. "Projected Entangled Pair States: Fundamental Analytical and Numerical Limitations". In: *Physical Review Letters* 125.21 (2020), p. 210504.

[11] Iztok Pižorn and Frank Verstraete. "Fermionic implementation of projected entangled pair states algorithm". In: *Physical Review B* 81.24 (2010), p. 245110.

[12] Philippe Corboz et al. "Simulation of strongly correlated fermions in two spatial dimensions with fermionic projected entangled-pair states". In: *Physical Review B* 81.16 (2010), p. 165104.

[13] Philippe Corboz, Jacob Jordan, and Guifré Vidal. "Simulation of fermionic lattice models in two dimensions with projected entangled-pair states: Next-nearest neighbor Hamiltonians". In: *Physical Review B* 82.24 (2010), p. 245119.

[14] Christina V Kraus et al. "Fermionic projected entangled pair states". In: *Physical Review A* 81.5 (2010), p. 052338.

[15] Pascual Jordan and Eugene Paul Wigner. "über das paulische äquivalenzverbot". In: *The Collected Works of Eugene Paul Wigner*. Springer, 1993, pp. 109–129.

[16] Sergey B Bravyi and Alexei Yu Kitaev. "Fermionic quantum computation". In: *Annals of Physics* 298.1 (2002), pp. 210–226.

[17] Paul Springer, Tong Su, and Paolo Bientinesi. "HPTT: A High-Performance Tensor Transposition C++ Library". In: *Proceedings of the 4th ACM SIGPLAN International Workshop on Libraries, Languages, and Compilers for Array Programming*. ARRAY 2017. Barcelona, Spain: ACM, 2017, pp. 56–62. ISBN: 978-1-4503-5069-3. DOI: 10.1145/3091966.3091968. URL: http://doi.acm.org/10.1145/3091966.3091968.

[18] Johnnie Gray. "quimb: A python package for quantum information and many-body calculations". In: *Journal of Open Source Software* 3.29 (2018), p. 819.

[19] Hong-Chen Jiang, Zheng-Yu Weng, and Tao Xiang. "Accurate determination of tensor network state of quantum lattice models in two dimensions". In: *Physical review letters* 101.9 (2008), p. 090603.

[20] Huanchen Zhai. *Block2*. 2021. URL: https://github.com/block-hczhai/block2-preview.

[21] Qiming Sun et al. "PySCF: the Python-based simulations of chemistry framework". In: *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 8.1 (2018), e1340.

[22] Johnnie Gray. *Cotengra*. 2021. URL: https://github.com/jcmgray/cotengra.

[23] Ling Wang and Frank Verstraete. “Cluster update for tensor network states”. In: *arXiv preprint arXiv:1110.4362* (2011).

[24] Jacob Jordan et al. “Classical simulation of infinite-size quantum lattice systems in two spatial dimensions”. In: *Physical review letters* 101.25 (2008), p. 250602.

Chapter 6

SUMMARY AND FUTURE OUTLOOK

When one looks back on the development of quantum many-body methods, we easily realize that it is more often a recursive rather than straight-forward trajectory. We devise numerical theories matching the state-of-art computation power. These theories help us gain a better understanding of the systems of interest, which are used to further refine the theories. From time to time, this recursive path is accompanied with paradigm shift that brings the entire field a big step forward. However, a paradigm shift almost never means that everything from the past is abandoned. In fact, more than often we find the “old" tools revitalized in another fashion. We are lucky enough to live in an age where all these intellectual developments are assisted with fast-growing computing power.

As I reflect back on my own research work over the years, the unified central focus has been on developing and investigating numerical tools that can potentially lead to a paradigm shift in quantum many-body simulations.

The first topic we are trying to address in Chapter 2 and Chapter 4 is what is the “next-generation" of simulation tools for electronic structure calculations of materials. Although the equations underpinning the properties of materials and molecules are the same, the majority of materials science and quantum chemistry communities have taken drastically different approaches to improve their theoretic framework. This can be partly traced to the slightly different focus of two communities and also the different computation cost. As we identify more and more exotic physics in the condensed phase, we come to realize that the wavefunction-based quantum chemical methods can potentially be the right candidate in this region.

In this context, we have considered two scenarios, one with correlated electron and the other with coupled electron and phonons in different regions. For the first case, we report the first unrestricted coupled cluster implementation for ground- and excited-state calculations on crystalline materials. Despite our coarse Brillouin zone sampling, we found much improved results compared with prevalent mean field methods. Our results clearly validate the huge potential of the coupled cluster framework in the correlated region of solid state. In the future, we expect more quantum chemical methods to be ported to the solid-state which can potentially yield more accurate properties and spectra. In this process, more electronic structure infrastructure is needed, and extrapolation toward thermodynamic limit must be properly addressed in order for the final outcome to carry predictive power. For the second case, we report the first attempt to use the coupled cluster method to study coupled electrons and phonons at the same footing. We developed this new machinery by combining electronic and vibrational coupled cluster. By comparison with other state-of-art numerical theories, we found satisfactory performance of our method in the weak to intermediate coupling region. Meanwhile, our first test case in the ab initio region suggests that the widely adopted first-order electron-phonon Hamiltonian may not be adequate to capture the effect of electron-phonon coupling. This is a delicate topic that requires more numerical study, and we expect our tools to be valuable benchmark tools.

The second topic described in Chapter 3 is related to how to achieve the best performance for our methods in the presence of symmetry groups. This topic is first motivated by our observation that cyclic group tensor contraction is the performance bottleneck in our coupled cluster implementation for the solid state. In fact, such kinds of contractions are ubiquitous in a vast array of many-body methods. We introduced irreducible representation alignment, an efficient scheme to store and contract these block sparse tensors. While our strategy does not work on the lowest level tensor contraction kernels, our algorithm transforms the problem into a set of batched matrix multiplication that can be efficiently handled by various math kernel libraries. This approach has allowed us to perform large-scale calculations using state-of-art tensor libraries under distributed parallel setting. In the future, we expect that more sophisticated numerical backend is needed to handle the numerous types of symmetry patterns in the ansatz. In the context of quantum many-body simulation, explicit use of such symmetry not only improves the numerical performance, but also helps constrain the solution in a subspace of interest.

In Chapter 5 we consider our last topic on how to represent fermion wavefunction efficiently using tensor network methods. Despite being a rather young research field, tensor network theory has demonstrated unlimited potential in correlated physics with the remarkable success of DMRG. As the research community delves into exploring the full capability of tensor network methods at higher dimension, piles of challenges arise, an important one being how to efficiently account for the fermion statistics in the ansatz. Our strategy to tackle this problem is to encode the anti-commutation rule directly in the tensor backend. In addition to preserving the intuitive diagrammatic representation, this scheme allows us to efficiently encode the states and operators under certain symmetry groups. Using this approach, we were able to able to directly inherit much of the pre-existing tensor network infrastructure towards fermion simulation with arbitrary graphs. As a result, we were able to leverage various approximate contraction methods to perform benchmark calculations on Hubbard models with different geometries. Our results indicate that in the strongly correlated region, tensor network methods can at least be as competitive as coupled cluster methods. Moving ahead, multiple questions remain to be addressed. For instance, what is the representability of each class of tensor network? How can we efficiently optimize the wavefunction tensors and perform approximate contraction with controlled accuracy? We believe our proposed approach will be an invaluable benchmark tool in addressing these questions.

Appendix A

APPENDIX FOR CHAPTER 3

Generalization to Higher-Rank Tensors

We now describe how to generalize the algorithm to tensors of arbitrary rank, including the more general symmetry conservation rules. We represent a rank N complex tensor with cyclic group symmetry as an rank $2N$ tensor, $\mathcal{T} \in \mathbb{C}^{n_1 \times H_1 \times \cdots \times n_N \times H_N}$ satisfying, modulus remainder $Z \in \{1 \dots G\}$ for coefficients $c_1 \dots c_N$ with $c_i = G/H_i$ or $c_i = -G/H_i$,

$$t_{i_1 I_1 \dots i_N I_N} = \begin{cases} r^{(T)}_{i_1 I_1 \dots i_N I_N} & : c_1 I_1 + \cdots + c_N I_N \equiv Z \pmod G \\ 0 & : \text{otherwise,} \end{cases} \tag{A.1}$$

where the rank $2N - 1$ tensor $\mathcal{R}^{(T)}$ is the *reduced form* of the cyclic group tensor $\mathcal{T}$.

For example, the symmetry conservation rules in the previous section follow Equation A.1 with coefficients that are either 1 or -1 ($G = H_i$).

Any cyclic group symmetry may be more generally expressed using a generalized Kronecker delta tensor with binary values, $\boldsymbol{\delta}^{(T)} \in \{0, 1\}^{H_1 \times \cdots \times H_N}$ as

$$t_{i_1 I_1 \dots i_N I_N} = r^{(T)}_{i_1 I_1 \dots i_N I_N} \delta^{(T)}_{I_1 \dots I_N}. \tag{A.2}$$

Specifically, the elements of the generalized Kronecker delta tensor are defined by

$$\delta^{(T)}_{I_1 \dots I_N} = \begin{cases} 1 & : c_1 I_1 + \cdots + c_N I_N \equiv Z \pmod G \\ 0 & : \text{otherwise.} \end{cases} \tag{A.3}$$

Using these generalized Kronecker delta tensors, we provide a specification of our approach for arbitrary tensor contractions (Figure A.1) in Algorithm 3. This algorithm performs any contraction of two tensors with cyclic group symmetry, written for some $s, t, v \in \{0, 1, \dots\}$, as

$$w_{i_1 I_1 \dots i_s I_s j_1 J_1 \dots j_t J_t} = \sum_{k_1 K_1 \dots k_v K_v} u_{i_1 I_1 \dots i_s I_s k_1 K_1 \dots k_v K_v} v_{k_1 K_1 \dots k_v K_v j_1 J_1 \dots j_t J_t}. \tag{A.4}$$

Algorithm 3 The irrep alignment algorithm for contraction of cyclic group symmetric tensors, for contraction defined as in Equation A.4.

```
Input two tensors $\boldsymbol{\mathcal{U}}$ of rank $s+v$ and $\boldsymbol{\mathcal{V}}$ of rank $v+t$ with symmetry conservation rules described using coefficient vectors $\boldsymbol{c}^{(U)}$ and $\boldsymbol{c}^{(V)}$ and remainders $Z^{(U)}$ and $Z^{(V)}$ as in Equation A.1.
Assume that these vectors share coefficients for contracted modes of the tensors, so that if $\boldsymbol{c}^{(U)} = \begin{bmatrix} \boldsymbol{c}_1^{(U)} \\ \boldsymbol{c}_2^{(U)} \end{bmatrix}$, then $\boldsymbol{c}^{(V)} = \begin{bmatrix} \boldsymbol{c}_2^{(U)} \\ \boldsymbol{c}_2^{(V)} \end{bmatrix}$.
Define new coefficient vectors, $\boldsymbol{c}^{(A)} = \begin{bmatrix} \boldsymbol{c}_1^{(U)} \\ 1 \end{bmatrix}$, $\boldsymbol{c}^{(B)} = \begin{bmatrix} \boldsymbol{c}_2^{(U)} \\ -1 \end{bmatrix}$, and $\boldsymbol{c}^{(C)} = \begin{bmatrix} \boldsymbol{c}_2^{(V)} \\ 1 \end{bmatrix}$.
Define generalized Kronecker deltas $\boldsymbol{\delta}^{(1)}$, $\boldsymbol{\delta}^{(2)}$, and $\boldsymbol{\delta}^{(3)}$ respectively based on the coefficient vectors $\boldsymbol{c}^{(A)}$, $\boldsymbol{c}^{(B)}$, $\boldsymbol{c}^{(C)}$ and remainders $Z^{(U)}$, 0, $Z^{(V)}$.
Let $\bar{\boldsymbol{\mathcal{R}}}^{(U)}$ and $\bar{\boldsymbol{\mathcal{R}}}^{(V)}$ be the given reduced forms for $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ (based on the generalized Kronecker deltas $\boldsymbol{\delta}^{(U)}$ and $\boldsymbol{\delta}^{(V)}$). Assume the reduced forms $\bar{\boldsymbol{\mathcal{R}}}^{(U)}$ and $\bar{\boldsymbol{\mathcal{R}}}^{(V)}$ for $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ do not store the last symmetry mode (other cases are similar). Compute the following new reduced forms $\boldsymbol{\mathcal{R}}^{(U)}$ and $\boldsymbol{\mathcal{R}}^{(V)}$, via contractions:
    $r^{(U)}_{i_1 I_1 \dots i_{s-1} I_{s-1} i_s k_1 K_1 \dots k_{v-1} K_{v-1} k_v Q} = \sum_{I_s K_v} \bar{r}^{(U)}_{i_1 I_1 \dots i_s I_s k_1 K_1 \dots k_{v-1} K_{v-1} k_v} \delta^{(1)}_{I_1 \dots I_s Q} \delta^{(2)}_{K_1 \dots K_v Q}$,
    $r^{(V)}_{k_1 K_1 \dots k_{v-1} K_{v-1} k_v j_1 J_1 \dots j_{t-1} J_{t-1} j_t Q} = \sum_{K_v J_t} \bar{r}^{(V)}_{k_1 K_1 \dots k_v K_v j_1 J_1 \dots j_{t-1} J_{t-1} j_t} \delta^{(2)}_{K_1 \dots K_v Q} \delta^{(3)}_{J_1 \dots J_t Q}$.
Compute
    $r^{(W)}_{i_1 I_1 \dots i_{s-1} I_{s-1} i_s J_1 J_1 \dots j_{t-1} J_{t-1} j_t Q} = \sum_{k_1 K_1 \dots k_{v-1} K_{v-1} k_v} r^{(U)}_{i_1 I_1 \dots i_{s-1} I_{s-1} i_s k_1 K_1 \dots k_{v-1} K_{v-1} k_v Q} r^{(V)}_{k_1 K_1 \dots k_{t-1} K_{v-1} k_v j_1 J_1 \dots j_{t-1} J_{t-1} j_t Q}$
If a standard output reduced form is desired, for example with the last mode of $\boldsymbol{\mathcal{W}}$ stored implicitly, then compute
    $\bar{r}^{(W)}_{i_1 I_1 \dots i_s I_s j_1 J_1 \dots j_{t-1} J_t j_t} = \sum_{Q} r^{(W)}_{i_1 I_1 \dots i_{s-1} I_{s-1} i_s J_1 J_1 \dots j_{t-1} J_{t-1} j_t Q} \delta^{(1)}_{I_1 \dots I_s Q}$.
    If we instead desire a reduced form with another implicit mode, it would not be implicit in $\boldsymbol{\mathcal{R}}^{(W)}$, so we would need to also contract with $\delta^{(3)}_{J_1 \dots J_t Q}$ and sum over the desired implicit mode.
```

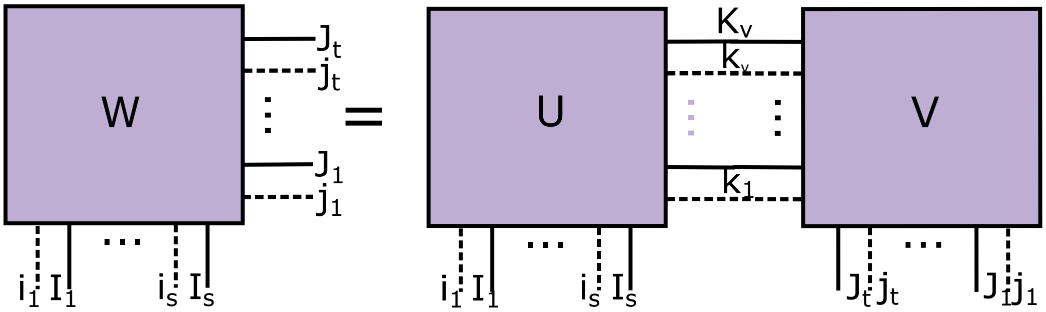

Figure A.1: A contraction of a tensor of rank $s + v$ with a tensor of rank $v + t$ into a tensor of rank $s + t$, where all tensors have cyclic group symmetry and are represented with tensors of twice the order. Note that unlike in the previous section, the lines are not labeled by arrows (denoting coefficients 1 or -1), but are associated with more general integer coefficients $c_i = \pm G/H_i$, to give symmetry conservation rules of the form Equation A.1.

The algorithm assumes the coefficients defining the symmetry of $\mathcal{U}$ and $\mathcal{V}$ match for the indices $K_1 \dots K_v$ (it is also easy to allow for the coefficients to differ by a sign, as is the case in the contraction considered in Chapter 3.3).

INDEX